Evolution d'une famille de protéines cruciales pour la minéralisation du squelette : les phosphoprotéines liant le calcium (SCPPs)

Sidney Delgado

Université Pierre et Marie Curie

Habilitation à Diriger les Recherches

Présentée Par

Sidney DELGADO
Maître de Conférences
Université Pierre et Marie Curie
Equipe Evolution et Développement du Squelette
UMR 7138 « Systématique Adaptation Evolution »

Evolution d'une famille de protéines cruciales pour la minéralisation du squelette : les phosphoprotéines liant le calcium (SCPPs)

30 Novembre 2012

Jury :

Françoise Bleicher, Professeur – Université Claude Bernard Lyon 1
Agnes Bloch-Zupan, Professeur et Vice Doyen - Université de Strasbourg
Jean-Yves Dubuisson, Professeur - Université Paris 6
Marc Girondot, Professeur - Université Paris XI
Michel Goldberg, Professeur émérite - Université Paris Descartes

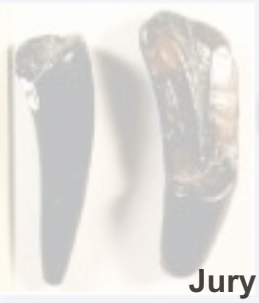
Université Pierre et Marie Curie

Habilitation à Diriger les Recherches

Présentée Par

Sidney DELGADO
Maître de Conférences
Université Pierre et Marie Curie
Equipe Evolution et Développement du Squelette
UMR 7138 « Systématique Adaptation Evolution »

Evolution d'une famille de protéines cruciales pour la minéralisation du squelette : les phosphoprotéines liant le calcium (SCPPs)



30 Novembre 2012

Jury :

Françoise Bleicher, Professeur – Université Claude Bernard Lyon 1
Agnes Bloch-Zupan, Professeur et Vice Doyen - Université de Strasbourg
Jean-Yves Dubuisson, Professeur - Université Paris 6
Marc Girondot, Professeur - Université Paris XI
Michel Goldberg, Professeur émérite - Université Paris Descartes

«A un certain moment, il se forma par accident une molécule particulièrement remarquable. Nous l'appellerons le Réplicateur. [...] Ils ont parcouru un long chemin, ces réplicateurs. On les appelle maintenant " gènes ", et nous sommes leurs machines de survie.»

Richard Dawkins - Le Gène égoïste (1976)

Remerciements

Je remercie, pour commencer, les personnes qui m'ont formé au métier de la recherche, et en tout premier lieu Jean-Yves Sire qui fut mon directeur de thèse et dirige aujourd'hui l'équipe « Evolution et Développement du Squelette » dans laquelle ces travaux de recherche ont été réalisés. De même, j'associe à ces remerciements Marc Girondot qui m'a appris énormément en biologie moléculaire et en évolution moléculaire durant mon Master 2 et une partie de ma thèse.

J'exprime toute ma gratitude au directeur de mon unité (UMR7138), Hervé Le Guyader qui, le premier, m'a fait comprendre que j'avais la possibilité de passer mon HDR et m'a poussé dans cette voie.

Je tiens à remercier chaleureusement deux professeurs en médecine dentaire, très humains, qui ont toujours soutenu mes recherches et m'ont encouragé, voyant l'intérêt qu'elles présentent pour leur propre domaine : Michel Goldberg et Henry Magloire.

Je remercie les personnes qui ont accepté de faire partie de la liste de mes rapporteurs potentiels : Michel Goldberg, Françoise Bleicher, Agnès Bloch-Zupan, Bernhard Ganss, Dan Deutsch, ainsi que les membres du Jury.

Au cours de mes recherches, j'ai eu la chance de rencontrer de nombreux étudiants, avec lesquels j'ai collaboré. Grâce à eux j'ai aussi « fait mes armes » comme encadrant. La recherche dépend beaucoup de leur travail et de leur enthousiasme : Merci à Nawfal Al-Hashimi (avec qui j'ai, par ailleurs, lié une amitié profonde), à Jérémie Silvent, à Barbara Gasse, à Claire Bardet, à Nathalie Assaraf-Weill et à Meriem Belheouane.

J'associe à ces remerciements, Ann Huysseune, Professeur à l'Université de Gand, Belgique. Elle a toujours apprécié mes recherches et m'a accueilli pendant un an dans son laboratoire pour un stage post-doctoral qui a beaucoup apporté à ma carrière.

J'ai, bien évidemment, une pensée émue pour ma famille qui est mon rayon de soleil, ma merveilleuse épouse Virginie qui m'a toujours encouragé pour que j'arrive à soutenir mon HDR, et mes deux fils Féodor et Virgil qui me soutiennent de leur amour.

Avant propos

L'HDR ne représente pas seulement un rite de passage vers de nouvelles responsabilités, c'est aussi le moment de faire le bilan de son parcours professionnel et personnel. C'est l'occasion de réfléchir aux travaux qui ont déjà été accomplis et à ceux qui devront l'être dans le futur. C'est un exercice parfois difficile tant la carrière d'un enseignant-chercheur est riche. Comment organiser de manière logique sur le papier tout ce qui a déjà été accompli?

Pour commencer, ma passion pour la génétique de l'évolution est née d'une rencontre avec un professeur de génétique de l'Université d'Orsay : Pierre-Henri Gouyon. J'ai tout de suite eu un engouement pour ses cours qui ne ressemblaient à aucun autre, évoquant des notions rarement vues ailleurs comme la mort, la reproduction, l'eugénisme... Sa façon de présenter les travaux de SJ Gould en évoquant la coupole de la cathédrale St Marc, ou la théorie du gène égoïste de Dawkins à travers les multiples conflits dans le génome, avait de quoi soulever l'enthousiasme. Cela rappelait en moi les mêmes émotions que, lorsque petit garçon, je lisais des livres sur les dinosaures ou sur le monde « sauvage ». C'est ainsi que je me suis retrouvé, quelques années plus tard, à étudier les gènes responsables de la mise en place des tissus minéralisés. Sujet d'une grande richesse mais qui, historiquement, a suivi un long chemin sinueux. Le laboratoire dans lequel j'ai commencé à travailler cherchait à l'origine à comprendre l'origine des tissus minéralisés présents dans les écailles de certains vertébrés terrestres et aquatiques. D'où l'idée d'introduire dans cet ancien laboratoire « d'Anatomie Comparée » la biologie moléculaire afin d'étudier les gènes et les protéines présents dans ces tissus. C'est de cette manière que finalement, on m'a proposé de travailler sur l'évolution d'une protéine de l'émail, l'amélogénine à l'occasion d'une thèse de Doctorat.

Depuis cette époque, « de l'eau est passée sous les ponts ». Les connaissances sur les protéines de minéralisation ont permis de découvrir une grande famille de gènes de minéralisation dont on ne soupçonnait pas l'existence. Les techniques de biologie moléculaires ont également beaucoup changé, permettant de connaître les gènes et leur organisation beaucoup plus rapidement que par le passé, parfois d'un simple clic sur un ordinateur. De nouvelles questions sont apparues, de nouvelles problématiques aussi. La manière de faire la recherche est différente mais, la curiosité et la volonté de trouver demeurent intactes en moi. Elles me poussent à continuer plus avant, à trouver de nouveaux outils, de nouveaux moyens pour tenter de répondre à ces questions passionnantes concernant l'évolution des vertébrés.

Sommaire

I- Parcours Personnel

1. Curriculum vitae

Nom, Prénoms:	DELGADO Christophe, Sidney
Date et lieu de naissance:	04 mai 1971 à Versailles (078)
Situation Familiale:	Marié, 2 enfants
Adresse professionnelle:	UMR 7138 SAE- Bâtiment A, 2ème étage, Boîte courrier 5, 7 quai Saint Bernard, 75005, PARIS
E-mail:	sidney.delgado@upmc.fr
www:	http://sites.google.com/site/delgadosidney

2. Titres Universitaires

2003 Qualification aux fonctions de maître de conférences en section 64 et 68 (“Biologie des organismes“ et “Biochimie, biologie moléculaire“)

1998-2002 Doctorat de Biodiversité : Génétique, histoire et mécanismes de l'évolution - Université Paris VII - Denis Diderot - « L’Amélogénine, protéine majeure de l'émail dentaire. Origine, analyses évolutive et phylogénétique chez les Amniotes et recherche de son expression lors de la formation des dents de Chalcides viridanus (Squamate, Scincidé) ».

1998 D.E.A. Biodiversité : Génétique, histoire et mécanismes de l'évolution Université Paris XI - Orsay

1997 Maîtrise de Génétique Moléculaire - Mention : Biologie Moléculaire et Génétique du Développement - Université Paris XI - Orsay

1995 Licence de Biochimie, Génétique Fondamentale, Biologie Moléculaire et Biologie Animale - Université Paris XI - Orsay

1994 DEUG B de Biologie - Université Paris XI - Orsay

3. Parcours

2007-2010 Maître de conférences titulaire - Section 68 - Université Paris 6
2006-2007 Maître de conférences stagiaire - section 68 - Université Paris 6

2006 Post doctorant - UMR 7138- Université Paris 6 - Bourse de recherche attribuée par “l’Institut Français pour la Recherche Odontologique“

2005 Post doctorant : *"Vertebrate Morphology and Developmental Biology laboratory“,* Ghent University - Bourse de la “Fondation pour la Recherche Médicale“

2003-2004 Post doctorant - Université Paris 6 - laboratoire FRE 2696
Bourse de recherche attribuée par “l’Institut Français pour la Recherche Odontologique“ (IFRO)

1998 -2002 Doctorant - Université Paris 7 – UMR 8570
Financement : "Collège de France"

4. Activités d'enseignements

4.1. Enseignement :

- Licence 1ère année - Parcours BCPG (cycle d'intégration)
LV 102 - Diversité du vivant

- Licence 2ème année – Mention : Sciences de la Vie
LV 201 - Organisation des métazoaires

- Licence 3ème année – Mention : Sciences de la Vie et Sciences de la Terre

LV 301 - Biologie comparée et évolution des animaux
LT 310 - Vie passée, Vie actuelle
LT 324 - Méthodes d'analyse de la Paléontologie

- Master 1ère : MSUE4207 – *Biominéralisations*

- *Master 2ème année :* UE: SEP 27- spécialité SEP - *Tissus squelettiques des Vertébrés*

4.2. Responsabilités:

Je suis également coresponsable de l'Unité d'Enseignement **LV102 "Diversité du Vivant"** constituée de 600 étudiants environ chaque année. Dans le cadre de ces responsabilités, je m'occupe des tâches suivantes :

- Gestion des plannings et réservations d'amphis et de salles de TP et de TD
- Gestion du site internet (plate-forme universitaire SAKAI)
- Participation à la mise en place de la charte des CME et aux sélections des CME
- Mise en place d'un tutorat en 2011
- Organisation des examens, consultations de copies et délibérations

5. Activités liées à l'administration

- (2009-2010) Président du comité de Thèse de David Marjanovic (Directeur de thèse M. Laurin, ED n° 392)

- (1999 - 2001) Membre élu du conseil scientifique de l'UFR de biologie de l'Université Paris VII - Denis Diderot : Représentant des étudiants de 3ème

cycle.

6. Activités liées à la recherche

1. Attribution de la **P**rime d'**I**nvestissement à la **R**echerche (campagne 2010).

2. Participation à une demande de financement ANR (2012). « Vers les origines de la minéralisation chez les vertébrés : apport du séquençage de transcriptomes à grande échelle » en Collaboration avec Frédéric DELSUC ((Institut des Sciences de l'Evolution - Université Montpellier 2).

3. *Editorial boards* :
    1. Dataset Papers in Bioinformatics
    2. Journal of Applied Ichthyology
    3. Frontiers in Evolutionary and Population Genetics

4. *Organisation de colloques* :

    Comité d'organisation des 13èmes Journées Françaises de Biologie des Tissus Minéralisés, Paris 25-27 mai 2011 (UPMC-MNHN).

5. Membre du groupement de recherche sur les biominéralisations et les tissus minéralisés. Ce réseau mis en place par le Professeur J Cubo (UPMC) rassemble tous les laboratoires travaillant sur les phénomènes de biominéralisation en Ile-De-France.

6. ACMO (Agent Chargé de la Mise en Oeuvre des règles d'hygiène et sécurité). Affecté au service de l'équipe "Evolution et Développement du Squelette".

7. Responsable du pôle "Biologie Moléculaire" de l'équipe "Evolution et Développement du squelette" dans l'UMR 7138.

7. Encadrements

Encadrements de Master 2

- **BELHEOUANE Myriem** (2011) – « Etude de l'évolution des protéines de l'émail chez les Tétrapodes ». (*encadrement*: 100 %)

  Poster:
  Belheouane M, Sire JY & **Delgado S** (2011). « Les gènes de minéralisation de l'émail chez les reptiles: Evolution et comparaison avec les Mammifères ». *13èmes Journées Françaises de Biologie des Tissus Minéralisés*, Paris 25-27 mai 2011.

- ASSARAF-WEILL Nathalie (2009) – « Approche évolutive du fonctionnement de l'améloblaste. Étude chez l'amphibien caudate*, Pleurodeles waltl* » (*Co-encadrement*: 10 %)

Poster:
Assaraf-Weill N, Al-Hashimi N, Bardet C, **Delgado S**, Sire JY et Davit-Béal T (2009). Caractérisation du gène de l'amélogénine chez *Pleurodeles waltl,* (Lissamphibia, Caudata) et son expression lors de l'odontogenèse. *11èmes Journées Françaises de Biologie des Tissus Minéralisés*, Nice, 19, 20, 21 Mars 2009.

- FROMENTIN Delphine (2004) - « Impact des modifications de la structure primaire de l'amélogénine sur la structure de l'émail dentaire des Mammifères». (*Co-encadrement*: 40 %)

Article :

Sire JY, **Delgado S**, Fromentin D & Girondot M (2005). Amelogenin: Lessons from Evolution. *Archives of Oral Biology 50:205-212.*

Encadrements de Thèse

- SILVENT Jérémie (2012) - « Morphologie, minéralisation et expression génique d'ostéoblastes primaires humains sur matrice dense de collagène ». (*Co-encadrement*: 10 %)

Article soumis:

Silvent J, Sire JY & **Delgado S** (2012). The dentin matrix acidic phosphoprotein 1 (DMP1) in the light of mammalian evolution.

- Al-HASHIMI Nawfal (2010) – « L'énaméline, la plus grande protéine de l'émail dentaire. Analyse évolutive chez les Amniotes ». (*Co-encadrement*: 70 %)

Articles:

Al-Hashimi N, Lafont AG, **Delgado S,** Kawasaki K & Sire JY (2010). The enamelin genes in lizard, crocodile and frog, and the pseudogene in the chicken provide new insights on enamelin evolution in tetrapods. *Mol Biol Evol. 2010 Sep;27(9):2078-94.*

Al-Hashimi N, Sire JY & **Delgado S** (2009). Evolutionary Analysis of Mammalian Enamelin, the Largest Enamel Protein, Supports a Crucial Role for the 32 kDa Peptide and Reveals Selective Adaptation in Rodents and Primates. *J Mol Evol. 2009 Dec;69(6):635-56.*

- Claire BARDET (2009) – « La Phosphoglycoprotéine de la Matrice Extracellulaire, MEPE. Origine, fonction et évolution » (*Co-encadrement*: 10 %)

Article:

Bardet C, **Delgado S** & Sire JY (2009). MEPE Evolution in Mammals Reveals Regions and Residues of Prime Functional Importance. *Cell Mol Life Sci. 2010 Jan;67(2):305-20*.

- DAVIT-BEAL Tiphaine (2006) – « Odontogenèse chez l'amphibien Caudate, *Pleurodeles waltl* » - (*Co-encadrement*: 20 %)

Articles:

Davit-Béal T, Chisaka H, **Delgado S** & Sire JY (2007). Amphibian teeth: current knowledge, unanswered questions, and some directions for future research. *Biological Reviews 2007 Feb;82(1):49-81.*

Sire JY, Davit-Béal T, **Delgado S,** Van Der Heyden C & Huysseune A (2002). First-generation teeth in non-mammalian lineages: Evidence for a conserved ancestral character ? *Microscopy Research and Technique 59 (5): 408-34.*

Encadrement de stage post-doctoral

- **LAFONT Anne-Gaëlle** (2009-2010) - (*Co-encadrement*: 10 %)

Article:

Al-Hashimi N, Lafont AG, **Delgado S,** Kawasaki & Sire JY (2010). The enamelin genes in lizard, crocodile and frog, and the pseudogene in the chicken provide new insights on enamelin evolution in tetrapods. *Molecular Biology and Evolution 2010 Sep;27(9):2078-94.*

- **CHISAKA Ideki** (2002-2003)- (*Co-encadrement*: 20 %)

Article:

Davit-Beal T, Chisaka H, **Delgado S** & Sire JY (2007). Amphibian teeth: current knowledge, unanswered questions, and some directions for future research. *Biological Reviews 2007 Feb;82(1):49-81.*

8. Publications

8.1. Articles dans des revues internationales avec comité de lecture

21. Le Roy N, Marie B, Gaume B, Guichard N, **Delgado S**, Zanella-Cléon I, Becchi M, Auzoux-Bordenave S, Sire JY & Marin F (2012,). Identification of two carbonic anhydrases in the mantle of the European abalone Haliotis tuberculata (Gastropoda, Haliotidae): phylogenetic implications. J Exp Zool B Mol Dev Evol. 2012 Jul;318(5):353-67. Impact Factor: 2,373

20. Sire JY, Huang WL, **Delgado S**, Goldberg M and Den Besten P (2012). Evolutionary story of mammalian-specific amelogenin exons 4, "4b", 8 and 9. Journal of Dental Research 2012 Jan;91(1):84-9. Epub 2011 Sep 26. Impact Factor: 3,496

19. Al-Hashimi N, Lafont AG, **Delgado S,** Kawasaki K, Sire JY (2010). The enamelin genes in lizard, crocodile and frog, and the pseudogene in the chicken provide new insights on enamelin evolution in tetrapods. *Mol Biol Evol. 2010 Sep;27(9):2078-94. Epub 2010 Apr 19.* Impact Factor: 7,28

18. Al-Hashimi N, Sire JY, **Delgado S** (2009). Evolutionary Analysis of Mammalian Enamelin, the Largest Enamel Protein, Supports a Crucial Role for the 32 kDa Peptide and Reveals Selective Adaptation in Rodents and Primates. *J Mol Evol. 2009 Dec;69(6):635-56.* Impact Factor: 3,234

17. Bardet C, **Delgado S**, Sire JY (2009). MEPE Evolution in Mammals Reveals Regions and Residues of Prime Functional Importance. *Cell Mol Life Sci. 2010 Jan;67(2):305-20*. Impact Factor: 5.511

16. Sire JY, **Delgado S,** Girondot M (2008). Hen's teeth with enamel cap: from dream to impossibility. *BMC Evolutionary Biology*: 8: 246. Impact Factor: 4,091

15. Delgado S, Vidal N, Veron G, Sire JY (2008). Amelogenin, the major protein of tooth enamel:a new phylogenetic marker for ordinal mammal relationships. *Mol phyl Evol 2008* Feb 2. Impact Factor: 3,994

14. Richard B, **Delgado S**, Gorry P, Sire JY (2007). A study of polymorphism in human AMELX. *Arch Oral Biol 2007 Jul 21*. Impact Factor: 1,554.

13. Sire JY, Davit-Béal T, **Delgado S**, Gu X (2007). The origine and evolution of enamel mineralization genes. *Cells Tissues Organs 186(1):25-48*. Impact Factor: 1,776

12. Delgado S, Ishiyama M & Sire JY (2007). Validation of Amelogenesis Imperfecta Inferred from Amelogenin Evolution. *J Dent Res 86(4):326-330.* Impact Factor: 3,496

11. Davit-Béal T, Chisaka H, **Delgado S**, Sire JY (2007). Amphibian teeth: current knowledge, unanswered questions, and some directions for future research. *Biological Reviews 2007 Feb;82(1):49-81.* Impact Factor: 8,833

10. Sire JY, **Delgado S** & Girondot M (2006). The amelogenin story: Origin and evolution. *European Journal of Oral Sciences 2006 May;114 Suppl 1:64-77; discussion 93-5, 379-80.* Impact Factor: 2,071

9. **Delgado S**, Couble ML, Magloire H & Sire JY (2006). Cloning, Sequencing and Expression of the Amelogenin Gene in two Scincid Lizards. *J Dent Res 85(2):138-143.* Impact Factor: 3,496

8. Huysseune A, **Delgado S** & Eckhard Witten P (2005). How to Replace a Tooth: Fish(ing) for Answers. *Oral Biosciences and Medicine Vol 2 (Issue 2/3): 75-81. (This journal has ceased publication)*

7. **Delgado S**, Girondot M & Sire JY (2005). Molecular evolution of amelogenin gene in mammals. *J Mol Evol. 60:12–30.* Impact Factor: 3,234

6. **Delgado S**, Davit-Béal T, Allizard F & Sire JY (2005). Tooth development in a scincid lizard, *Chalcides viridanus,* (Squamata), with particular attention paid on enamel formation. *Cell and Tissue Research 319: 71–89.* Impact Factor: 2,613

5. Sire JY, **Delgado S**, Fromentin D & Girondot M (2005). Amelogenin: Lessons from Evolution. *Archives of Oral Biology 50:205-212.* Impact Factor: 1,554

4. **Delgado S,** Davit-Béal T & Sire J-Y (2003). The dentition and tooth replacement pattern in *Chalcides* (Squamata; Scincidae). *Journal of Morphology 256(2):146-59.* Impact Factor: 1,621

3. Sire JY, Davit-Béal T, **Delgado S,** Van Der Heyden C & Huysseune A (2002). First-generation teeth in non-mammalian lineages: Evidence for a conserved ancestral character? *Microscopy Research and Technique 59 (5): 408-34.* Impact Factor: 1,644

2. **Delgado S**, Casane D, Bonnaud L, Laurin M, Sire JY & Girondot M (2001). Molecular Evidence for Precambrian origin of amelogenin, the major protein of vertebrate enamel. *Molecular Biology and Evolution 18:2146-2153.* Impact Factor: 6,438

1. Girondot M, **Delgado S** & Laurin M (1998). Evolutionary analysis of "hagfish amelogenin". *Anatomical Records 252:608-611.* Impact Factor: 1,801

8.2. Articles dans des revues nationales avec comité de lecture

Delgado S, Davit-Béal T, Al Hashimi N & Sire J-Y (2007). Analyse évolutive de l'énaméline chez les Tétrapodes : mise en évidence de régions fonctionnelles et aide à la validation de mutations conduisant à l'amélogenèse imparfaite de type 2. *Les Cahiers de l'ADF 22-23(vol. 10):33-42.*

Delgado S, Davit-Béal T & Sire J-Y (2005). L'analyse évolutive moléculaire: un outil pour le diagnostic des pathologies génétiques héréditaires liées aux protéines dentaires. *Les Cahiers de l'ADF 18-19(vol. 8):34-42.*

8.3. Articles résumés de conférences internationales avec comité de lecture

Delgado S, Al-Hashimi N & Sire J-Y (2007). Evolutionary analysis of DMP1. *European Cells and Materials Vol. 14. Suppl. 2, 2007 (page 10)*

Delgado S, Ishiyama M, Mikami M, Imai A, Shimomura H & Sire JY (2001). Evolutionnary and phylogenetic analyses of amelogenin genes in amniotes. *Connect Tissue Res Vol 43, Number 2-3*

Delgado S, Sire JY & Girondot M (1998). Evolutionnary Analysis of Non-mammalian Amelogenin Genes. *Chemistry and Biology of Mineralized Tissues - American Academy of Orthopaedic Surgeons (1998 - Page 402).*

8.4. Communications orales dans des conférences nationales & internationales

Le Roy N, Gaume B, Marie B, Guichard N, **Delgado S**, Zanella-Cleon I, Becchi M, Auzoux-Bordenave S, Sire JY. & Marin F (2011). Identification de deux anhydrases carboniques dans le manteau de l'ormeau européen *Haliotis tuberculata* (Gastropoda, Haliotidae): Implications phylogénétiques. *13èmes Journées Françaises de Biologie des Tissus Minéralisés*, Paris 25-27 mai 2011

Delgado S, Al-Hashimi N, Lafont AG, Kawasaki K & Sire JY (2010). Evolutionary analysis of enamelin in mammals, sauropsids and amphibians provides new insights on its function. *10th International Conference on Tooth Morphogenesis and Differentiation* – Berlin, Allemagne.

Delgado S, Al-Hashimi N & Sire JY (2007). Evolutionary analysis of DMP1. *9th International Conference on Tooth Morphogenesis and Differentiation* - Zurich, Suisse.

Al-Hashimi N, **Delgado S** & Sire JY (2007). L'énaméline des Mammifères : mise en évidence de régions conservées, mode d'évolution et aide à la validation de mutations conduisant à l'amélogenése imparfaite. *10èmes Journées Françaises de Biologie des Tissus Minéralisés*, St Valéry sur Somme, 24-26 mai 2007.

Sire JY, **Delgado S** & Girondot M (2005). The amelogenin story: Origin and evolution. *7th International Symposium on the Composition, Properties and Fundamental Structure of Tooth Enamel*.

Delgado S, Davit-Béal T & Sire JY (2005). L'analyse évolutive moléculaire: un outil pour le diagnostic des pathologies génétiques héréditaires liées aux protéines dentaires. *Congrès de l'Association Dentaire Française, Session de l'IFRO, 2004, Paris.*

Delgado S & Sire JY (2002). Analyse évolutive de l'amélogénine chez les Mammifères. *5émes Journées Françaises de Biologie des Tissus Minéralisés* - Faculté de Chirurgie Dentaire, Université Paris 5, Montrouge.

Delgado S, Ishiyama M, Mikami M, Imai A, Shimomura H & Sire JY (2001). Evolutionnary and phylogenetic analyses of amelogenin genes in amniotes. *7th International Conference on Tooth Morphogenesis and Differentiation* - La Londe-les-Maures, France.

Delgado S, Girondot M & Sire JY (1999). Evolutionnary analysis of non-mammalian amelogenin genes. *COST Action B8*, *Odontogenesis* - Ghent University, Belgique.

8.5. Posters dans des conférences nationales & internationales

Belheouane M, Sire J.Y. & **Delgado S** (2011). Les gènes de minéralisation de l'émail chez les reptiles : Evolution et comparaison avec les Mammifères. *13èmes Journées Françaises de Biologie des Tissus Minéralisés*, Paris 25-27 mai 2011.

Assaraf-Weill N, Al-Hashimi N, Bardet C, **Delgado S**, Sire JY et Davit-Béal T (2009). Caractérisation du gène de l'amélogénine chez *Pleurodeles waltl, (Lissamphibia*, Caudata) et son expression lors de l'odontogenèse. *11èmes Journées Françaises de Biologie des*

Tissus Minéralisés, Nice, 19, 20, 21 Mars 2009.

Al-Hashimi N, **Delgado S** & Sire JY (2007). Mammalian Enamelin: Identification of conserved regions, evolution mode and made use of for validation of mutations leading to amelogenesis imperfecta. *9th International Conference on Tooth Morphogenesis and Differentiation – Zurich, Suisse.*

Delgado S, Richard B & Sire JY (2007). Étude du polymorphisme de l'Amélogénine et utilisation de l'analyse évolutive pour valider des mutations conduisant à l'amélogenèse imparfaite. *10èmes Journées Françaises de Biologie des Tissus Minéralisés*, St Valéry sur Somme, 24-26 mai 2007.

Delgado S, Al Hashimi N, Davit-Béal T & Sire JY (2006). A spotlight on Enamelin: evolutionnary genetics brought new insights into the role of different sites of the protein. *Comparative Evolution, Development and Regeneration of Epidermal-Mesenchymal Organs meeting; COST Action B23, Oral facial development and regeneration* - Université Paris 5, France, 07-08 dec. 2006.

Soenens M, **Delgado S** & Huysseune A (2005). Wnt overexpression and tooth development in the zebrafish. *4th European Zebrafish Development and Genetics meeting,* Dresden, Germany, 13-16 July, 2005.

Delgado S, Davit-Béal T, Allizard F & Sire JY (2004). Tooth development in a scincid lizard, *Chalcides viridanus* (Squamata), with particular attention paid on enamel formation. *7th International Congress of Vertebrate Morphology* - Boca-Raton, Florida (USA), 26-31 Juillet 2004.

Delgado S, Davit-Béal T, Couble M-L, Allizard F & Sire JY (2004). Dentition pattern, tooth development and amelogenin expression in the lizard, *Chalcides viridanus* (Scincidae, Squamata). *8th Int. Conf. Tooth Morphogenesis and Differentiation* - York, UK, Juillet 2004.

Delgado S, Sire JY & Girondot M (1999). Apports phylogénétiques à l'analyse structurale et fonctionnelle des Amélogénines. *2èmes Journées Françaises de Biologie des Tissus Minéralisés*, INRA - Versailles.

Delgado S, Sire JY & Girondot M (1998). Evolutionnary Analysis of Non-mammalian Amelogenin Genes. *6th Inter. Conference on the Chemistry and Biology of Mineralized Tissues* - Vittel.

Sire JY, **Delgado S** & Girondot M (1999). L'évolution du squelette dermique des vertébrés. L'apport des études comparatives, structurales et développementales à la reconnaissance d'homologies. *2èmes Journées Françaises de Biologie des Tissus Minéralisés*, INRA – Versailles.

Girondot M, **Delgado S**, Sire JY & Laurin M (1999). Datation moléculaire de l'acquisition de matrices des tissus minéralisés chez les vertébrés. *2èmes Journées Françaises de Biologie des Tissus Minéralisés*, INRA - Versailles.

8.6. Séquences d'ADN publiées dans Genbank

NM_001098513 : *Macaca mulatta* amelogenin (AMELX), mRNA
gi|148612850|ref|NM_001098513.1|[148612850]

EF537873 : *Tarsius syrichta* amelogenin (AMELX) gene, partial cds
gi|147743852|gb|EF537873.1|[147743852]

EF537872 : *Callithrix jacchus* amelogenin (AMELX) gene, complete cds
gi|147743850|gb|EF537872.1|[147743850]

EF537871 : *Macaca mulatta* amelogenin (AMELX) gene, complete cds
gi|147743848|gb|EF537871.1|[147743848]

EF537870 : *Pongo pygmaeus* amelogenin (AMELX) gene, partial cds
gi|147743846|gb|EF537870.1|[147743846]

EF537869 : *Pan troglodytes* amelogenin (AMELX) gene, complete cds
gi|147743844|gb|EF537869.1|[147743844]

AY788990 : *Loxodonta africana* amelogenin (AMEL) gene, exon 6 and partial cds
gi|55709878|gb|AY788990.1|[55709878]

AY787744 : *Tursiops truncatus* amelogenin (AMEL) gene, exon 6 and partial cds
gi|55501314|gb|AY787744.1|AY787743S2[55501314]

AY787743 : *Tursiops truncatus* amelogenin (AMEL) gene, exon 5
gi|55501313|gb|AY787743.1|AY787743S1[55501313]

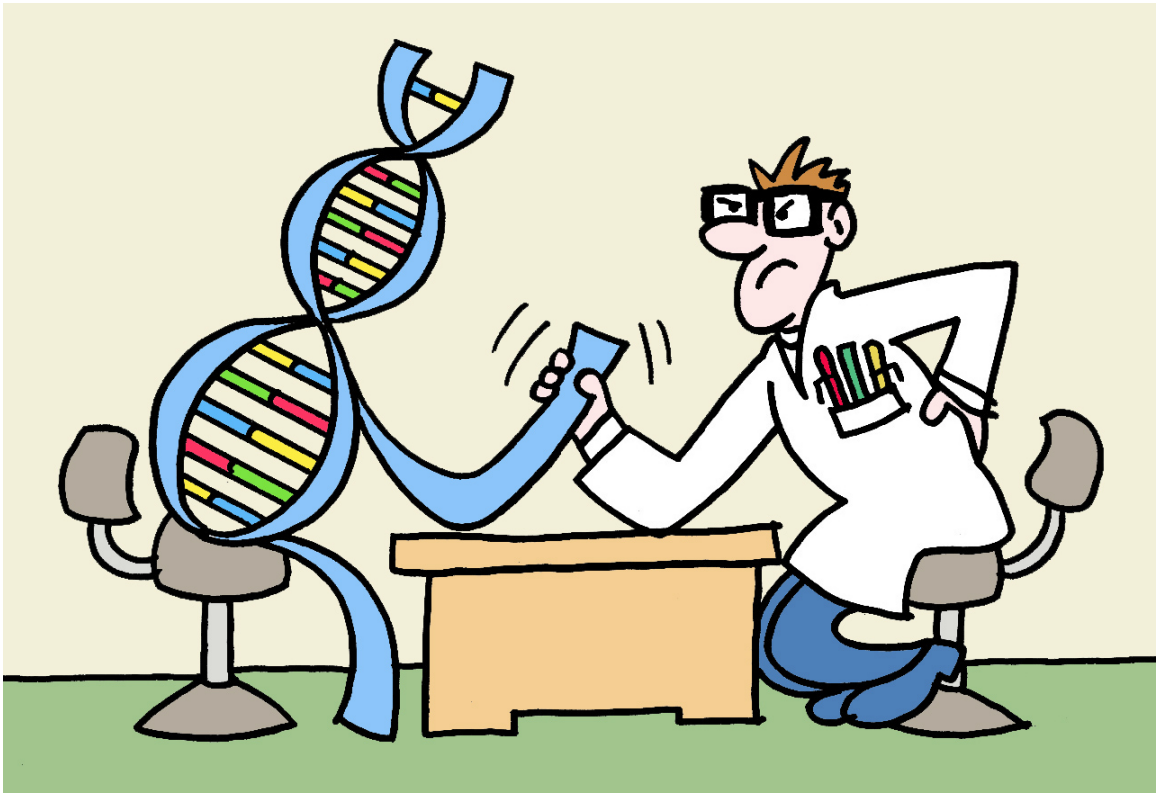
AH014446 : *Tursiops truncatus* amelogenin (AMEL) gene, exons 5, 6 and partial cds
gi|55501312|gb|AH014446.1|SEG_AY787743S[55501312]

AY787742 : *Hexaprotodon liberiensis* amelogenin (AMEL) gene, exon 6 and partial cds
gi|55501293|gb|AY787742.1|[55501293]

DQ364453 : *Chalcides sexlineatus* amelogenin (AMEL) mRNA, partial cds.
gi|86450315|gb|DQ364453.1|[86450315]

DQ364454 : *Chalcides viridanus* amelogenin (AMEL) mRNA, partial cds
gi|86450317|gb|DQ364454.1|[86450317]

II- Activités de recherche



II- Activités de recherche
Synthèse des travaux et description des principaux résultats

Origine et évolution des protéines minéralisantes

Introduction

L'invention des tissus minéralisés représente un moment important dans l'évolution des vertébrés; en effet, ces tissus furent à l'origine d'importantes adaptations phénotypiques comme l'armure corporelle pour la protection, les dents pour la prédation ou encore le squelette pour la locomotion.

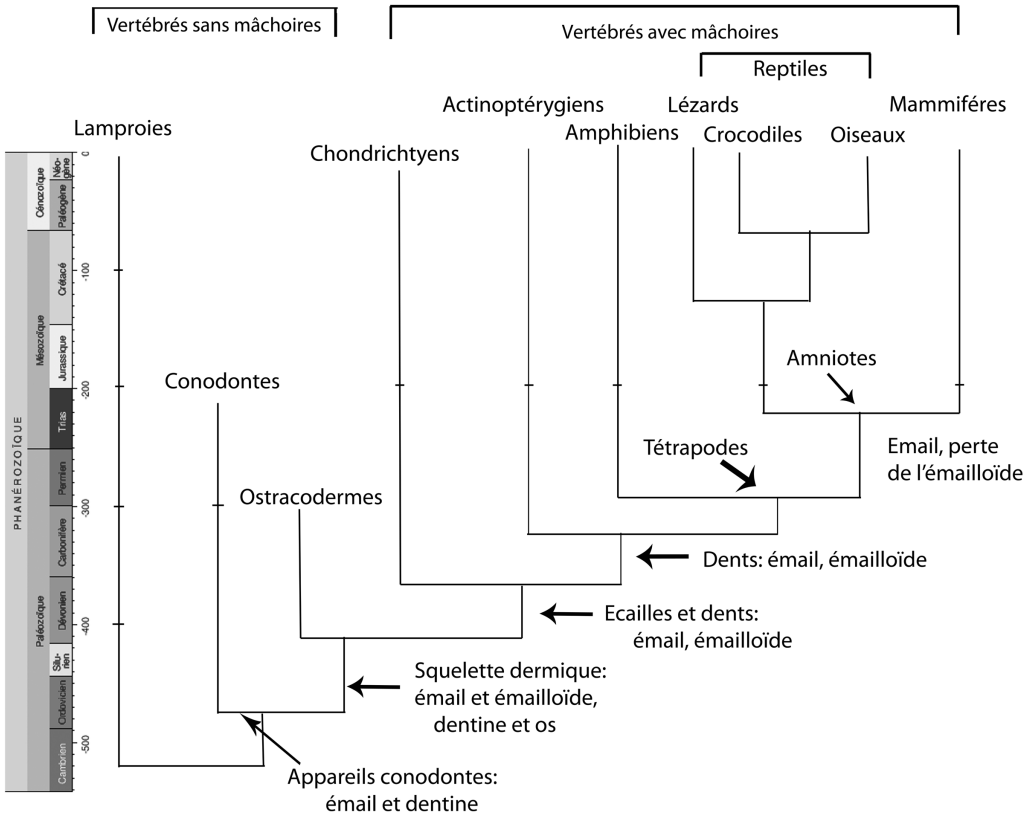
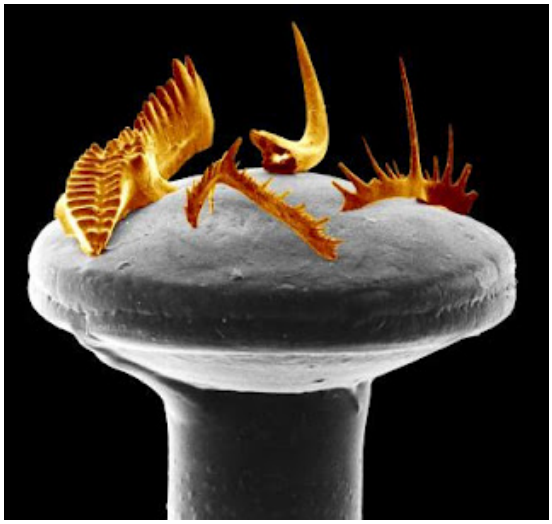


Fig. 1. Evolution des tissus minéralisés des vertébrés (d'après Al-Hashimi, 2010)

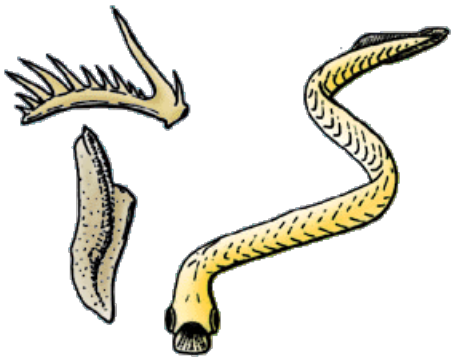
L'origine des tissus minéralisés a longtemps été au cœur d'un débat. Les premiers éléments minéralisés apparentés aux dents qui apparaissent dans le registre fossile se trouvent chez des vertébrés primitifs sans mâchoires, les Conodontes. Ils possédaient un alignement complexe d'éléments dentaires qu'on appelle "appareil conodonte" et qui occupait une position interne (bucco pharyngienne). Ces éléments sont visibles à partir du Cambrien

moyen (500 millions d'années, Ma) jusqu'à la fin du trias et sont composés de cristaux de phosphate de calcium, un composé chimique qui n'est pas très répandu chez les êtres vivants. Il compose notamment l'os et les dents des vertébrés.



Quatre types différents de conodontes (MEB). (Source : Mark Purnell, University of Leicester).

Cet "animal conodonte" est un peu devenu un symbole de la paléontologie : un organisme dont on connaissait l'existence (puisqu'on connaissait les appareils conodontes), mais sans l'avoir jamais trouvé. Du fait de ce questionnement et de leur importance, la nature des conodontes fut longtemps l'un des plus grands mystères de la paléontologie. Il a fallu attendre…1983 pour que ce Saint Graal soit découvert au fond de la collection d'un musée !



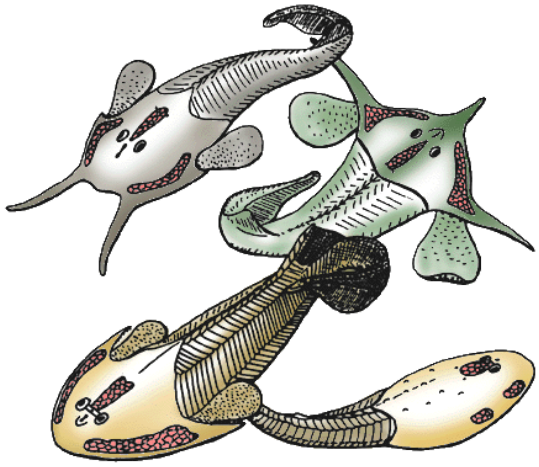
Reconstitution d'un conodonte (à droite) et appareil conodonte (à gauche) (© Philippe Janvier, 1997)

Aujourd'hui, la majorité des spécialistes les considèrent comme des chordés. Cependant, leur position phylogénétique au sein des chordés est toujours objet de débats car ils ne possédaient pas certaines caractéristiques des vertébrés, comme un squelette minéralisé ou des nageoires paires (Donoghue et Sansom, 2002). De plus, l'os semble absent de "l'appareil

conodonte" (Donoghue et al., 2006). Cependant, de plus en plus d'auteurs les considèrent comme des vertébrés (Janvier, 2006, 2007, 2008; Aldridge & Briggs 2009), et même comme des Gnathostomes (Donoghue et al. 2000, 2008).

D'après des travaux récents (Jones et al. 2012) on a pu établir grâce à l'usure et à la forme des différentes dents, qu'elles fonctionnaient par paires, l'une en face de l'autre, exerçant des mouvements de rotation pour broyer la nourriture. C'est grâce à ces mouvements que cette espèce aurait très bien pu se passer de mâchoires. Les éléments dentaires des Conodontes sont maintenant considérés comme étant composés de dentine et d'émail, comme les dents des vertébrés actuels.

Les fossiles de vertébrés qui montrent les premières traces d'os cellulaire sont les Ostracodermes (des vertébrés fossiles sans mâchoires, ou *agnathes*). Ce tissu est présent simultanément dans le dermosquelette et le neurocrâne (Donoghue et al., 2006) de nombreux fossiles. Email et émailloïde (un tissu semblable à de l'émail par sa localisation et sa minéralisation mais contenant du collagène) sont également observés dans le squelette de ces premiers vertébrés sans mâchoires et plus généralement dans le groupe de Ptéraspidomorphes (qui comprend les Ostracodermes) (Donoghue et al., 2006). Email et émailloïde sont considérés généralement comme deux types de tissus hyperminéralisés qui ont évolué indépendamment (Donoghue et Sansom, 2002).



(Dessins de plusieurs Ostracodermes *: Tauraspis, Hoelaspis, Trematapis, Zenaspis;* © 1997 Philippe Janvier)

Chez les vertébrés à mâchoires, les chondrichtyens (requins et raies) actuels et fossiles présentent un dermosquelette composé de denticules dermiques microscopiques, identiques à des dents et qui se développent à partir d'une papille dentaire. Chaque denticule est composé de dentine et d'os cellulaire ainsi que d'émail ou d'émailloïde selon les taxons (Donoghue et Sansom, 2002). Enfin, chez les Actinoptérygiens on peut trouver à la fois de l'émail "vrai" et

de l'émailloïde (Smith, 1992) et seulement de l'émail "vrai" chez les Sarcoptérygiens, recouvrant les dents et les écailles (Smith, 1989, 1992).

Au niveau des gènes…

La recherche des gènes impliqués dans la formation des tissus dentaires et osseux a permis la découverte d'une famille de gènes résultant de duplications génétiques successives et qui, malgré des divergences importantes, ont conservé des aspects fonctionnels et séquentiels similaires. Cette famille a été appelée SCPPs (pour Secretory Calcium-binding PhosphoProtein) suite aux travaux de Kawasaki et Weiss (2003). Les SCPPs (plus d'une vingtaine de membres chez l'homme) et qui représentent plus de 90% des protéines "minéralisantes", ont été créées par duplications successives à partir d'un ancêtre commun; les plus anciennes d'entre elles étaient probablement déjà présentes lors de l'apparition des structures minéralisées (os, dentine et émail) chez les tout premiers vertébrés, il y a au moins 450 millions d'années (*voir introduction*). Cependant, dès 2001 nous avons découvert que les gènes de minéralisation étaient probablement apparentés en montrant que l'amélogénine, le gène majeur de la mise en place de l'émail dentaire, était apparenté au gène SPARC (Secreted Protein, Acidic Cystein Rich; aussi appelée ostéonectine), qui joue un rôle vital dans la minéralisation de l'os (Delgado et al., 2001).

C'est donc sur cet axe de recherche que j'ai travaillé depuis plusieurs années dans le but de tenter de retracer l'histoire de ces protéines impliquées dans la minéralisation des tissus du squelette et des dents (dentine et émail) SCPPs.

Pour chacune de ces cibles, il s'agissait pour moi (1) d'identifier ses patrons évolutifs et de définir à quelle époque les caractéristiques propres ont été sélectionnées, (2) de mettre en évidence ses régions fonctionnellement très importantes (conservation de séquences), et (3) de définir ses liens de parenté avec les autres membres de la famille des SCPPs et son origine évolutive. Ces analyses évolutives ont permis, grâce à l'identification de résidus conservés pendant des centaines de millions d'années, de prédire certaines maladies génétiques associées à des mutations et de valider des mutations identifiées chez des patients.

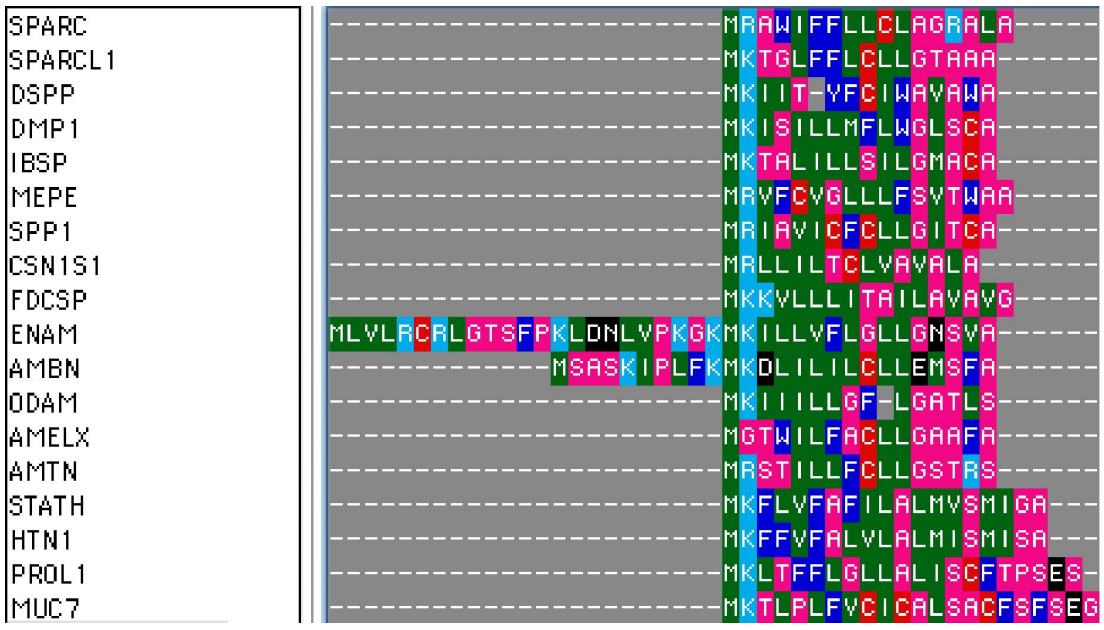
Dans la suite de ce manuscrit, j'ai choisi de présenter les connaissances actuelles sur cette famille génique en incluant les résultats de mes travaux de recherche.

1. Caractéristiques générales des gènes des SCPPs

Comme je l'ai déjà expliqué plus haut, des séries de duplications de gènes ont donné naissance à la famille des gènes des SCPPs. Même si les protéines de cette famille, avec quelques exceptions cependant, n'ont pratiquement pas d'homologie de séquences en dehors de leur peptide signal, les indices suivants montrent leur origine commune (voir Tableau 1).

Une caractéristique biochimique commune à tous les SCPPs : le peptide signal

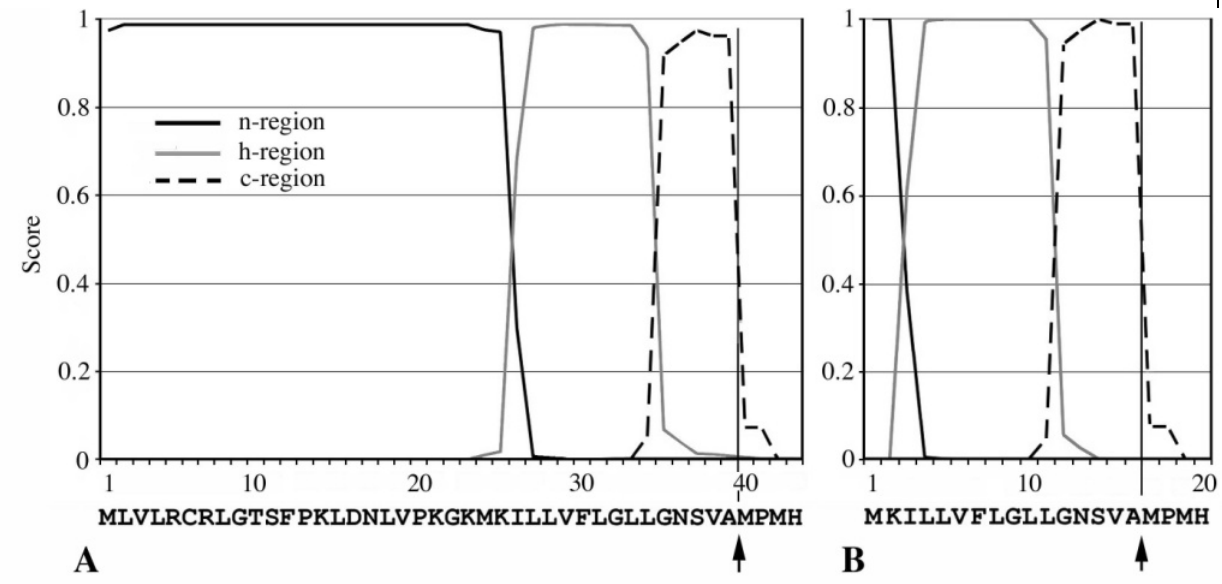
Tous les SCPPs ont un peptide signal (PS) et sont donc des protéines sécrétées. Le peptide signal est une séquence essentielle permettant, entre autres, aux protéines destinées à être secrétées à se retrouver dans une vésicule de sécrétion. Avant d'entrer dans la vésicule, le peptide signal est coupé et ne se retrouve pas à l'extérieur de la cellule, il ne joue donc aucun rôle direct dans la minéralisation. Par ailleurs, On observe une très bonne conservation des séquences des peptides signaux des SCPPs. C'est grâce à cette caractéristique que le peptide signal des SCPPs a été le principal outil utilisé par la recherche afin d'identifier de nouveaux gènes appartenant à ce groupe.



Peptides signaux des SCPPs et de SPARC et SPARCL1 chez l'homme (D'après la thèse de Doctorat de Al-Hashimi, 2010).

Histoire de la découverte d'un peptide signal inhabituel...

Tous les membres de la famille des SCPPs (22 chez l'homme) possèdent un peptide signal (PS) allant de 17 à 26 acides aminés, excepté ENAM, dans lequel le PS est exceptionnellement grand (39 résidus chez l'homme : Hu et al. 2000; 38 aa chez le porc : Hu et al. 1997). Un peptide signal aussi grand est rare chez les Eucaryotes; l'analyse comparative d'un grand nombre de PS d'Eucaryotes montre une longueur de 20-25 acides aminés en moyenne (von Heijne 1985; Martoglio et Dobberstein 1998). Ce grand PS est codé par l'exon 3 (dans lequel est situé le site d'initiation de la traduction (TIS) qui est bien conservé chez les Mammifères), et l'exon 4 qui code les résidus bien conservés (région riche en leucines) composant la région hydrophobe du peptide (*région-h*) exigée pour l'adressage des protéines et leur insertion dans les membranes (von Heijne 1985). Par contre, la grande *région n* (de fonction inconnue) codée par l'exon 3 (la région N-ter de ce PS) n'est pas soumise à des pressions sélectives fortes, bien que présente dans tous les ENAM des Mammifères. En comparant ce grand PS avec ceux des autres SCPPs, on constate que la *région-h* est homologue aux PS des autres SCPPs, alors que la région n est propre à ENAM.



Prédiction des peptides signaux (PS) de ENAM chez l'homme, codés soit par l'exon 3 et 4 (A) soit par l'exon 4 seul (B). Le site de coupure du peptide signal est indiqué par une flèche. Les deux PS de l'homme ont des scores élevés (0.973 et 1.000, respectivement) indiquant une probabilité identique d'utilisation.

Le peptide signal d'ENAM possède donc deux peptides signaux, un grand et un court, ce dernier étant inclus dans le premier. En conséquence, la région signal de ENAM possède également 2 sites d'initiation de la traduction (TIS), l'un codé par l'exon 3, l'autre codé par l'exon 4. Il est probable que le premier TIS, codé par l'exon 3 se soit rajouté au cours de l'évolution probablement par le phénomène « d'exon shuffling » (Patthy, 1999). Cet événement s'est produit chez un ancêtre commun aux Mammifères car les reptiles ne possèdent pas de grand PS. La conservation de ce nouvel exon codant, dans le gène de ENAM, signifie que la présence d'un grand PS (avec une grande région-*n*) a été sélectionnée positivement. L'utilisation de l'un ou l'autre des PS par ENAM se fait probablement par épissage alternatif impliquant l'apparition d'au moins deux isoformes de la protéine, l'un possédant le PS court (homologue à celui d'autres SCPPs) et l'autre le PS long. La présence de deux PS dans ENAM n'est pas unique et l'utilisation d'un PS alternatif est un dispositif commun à plusieurs protéines (Davis et al., 2006). Par exemple, l'interleukin-15 présente un PS court et un long indiquant qu'il existe des voies complexes pour le trafic intracellulaire de cette protéine (Kurys et al., 2000). En effet, généralement les propriétés du peptide signal sont en rapport avec l'adressage cellulaire qui est multiple : sécrétion de protéine à l'extérieur de la cellule, localisation dans le cytoplasme, adressage vers la mitochondrie, protéine transmembranaire (Hiss et al. 2008; Davis et al. 2006).

Malheureusement, les études sur les différents types de peptides signaux manquent pour déterminer le rôle exact des PS de ENAM à partir de la seule connaissance des acides aminés. Cependant, de la structure de ces deux PS nous pouvons tout de même conclure qu'ils sont employés uniquement pour la sécrétion de protéine dans la matrice extracellulaire et pas pour d'autres voies cellulaires. Par contre on peut supposer que les deux peptides signaux ont une efficacité d'exportation différente comme cela a déjà été montré pour le PS de la protéine Shrew-1 (Hiss et al. 2008). En effet, l'efficacité d'exportation semble être corrélée avec l'existence et l'intégrité de la zone séparant les régions -*n* et -*c*. Cette région appelée la « zone de transition » existe dans beaucoup de longs peptides signaux et elle est caractérisée par 4-7 acides aminés comprenant une glycine (G). Les auteurs cités ci-dessus ont montré que des mutations contrôlées, à l'intérieur de cette région, diminuent la quantité de protéine secrétée. Ainsi, l'utilisation alternative du PS court et du PS long chez ENAM pourrait moduler de manière très précise la quantité de protéines sécrétées par la cellule, ce phénomène se rajoutant aux contrôles habituels de l'expression des gènes. Ce système assez souple pourrait permettre un contrôle très précis de la minéralisation à travers la quantité de protéines versées dans la matrice. En effet, la diminution de la quantité de ENAM dans la matrice

extracellulaire durant les phases de transition et de maturation de l'émail pourrait faciliter le remplacement ordonné de la partie organique par les cristaux d'émail (Lu et al. 2008).

Le grand peptide signal de ENAM possède également une zone de transition au début de l'exon 4, et ce qui est très intéressant, c'est que cette région est variable chez les Mammifères sauf la glycine en position 22 qui est essentielle. Ceci suggère une régulation très complexe de la sécrétion de ENAM, sécrétion qui diffère en fonction des groupes de Mammifères. Chaque lignée de mammifère semble donc posséder ses propres caractéristiques de sécrétion. C'est en observant ces différences entre les PS de ENAM des Mammifères que nous avons remarqué une anomalie intéressante: le calcul de la probabilité du grand peptide signal du dauphin est très faible. Par conséquent, ou ce grand PS n'est pas fonctionnel ou il pourrait jouer une autre fonction. Nous pouvons également nous demander si cette différence comparée aux autres ENAMs des Mammifères pourrait être liée à la dentition homodonte (dents identiques) des cétacés odontocètes.

Autres caractéristiques communes…

Les SCCPs de la dentine et de l'os (= SIBLINGs) sont riches en acides aminés acides (Glu et Asp) tandis que les autres (SCPPs du lait, de la salive et de l'émail) sont riches en Pro et Gln mais pauvres en cystéines. Chez les Téléostéens, les SCPPs responsables de la formation de l'émailloïde sont riches en Pro et Gln. Cela confirmerait aussi l'hypothèse de la dérive phylogénétique de deux lignées différentes de SCPPs (chez les Actinoptérygiens et les Sarcoptérygiens).

La plupart des SCPPs ont des motifs Ser-Xaa-Glu (SXE) dans lesquels Ser est phosphorylée (Xaa représente n'importe quel acide aminé et l'asparagine ou la phosphosérine peut remplacer l'acide glutamique).

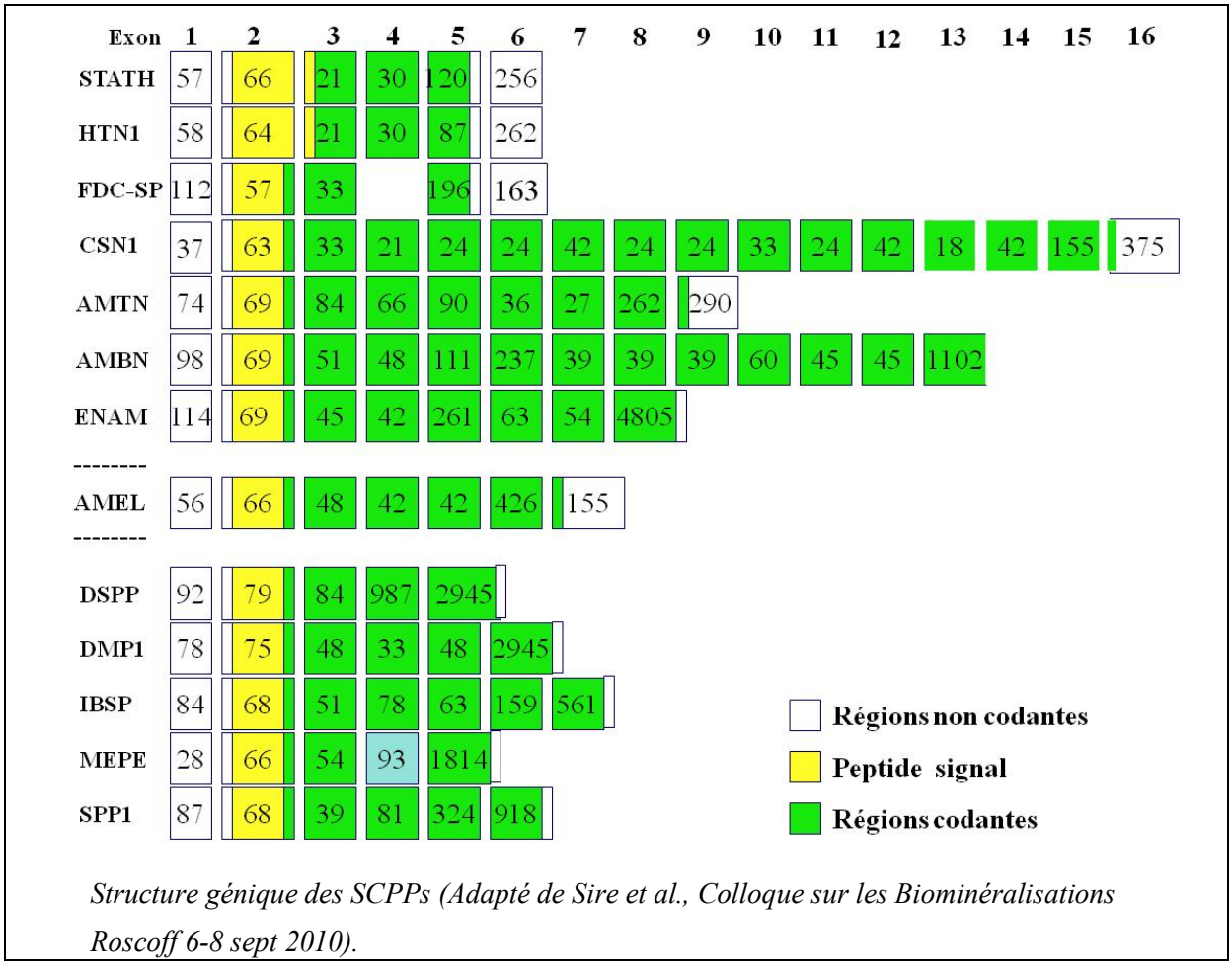
Bien que SPARC soit une protéine qui se lie au calcium, elle ne dispose pas de motif SXE alors que SPARCL1 (SPARC-like1) présente 11 à 14 de ces motifs (Tableau 1). Il semble que le motif SXE soit initialement développé chez SPARCL1 puis ait été transmis aux SCPPs au fur et à mesure de leur apparition (Kawasaki et al., 2004). Par ailleurs, le motif SXE des SCPPs semble se situer majoritairement dans la partie 3' de l'exon 3.

Enfin, ces protéines ainsi que leurs supposés ancêtres SPARC et SPARCL1 se lient aux ions de calcium via leurs acides aminés ayant un caractère acide et leurs motifs SXE. Pour cette raison, ces protéines sont appelées « phosphoprotéines secrétées se liant au calcium ».

Structure génétique similaire

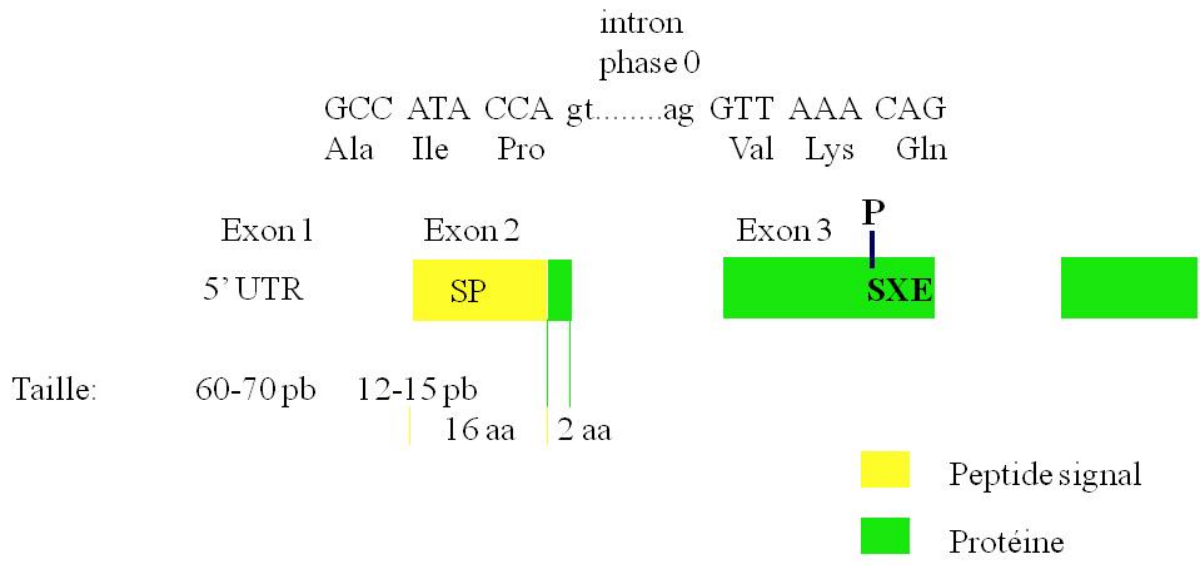
Une deuxième caractéristique commune concerne la structure génique des SCPPs.

Premièrement, les introns sont exclusivement en phase-0 ce qui signifie qu'ils sont situés entre deux codons adjacents au lieu d'altérer un codon (Kawasaki et Weiss, 2006). Cette distribution non arbitraire de phase d'introns est principalement due au phénomène de duplication d'exons (« exon suffling »).



Structure génique des SCPPs (Adapté de Sire et al., Colloque sur les Biominéralisations Roscoff 6-8 sept 2010).

Deuxièmement, la totalité de l'exon 1 et l'extrémité 5' de l'exon 2 constituent la région 5' non traduite (5'UTR), cette caractéristique est commune à tous les SCPPs et à SPARC et SPARCL1.



Points communs de SCPPs au niveau des exons.

(Adapté de Sire et al., Colloque sur les Biominéralisations, Roscoff 6-8 sept 2010).

Chez l'homme tous les gènes des SCPPS sont situés sur le chromosome 4 et regroupés en deux régions, l'une pour les gènes des SCPPs de la dentine et de l'os et l'autre pour les gènes des SCPPS de l'émail, de la salive et du lait. La seule exception est *AMEL* qui est situé sur les chromosomes sexuels chez les Mammifères placentaires.

SPARCL1 est situé sur le chromosome 4 ce qui soutient l'hypothèse que les gènes des SCPPs sont des duplications en tandem à partir de SPARCL1.

Caractéristiques fonctionnelles similaires

Toutes les SCPPs sont impliquées dans des processus de minéralisation soit dans l'os et la dentine, soit dans l'émail et l'émailloïde (SCPPs des Téléostéens), excepté les SCPPs de la salive et du lait qui conservent toutefois la fonction principale de se lier au calcium (Ca-binding) mais restent fortement impliquées dans la régulation du calcium au niveau de la salive et du lait. Quelques éléments de ce groupe ont complètement changé de fonction comme l'histatine (HTN3) qui agit comme agent antibactérien.

		Peptides signaux	Motif SXE	Chr. 4 (homme)	Riche en Pro/Qln	Acidité (Glu/Asp)	Motif RGD	Introns en phase 0
SPARC		✓	✓	×(1)	×	✓ (Glu)	×	✓
SPARC L1		✓	11 à 14	✓	✓	✓ (Glu)	×	✓
Os et Dentine	**DSPP**	✓	✓	✓	×	✓	✓	✓
	DMP1	✓	✓	✓	×	✓	✓	✓
	IBSP	✓	✓	✓	×	✓	✓	✓
	MEPE	✓	✓	✓	×	×	✓	✓
	SPP1	✓	✓	✓	×	✓	✓	✓
Email (EMPs)	**AMEL**	✓	✓	×(2)	✓	×	×	✓
	AMBN	✓	✓	✓	✓	×	×	✓
	ENAM	✓	✓	✓	✓	×	✓	✓
	AMTN	✓	×	✓	✓	×	×	✓
	ODAM	✓	✓	✓	✓	×	×	✓
Salive et Lait	**CSN1S1**	✓	✓	✓	✓	×	×	✓
	CSN2	✓	✓	✓	✓	×	×	✓
	CSN3	✓	✓	✓	✓	×	×	✓
	FDCSP	✓	✓	✓	✓	×	×	✓
	STATH	✓	✓	✓	✓	×	×	✓
	HTN1	✓	✓	✓	✓	×	×	✓
	HTN3	✓	✓	✓	×	×	×	✓
	PROL1	✓	✓	✓	×	×	×	✓
	PROL2	✓	✓	✓	✓	×	×	✓
	PROL3	✓	✓	✓	✓	×	×	✓
	MUC7	✓	✓	✓	✓	×	×	✓

Tableau 1- caractéristiques générales des SCCPs. (1) chez l'homme, le gène SPARC est situé sur le chromosome 5; (2) chez l'homme, le gène de l'amélogénine est situé sur les chromosomes sexuels X et Y. (D'après la thèse de Al-Hashimi, 2010).

2. Origine et évolution de la famille des SCPPs

Il y a deux hypothèses concernant l'origine des SCPPs qui diffèrent sur la date exacte du premier événement de duplication (Fig. 2). Selon nos travaux antérieurs (Delgado et al. 2001), et d'après l'observation d'une homologie de séquence nucléotidique entre l'exon 2 d'*AMEL*, de *SPARC* (ostéonectine) et *SPARCL1*, AMEL dériverait de SPARC/SPARC-L1 et leur date de divergence se situerait il y a environ 600 Ma. Ces travaux ont été complétés par la suite (Sire et al., 2007) et il a été montré qu'en fait *AMEL* (et *AMBN)* avaient été créées par duplication à partir d'un gène ancestral, apparenté à *ENAM*.

Kawasaki et Weiss (2003) ont proposé une autre hypothèse selon laquelle le tout premier événement, la duplication de *SPARC* en *SPARCL1*, est estimé à environ 531 Ma, soit bien après la date de 600 Ma estimée par nous (Delgado et al., 2001; Sire et al., 2007).

SPARC est probablement l'une des premières protéines impliquées dans la minéralisation car elle contient un large domaine acide (Kawasaki et al., 2006). Cette hypothèse est en accord avec l'observation de l'expression de SPARC dans les os et écailles de poissons Téléostéens.

Chez l'homme, le gène *SPARC* est situé sur le chromosome 5q31.3-q32 tandis que *SPARCL1* est localisé sur le chromosome 4q22.1. Chez les autres Mammifères, les oiseaux et les Téléostéens, ces deux gènes sont également situés sur deux chromosomes différents (Kawasaki et Weiss, 2006). Les protéines de SPARC et SPARCL1 se composent d'un peptide signal et de trois modules fonctionnels appelés domaine-I (acide), domaine-II ("follistatin like"), domaine-III (extracellulaire se liant au calcium). Le domaine 1 possède la structure caractéristique des gènes de la famille des SCPPs.

SPARC et *SPARCL1* ont divergé après une duplication d'une grande région du génome (WGD ou Wide Genome Duplication) qui a eu lieu chez l'un des tous premiers vertébrés à mâchoires. On pense que *SPARCL1* a surtout conservé la région codant le domaine 1 de *SPARC* (Fig. 2) qui est devenu plus long et plus acide. L'apparition de *SPARCL1* coïncide avec celle du squelette minéralisé (Kawasaki et al., 2007).

Le domaine I est radicalement différent entre SPARC et SPARCL1. Chez l'homme, celui de SPARC est constitué de 52 acides aminés tandis que celui de SPARCL1 contient 414 acides aminés parmi lesquels 339 résidus codés par un grand exon présent seulement chez *SPARCL1*. De plus, il est important de souligner que le domaine 1 de SPARC contient 18 acides aminés acides mais aucun acide aminé basique alors que celui de SPARCL1 possède 104 acides aminés acides et 55 acides aminés basiques. On retrouve ces différences dans SPARC et SPARCL1 de tous les Téléostéens et les Tétrapodes étudiés jusqu'à aujourd'hui. La conservation phylogénétique de ces caractères biochimiques suggère que le taux des acides aminés acides et basiques et le regroupement des résidus acides, plutôt que la séquence primaire, sont importants pour la cristallisation de l'hydroxyapatite (Kawasaki et al., 2004). Un autre

point qui doit être souligné dans ce contexte est que le domaine I n'est pas incorporé dans une structure globulaire rigide et ne contient pas de cystéine comme c'est le cas dans la plupart des SCPPs qui contiennent peu voire aucune cystéine et ne peuvent donc pas stabiliser des structures globulaires fixes à travers un ou plusieurs ponts disulfures intramoléculaires.

Le deuxième événement dans l'évolution des SCPPs est l'apparition des premiers gènes des SCPPs à la suite de duplications en tandem à partir de *SPARCL1,* avant la divergence des Actinoptérygiens et des Sarcoptérygiens. Ces nouveaux gènes n'ont pas conservé les exons codant les domaines II et III qu'on retrouve chez SPARC et SPARCL1 (Fig. 1).

2.1 Dérive phénogénétique : un point crucial dans l'évolution des SCPPs

Le troisième événement dans l'évolution des SCPPs est caractérisé par une dérive phénogénétique qui peut être définie comme une modification du génotype alors que durant cette période, le phénotype (la minéralisation des dents) est resté stable (Kawasaki et al., 2005). En effet, les gènes des premières SCPPs ont subi une duplication indépendante et parallèle dans les deux grandes lignées d'Ostéichtyens. Elle a donné naissance à deux lignées spécifiques de gènes des SCPP, alors que l'ancêtre commun du gène a été supprimé du génome.

Ces processus de naissance et de mort, ainsi que l'évolution du nombre d'exons et de la taille des SCPPs ont engendré des SCPPs qui ont évolué indépendamment chez les Actinoptérygiens et les Sarcoptérygiens (Fig. 2). On peut encore retrouver le gène SPP1 dans ces deux lignées alors qu'il a été, semble t-il, perdu chez les amphibiens et le fugu (Kawasaki et Weiss, 2008).

Néanmoins, les SCPPs impliquées dans le développement des dents démontrent l'existence d'aspects biochimiques similaires entre les protéines de fonctions correspondantes. À titre d'exemple, l'émail des Tétrapodes est un tissu dur et spécialisé à caractère unique formé à partir de SCPPs distinctes sécrétées par les améloblastes, alors que l'émailloïde des Téléostéens se développe à partir du collagène 1 sécrété à la fois par les améloblastes et les odontoblastes.

2.2 Les SCPPs des Actinoptérygiens

Bien que je n'aie jamais eu encore l'occasion de travailler sur les gènes des SCPPs dans cette lignée de vertébrés, il était important de mentionner leur existence.

Les gènes des SCPPS sont connus uniquement chez quelques Téléostéens. Ils constituent un groupe indépendant des SCPPs des Tétrapodes mais ils ont cependant des traits communs avec ceux-ci. Dans ce groupe, on trouve huit gènes de SCPPs chez le fugu (SPP1, SCPP1, SCPP2, SCPP3A, SCPP3B, SCPP3C, SCPP4, SCPP5) et quatre chez le poisson-zèbre (SPP1, SCPP1, SCPP2, SCPP5) (Kawasaki et al., 2008).

Alors que ces SCPPs ont évolué indépendamment de ceux des Tétrapodes et qu'il n'y a aucune homologie de séquences entre eux, ils partagent néanmoins quelques caractéristiques communes à tous les gènes des SCPPs (5' UTR similaires, introns en phase 0, motifs SXE et région riche en proline et glutamine).

Certaines SCPPs des Téléostéens sont impliquées dans la formation de l'émailloïde comme le sont les protéines de l'émail des Tétrapodes dans la formation de l'émail.

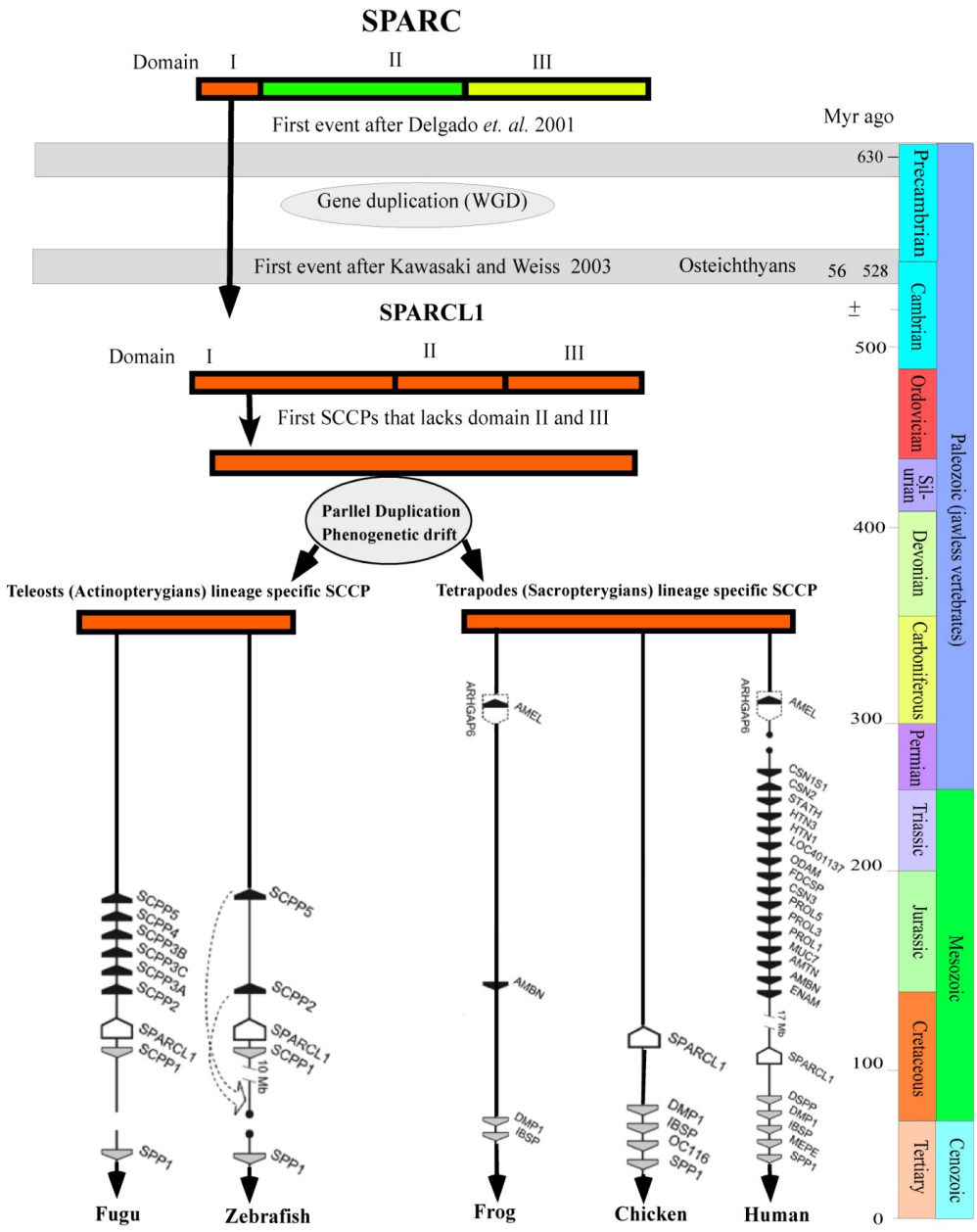


Figure 1. Evolution des SCPPs. A partir de SPARC, des événements de duplications de gènes ou de génomes ont amené à la formation de SPARCL1, puis de l'ancêtre des SCPPs puis de tous les SCPPs actuels. Ces derniers événements se sont déroulés indépendamment dans la lignée des Actinoptérygiens et dans celle des Sarcoptérygiens. (D'après la thèse de Al-Hashimi, 2010).

2.3 Les SCPPs des Sarcoptérygiens

Il n'a pas été trouvé, à ce jour, de SCPPs chez les deux lignées basales des Sarcoptérygiens, les cœlacanthes ou les dipneustes. Les connaissances se limitent donc aux Tétrapodes et encore, surtout chez les Mammifères grâce au grand nombre de génomes séquencés dans cette lignée. Chez les Mammifères justement, on peut identifier trois sous-groupes principaux de SCPPs:

1) les SCPPs impliquées dans la minéralisation de la dentine et de l'os, encore appelées SIBLINGs (pour Small Integrin-Binding Ligand N-linked Glycoproteins).

2) les SCPPs impliquées dans la minéralisation de l'émail, ou EMPs (pour Enamel Matrix Proteins).

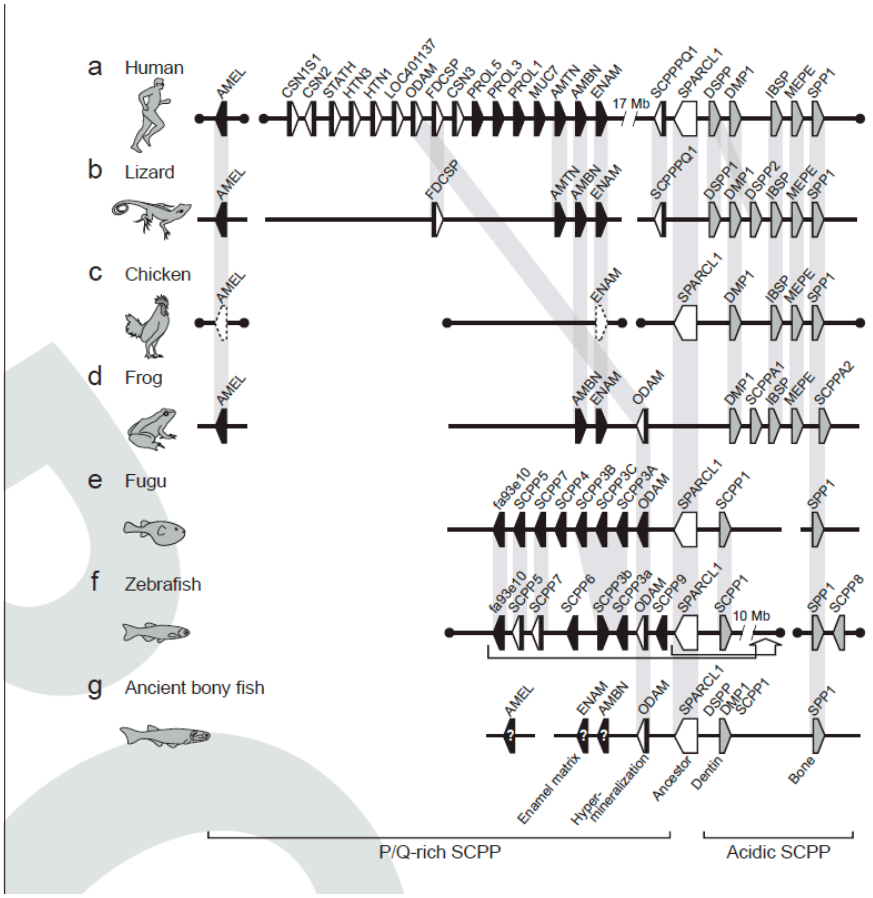
3) les SCPPs constituant les protéines du lait et de la salive. On retrouve les deux premiers sous-groupes chez les Sauropsides et les Amphibiens.

2.4 Evolution des SCPPs

On estime que les SCPPs impliquées dans la minéralisation de la dentine et de l'os constituent le groupe qui s'est formé le premier au cours de l'évolution, avant celui des EMPs et des protéines du lait et de la salive. Cette hypothèse est consolidée par la présence du gène de la SPP1 (Bone sialoprotein-1) aussi bien dans le génome des Téléostéens que dans celui des Tétrapodes, alors que l'on n'a pas retrouvé de gène d'EMPs chez les Téléostéens. Pourtant, les travaux menés chez le Fugu (Kawasaki et al., 2008), ont permis d'analyser tous les gènes exprimés au niveau des dents mais ceux-ci sont les gènes SCPPs décrits au chapitre 1.2 qui n'ont aucune homologie avec les gènes des EMPs des Tétrapodes.

L'apparition des protéines du lait et de la salive constituent une innovation majeure chez les Mammifères rendue possible grâce à la création d'un nouveau groupe de SCPPs ayant de nouvelles fonctions. Les caséines du lait fournissent du phosphate de calcium pour les nouveaux-nés des Mammifères, ce qui aide le développement de leurs os et de leurs dents. Les protéines de la salive protègent les dents en régulant la précipitation des sels de phosphate de calcium à la surface de l'émail. Ce nouveau groupe de SCPPs du lait et de la salive a probablement évolué à partir de gènes de SCPPs, et l'hypothèse la plus récente désigne ODAM comme le gène le plus proche de ce groupe de SCPPs (Kawasaki et al., 2011). Si les caséines sont connues chez tous les Mammifères, les protéines de la salive n'ont été identifiées que chez quelques Mammifères. Par exemple, un pseudogène de stathérine (STATH) mais pas d'histatine (HTN) a été identifié dans le génome de la souris. L'arbre phylogénétique basé sur les séquences du dernier exon de ces gènes montre que STATH et HTN proviennent des caséines CSN1S2 (Kawasaki et al., 2003). Le gène STATH aurait été créé à partir de *CSN1S2* avant la divergence des rongeurs, il y a 96 millions d'années (Kawasaki et al., 2003). La duplication de HTN1-HTN3 a eu lieu

entre 15 à 30 millions d'années d'après l'analyse des taux de substitution (Fig. 2). Tous ces éléments suggèrent une origine récente des protéines salivaires.



Les SCPPs dans différentes lignées de vertébrés. D'après Kawasaki, 2011.

Finalement, le bilan des données disponibles sur la famille des SCPPs, permet de reconstituer le scénario suivant : actuellement, nous pensons que les SCPPs se sont séparées rapidement en deux groupes distincts à partir d'un gène ancestral des SCPPs, donnant le groupe des protéines de l'os et de la salive (Siblings), d'une part et le groupe des protéines de l'émail (EMPs) d'autre part. C'est à partir d'un gène de l'émail identifié comme ODAM que sont apparues les protéines du lait et de la salive. Ce scénario reste à préciser. En effet, bien que les relations au sein des EMPs soient connues aujourd'hui suite à nos travaux (voir partie suivante), les relations au sein des Siblings sont largement méconnues. De plus, de nouveaux gènes appartenant à cette famille continuent d'être découverts, comme SCPPPQ1 et FDCSP par exemple. On suppose qu'ils pourraient provenir de duplications de ODAM, ce qui pourrait encore modifier nos phylogénies.

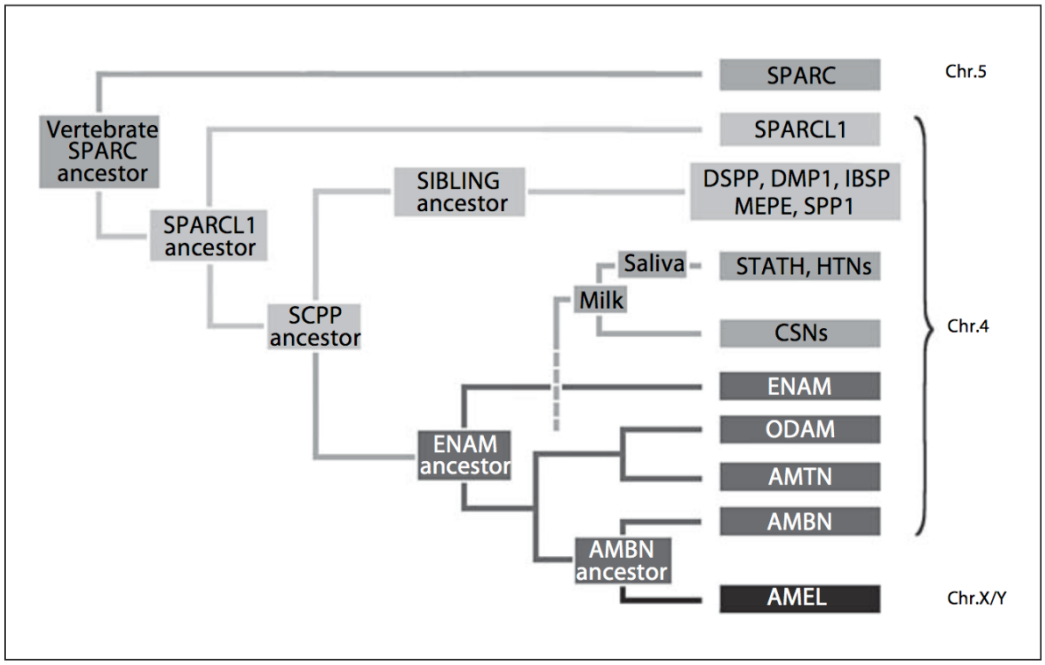


Fig. 2. Relations de parenté des SCPPs. D'après Sire et al., 2007.

3. Recherches sur les protéines de l'émail

3.1. AMEL

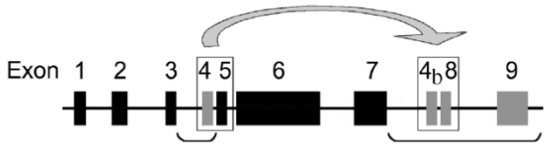
Parmi les SCPPs, un groupe de protéines m'a particulièrement intéressé: les protéines de l'émail, en partie à cause de leur intérêt médical. La première à avoir été étudiée est l'amélogénine (AMEL), la protéine majeure de ce tissu. Son abondance dans la matrice amélaire (plus de 90 % des protéines) en faisait le sujet d'étude le plus intéressant. Toutefois, peu de choses étaient connues sur le rôle d'AMEL et, de manière plus générale, sur les SCPPs, en particulier leur interaction avec les cristaux de minéral et leur rôle dans l'orientation et le développement des cristaux en formation. Pour cette raison, différentes études ont vu le jour dans l'équipe Evolution et Développement du Squelette (EDS) dont la ligne directrice majeure était de pouvoir comparer les séquences protéiques des SCPPs obtenues dans de nombreuses lignées animales afin de mieux identifier les régions fonctionnelles. Les premiers travaux que j'ai menés à partir de ma thèse de doctorat sur l'amélogénine, ont été très fructueux (Delgado et al., 2001, 2005, 2006, 2007; Sire et al., 2005, 2006) et l'idée d'étendre les connaissances de cette protéine aux reptiles a permis d'apprécier les variations de la séquence de cette protéine au cours de l'évolution des Amniotes, sur une période de plus de 300 Ma. L'une des retombées médicales de ces travaux a été d'établir une carte des acides aminés susceptibles de conduire à une amélogenèse imparfaite (AIH1) en cas de mutation (Delgado et al., 2005). Ce résultat a été obtenu grâce à un grand nombre de séquences d'AMEL représentatives des lignées de Mammifères et de reptiles; cette base de données nous a permis, entre

autres, d'identifier les résidus conservés pendant 300 millions d'années en raison d'une importante pression sélective, preuve du rôle majeur que jouent certainement ces acides aminés dans la fonction de la protéine. La relation entre pression sélective et fonction cruciale de certains résidus est d'ailleurs bien confortée d'une part, par une étude montrant que 95% des substitutions d'acides aminés conduisant à une maladie génétique ont lieu sur des résidus conservés durant 100 million d'années d'évolution (Subramanian et Kumar, 2006) et d'autre part, par les quelques cas connus d'AIH1 concernant nos gènes dans lesquels les acides aminés substitués impliqués dans cette maladie font effectivement partie des résidus conservés au cours de l'évolution. Cette étude évolutive sur l'amélogénine (Delgado et al., 2005) a montré que plus de 30 acides aminés étaient de bons candidats pour conduire à une AIH1 s'ils étaient mutés.

Ce qui m'a encouragé à continuer dans cette voie est que toutes les mutations conduisant à une amélogenèse imparfaite découvertes par la médecine dentaire depuis ces recherches en 2005 sont situées dans les régions mises en évidence dans cet article.

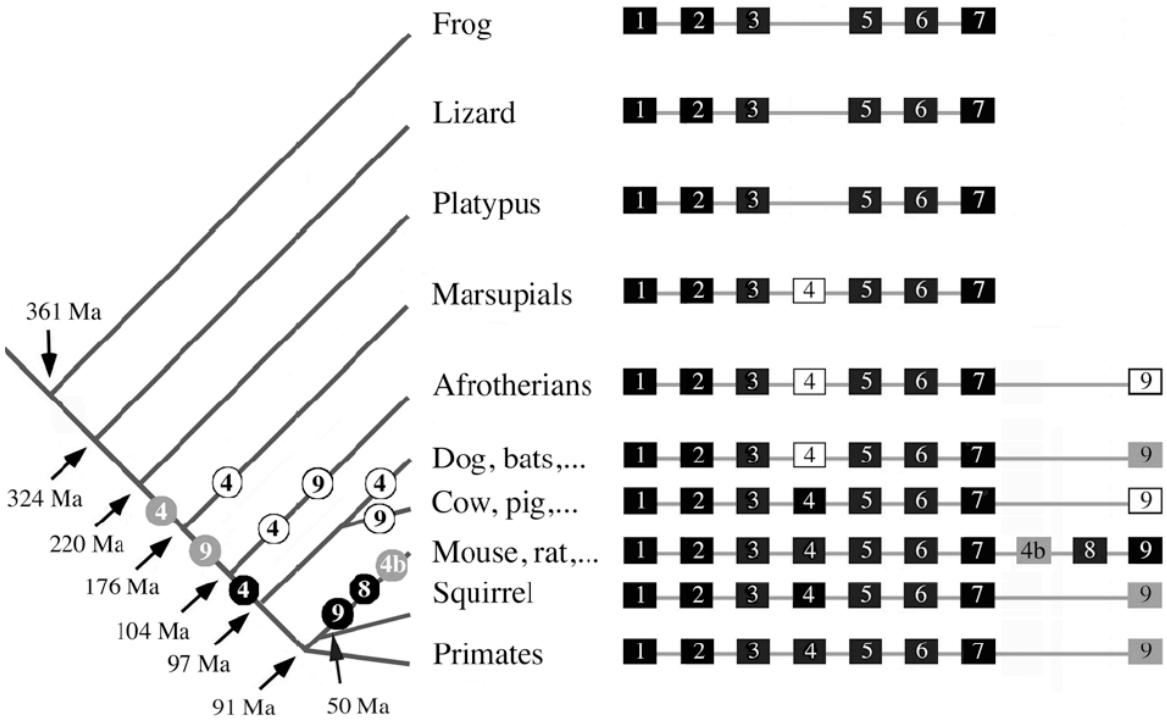
En dehors du rôle de AMEL dans la minéralisation, j'ai découvert au cours de mes recherches que AMEL est un marqueur phylogénétique intéressant pour reconstituer les relations de parenté chez les Mammifères (Delgado et al., 2007). Cette découverte inattendue m'a permis de revenir, le temps d'un article, à ma spécialité de Master de génétique de l'évolution : la phylogénie moléculaire.

Je suis ensuite passé à d'autres sujets d'étude, d'autres gènes de minéralisation, cependant récemment, j'ai eu l'occasion de revenir à mes « premières amours ». En effet, il restait certaines questions en suspens à la fin de ma thèse auxquelles j'ai toujours souhaité répondre. Parmi celles-ci, la présence d'un exon 4 chez certains Mammifères seulement et la présence de 2 exons terminaux (exon 8 et 9) offrant une terminaison alternative aux ARNm de l'homme et de la souris étaient assez intriguant. Grâce à une étude menée conjointement avec des collègues de l'Université de Californie à San Francisco, et du Laboratoire « Différenciation de Cellules Souches et Prions », de l'Université Paris Descartes (Sire et al., 2012), il a été possible de découvrir l'origine de ces exons, de montrer par exemple qu'il y a eu une duplication des exons 4 et 5 pour donner un exon 4b et l'exon 8 (voir figure ci-dessous).



(Sire et al., 2012)

De plus, notre étude a permis de reconstituer le scénario de l'apparition de ces différents exons chez les Mammifères. Ainsi, c'est la genèse de nouvelles régions codantes, le phénomène de naissance et de mort des exons, qui est visible à travers l'étude de l'évolution des exons d'AMEL. Reste à connaitre plus précisément le rôle de ces exons supplémentaires chez les Mammifères qui les ont gardés fonctionnels. Ces exons font partie de transcrits minoritaires ce qui a rendu la connaissance de leur existence tardive. Cependant, ils ont trouvé leur place au sein du processus de minéralisation de l'émail, il convient donc de savoir à quel moment et pourquoi ils sont exprimés.



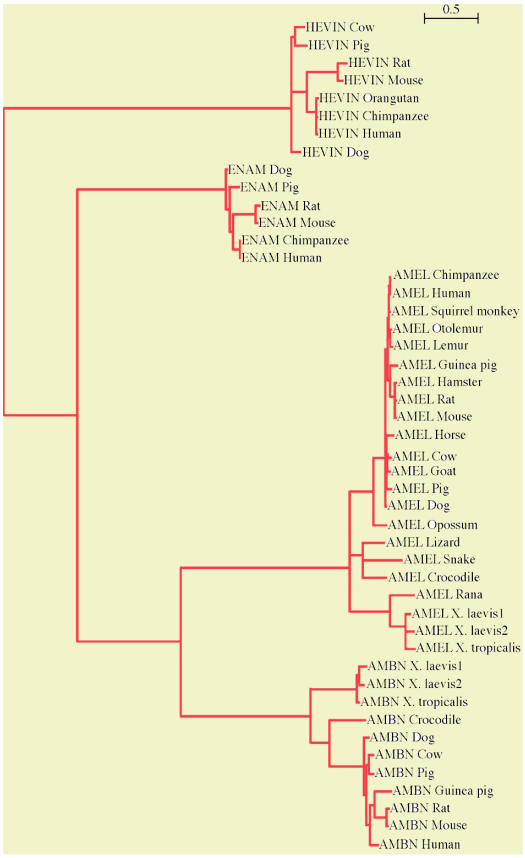
Evolution des exons 4, 8 et 9 chez les Mammifères. Les exons en noir sont codants, ceux en blanc sont des pseudo-exons et les exons en gris sont supposés codants. (Sire et al., 2012)

3.2. ENAM

La découverte des relations de parenté entre l'amélogénine et les autres protéines de l'émail (Enamel Matrix Proteins: EMPs), améloblastine (AMBN) et énaméline (ENAM), m'a amené à considérer que la seule étude d'AMEL n'était pas suffisante pour comprendre la biologie de l'émail et les interactions possibles de cette protéine avec les autres EMPs et, également, d'autres SCPPs (Sire et al., 2007).

Je me suis alors intéressé à ENAM car cette protéine est la plus ancienne des EMPs. Plus exactement, les trois EMPs connues à ce jour, AMEL, AMBN et ENAM, proviennent de la duplication du gène ancestral de d'ENAM (Sire et al., 2005, 2006, 2007). Comme je l'ai expliqué dans le chapitre précédent, nos travaux ont suggéré que l'ancêtre des EMPs pourrait être apparu à la suite de la duplication

de SPARC (anciennement connue sous le nom d'ostéonectine), il y a plus de 600 Ma (Delgado et al., 2001; Sire et al., 2007).



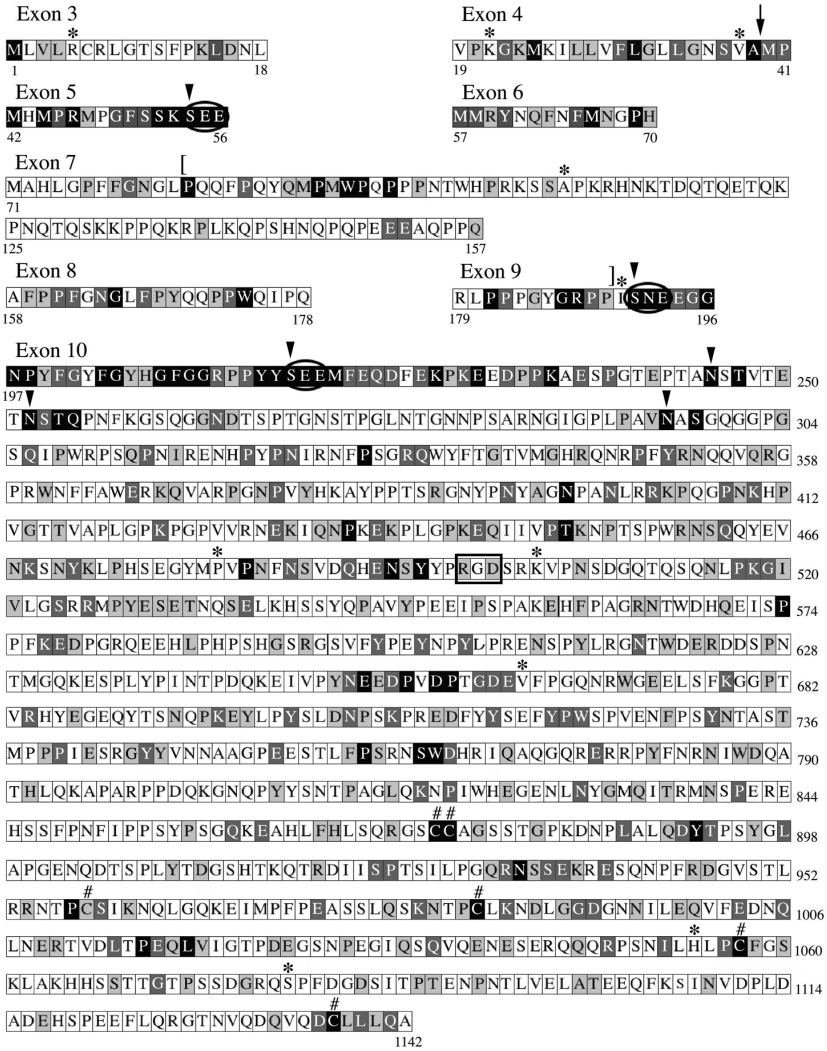
Relations phylogénétiques entre les gènes des protéines de l'émail ; D'après Sire et al., 2007.

Cette place particulière d'ENAM à l'origine des EMPs des vertébrés amenait plusieurs questions et la recherche des réponses a débouché sur une thèse de doctorat que j'ai co-encadrée et qui s'est terminée en 2010 (Al-Hashimi, 2010). Cette thèse a apporté de nombreuses connaissances sur ce gène crucial dans la minéralisation des dents. Quand nous avons commencé à étudier ENAM les données bibliographiques indiquaient que son rôle était incertain.

L'étude comparative des séquences d'ENAM de 36 Mammifères dont les génomes sont disponibles dans les banques de données a permis de mesurer les pressions de sélection au niveau des acides aminés et de détecter les régions conservées.

La découverte la plus importante résultant de cette analyse est la mise en évidence du rôle central joué par le fragment de 32 kDa d'ENAM. En effet, comme toutes les protéines de la matrice de l'émail, ENAM subit des coupures protéolytiques, au court de son cycle, qui génèrent des polypeptides de différentes tailles. Ainsi, la littérature décrit pour ENAM des peptides de 155, 142, 89, 34, 32 et 25-kDa (Fukae and Tanabe 1985, 1987; Uchida et al. 1991; Tanabe et al. 1994; Fukae et al. 1996). Malheureusement, il est toujours difficile de connaître le rôle exact de ces polypeptides. D'autant que parfois, ces fragments ne sont que des résidus de coupure qui n'ont en réalité pas de rôles actifs dans la

minéralisation. Notre travail a permis de montrer que le fragment de 32 kDa d'ENAM dont l'existence est connue depuis longtemps, mais dont le rôle était encore incertain, est la "clef de voûte" dans la minéralisation de l'émail. Sans le fragment de 32 kDa, il n'y a pas d'initiation de la minéralisation, donc pas d'émail du tout. A partir de cette dernière découverte et en compilant nos résultats avec les données de la littérature, nous avons pu proposer un nouveau modèle de distribution des protéines au cours de la minéralisation de la matrice amélaire.



Séquence humaine de ENAM. Elle est composée de 1142 résidus incluant le peptide signal supposé codé par les exons 3 et 4. La flèche indique le site de clivage du peptide signal. Les positions conservées, i.e., sujettes à la sélection purifiante chez les Mammifères (durant 160 million d'années d'évolution), sont indiquées en noir et gris foncé (plus forte sélection) et en gris clair (sélection plus faible). (D'après Al-Hashimi et al., 2009).

Juste après, nous avons étendu cette étude aux reptiles et aux amphibiens (Al-Hashimi et al., 2010) ce qui a permis de montrer que le peptide signal original de ENAM découvert précédemment est bien une innovation apparue chez les Mammifères. Ce travail a aussi permis de montrer que le poulet possède les traces encore visibles dans son génome de la présence du gène de l'émail ENAM mais sous forme d'un

pseudogène désormais non fonctionnel. Cette découverte, n'était pas très surprenante puisque nous avions déjà découvert le pseudogène de l'amélogénine (autre gène majeur de l'émail) il y a quelques années (Sire et al., 2008). Ces travaux avaient fait suite à certaines expériences ambigües qui avait permis de faire développer des dents ou des bourgeons de dents chez le poulet sans prouver formellement la présence d'émail (Kollar et Fisher, 1980; Chen et al., 2000; Mitsiadis et al., 2003). Certains auteurs, se demandaient alors si les oiseaux n'auraient pas gardé potentiellement les outils génétiques pour fabriquer des dents. L'invalidation des gènes dentaires prouve ainsi le contraire.

4. Recherches sur les protéines de la dentine et de l'os

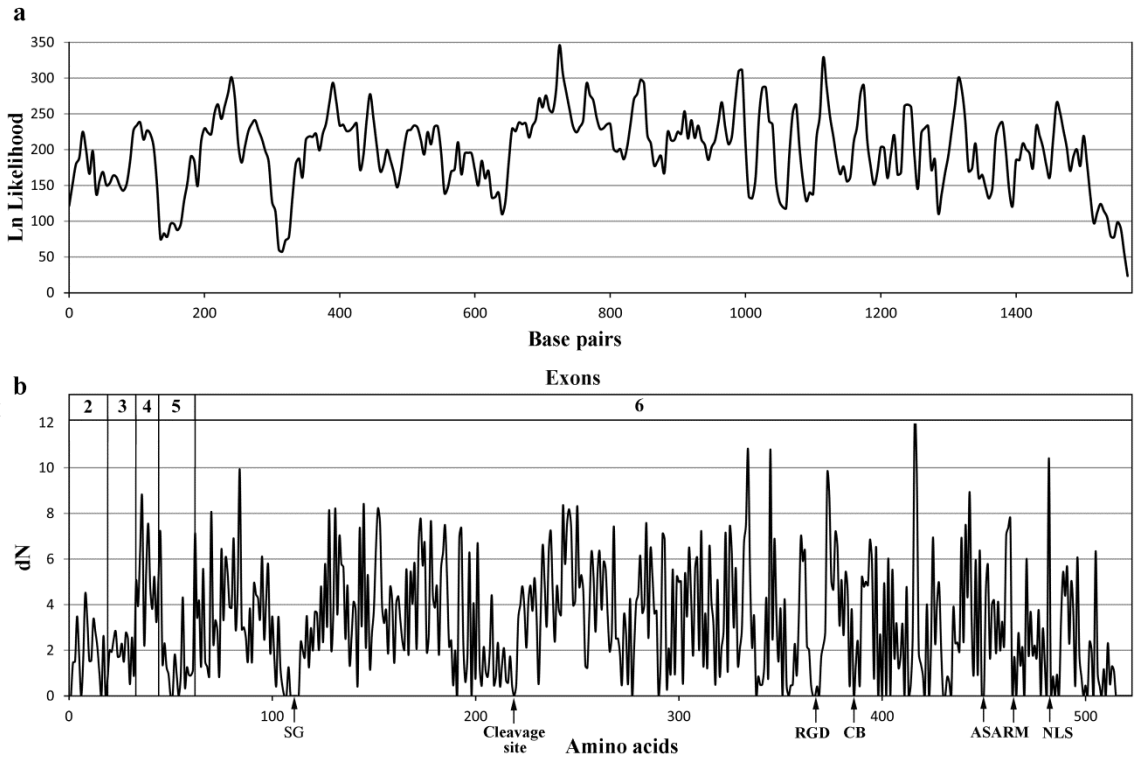
4.1. DMP1

Ce travail fait partie d'une des dernières thèses en date dans notre laboratoire, réalisée par Jérémie Silvent. Cet encadrement a abouti à l'écriture d'un article qui vient d'être soumis à une revue, sous le titre « The dentin matrix acidic phosphoprotein 1 (DMP1) in the light of mammalian évolution » (par Jérémie Silvent, Jean-Yves Sire et Sidney Delgado).

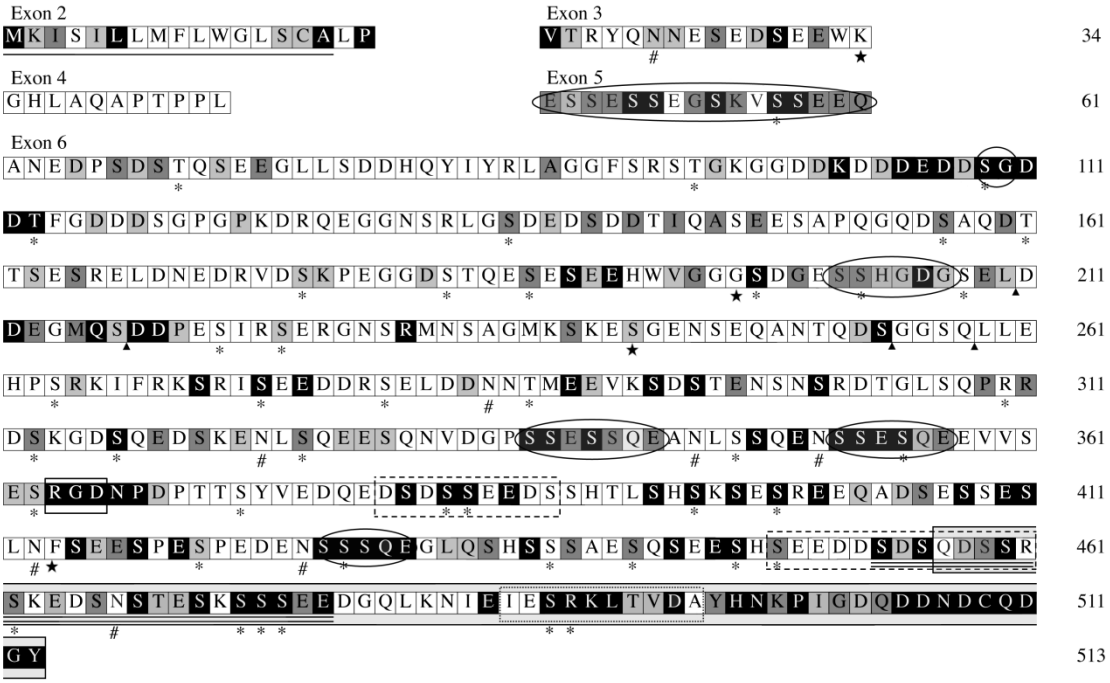
DMP1 m'a donné l'occasion de travailler pour la première fois sur une des protéines de la dentine et de l'os (Siblings). Parmi toutes les Siblings, DMP1 semblait une cible d'étude très intéressante pour plusieurs raisons. En premier lieu à cause de son importance dans la minéralisation de la dentine et de l'os. Cette importance a été mise en évidence par l'existence de certaines maladies génétiques chez l'homme et par des expériences d'invalidation du gène chez la souris. Les expériences de Knockout chez la souris montrent des défauts de minéralisation de l'os et de la dentine (Ye et al. 2005; Ye 2004). Chez l'homme, des mutations de DMP1 ont été associées avec divers syndromes comme une forme autosomale récessive de rachitisme hypophosphatémique (ARHR) (Feng et al. 2006; Lorenz-Depiereux et al. 2006; Mäkitie et al. 2010; Turan et al. 2010), de l'ostéomalacie (Feng et al. 2006), ou encore des défauts de maturation de la dentine (Ye 2004). Une autre raison de s'intéresser à DMP1 est la diversité de ses fonctions. En plus de son rôle dans la nucléation et la régulation de la minéralisation au niveau de la matrice de collagène, DMP1 peut agir dans l'ostéogenèse comme facteur de transcription en activant certains gènes spécifiques des cellules osseuses (ostéoblastes) comme par exemple l'ostéocalcine (Narayanan et al. 2003) ou DSPP (Lu et al. 2007; Ye 2004). Cette fonction est possible grâce à la présence d'un motif (NLS) permettant l'importation de DMP1 dans le noyau. Récemment il a aussi été suggéré que DMP1 pourrait jouer un rôle dans l'angiogenèse (Pirotte et al. 2011) ou agir dans le turnover des protéines de la matrice extracellulaire endommagées par oxydation en formant un complexe avec MMP-9 (Ogbureke and Fisher 2006; Ogbureke and Fisher 2005; Ogbureke and Fisher 2004). Une telle palette d'action soulevait les questions suivantes : Où se localisaient ces différentes fonctions sur la protéine ? Et comment les motifs potentiellement associés à ces fonctions ont-ils évolué ?

Nous avons donc étudié l'évolution de DMP1 chez les Mammifères dans le but 1) d'identifier des acides aminés ou des régions qui pourraient jouer un rôle important pour la fonction de DMP1, 2) de définir de manière plus précise les régions fonctionnelles précédemment décrites et 3) d'établir l'emplacement potentiel de mutations associées à des maladies génétiques humaines.

En plus des sites déjà décrits dans la littérature et qui sont, comme prédit, conservés durant 200 Ma d'évolution (voir les deux figures ci-dessous), nous avons effectivement découvert des motifs inconnus dont la fonction reste à définir, ce qui n'est guère surprenant étant donné la gamme étendue d'actions de DMP1. Le plus frappant dans cette étude est d'avoir pu constater que les limites de ces motifs protéiques précédemment décrites dans la littérature ont été redéfinies par nos calculs de conservation. Lorsque l'on étudie de plus près les articles précédents concernant DMP1, on constate que les limites des motifs sont parfois définies par des techniques consistant à couper la protéine en de nombreux morceaux pour ensuite trier les peptides en fonction de leur rôle. Dans ces conditions, les limites des motifs sont confondues avec celles des peptides qui les portent. Parfois, certains motifs sont définis par homologie avec des motifs identifiés sur d'autres protéines, les limites en sont donc putatives. Les comparaisons de séquences que nous avons réalisées ont donc redéfini les contours des régions fonctionnelles de DMP1 parfois de manière importante, par exemple pour le motif de régulation de DSPP.



Analyse évolutive de DMP1 à partir de 41 séquences de Mammifères. « Sliding Window » (fenêtre de 15 pb et déplacement de 5 pb). Logarithme du taux de substitution par site (b) Taux de substitution non synonyme (dN) le long de DMP1. Les fortes contraintes sélectives correspondant aux faibles taux de substitution. Les zones connues pour avoir un rôle fonctionnel sont indiquées en abscisse (RGD, liaison au collagène, ASARM, NLS (motif d'adressage dans le noyau)).



Séquence en acides aminés de DMP1 de l'homme sur laquelle est indiquée le résultat de l'analyse évolutive des conservations de séquences. La couleur la plus sombre représente le niveau de conservation le plus élevé. Les régions déjà connues pour leur rôle dans DMP1 sont indiquées ainsi : signal peptide (souligné); motif RGD (carré); site SG (ovale); Liaison collagène (carré contours grisés); peptide ASARM souligné deux fois; Site de localisation nucléaire (carré avec ligne pointillée); liaison DSPP (carré grisé). Pointe de flèche : 4 sites de clivage connus chez le rat. Etoiles : résidus sélectionnés positivement. (*) : Résidus phosphorylés; (#) : Résidus N-glycosylés. Ovales contours grisés : sites inconnus supposés avec une fonction importante pour DMP1.

4.1. MEPE

Dans le cadre d'une thèse de Doctorat effectuée dans notre laboratoire et consacrée à MEPE (Bardet, 2009) j'ai eu l'occasion de travailler sur une autre protéine importante dans la minéralisation de l'os et de la dentine. Avec la doctorante chargée de ce sujet, nous avons essayé de mettre à jour les patrons d'évolution de cette protéine chez les Mammifères. Ma contribution à ces travaux étant plus modeste que dans les chapitres précédents, je n'entrerai donc pas dans les détails de cet article, mais je souhaiterai insister sur le fait que cette fois encore, les résultats se sont avérés intéressants dans la compréhension du fonctionnement de MEPE. Cette protéine est non seulement connue pour être impliquée dans la minéralisation de l'os et de la dentine mais également dans la minéralisation de la coquille d'œuf chez les oiseaux. Ainsi on suppose que la fonction primordiale de MEPE serait la minéralisation de l'os, ensuite se serait ajoutée la minéralisation de la coquille d'œuf chez les Tétrapodes ovipares. Puis cette fonction se serait perdue secondairement chez les Mammifères placentaires. Il n'a malheureusement pas été possible de découvrir le gène MEPE chez l'ornithorynque qui est un mammifère ovipare. La comparaison avec les autres Mammifères aurait été intéressante.

Avant cette étude, la littérature décrivait la région « dentonine » de MEPE avec ses deux motifs de liaison à la matrice (RGD et SGDG) comme étant une caractéristique ancienne des Sibling (Fisher et Fedarko, 2003). Or nous avons pu montrer qu'au contraire, ces deux motifs sont une particularité apparue chez MEPE dans la lignée placentaire.

III - Projets de recherches

Plusieurs voies possibles de recherches se sont dégagées à partir des études que j'ai menées ces dernières années.

Premièrement, je souhaite m'intéresser à ce groupe de protéines, les *Siblings* dont j'ai eu l'occasion d'exposer les caractéristiques dans le chapitre précédent. Ce groupe de protéines est très intéressant car assez diversifié, elles sont principalement impliquées dans la minéralisation de tissus comme l'os ou la dentine, mais aussi dans la minéralisation de la coquille d'œuf chez les oiseaux. De plus, des mutations dans ces gènes semblent avoir un large spectre d'effets phénotypiques : par exemple, des mutations dans le gène DMP1 sont impliquées dans certaines formes de rachitisme. L'étude individuelle de ces gènes permettra donc de mieux comprendre leurs rôles et leurs fonctions.

Par ailleurs, les relations de parenté au sein de cette famille de gènes sont mal connues alors que tous possèdent un ancêtre commun. Le but des recherches que je souhaite entreprendre est de comprendre l'histoire évolutive de cette famille afin d'imaginer un scénario évolutif concernant l'apparition des tissus qu'ils forment. Plusieurs questions se posent auxquelles je souhaite apporter une réponse : les Siblings étaient-elles toutes présentes chez l'ancêtre commun des Sarcoptérygiens ou bien certaines ont-elles été recrutées plus tardivement? Comment ont été acquises la structure des gènes ainsi que les caractéristiques propres à chaque protéine ? Celles-ci différent-elles dans les différentes lignées? Quelles sont les relations de parenté entre les protéines de la dentine et de l'os.

Le gène MEPE est connu pour intervenir dans la minéralisation de l'os et de la dentine mais également dans la minéralisation de la coquille de l'eau des oiseaux. Comment ce gène s'est adapté à ces différentes fonctions au cours de l'évolution ? Quels ont été les changements de structure de la protéine au cours de la transition vers les amniotes et ensuite vers les placentaires ?

Pour cela il serait nécessaire de compléter les données sur ces protéines en étendant les connaissances aux reptiles, aux amphibiens et Sarcoptérygiens aquatiques (cœlacanthe et dipneuste), afin d'avoir une idée de l'évolution des gènes concernés sur une période au moins 450 Ma.

Deuxièmement, je souhaite développer un axe de recherche consacré à une autre famille de gènes : les EMP (Enamel Matrix Proteins) qui participent à la formation de l'émail. Nous travaillons sur les EMPs depuis plusieurs années dans notre équipe (voir chapitre précédent). Cependant, comme pour les Siblings, les données sur ces gènes sont faibles dès que l'on cherche hors du groupe des Mammifères, excepté les celles issues de nos précédents travaux (delgado et al., 2003 ; Bardet et al., 2010). Il n'existe aucune donnée sur l'expression de ces gènes dans les deux lignées de Tétrapodes autres que les Mammifères : les reptiles et les amphibiens. Enfin, étonnamment, on a montré un lien de parenté entre les gènes de l'émail et les protéines de la salive et du lait (Kawasaki et al., 2011) qui ne sont pourtant pas

impliqués dans des processus de minéralisation. Il sera nécessaire d'inclure ces gènes dans des analyses phylogénétiques afin de comprendre l'histoire évolutive de cette grande famille. En ce qui concerne le protéines du lait et de la salive, on peut supposer qu'un tel changement de fonction au cours de l'évolution s'est traduit par des changements important des pressions de sélection qu'il serait intéressant d'analyser.

De plus, plusieurs gènes de la famille des EMPs (ODAM, AMTN, SCPP-PQ1) ont été découverts très récemment et on connaît peu ou pas de choses sur leurs fonctions. Interviennent-ils exclusivement dans la minéralisation ? Quand ont-ils été crées ? Quel est leur rôle spécifique dans la minéralisation ? Ils représentent donc un nouveau champ d'étude à explorer.

Plusieurs gènes de EMPs seront donc mes prochaines cibles de recherche, car il existe de nombreuses « zones d'ombre » les concernant :

1. L'Améloblastine (AMBN)

L'améloblastine a été découverte simultanément par trois groupes, deux travaillant sur les incisives du rat (Cerny et al., 1996; Fong et al., 1996; Krebsbach et al., 1996) et l'autre sur les dents du porc (Hu et al., 1997a). L'expression d'*AMBN* est restreinte à l'améloblaste, comme pour l'amélogénine, mais cette expression diminue durant la maturation de l'émail (Cerny et al., 1996; Fong et al., 1996; Krebsbach et al., 1996; Uchida et al., 1997, 1998; Fukumoto et al., 2004). Seules de petites quantités d'AMBN ont été détectées, représentant moins de 5% des protéines totales de la matrice de l'émail (Krebsbach et al., 1996). AMBN est, comme AMEL, une protéine spécifique des dents. Elle est exprimée par toutes les cellules de la couche basale de l'organe de l'émail (gaine épithéliale d'Hertwig, améloblastes en pré-sécrétion, en sécrétion et en maturation) et, de façon transitoire, au niveau des odontoblastes. De plus, les régions C-ter et N-ter montrent des localisations différentes au cours du développement de la matrice de l'émail. En effet, contrairement à ce qui a été observé pour la région N-ter dans les expériences d'immuno-marquage d'Uchida et al. (1991) et de Fukae et al. (1993) (voir plus haut), la région C-ter est fortement concentrée à l'intérieur des 2 μm des fibres de Tomes. On la retrouve ensuite de moins en moins concentrée sur une profondeur de 50 μm avec un patron d'immuno-marquage en forme de "ruche d'abeille" inversée. Elle ne montre aucun marquage dans l'émail plus profond.

Le rôle d'AMBN reste pour l'instant mal connu, mais certains auteurs ont suggéré que la protéine pourrait jouer un rôle important dans le contrôle de la croissance des cristaux d'émail et dans la détermination de la structure prismatique (Robinson et al., 1998). Hu et al. (1997) pensent qu'AMBN prévient la croissance des cristallites dans le manteau de l'émail durant la phase de sécrétion et maintient ouvert un chemin par lequel les protéines de l'émail profond peuvent s'échapper durant la phase de maturation.

Et puis, il reste un fait intriguant : bien que les études citées ci-dessus ont démontré que l'améloblastine joue un rôle important dans l'amélogenèse, c'est le seul gène de l'émail pour lequel aucune mutation ponctuelle provoquant une maladie génétique (Amélogenèse Imparfaite) n'a encore été découverte.

2. L'Amélotine (AMTN)

Une étudiante en thèse travaille actuellement sur ce sujet et une collaboration avec elle devrait probablement voir le jour très bientôt.

L'amélotine s'est ajoutée récemment à la famille des EMPs; elle a été initialement identifiée au niveau des améloblastes d'incisives de souris par la technique de DD-PCR (*Differential Display Polymerase Chain Reaction*) (Iwasaki et al., 2005). Le clonage des ADNc d'*AMTN* du rat a donné trois différents produits de transcription, le plus long contenant 1032 nucléotides avec 9 exons (Moffatt et al., 2006). Les deux autres produits de transcription représentent des variantes d'épissage : le produit de transcription 2 ne contient pas l'exon 7 alors que le produit de transcription 3 ne contient pas les exons 3 à 7. Les sept premiers exons codants sont en phase 0. L'amélotine est une protéine sécrétée caractérisée par une abondance égale de proline, leucine, glutamine et thréonine et par l'absence de cystéine. Des modifications post-traductionnelles ont révélé la présence de motifs SXE très conservés chez sept Mammifères (humains, rats, souris, chimpanzés, macaques, chiens et opossums) ainsi que de nombreux sites potentiels d'O-glycosylation. Étonnamment, les produits de transcription 2 et 3 ne possèdent pas de motifs SXE, ce qui peut révéler des différences fondamentales dans les fonctions des différentes isoformes de l'AMTN (Moffatt et al., 2006). L'expression d'*AMTN* est restreinte au stade de maturation des améloblastes des molaires et des incisives de souris en développement (Iwasaki et al., 2005). Cette découverte est confirmée par Northern Blot (Moffatt et al., 2006) qui révèle une expression importante d'*AMTN* dans les stades de maturation de l'organe de l'émail. Cependant, cette expression d'*AMTN* dans les stades de maturation est différente de celles d'*AMBN* et d'*ENAM* qui sont principalement exprimés dans les stades de sécrétion de l'améloblaste (Smith, 1998). Il convient aussi de noter qu'AMTN est exprimée à de faibles niveaux dans d'autres tissus comme les ligaments parodontaux, les poumons, la gencive et le thymus. La localisation immuno-histochimique d'AMTN a révélé une localisation unique à l'interface entre l'extrémité apicale des améloblastes et la surface de l'émail au cours de la maturation dans la lame basale. Cette localisation cellulaire particulière suggère un rôle dans l'adhésion des cellules. Néanmoins, aucun motif RGD connu pour se lier aux intégrines de la membrane cellulaire n'est présent dans l'AMTN (Moffatt et al., 2006). Iwasaki et al. (2005) ont aussi indiqué qu'il est peu probable que les améloblastes sécrètent des quantités importantes d'AMTN dans un environnement fortement protéolytique, à moins que la protéine elle-même ne soit une protéase.

3. Odontogenic, ameloblast-associated protein (ODAM)

Cette protéine a été initialement identifiée par Solomon et al. (2003) dans des dépôts d'amyloïde d'échantillons de tumeurs odontogénétiques humaines; ils l'ont appelée Apin.

Une recherche intensive des gènes de protéines exprimées par les améloblastes à partir de banques d'ADNc provenant d'incisives de rat a récemment permis l'identification d'*ODAM* en même temps qu'*AMTN*. Ces deux gènes sont en effet exprimés par les améloblastes durant le stade de maturation de l'émail (Moffatt et al., 2006). Cependant, Park et al. (2007) ont aussi détecté une faible expression d'*ODAM* par les améloblastes au stade de sécrétion. Étant donné la localisation d'*ODAM* sur le chromosome 4 chez les humains et son architecture, le gène a été inclus dans la famille des SCPPs (Kawasaki et Weiss, 2003). ODAM contient un peptide signal riche en glutamine et proline (Moffatt et al., 2006), aussi faut-il souligner qu'*ODAM* a quelques autres caractéristiques des SCPPs, par exemple des introns en phase 0 et la présence d'un motif SXE codé par l'extrémité de l'exon 3. Chez l'homme, *ODAM* semble avoir une variante d'épissage qui ne dispose pas de l'exon 2, mais on ne retrouve pas cette caractéristique chez les rongeurs (Moffatt et al., 2006). *ODAM* est surexprimée dans certaines cellules cancéreuses (cancer du col de l'utérus: Rosty et al., 2005 ; quelques cancers gastriques: Aung et al., 2006). L'expression excessive d'ODAM dans plusieurs formes de cancers reste encore un phénomène incompris.

Comme *AMTN*, *ODAM* est détecté au début du stade de la maturation dans la région de Golgi et dans la région apicale des améloblastes (Moffatt et al., 2008). Le marquage immunologique dans le Golgi pour ODAM persiste tout au long du stade de maturation, alors que celui d'AMTN devient faible juste après la progression des améloblastes dans les stades de maturation. Une autre différence réside dans l'observation d'une coloration plus diffuse d'ODAM le long de la surface apicale de l'améloblastes alors que celle d'AMTN est localisée précisément au niveau de la région de la lame basale faisant la connexion entre l'organe de l'émail et la surface de l'émail (Moffatt et al., 2006). L'autre site d'expression d'ODAM et d'AMTN est l'épithélium de jonction qui entoure les dents en éruption (Park et al., 2007; Moffatt et al., 2008).

La localisation d'ODAM et d'AMTN indique que ces deux protéines jouent peut-être un rôle favorisant l'adhésion de l'épithélium à la surface des dents et cela suggère aussi une interaction possible entre les deux protéines qui est nécessaire à l'accomplissement de cette fonction.

Cependant, Park et al. (2007) ont suggéré qu'ODAM pourrait aussi être impliquée dans la minéralisation et la maturation de l'émail, mais probablement de façon indirecte en régulant l'expression de la métalloprotéinase MMP 20 et de la tufteline.

MMP20 et KLK4

La minéralisation de l'émail est sous le contrôle des EMPs. Cependant, cela fait longtemps que l'on soupçonne un rôle prédominant des protéases dans la mise en place correcte de l'émail. En effet, l'émail est un tissu unique dans le monde animal car il est hyperminéralisé ce qui empêche la présence de protéines comme le collagène qui servent habituellement de soutien, de trame de minéralisation dans les tissus comme l'os. Les EMPs, et particulièrement l'amélogénine, jouent ce rôle clef de trame de minéralisation mais avec une contrainte de taille : cette trame doit être ôtée de la matrice durant la phase de minéralisation afin d'obtenir un tissu minéralisé à plus de 98%. C'est ici qu'interviennent les protéases.

La protéase prédominante dans la matrice de l'émail au cours de l'étape de sécrétion est l'énamélysine (MMP20) (Li et al., 1999 ; Ryu et al., 1999). MMP20 est supposée être une protéase qui clive rapidement AMBN et ENAM, ainsi qu'AMEL, juste après leur sécrétion par les améloblastes et qu'elle dégrade ensuite sélectivement certains de leurs produits de clivage, alors que d'autres au contraire s'accumulent, faute d'être protéolysés, comme le fragment d'ENAM de 32 kDa (Yamakoshi et al., 2006).

L'inactivation de MMP20 chez la souris par la technique du "*knock-out"* entraîne de profonds défauts de l'émail, ce qui confirme son importance majeure dans la formation de l'émail (Caterina et al., 2002). Cependant, Yamakoshi et al. (2006) ont montré que MMP20 ne peut pas cliver le fragment de 32 kDa d'ENAM in vitro car les glycosylations protègent ce peptide des dégradations. En revanche, il a été démontré que KLK4 (kallikrein 4), une protéase plus petite que MMP20 (Scully et al., 1998) est capable de dégrader le fragment de 32 kDa (Yamakoshi et al., 2006). Cinq sites majeurs de clivage par la KLK4 ont été identifiés chez ENAM. Il a également été montré que MMP20 et KLK4 clivent les protéines de l'émail de manière différente, suggérant qu'elles jouent un rôle complémentaire dans l'amélogenèse (Yamakoshi et al., 2006).

Ces protéases ont donc probablement évolué conjointement avec les EMPs pour assurer la minéralisation de l'émail. L'émail a dû être un tissu qui est devenu de plus en plus minéralisé au cours de l'histoire des vertébrés et cette évolution s'est réalisée à la fois par la spécialisation de protéines de minéralisation comme les EMPs et aussi par le recrutement de protéases assurant à l'arrivée la construction d'un tissu presque entièrement minéral. C'est pourquoi, l'étude de l'évolution de l'émail doit passer par l'étude de l'évolution des protéases de la matrice amélaire.

miR-1304

Des travaux récents (Lopez-Valenzuela et al., 2012) ont montré qu'un MicroARNs (miR-1304) régule les gènes responsables de la formation de l'émail comme ENAM et AMTN. Une version ancestrale de ce MicroARN présente chez Neandertal est responsable d'une réduction de 50% de l'expression de ces gènes. Il existe des différences évidentes de morphologie de l'émail entre l'Homme moderne, comme par exemple une plus grande finesse de l'émail chez ce dernier. Les auteurs concluent que ces différences sont le résultat d'une mutation sur miR-1304 survenue sur la lignée *d'Homo sapiens*.

Il serait intéressant d'étudier l'évolution de la séquence de ce MicroARN ou de sa région de liaison située dans les partie 3'-UTR de ENAM et AMTN, surtout chez les Mammifères qui montrent une diversité de morphologie de l'émail très importante.

Perspectives

A ce jour nous ne savons pas quels membres des SCPPs étaient présents lors de la minéralisation des tissus squelettiques des premiers vertébrés. Tout au plus, les études d'évolution moléculaire menées dans notre équipe ont indiqué que les gènes de l'émail étaient présents à cette époque (Delgado et al., 2001 ; Sire et al., 2007). De par leur nature non structurée (Tompa 2002), les SCPPs possèdent des régions variables qui limitent beaucoup l'identification de leurs gènes par les méthodes de PCR classiques, surtout chez des espèces phylogénétiquement éloignées. Pour cette raison, les connaissances sont très dépendantes du séquençage de génomes et elles ont beaucoup plus progressé chez les Mammifères (45 génomes disponibles) que dans les autres lignées sarcoptérygiennes où elles sont limitées à quelques espèces de Tétrapodes non mammaliens (3 génomes d'oiseaux, 1 génome de lézard, 1 génome d'amphibien) chez lesquelles les SCPPs sont assez bien connues. L'étude de l'origine, des liens de parenté, des patrons évolutifs des SCPPs et des régions fonctionnelles se heurte donc aujourd'hui à ces données fragmentaires chez les Sauropsides et les amphibiens, et à leur totale absence chez les représentants actuels des lignées basales des Sarcoptérygiens, dipneustes et cœlacanthes.

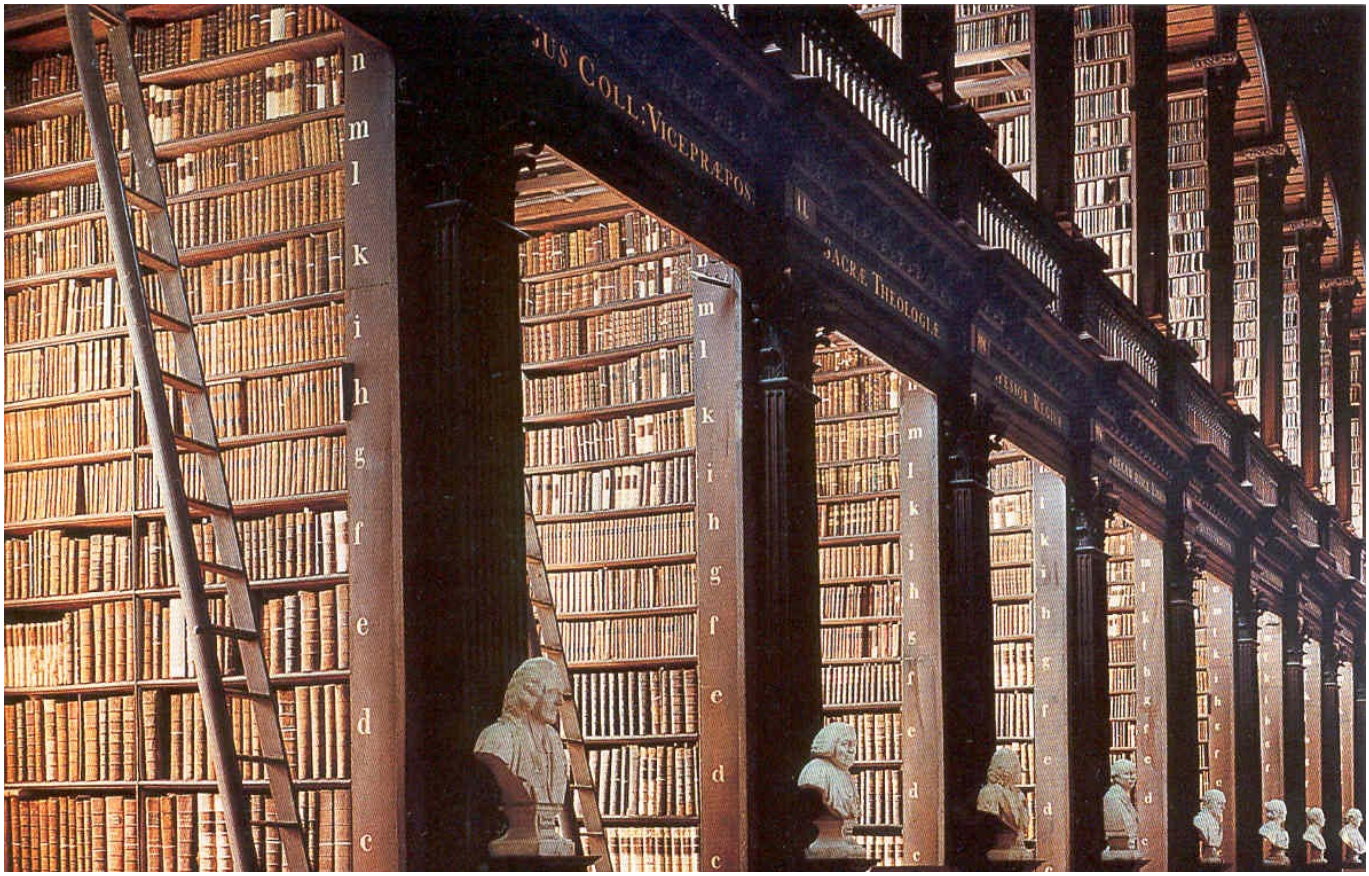
J'ai donc commencé par accueillir une étudiante sur un sujet de Master 2 et bien qu'elle n'ait pu poursuivre nos recherches en Thèse de Doctorat, son travail a permis d'obtenir les premières données sur les EMPs chez les reptiles. D'un point de vue personnel, cette expérience d'encadrement en Master a également été très enrichissante, mais le fait de devoir s'arrêter après quelques mois fructueux m'a paru assez frustrant et m'a incité à aller plus loin, vers la thèse de doctorat. Pour poursuivre ces projets, j'ai donc écrit un sujet de thèse pour atteindre un double objectif : tout d'abord, acquérir de nouvelles données sur ces gènes chez les reptiles archosauriens (crocodiliens), chez les amphibiens (pleurodèle et gymnophiones), chez les dipneustes et les cœlacanthes, afin de faire l'analyse évolutive des SCPPs chez les Sarcoptérygiens.

La séquence d'un gène ou d'une protéine apporte évidemment énormément d'informations sur un phénomène biologique comme la biominéralisation, mais elle ne doit jamais être déconnectée de l'aspect fonctionnel et physiologique. Aussi, le deuxième objectif d'une thèse sous ma direction serait d'étudier l'expression des SCPPs chez le pleurodèle (en élevage au laboratoire) et chez un gymnophione (facile à obtenir dans le commerce spécialisé) au cours du développement des dents et des os (mâchoire) et permettrait de comprendre comment les fonctions de ces protéines ont été acquises dans les différentes lignées sarcoptérygiennes. Une étude faisant appel à l'hybridation *in situ* aiderait certainement à répondre aux question suivantes : l'expression de ces gènes est-elle semblable ou différente dans le temps et l'espace chez ces diverses espèces? Est-elle étendue à d'autres tissus, y compris non minéralisés, ou au contraire la spécificité de ces protéines était-elle déjà bien établie chez l'ancêtre des Sarcoptérygiens, il y a environ 450 Ma?

Comme je viens de l'exposer, il manque des données moléculaires cruciales que je me propose d'obtenir dans les années à venir. Pour combler ces manques, je peux également m'appuyer sur certains génomes incomplets en cours de réalisation. Celui du cœlacanthe est le plus prometteur et permettra de compenser la rareté de son ADN ! Cependant, afin de ne pas dépendre de l'avancée aléatoire d'autres recherches, nous nous sommes lancés récemment dans le séquençage de transcriptomes de mâchoires sur les modèles qui nous intéressent. Ainsi, nous avons réalisé le séquençage de transcriptomes de mâchoires d'un crocodile et d'un dipneuste. Malgré cela, à ce jour ces nouvelles données n'ont pu être exploitées faute d'un étudiant en thèse.

De plus, nous prévoyons dans le futur d'obtenir d'autres transcriptomes comme ceux d'un gymnophione et d'un pleurodèle. Ces projets dépendent évidemment de financements car le séquençage à grande échelle reste une technique onéreuse. C'est dans ce cadre que nous avons fait une demande de financement ANR en 2012.

Bibliographie



Bibliothèque du « Trinity College », Dublin

Aldridge RJ, Briggs DEG (2009).The discovery of the conodont anatomy and its importance for understanding the early history of vertebrates, in Sepkoski D & Ruse M (eds), The Paleobiological Revolution. Essays on the Growth of Modern Paleontology. University of Chicago Press, Chicago & London: 73-88.

Al-Hashimi N, Sire JY, Delgado S (2009). Evolutionary Analysis of Mammalian Enamelin, the Largest Enamel Protein, Supports a Crucial Role for the 32 kDa Peptide and Reveals Selective Adaptation in Rodents and Primates. *J Mol Evol. 2009 Dec;69(6):635-56.*

Al-Hashimi N, Lafont AG, Delgado S, Kawasaki K, Sire JY (2010). The enamelin genes in lizard, crocodile and frog, and the pseudogene in the chicken provide new insights on enamelin evolution in tetrapods. Mol Biol Evol. 2010 Sep;27(9):2078-94.

Al-Hashimi N (2010). L'énaméline, la plus grande protéine de l'émail dentaire. Analyse évolutive chez les amniotes. Thèse de Doctorat de l'Université Pierre et Marie Curie, 18 juin 2010.

Aung PP, Oue N, Mitani Y, Nakayama H, Yoshida K, Noguchi T, et al. (2006). Systematic search for gastric cancer-specific genes based on SAGE data: melanoma inhibitory activity and matrix metalloproteinase-10 are novel prognostic factors in patients with gastric cancer. Oncogene 25(17):2546-57.

Bardet C, Vincent C, Lajarille MC, Jaffredo T, Sire JY (2010). OC-116, the chicken ortholog of mammalian MEPE found in eggshell, is also expressed in bone cells. J Exp Zool B Mol Dev Evol. Dec 15;314(8):653-62.

Bardet C, Delgado S, Sire JY (2010). MEPE evolution in mammals reveals regions and residues of prime functional importance. Cell Mol Life Sci. Jan;67(2):305-20.

Cerny R, Slaby I, Hammarstrom L, Wurtz T (1996). A novel gene expressed in rat ameloblasts codes for proteins with cell binding domains. J Bone Miner Res 11(7):883-91.

Caterina JJ, Skobe Z, Shi J, Ding Y, Simmer JP, Birkedal-Hansen H, et al. (2002). Enamelysin (matrix metalloproteinase 20)-deficient mice display an amelogenesis imperfecta phenotype. J Biol Chem 277(51):49598-604.

Chen Y, Zhang Y, Jiang T-X, Barlow AJ, St Amand TR, Hu Y, Heaney S, Francis-West P, Chuong C-M, Maas R: Conservation of early odontogenic signaling pathways in Aves. Proc Nat Acad Sci USA 2000, 97:10044-10049.

David J, Alistair R. Evans, Karen K. W. Siu, Emily J. Rayfield and Philip C. J. Donoghue (2012). The sharpest tools in the box? Quantitative analysis of conodont element functional morphology. Proc Biol Sci. 2012 Jul 22;279(1739):2849-54

Davis MJ, Hanson KA, Clark F, Fink JL, Zhang F, Kasukawa T, et al. (2006). Differential use of signal peptides and membrane domains is a common occurrence in the protein output of transcriptional units. PLoS Genet 2(4):e46.

Delgado S, Casane D, Bonnaud L, Laurin M, Sire JY, Girondot M (2001). Molecular evidence for precambrian origin of amelogenin, the major protein of vertebrate enamel. Mol Biol Evol 18(12):2146-53.

Delgado S, Couble ML, Magloire H, Sire JY (2006). Cloning, sequencing, and expression of the amelogenin gene in two scincid lizards. J Dent Res 85(2):138-43.

Delgado S, Davit-Béal T & Sire JY (2003). The dentition and tooth replacement pattern in Chalcides (Squamata; Scincidae). Journal of Morphology 256(2):146-59.

Delgado S, Girondot M, Sire JY (2005). Molecular evolution of amelogenin in mammals. J Mol Evol 60(1):12-30.

Delgado S, Ishiyama M, Sire JY (2007). Validation of amelogenesis imperfecta inferred from amelogenin evolution. J Dent Res 86(4):326-30.

Donoghue PC, Sansom IJ (2002). Origin and early evolution of vertebrate skeletonization. Microsc Res Tech 59(5):352-72.

Donoghue PC, Sansom IJ, Downs JP (2006). Early evolution of vertebrate skeletal tissues and cellular interactions, and the canalization of skeletal development. J Exp Zool B Mol Dev Evol 306(3):278-94.

Donoghue PC, Forey PL & Aldridge RJ (2000). Conodont affinity and chordate phylogeny. Biological Reviews 75: 191-251.

Donoghue PC, Purnell MA, Aldridge RJ, Zhang SX (2008). The interrelationships of "complex" conodonts (Vertebrata). Journal of Systematic Palaeontology 6 (2): 119-153.

Fisher LW, Fedarko NS (2003). Six genes expressed in bones and teeth encode the current members of the SIBLING family of proteins. Connect Tissue Res 44: Suppl 1:33-40.

Feng JQ, Ward LM, Liu S, Lu Y, Xie Y, Yuan B, *et al.* (2006). Loss of DMP1 causes rickets and osteomalacia and identifies a role for osteocytes in mineral metabolism. *Nat. Genet.* 38: 1310–1315.

Fong CD, Slaby I, Hammarstrom L (1996). Amelin: an enamel-related protein, transcribed in the cells of epithelial root sheath. J Bone Miner Res 11(7):892-8.

Fukae M, Tanabe T (1985). Separation of non-amelogenin component from purified amelogenin preparation of immature porcine enamel. Jpn J Oral Biol 27(1249-1251.

Fukae M, Tanabe T (1987). Nonamelogenin components of porcine enamel in the protein fraction free from the enamel crystals. Calcif Tissue Int 40(5):286-93.

Fukae M, Tanabe T, Uchida T, Yamakoshi Y, Shimizu M (1993). Enamelins in the newly formed bovine enamel. Calcif Tissue Int 53(4):257-61.

Fukae M, Tanabe T, Murakami C, Dohi N, Uchida T, Shimizu M (1996). Primary structure of the porcine 89-kDa enamelin. Adv Dent Res 10(2):111-8.

Fukumoto S, Kiba T, Hall B, Iehara N, Nakamura T, Longenecker G, et al. (2004). Ameloblastin is a cell adhesion molecule required for maintaining the differentiation state of ameloblasts. J Cell Biol 167(5):973-83.

Hedges SB (2002). The origin and evolution of model organisms. Nat Rev Genet. 3:838–849.

Hiss JA, Resch E, Schreiner A, Meissner M, Starzinski-Powitz A, Schneider G (2008). Domain organization of long signal peptides of single-pass integral membrane proteins reveals multiple functional capacity. PLoS One 3(7):e2767.

Hu CC, Fukae M, Uchida T, Qian Q, Zhang CH, Ryu OH, et al. (1997). Sheathlin: cloning, cDNA/polypeptide sequences, and immunolocalization of porcine enamel sheath proteins. J Dent Res 76(2):648-57.

Hu CC, Hart TC, Dupont BR, Chen JJ, Sun X, Qian Q, et al. (2000). Cloning human enamelin cDNA, chromosomal localization, and analysis of expression during tooth development. J Dent Res 79(4):912-9.

Iwasaki K, Bajenova E, Somogyi-Ganss E, Miller M, Nguyen V, Nourkeyhani H, et al. (2005). Amelotin, a Novel Secreted, Ameloblast-specific Protein. J Dent Res 84(12):1127-32.

Janvier P (2006). Modern look for ancient lamprey. Nature 443: 921-924.

Janvier P (2007). Living primitive fishes and fishesfrom deep time, In Mc Kenzie D. J., Farrel A. P. & Brauner C. J.(eds), Primitive Fishes. Fish Physiology 26. Elsevier, Amsterdam: 1-51

Janvier P (2008). The brain in the early jawless vertebrates: evolutionary information from an empty nutshell. Brain Research Bulletin 75: 314-318.

Kollar EJ, Fisher C (1980) Tooth induction in chick epithelium: expression of quiescent genes for enamel synthesis. Science, 207:993-995.

Kawasaki K, Weiss KM (2003). Mineralized tissue and vertebrate evolution: the secretory calcium-binding phosphoprotein gene cluster. Proc Natl Acad Sci U S A 100(7):4060-5.

Kawasaki K, Weiss KM (2006). Evolutionary genetics of vertebrate tissue mineralization: the origin and evolution of the secretory calcium-binding phosphoprotein family. J Exp Zoolog B Mol Dev Evol 306(3):295-316.

Kawasaki K, Buchanan AV, Weiss KM (2007). Gene duplication and the evolution of vertebrate skeletal mineralization. Cells Tissues Organs 186(1):7-24.

Kawasaki K, Suzuki T, Weiss KM (2005). Phenogenetic drift in evolution: the changing genetic basis of vertebrate teeth. Proc Natl Acad Sci U S A 102(50):18063-8.

Kawasaki K, Suzuki T, Weiss KM (2004). Genetic basis for the evolution of vertebrate mineralized tissue. Proc Natl Acad Sci U S A 101(31):11356-61.

Kawasaki K, Weiss KM (2008). SCPP Gene Evolution and the Dental Mineralization Continuum. J Dent Res 87(6):520-31.

Kawasaki K, Lafont AG, Sire JY (2011). The evolution of milk casein genes from tooth genes before the origin of mammals. Mol Biol Evol. 2011 Jul;28(7):2053-61.

Kawasaki K (2011). The SCPP Gene Family and the Complexity of Hard Tissues in Vertebrates. Cells Tissues Organs. 2011;194(2-4):108-12.

Kurys G, Tagaya Y, Bamford R, Hanover JA, Waldmann TA (2000). The long signal peptide isoform and its alternative processing direct the intracellular trafficking of interleukin-15. J Biol Chem 275(39):30653-9.

Krebsbach PH, Lee SK, Matsuki Y, Kozak CA, Yamada KM, Yamada Y (1996). Full-length sequence, localization, and chromosomal mapping of ameloblastin. A novel tooth-specific gene. J Biol Chem 271(8):4431-5.

Li W, Machule D, Gao C, DenBesten PK (1999). Activation of recombinant bovine matrix metalloproteinase-20 and its hydrolysis of two amelogenin oligopeptides. Eur J Oral Sci 107(5):352-9.

Lorenz-Depiereux B, Bastepe M, Benet-Pagès A, Amyere M, Wagenstaller J, Müller-Barth U, *et al.,* (2006). DMP1 mutations in autosomal recessive hypophosphatemia implicate a bone matrix protein in the regulation of phosphate homeostasis. *Nat. Genet.* 38: 1248–1250.

Lopez-Valenzuela M, Ramírez O, Rosas A, García-Vargas S, de la Rasilla M, Lalueza-Fox C, Espinosa-Parrilla Y. (2012) An ancestral miR-1304 allele present in Neanderthals regulates genes involved in enamel formation and could explain dental differences with modern humans. Mol Biol Evol. 2012 Jul;29(7):1797-806.

Lu Y, Ye L, Yu S, Zhang S, Xie Y, McKee MD, *et al.* (2007). Rescue of odontogenesis in Dmp1-deficient mice by targeted re-expression of DMP1 reveals roles for DMP1 in early odontogenesis and dentin apposition in vivo. *Dev. Biol.* **303**: 191–201.

Lu Y, Papagerakis P, Yamakoshi Y, Hu JC, Bartlett JD, Simmer JP (2008). Functions of KLK4 and MMP-20 in dental enamel formation. Biol Chem 389(6):695-700.

Mäkitie O, Pereira RC, Kaitila I, Turan S, Bastepe M, Laine T, *et al.* (2010). Long-term clinical outcome and carrier phenotype in autosomal recessive hypophosphatemia caused by a novel DMP1 mutation. *J. Bone Miner. Res.* **25**: 2165–2174.

Martoglio B, Dobberstein B (1998). Signal sequences: more than just greasy peptides. Trends Cell Biol 8(10):410-5.

Mitsiadis TA, Chéraud Y, Sharpe P, Fontaine-Pérus J (2003) Development of teeth in chick embryos after mouse neural crest transplantations. Proc Nat Acad Sci USA, 100:6541-6545

Moffatt P, Smith CE, Sooknanan R, St-Arnaud R, Nanci A (2006). Identification of secreted and membrane proteins in the rat incisor enamel organ using a signal-trap screening approach. Eur J Oral Sci 114 Suppl 1(139-46); discussion 164-5, 380-1.

Moffatt P, Smith CE, St-Arnaud R, Nanci A (2008). Characterization of Apin, a secreted protein highly expressed in tooth-associated epithelia. J Cell Biochem 103(3):941-56.

Narayanan K, Ramachandran A, Hao J, He G, Park KW, Cho M, *et al.* (2003). Dual functional roles of dentin matrix protein 1. Implications in biomineralization and gene transcription by activation of intracellular Ca2+ store. *J. Biol. Chem.* 278: 17500–17508.

Ogbureke KUE. & Fisher LW (2004). Expression of SIBLINGs and their partner MMPs in salivary glands. *J. Dent. Res.* 83: 664–670.

Ogbureke KUE. & Fisher LW (2005). Renal expression of SIBLING proteins and their partner matrix metalloproteinases (MMPs). *Kidney Int.* 68: 155–166.

Ogbureke KUE. & Fisher LW (2006). SIBLING Expression Patterns in Duct Epithelia Reflect the Degree of Metabolic Activity. *J. Histochem. Cytochem.* 55: 403–409.

Park JC, Park JT, Son HH, Kim HJ, Jeong MJ, Lee CS, et al. (2007). The amyloid protein APin is highly expressed during enamel mineralization and maturation in rat incisors. Eur J Oral Sci 115(2):153-60.

Patthy L (1999). Genome evolution and the evolution of exon-shuffling - a review. Gene 238(1):103-14.

Pirotte S., Lamour V, Lambert V, Alvarez Gonzalez ML, Ormenese S, Noël A., *et al.* (2011). Dentin matrix protein 1 induces membrane expression of VE-cadherin on endothelial cells and inhibits VEGF-induced angiogenesis by blocking VEGFR-2 phosphorylation. *Blood* 117: 2515–2526.

Smith (1992). Microstructure and evolution of enamel amongst osteichthyan and early tetrapods. Jerusalem, Israel. Freund publishing house. p. 125-150.

Smith (1989). Distribution and variation in enamel structure in the oral teeth of Sarcopterygians: its significance for the evolution of a protoprismatic enamel. Hist Biol 3(97-126).

Robinson C, Brookes SJ, Shore RC, Kirkham J (1998). The developing enamel matrix: nature and function. Eur J Oral Sci 106 Suppl 1(282-91).

Rosty C, Aubriot MH, Cappellen D, Bourdin J, Cartier I, Thiery JP, et al. (2005). Clinical and biological characteristics of cervical neoplasias with FGFR3 mutation. Mol Cancer 4(1):15.

Ryu OH, Fincham AG, Hu CC, Zhang C, Qian Q, Bartlett JD, et al. (1999). Characterization of recombinant pig enamelysin activity and cleavage of recombinant pig and mouse amelogenins. J Dent Res 78(3):743-50.

Scully JL, Bartlett JD, Chaparian MG, Fukae M, Uchida T, Xue J, et al. (1998). Enamel matrix serine proteinase 1: stage-specific expression and molecular modeling. Connect Tissue Res 39(1-3):111-22; discussion 141-9.

Sire JY, Davit-Beal T, Delgado S, Gu X (2007). The origin and evolution of enamel mineralization genes. Cells Tissues Organs 186(1):25-48.

Sire JY, Delgado S, Fromentin D, Girondot M (2005). Amelogenin: lessons from evolution. Arch Oral Biol 50(2):205-12.

Sire JY, Delgado S, Girondot M (2006). The amelogenin story: origin and evolution. Eur J Oral Sci 114 Suppl 1(64-77); discussion 93-5, 379-80.

Sire JY, Delgado S, Girondot M (2008). Hen's teeth with enamel cap: from dream to impossibility. *BMC Evolutionary Biology*: 8: 246.

Smith CE (1998). Cellular and chemical events during enamel maturation. Crit Rev Oral Biol Med 9(2):128-61.

Solomon A, Murphy CL, Weaver K, Weiss DT, Hrncic R, Eulitz M, et al. (2003). Calcifying epithelial odontogenic (Pindborg) tumor-associated amyloid consists of a novel human protein. J Lab Clin Med 142(5):348-55.

Subramanian S, Kumar S (2006). Evolutionary anatomies of positions and types of disease-associated and neutral amino acid mutations in the human genome. BMC Genomics 7(306).

Tanabe T, Fukae M, Shimizu M (1994). Degradation of enamelins by proteinases found in porcine secretory enamel in vitro. Arch Oral Biol 39(4):277-81.

Tompa P (2002). Intrinsically unstructured proteins. Trends Biochem Sci 27(10):527-33.

Turan S, Aydin C, Bereket A, Akcay T, Güran T, Yaralioglu BA, *et al.* (2010). Identification of a novel dentin matrix protein-1 (DMP-1) mutation and dental anomalies in a kindred with autosomal recessive hypophosphatemia. Bone 46: 402–409.

Uchida T, Tanabe T, Fukae M, Shimizu M, Yamada M, Miake K, et al. (1991). Immunochemical and immunohistochemical studies, using antisera against porcine 25 kDa amelogenin, 89 kDa enamelin and the 13-17 kDa nonamelogenins, on immature enamel of the pig and rat. Histochemistry 96(129-138).

Uchida T, Murakami C, Dohi N, Wakida K, Satoda T, Takahashi O (1997). Synthesis, secretion, degradation, and fate of ameloblastin during the matrix formation stage of the rat incisor as shown by immunocytochemistry and immunochemistry using region-specific antibodies. J Histochem Cytochem 45(10):1329-40.

Uchida T, Murakami C, Wakida K, Dohi N, Iwai Y, Simmer JP, et al. (1998). Sheath proteins: synthesis, secretion, degradation and fate in forming enamel. Eur J Oral Sci 106 Suppl 1(308-14.

Von Heijne G (1985). Signal sequences. The limits of variation. J Mol Biol 184(1):99-105.

Van Rheede T, Bastiaans T, Boone DN, Hedges SB, de Jong WW, Madsen O (2006). The platypus is in its place: nuclear genes and indels confirm the sister group relation of monotremes and therians. Mol Biol Evol. 23:587–597.

Ye L (2004). Deletion of Dentin Matrix Protein-1 Leads to a Partial Failure of Maturation of Predentin into Dentin, Hypomineralization, and Expanded Cavities of Pulp and Root Canal during Postnatal Tooth Development. J. Biol. Chem. 279: 19141–19148.

Ye L, Mishina Y, Chen D, Huang H, Dallas SL, Dallas MR, et al. (2005). Dmp1-deficient mice display severe defects in cartilage formation responsible for a chondrodysplasia-like phenotype. J. Biol. Chem. 280: 6197–6203.

Yamakoshi Y, Hu JC, Fukae M, Yamakoshi F, Simmer JP (2006). How do enamelysin and kallikrein 4 process the 32-kDa enamelin ? Eur J Oral Sci 114 Suppl 1(45-51); discussion 93-5, 379-80.

IV – Sélection d'articles

1. Sire JY. Huang WL, **Delgado S**, Goldberg M, Den Besten P (2012). Evolutionary story of mammalian-specific amelogenin exons 4, "4b", 8 and 9. Journal of Dental Research. J Dent Res. 2012 Jan;91(1):84-9.

2. Al-Hashimi N, Lafont AG, **Delgado S**, Kawasaki K, Sire JY (2010). The enamelin genes in lizard, crocodile and frog, and the pseudogene in the chicken provide new insights on enamelin evolution in tetrapods. Molecular Biology and Evolution. Mol Biol Evol. 2010 Sep;27(9):2078-94.

3. Al-Hashimi N, Sire JY, **Delgado S** (2009). Evolutionary Analysis of Mammalian Enamelin, the Largest Enamel Protein, Supports a Crucial Role for the 32 kDa Peptide and Reveals Selective Adaptation in Rodents and Primates. J Mol Evol. 2009 Dec;69(6):635-56.

4. Sire JY, **Delgado S**, Girondot M (2008). Hen's teeth with enamel cap: from dream to impossibility. BMC Evolutionary Biology: 8: 246.

5. **Delgado S**, Vidal N, Veron G, Sire JY (2008). Amelogenin, the major protein of tooth enamel: a new phylogenetic marker for ordinal mammal relationships. Mol phyl Evol 2008 Feb 2.

6. Sire JY, Davit-Beal T, **Delgado S**, Gu X (2007). The origine and evolution of enamel mineralization genes. Cells Tissues Organs 186(1):25-48.

7. **Delgado S**, Ishiyama M & Sire J-Y (2007). Validation of Amelogenesis Imperfecta Inferred from Amelogenin Evolution. J Dent Res 86(4):326-330.

RESEARCH REPORTS
Biological

J.-Y. Sire[1]*, Y. Huang[2,3], W. Li[2], S. Delgado[1], M. Goldberg[4], and P.K. DenBesten[2]

[1]Evolution & Développement du squelette, UMR 7138, Université Pierre et Marie Curie, 7 Quai Saint-Bernard, Bat A2, Case 5, 75005 Paris, France; [2]Department of Orofacial Sciences, University of California at San Francisco, San Francisco, CA, USA; [3]Guanghua School of Stomatology, Sun Yat-sen University, Guangzhou, Guangdong, P.R. China; and [4]Laboratoire Différenciation de Cellules Souches et Prions, U747, Université Paris Descartes, Paris, France; *corresponding author, jean-yves.sire@upmc.fr

J Dent Res 91(1):84-89, 2012

Evolutionary Story of Mammalian-specific Amelogenin Exons 4, "4b", 8, and 9

ABSTRACT

Amelogenin gene organization varies from 6 exons (1,2,3,5,6,7) in amphibians and sauropsids to 10 in rodents. The additional exons are exons 4, 8, 9, and "4b", the latter being as yet unidentified in *AMELX* transcripts. To learn more about the evolutionary origin of these exons, we used an *in silico* approach to find them in 39 tetrapod genomes. *AMEL* organization with 6 exons was the ancestral condition. Exon 4 was created in an ancestral therian (marsupials + placentals), then exon 9 in an ancestral placental, and finally exons "4b" and 8 in rodents, after divergence of the squirrel lineage. These exons were either inactivated in some lineages or remained functional: Exon 4 is functional from artiodactyls onward; exon 9 is known, to date, only in rodents, but could be coding in various mammals; and exon "4b" was probably coding in some rodents. We performed PCR of cDNA isolated from mouse and human tooth buds to identify the presence of these transcripts. A sequence analogous to exon "4b", and to exon 9, could not be amplified from the respective tooth cDNA, indicating that even though sequences similar to these exons are present, they are not transcribed in these species.

KEY WORDS: amelogenin, small exons, evolutionary origin, PCR, enamel, tetrapods.

DOI: 10.1177/0022034511423399

Received July 5, 2011; Last revision August 19, 2011; Accepted August 23, 2011

A supplemental appendix to this article is published electronically only at http://jdr.sagepub.com/supplemental.

INTRODUCTION

Amelogenin, the major protein of forming enamel, mainly plays a role in crystal growth (Robinson *et al.*, 1996; Beniash *et al.*, 2005). Its encoding gene (*AMELX* = *AMEL* in non-mammalian species) is composed of 7 exons in mammals, except in rats and mice, in which 2 additional exons (8 and 9) have been found (R Li *et al.*, 1995; W Li *et al.*, 1998). These exons and exon 4 are not present in non-mammalian *AMEL*. *AMELX* is subjected to alternative splicing, giving rise to several transcripts and various isoforms. Some of them might possess signaling capabilities. In the mouse, 17 *AMELX* transcripts have been identified, among which 7 lack exon 7 and end with exons 8 and 9 (R Li *et al.*, 1995; W Li *et al.*, 1998; Bartlett *et al.*, 2006). In 2006, when analyzing the genomic region containing *AMELX* exon 8, Bartlett and colleagues revealed that this exon was homologous to exon 5. In addition, they found that a small sequence located immediately upstream of exon 8 was identical to the exon 4 sequence, and they referred to this as a putative exon 4b. Exons 4b and 8 were, therefore, generated from the duplication of a gDNA segment containing exons 4 and 5, which was translocated downstream of exon 7. Surprisingly, exon 4 was never found in *AMELX* transcripts identified in rodent cDNA, and hence hereafter are marked "4b".

Using *in silico* investigations, we traced the origin of mammalian-specific *AMELX* exons 4, "4b", 8, and 9 through tetrapod evolution. Our analyses provide information on the birth of these exons and led us to wonder whether *AMELX* exon "4b" is coding in the mouse, and whether exon 9 is present in human *AMELX* and *AMELY* transcripts. We addressed these questions using PCR.

MATERIALS & METHODS

In silico Searches

In total, 39 sequenced tetrapod genomes [37 mammals, a lizard (*Anolis carolinensis*), and a frog (*Xenopus tropicalis*)] were explored for *AMEL* exons 4, "4b", 8, and 9. Published sequences were used as a template for localization of the sequences in the genomes by BLAST. The regions potentially housing *AMEL* exon 4, 1.5 kb between exons 3 and 5, and *AMEL* exons "4b", 8, and 9, 20 kb downstream of exon 7, were extracted from each genome (Fig. 1). These regions were screened with UniDPlot (Girondot and Sire, 2010), with human and rodent sequences of the targeted exons as a template. The sequences were validated by means of alignment with human and murine sequences with Se-Al 2.0 (Rambaut, 1996).

References of the studied genomes and *AMEL* sequences are listed in Appendix 1.

Selective Pressure Analysis by the Hyphy Method

Hyphy software (http://hyphy.org; Pond *et al*., 2005) was used in the search for selective pressures that acted on exon 4 during evolution (Appendix 2).

PCR

Primers were designed with Primer Premier 5.0 software (PREMIER Biosoft International, Palo Alto, CA, USA) (Appendix 3). PCRs were performed on mouse genomic DNA (gDNA) with primer pair M1, to obtain an accurate sequence of the non-coding genomic region located between exons 7 and 8. This aimed to check whether *AMELX* exon "4b" sequence was really present in the mouse genome or was an artifact resulting from an incorrect computer-predicted sequence assembly. PCRs were also done on a mouse tooth bud cDNA library with primer pairs M2, M3, and M4, to find transcripts possessing the putative exon "4b". In humans, utilizing primer pairs H1 and H2, we performed PCR on cDNAs for putative transcripts ending with exon 9 (including a human fetal tooth cDNA library and cDNA samples freshly prepared from fetal human tooth buds, collected under the guidelines of the University of California Committee on Human Research), by using the RNeasy Mini Kit and SuperScript III Reverse Transcriptase Kit. Both in humans and mice, the primers designed for PCRs on cDNA were used to amplify gDNA, to demonstrate the effectiveness of the primer sets.

RESULTS

AMELX Exon 4

Exon 4 was found in all mammalian genomes, except in the monotreme platypus (Fig. 2). Exon 4 is absent in lizard and frog gDNA. In several mammals [a primate (marmoset) out of 11 species studied, 7 laurasiatherians out of 11, the 2 xenarthrans, the 3 afrotherians, and the 2 marsupials], our analysis indicated that the putative exon 4 was inactivated (see Fig. 2). In contrast, it is possibly coding in primates, the tree shrew, rodents, artiodactyls (cow, alpaca, and pig), the cat, and in human *AMELY*, in which it possesses correct intronic splice sites and no deleterious mutation (Fig. 2). Selective pressure analysis identified the second residue encoded by exon 4 as being negatively selected (*i.e.*, conserved) and residues 3, 7, 8, and 13, and the intronic splicing site as being positively selected (see Appendix 2).

AMELX Exon "4b"

In the mouse, sequencing the 2.0-kb genomic region separating exons 7 and 8 provided a sequence identical (not shown) to that available in databases, which means that the gDNA sequence containing exon "4b", and identical to the sequence around exon 4, is a valid sequence and not an artifact generated during computer-aided assembly of this region.

A search for exon "4b" downstream of *AMEL* exon 7 in all genomic sequences revealed that a homologous sequence is present in murids (mouse, rat, and the 2 related species) and in caviids (guinea pig), while it is absent in sciurids (squirrel) (Fig. 3A). It could be coding in the guinea pig, rat, deer mouse, and mouse, while not coding in the kangaroo rat, in which the splice acceptor site is mutated. The inactivating mutation of exon "4b" in this species could have favored the accumulation of numerous substitutions, as observed in its sequence.

However, we failed to PCR amplify exon "4b" from a mouse tooth bud cDNA library.

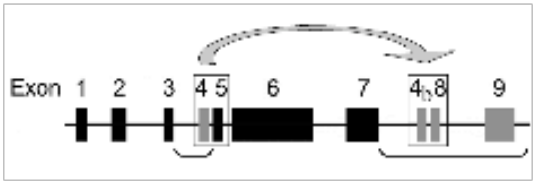


Figure 1. *AMELX* organization in the mouse genome showing the 10 exons: exons 1 to 9 and the putative exon 4b. Exon 4b and exon 8 were created after the DNA region containing exon 4 and exon 5 was duplicated and translocated downstream of exon 7 (arrow). The regions that were explored in various genomes for exons 4, 4b, 8, and 9 (gray blocks) are indicated with brackets.

AMELX Exon 8

Genome exploration downstream of *AMEL* exon 7 revealed that exon 8 is present, in addition to the mouse and rat, in only 3 other rodent *AMELX* (Fig. 3B). It was not found in the squirrel and in all non-rodent species, even as pseudogenic. Sequence analysis pointed to several substitutions when compared with the sequences of exon 5, from which it is derived. However, with the intronic splice sites being correct and no deleterious mutation being observed, there is no reason to think that *AMELX* exon 8 was not coding in the deer mouse, kangaroo rat, and guinea pig (Fig. 3B).

AMELX Exon 9

Screening the region downstream *AMEL* exon 8 in all genomes not only revealed the presence of exon 9 in all rodent sequences possessing exon 8, but also demonstrated the presence of a sequence with a high percentage of nucleotide identity in all placental *AMELX*, including human *AMELX* and *AMELY*. No sequence similarity was found in marsupial, monotreme, and non-mammalian gDNA.

Alignment of all putative exon 9 against murine sequences showed that most sequences, including human *AMELX* and *AMELY* exon 9, possessed a correct intron splice donor site at their 5′ side, along with a stop codon (Fig. 3C). In a few sequences only (*e.g.*, kangaroo rat), the putative intron splice site is mutated, which indicates either that exon 9 is not coding (*i.e.*, independently inactivated in a few species) or that another putative intron splice site is located upstream but is hard to define. The sequence length of exon 9 is variable (*e.g.*, 15 bp in guinea pig, 27 in mouse, 60 in human *AMELX*, but 27 in *AMELY*, 69 in shrew), but the nucleotide identity in homologous regions indicates that this exon is probably coding. A putative polyadenylation signal is also

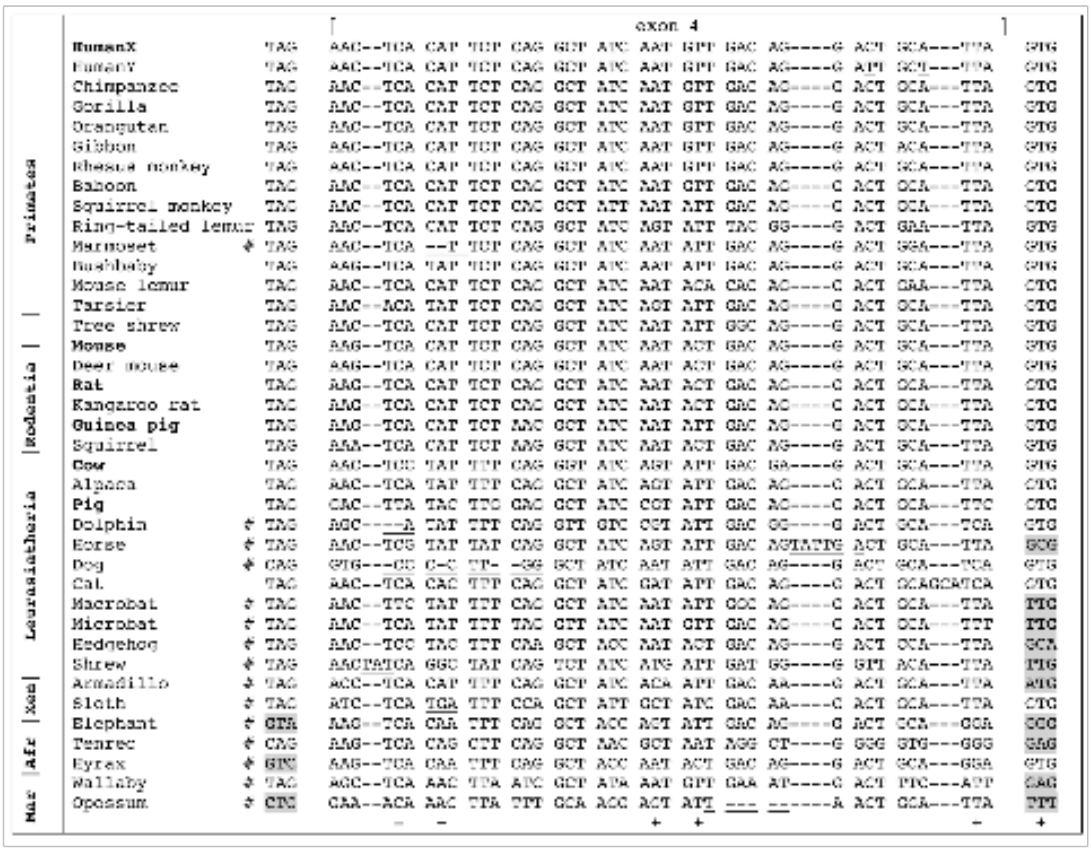


Figure 2. Alignment of 37 nucleotide sequences of either functional (*i.e.*, found in cDNA, species indicated in bold), or putatively functional (but no cDNA data), or inactivated (see below) *AMELX* exon 4 recovered in the genome of representative species of mammalian lineages. Human *AMELY* exon 4 was included in this alignment; it exhibits only 2 nucleotide substitutions (underlined) and looks functional. The important nucleotides of the donor (left) and acceptor (right) intron splices are indicated on both sides of the alignment. Exon 4 is assumed to be functional in all primates except in the marmoset, in the tree shrew, in rodents, and in a few laurasiatherians. In contrast, exon 4 is inactivated (# = pseudogenic) in numerous lineages; it shows either splice-site-mutated (in gray background) or deleterious mutations (reading frame shift or stop codon, underlined). Selective pressure analysis (Hyphy method, see Appendix 2) identified 1 negatively (-) and 5 positively (+) selected sites. Latin names of species and accession numbers of sequences are indicated in Appendix 1. Afr = afrotherians; Mar = marsupials; Xen = xenarthrans.

at the correct location as compared with rodent sequences (not shown). Given the similarity between *AMELX* exon 9 sequence in representatives of most placental mammal lineages, *i.e.*, covering a period of 104 million years (Ma) of evolution (Fig. 4), we believe that exon 9 is coding; otherwise, mutations would have accumulated at random.

However, we have not been able to PCR amplify exon 9 from human fetal tooth cDNA.

DISCUSSION

AMELX Exons 4, "4b", 8, and 9 are Mammalian Innovations

The exploration of 39 tetrapod genomes allowed us to trace the origin of *AMEL* exons 4, "4b", 8, and 9 in tetrapod evolution and to demonstrate that they are mammalian-specific *AMELX* exons. Indeed, these exons are not present in lizard and frog gDNA, confirming previous published *AMEL* transcript sequences in reptiles (Ishiyama *et al.*, 1998; Delgado *et al.*, 2006; Wang *et al.*, 2006) and amphibians (Toyosawa *et al.*, 1998; Diekwisch *et al.*, 2009). None of these exons was found in the platypus genome, confirming previously sequenced *AMEL* transcripts (Toyosawa *et al.*, 1998). This means that *AMELX* was composed of 6 functional exons (1-3, 5-7) in the last common ancestor of therian mammal (placentals + marsupials), *i.e.*, *circa* 176 Ma ago.

Exon 4 appeared in an ancestral therian, *i.e.*, between 220 and 176 Ma ago, but was not functional, confirming previous cDNA sequencing in the opossum (Hu *et al.*, 1996). It was retained as a functional exon only later in placental evolution (thus, 7 exons for *AMELX*). Then, exon 9 was recruited in an ancestral placental, but was not retained in several lineages.

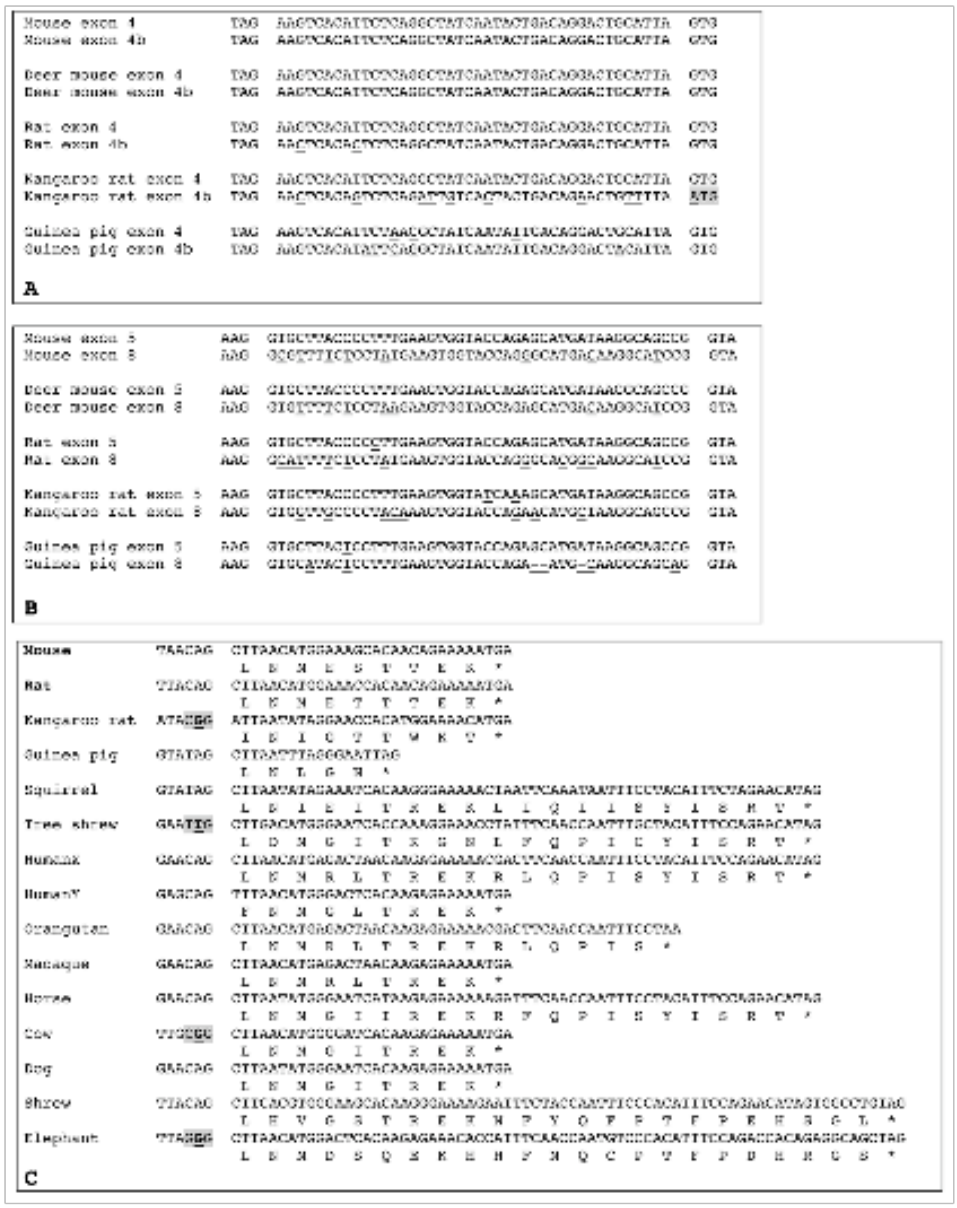


Figure 3. Alignments of amelogenin exons 4 and 4b, exons 5 and 8, and exon 9. **(A)** Alignment of *AMELX* exon 4 with the putative exon "4b" sequence found in 5 rodent genomes. In 4 species, exon "4b" is putatively functional: correct donor (left) and acceptor (right) intron splices and sequence either identical or close to that of exon 4. In the kangaroo rat, exon "4b" is not coding, as indicated by the mutation of the acceptor intron splice (gray background); note that this exon "4b" sequence shows more nucleotide substitutions than in, *e.g.*, mouse and rat sequences. Nucleotide differences between exon 4 and exon "4b" are underlined. Latin names of species are indicated in Appendix 1. **(B)** Alignment of *AMELX* exon 5 with exon 8 sequences found in 5 rodent genomes. Mouse and rat sequences are functional. The 3 other exons 8 are putatively functional: correct donor (left) and acceptor (right) intron splices, no deleterious mutations, and sequence close to that of mouse exon 8. Nucleotide differences between exon 5 and exon 8 are underlined. Latin names of species are indicated in Appendix 1. **(C)** Alignment of functional (*i.e.*, found in mouse and rat *AMELX* transcripts, in bold), putatively functional (but no cDNA data), or non-coding (*i.e.*, putative intron splice mutated, gray background) *AMELX* exon 9 (nucleotide and protein sequences) of species representative of various mammalian lineages. Several sequences are putatively functional: correct donor intron splice (shown on the left) and beginning of the sequence similar to that of mouse and rat exon 9. Note the remarkable conservation of the sequence from elephant to mice, and the differences between human *AMELX* and *AMELY* sequences. * = stop codon. Latin names of species are indicated in Appendix 1.

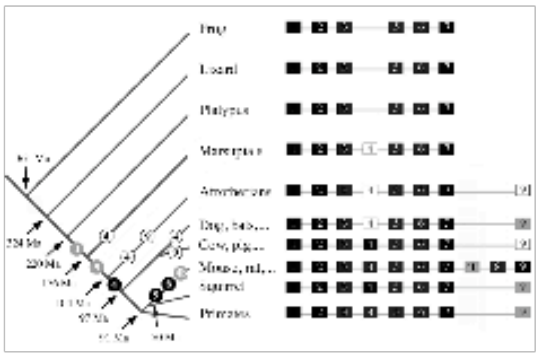


Figure 4. A summary of the story of AMELX small exons during mammalian evolution. **On the left:** putative location of the recruitment (numbered gray circles) then fixation (numbered black circles) or inactivation (numbered white circles) of *AMELX* exons 4, "4b", 8, and 9 during evolution. *AMELX* exon 4 was recruited first, in an ancestral therian, then exon 9 in an ancestral placental mammal. Exon "4b" and exon 8 were recruited in the lineage leading to murid rodents, around 50 million years (Ma) ago. Once created, these 4 small, mammalian-specific *AMELX* exons were conserved as either functional (fixed) or not (pseudogenic) in all subsequent mammalian lineages. **On the right:** For each lineage, *AMELX* organization is shown with functional exons as black blocks, pseudo-exons as white blocks, and putatively coding exons as gray blocks. Estimated times of divergence: tetrapods from Hedges (2009), amniotes from Shedlock and Edwards (2009), mammals from Madsen (2009), placental mammals from Murphy and Eizirik (2009), and rodents from Adkins *et al.* (2001) and Huchon *et al.* (2002).

Eventually, exons "4b" and 8 were created in an ancestral rodent.

Coding or Not Coding Mammalian-specific Exons

In addition to providing information on the timing of the recruitment of these small exons during evolution, our analyses indicated whether these exons are putatively coding in representatives of the mammalian lineages. Indeed, current data concerning *AMELX* transcript sequences are available in a limited species only.

Exon 4

Exon 4 was identified in the first published human, mouse, rat, cow, and pig *AMELX* cDNAs (Gibson *et al.*, 1991; Salido *et al.*, 1992; Simmer, 1995). However, this exon is not encoded in the major AMELX isoform (known as A-4) (Veis, 2003). In rodents, the isoform containing the region encoded by exon 4 (called A+4) could have a different function compared with A-4, as suggested by bead implantation in the exposed pulp of rat molars. A-4 induced closure of the root canal, formation of a reparative dentinal bridge, and diffuse mineralization in the mesial part of the pulp chamber. The reaction was weaker after A+4 implantation (Six *et al.*, 2004; Jegat *et al.*, 2007). The positive selection of 5 residues during mammalian evolution means that these residues were fixed (no longer subjected to substitution) during mammalian evolution, suggesting that they have acquired a function important for this region of the protein.

Our results support the following scenario:

(1) Exon 4 appeared in an ancestral therian after a DNA region containing a similarly sized coding exon was duplicated. *AMELX* exon 5 is the most probable candidate, being close to exon 4 and having the same size. Additional exons are mostly recruited through the duplication of a DNA region within the same gene, as, *e.g.*, for the creation of exons "4b" and 8 in rodent *AMELX* (see below). A vast majority of such tandem duplications are likely to be involved in mutually exclusive alternative splicing events (Kondrashov and Koonin, 2001; Letunic *et al.*, 2002). Such an alternative splicing is known for exon 4.

(2) Once created from exon 5, the peptide encoded by exon 4 did not improve protein function (as redundant peptide), and it accumulated numerous substitutions until a functional copy was retained by natural selection in, *e.g.*, murids, primates, and artiodactyls; this process is expected after exon duplication, and could explain why sequence homology with exon 5 is no longer recognizable in all species possessing AMELX exon 4.

(3) Additional mutations occurred independently in therian lineages. Mutations affected the splice donor site, resulting in exon 4 inactivation in marsupials, afrotherians, xenarthrans, and some laurasiatherians. Given the short evolutionary period since exon 4 was inactivated, the sequence is still easily recognizable as pseudogenic.

(4) In some placental mammal lineages, exon 4 mutation resulted in a protein sequence somewhat useful for protein function (*e.g.*, useful when included in some particular transcripts), and these changes were fixed after positive selection, which occurred in an ancestor of primates, rodents, and artiodactyls, as revealed by the conserved exon 4 sequence in these species.

Exons "4b" and 8

As mentioned before, exons "4b" and 8 are homologous to exons 4 and 5, respectively (Bartlett *et al.*, 2006). However, exon "4b" has not been identified in the various *AMELX* transcripts identified thus far in murids.

In the mouse, by sequencing the genomic region separating exons 7 and 8, we answered "no" to the question of exon "4b" being a possible artifact generated during computer-aided assembly of this genomic region. But, how to explain the identical sequences of exons "4b" and 4, and the absence of exon "4b" in *AMELX* transcripts? If the duplication/translocation event had occurred recently, *i.e.*, in the murine lineage, the resulting exon "4b" could have been non-coding without accumulating mutations during such a short time. However, we found that exons "4b" and 8 were created earlier, after the squirrel lineage diverged, *i.e.*, approximately 50 Ma ago. This is sufficiently long for mutations accumulating at random; otherwise, sequence conservation means that it is subjected to functional constraints, *i.e.*, it is coding. This finding is additionally supported when one considers the few nucleotide substitutions observed in exon "4b" compared with that observed in exon 8...that it is coding; both sequences were duplicated at the same time, but only exon 8 was found in several transcripts. Also, in the kangaroo rat, numerous substitutions were observed, while exon "4b" was inactivated.

However, although these findings indicate that exon "4b" should be coding, at least in rodents, we failed to find *AMELX* transcripts containing exon "4b" in a cDNA library of murine

tooth buds. Either such a transcript is stage-specific during enamel formation or it is to be found in other loci in the mouse tooth. To date, the presence of exon "4b" in rodent *AMELX* gDNA remains an enigma.

Exon 9

In rodents, *AMELX* exon 9 encodes 9 residues and a stop codon. It was believed that the translocation of the gDNA region containing exons 4 and 5 downstream of exon 7 has triggered the activation of a downstream sequence, resulting in the expression of exon 9 (Bartlett *et al*., 2006). Here, we show that the exon 9 sequence was recruited long before rodent differentiation, in a placental mammal ancestor, 176-104 Ma ago. As discussed for exon 4, a gDNA region containing a coding exon was duplicated/translocated downstream of exon 7, mutations were accumulated at random, and fixation or inactivation occurred depending on whether the sequence was subjected to functional constraints.

However, why was exon 9 not found in mammalian *AMELX* transcripts other than in murids? The high similarity of exon 9 sequence in various mammals indicates that it should be coding, but our attempts to find transcripts including exon 9 in human tooth germs were unsuccessful. It is possible that such transcripts are stage-specific or expressed in other loci, or that human tooth enamel no longer requires the encoded peptide. Our cDNAs were prepared from human fetal tissue younger than 23 wks, when the tooth enamel was still in pre-secretory and early secretory stages. It is possible that exon 9 may not express at a detectable level at these stages. To date, the presence of an exon 9 sequence in non-murine *AMELX* gDNA also remains an enigma.

ACKNOWLEDGMENTS

This study has been financially supported by grants from the Centre National de la Recherche Scientifique and the Université Pierre et Marie Curie (UMR7138). The authors declare no potential conflicts of interest with respect to the authorship and/or publication of this article.

REFERENCES

Adkins RM, Gelke EL, Rowe D, Honeycutt RL (2001). Molecular phylogeny and divergence time estimates for major rodent groups: evidence from multiple genes. *Mol Biol Evol* 18:777-791.

Bartlett JD, Ball RL, Kawai T, Tye CE, Tsuchiya M, Simmer JP (2006). Origin, splicing, and expression of rodent amelogenin exon 8. *J Dent Res* 85:894-899.

Beniash E, Simmer JP, Margolis HC (2005). The effect of recombinant mouse amelogenins on the formation and organization of hydroxyapatite crystals in vitro. *J Struct Biol* 149:182-190.

Delgado S, Couble ML, Magloire M, Sire JY (2006). Cloning, sequencing and expression of the amelogenin gene in two scincid lizards. *J Dent Res* 85:138-143.

Diekwisch TG, Jin T, Wang X, Ito Y, Schmidt M, Druzinsky R, *et al.* (2009). Amelogenin evolution and tetrapod enamel structure. *Front Oral Biol* 13:74-79.

Gibson C, Golub E, Herold R, Risser M, Ding W, Shimokawa H, *et al.* (1991). Structure and expression of the bovine amelogenin gene. *Biochemistry* 30:1075-1079.

Girondot M, Sire JY (2010). UniDPlot: a software to detect weak similarities between two DNA sequences. *J Bioinform Seq Anal* 2:69-74.

Hedges SB (2009). Vertebrates (Vertebrata). In: The timetree of life. Hedges SD, Kumar S, editors, New York, NY: Oxford University Press, pp. 309-314.

Hu CC, Zhang C, Qian Q, Ryu OH, Moradian-Oldak J, Fincham AG, *et al.* (1996). Cloning, DNA sequence, and alternative splicing of opossum amelogenin mRNAs. *J Dent Res* 75:1728-1734.

Huchon D, Madsen O, Sibbald MJ, Ament K, Stanhope MJ, Catzeflis F, *et al.* (2002). Rodent phylogeny and a timescale for the evolution of Glires: evidence from an extensive taxon sampling using three nuclear genes. *Mol Biol Evol* 19:1053-1065.

Ishiyama M, Mikami M, Shimokawa H, Oida S (1998). Amelogenin protein in tooth germs of the snake *Elaphe quadrivirgata*, immunohistochemistry, cloning and cDNA sequence. *Arch Histol Cytol* 61:467-474.

Jegat N, Septier D, Veis A, Poliard A, Goldberg M (2007). Short-term effects of amelogenin gene splice products A+4 and A-4 implanted in the exposed rat molar pulp. *Head Face Med* 3:40.

Kondrashov FA, Koonin EV (2001). Origin of alternative splicing by tandem exon duplication. *Hum Mol Genet* 10:2661-2669.

Letunic I, Copley RR, Bork P (2002). Common exon duplication in animals and its role in alternative splicing. *Hum Mol Genet* 11:1561-1567.

Li R, Li W, DenBesten PK (1995). Alternative splicing of amelogenin mRNA from rat incisor ameloblasts. *J Dent Res* 74:1880-1885.

Li W, Mathews C, Gao C, DenBesten PK (1998). Identification of two additional exons at the 3' end of the amelogenin gene. *Arch Oral Biol* 43:497-504.

Madsen O (2009). Mammals (Mammalia). In: The timetree of life. Hedges SD, Kumar S, editors, New York, NY: Oxford University Press, pp. 459-461.

Murphy WJ, Eizirik E (2009). Placental mammals (Eutheria). In: The timetree of life. Hedges SD, Kumar S, editors, New York, NY: Oxford University Press, pp. 471-474.

Pond SLK, Frost SD, Muse SV (2005). HyPhy: hypothesis testing using phylogenies. *Bioinformatics* 21:676-679.

Rambaut A (1996). Se-Al sequence alignment editor. Found at http://tree.bio.ed.ac.uk/software/seal

Robinson C, Brookes SJ, Kirkham J, Bonass WA, Shore RC (1996). Crystal growth in dental enamel: the role of amelogenins and albumin. *Adv Dent Res* 10:173-179.

Salido EC, Yen PH, Koprivnikar K, Yu LC, Shapiro LJ (1992). The human enamel protein gene amelogenin is expressed from both the X and the Y chromosomes [see comments]. *Am J Hum Genet* 50:303-316.

Shedlock AM, Edwards SV (2009). Amniotes (Amniota). In: The timetree of life. Hedges SD, Kumar S, editors, New York, NY: Oxford University Press, pp. 375-379.

Simmer JP (1995). Alternative splicing of amelogenins. *Connect Tissue Res* 32:131-136.

Six N, Tompkins K, Septier D, Veis A, Goldberg M (2004). Recruitment and characterization of the cells involved in reparative dentin formation in the exposed rat molar pulp after implantation of amelogenin gene splice products A+4 and A-4. *Oral Biosci Med* 1:35-44.

Toyosawa S, O'Huigin C, Figueroa F, Tichy H, Klein J (1998). Identification and characterization of amelogenin genes in monotremes, reptiles, and amphibians. *Proc Natl Acad Sci USA* 95:13056-13061.

Veis A (2003). Amelogenin gene splice products: potential signaling molecules. *Cell Mol Life Sci* 60:38-55.

Wang X, Fan JL, Ito Y, Luan X, Diekwisch TG (2006). Identification and characterization of a squamate reptilian amelogenin gene: *Iguana iguana*. *J Exp Zool (Mol Dev Evol)* 306(B):393-406.

Molecular Biology and Evolution

The Enamelin Genes in Lizard, Crocodile, and Frog and the Pseudogene in the Chicken Provide New Insights on Enamelin Evolution in Tetrapods

Nawfal Al-Hashimi,†[1] Anne-Gaelle Lafont,†[1] Sidney Delgado,[1] Kazuhiko Kawasaki,[2] and Jean-Yves Sire*[1]

[1]Université Pierre et Marie Curie, UMR 7138-Systématique-Adaptation-Evolution, Paris, France

[2]Department of Anthropology, Pennsylvania State University

†Both should be considered as first authors.

***Corresponding author:** E-mail: jean-yves.sire@upmc.fr.

Associate editor: Naoko Takezaki

Abstract

Enamelin (ENAM) has been shown to be a crucial protein for enamel formation and mineralization. Previous molecular analyses have indicated a probable origin early in vertebrate evolution, which is supported by the presence of enamel/enameloid tissues in early vertebrates. In contrast to these hypotheses, *ENAM* was only characterized in mammals. Our aims were to 1) look for *ENAM* in representatives of nonmammalian tetrapods, 2) search for a pseudogene in the chicken genome, and 3) see whether the new sequences could bring new information on *ENAM* evolution. Using in silico approach and polymerase chain reaction, we obtained and characterized the messenger RNA sequences of *ENAM* in a frog, a lizard, and a crocodile; the genomic DNA sequences of *ENAM* in a frog and a lizard; and the putative sequence of chicken *ENAM* pseudogene. The comparison with mammalian *ENAM* sequences has revealed 1) the presence of an additional coding exon, named exon 8b, in sauropsids and marsupials, 2) a simpler 5′-untranslated region in nonmammalian *ENAMs*, 3) many sequence variations in the large exons while there are a few conserved regions in small exons, and 4) 25 amino acids that have been conserved during 350 million years of tetrapod evolution and hence of crucial biological importance. The chicken pseudogene was identified in a region that was not expected when considering the gene synteny in mammals. Together with the location of lizard *ENAM* in a homologous region, this result indicates that enamel genes were probably translocated in an ancestor of the sauropsid lineage. This study supports the origin of *ENAM* earlier in vertebrate evolution, confirms that tooth loss in modern birds led to the invalidation of enamel genes, and adds information on the important role played by, for example, the phosphorylated serines and the glycosylated asparagines for correct ENAM functions.

Key words: lizard, crocodile, clawed toad, chicken, dental proteins, enamelin, pseudogene, evolution.

Introduction

Enamelin (ENAM), ameloblastin (AMBN), and amelogenin (AMEL) constitute the enamel matrix protein (EMP) family, a group of proteins that belongs to the large family of secretory calcium–binding phosphoproteins (SCPP) recently identified by Kawasaki and Weiss (2003). It is well established that these three constitutive proteins play an essential role during enamel matrix formation, organization, and mineralization. ENAM, the largest protein in the enamel matrix of developing teeth, comprises only 5% of the total EMPs (Termine et al. 1980) but is probably a crucial member of the family. Indeed, in the *ENAM* $^{-/-}$ mice, the mineral that forms on dentin is not true enamel and easily crumbles as also described in *AMBN* $^{-/-}$ mice (Hu et al. 2008; Smith et al. 2009). In contrast, enamel is present in *AMEL* $^{-/-}$ mice, although it displays severe hypoplasia (Gibson et al. 2001; Fukumoto et al. 2004; Smith et al. 2009). In humans, nine autosomal-dominant or -recessive mutations of *ENAM* were reported to lead to a genetic disease, amelogenesis imperfecta (Hart et al. 2003; Kim et al. 2005; Kang et al. 2009). A recent evolutionary analysis of ENAM in mammals, that is, covering approximately 200 million years (My) of evolution, enlighted several well-conserved residues and motifs, which indicates important functions resulting from long-lasting natural selection (Al-Hashimi et al. 2009).

Recent knowledge of the relationships among EMP genes has brought additional support to *ENAM* as being probably the oldest member of the family. Comparative studies of EMPs in tetrapods, performed in order to trace back the EMP origins, relationships, and mode of evolution, have suggested that *AMEL* was derived from a duplication of *AMBN* and that the latter was created after a duplication of *ENAM* (Sire et al. 2005, 2006, 2007). In addition, molecular data also suggested that at least one EMP was present by the end of the Precambrian period, 600–550 million years ago (Ma) (Delgado et al. 2001; Sire et al. 2007); if these assumptions are correct, this EMP, therefore, should be ENAM, and this would mean that *ENAM* differentiation occurred probably long before the early jawless vertebrates acquired a mineralized skeleton. This first *EMP* was probably created from a duplication of *SPARC-L1*, itself derived from *SPARC* (Delgado

et al. 2001; Kawasaki and Weiss 2003, 2006; Kawasaki et al. 2004, 2005; Sire et al. 2007; Kawasaki 2009).

Enameloids and/or enamels, the highly mineralized tissues that protect tooth-like elements, such as odontodes, denticles, and a variety of scales, were identified in the dermal skeletal elements of jawless and jawed vertebrates that have lived approximately 450 My (Janvier 1996; Donoghue and Sansom 2002; Donoghue et al. 2006; Sire et al. 2009). These hypermineralized fossilized tissues display a characteristic structure, highly reminiscent of that of enameloids and/or enamels in extant species. We know, for instance, that the forming enameloid matrix in teleost fish contains collagen type I, which is synthesized both by the odontoblasts and by the ameloblasts (Kawasaki et al. 2005; Huysseune et al. 2008). Then, the enameloid matrix is mineralized as enamel through a process of maturation. Such a structural similarity allows to infer that the enamel matrix in early osteichthyans was 1) composed with the same proteins, 2) deposited by similar differentiated cells (ameloblasts), and 3) built through similar spatiotemporal processes as described in living species. Therefore, the forming enamel matrix of these ancestral osteichthyans certainly consisted of a combination of EMPs, especially when considering that the EMPs are enamel-specific proteins (Deméré et al. 2008; Sire et al. 2008; Meredith et al. 2009). The history of enamels, and probably of enameloids, started when the EMPs (or at least one of them) were recruited to build these tissues in early vertebrates.

In contrast to these findings that support an ancient origin for the *EMP* genes, and in particular of *ENAM* as being the ancestor of the family, EMP genes have only been characterized in the tetrapod lineages, that is, mammals, reptiles, and amphibians (Toyosawa et al. 1998; Shintani et al. 2002, 2003; Hu and Yamakoshi 2003; Al-Hashimi et al. 2009). The presence of *AMEL* and *AMBN* in all extant tetrapod lineages indicates that these EMP genes at least existed in a common ancestor of the tetrapod lineages and that their recruitment predated the divergence between the amphibian and amniote (mammals, reptiles, and birds) lineages, which occurred approximately 350 Ma (Hedges 2002).

The fact that *ENAM* was only characterized in mammals seemed to contradict our hypothesis that ENAM is the oldest and most important EMP. In order to test the hypothesis that *ENAM* was present in nonmammalian tetrapods, we looked for this gene in genome sequences of both an amphibian, *Xenopus (Silurana) tropicalis*, and a lizard, *Anolis carolinensis*. We fulfilled this objective, and two complementary issues appeared when obtaining these sequences. First, we were able to obtain messenger RNA (mRNA) sequence of *ENAM* in a crocodile as well as in these two species. The second issue concerned chicken *ENAM*. In a previous study using in silico approaches, we localized *AMEL* pseudogene (ψ) in the chicken genome, but all attempts to look for ψ*ENAM* were unsuccessful. We concluded that after being invalidated, *ENAM* probably disappeared from the genome after chromosomal rearrangement (Sire et al. 2008). However, by using an in silico approach to localize the target region on chicken chromosomes, we found the chicken ψ*ENAM* in an unexpected region of the chicken genome compared with *ENAM* location in mammalian genomes.

Materials and Methods

Biological Materials

A 1-month-old *Crocodylus niloticus* (Crocodylidae; the Nile crocodile, hereafter referred as crocodile), a juvenile *Anolis carolinensis* (Iguanidae; the green anole, hereafter referred as lizard), and a young adult *X. (Silurana) tropicalis* (the Western clawed frog, hereafter referred as frog) were used.

The animals were sacrificed according to the guidelines of ethics committees. Immediately after dissection, the jaws were immersed in liquid nitrogen and reduced to a thin powder. Total RNA was purified (Rneasy Midi; Qiagen S.A.), mRNAs were isolated (Oligotex; Qiagen S.A.), and aliquoted.

Search in Databases

Xenopus tropicalis ENAM

The fourth assembly of the frog genome (*X. tropicalis* 4.1) was searched for *ENAM* in *Ensembl* (http://www.ensembl.org/Xenopus_tropicalis/Info/Index). Blasting the frog genome using mammalian *ENAM* sequences provided no results. Therefore, we proceeded using gene synteny. First, we found *AMBN* (ENSXETT00000000694) in scaffold 392 of the frog genome sequence. Because *AMBN* is always the closest gene upstream to *ENAM* in mammalian chromosomes, we extracted 200 kilobases (kb) of genomic DNA (gDNA) downstream *AMBN*. Then, the target region was explored with UniDPlot, a software package designed to screen DNA regions showing a weak sequence similarity (http://www.ese.u-psud.fr/epc/conservation/UniDPlot/) (Sire et al. 2008). The first BLAST search was performed using 100 base pairs (bp) of the conserved 5′ region of the putative ancestral sequence of mammalian *ENAM* exon 10 (Al-Hashimi et al. 2009). This led to one hit in the target region. This short sequence was translated into an amino acid (aa) sequence and identified as frog ENAM by means of alignment with mammalian sequences using Se-Al v2.0a11 software (http://tree.bio.ed.ac.uk/software/seal) (Rambaut 1996). Then, the gDNA region potentially housing *ENAM* (25 kb on both sides of the first hit) was explored with UniDPlot, using each exon of the ancestral mammalian *ENAM* as template. Most of the frog *ENAM* sequence was identified, including the exon–intron boundaries: downstream, *ENAM* exon 10 was completed up to the stop codon, and upstream, from the beginning of exon 10 to exon 5. These sequences were translated into amino acid sequences then validated by means of alignment with mammalian ENAMs.

The full-length gDNA *ENAM* sequence of the frog was similarly recovered using the cDNA sequences obtained with Rapid Amplification of cDNA Ends–polymerase chain reaction (RACE-PCR) (see below).

Anolis carolinensis ENAM

The first assembly of lizard genome (AnoCar1.0) was searched for *ENAM* (http://www.ensembl.org/Anolis_carolinensis/

info/index) as described above for the frog *ENAM*. *ENAM* was found in scaffold 312.

Mammalian ENAMs

The sequences of mammalian *ENAMs* available in GenBank were used to look for the presence of an additional exon 8b within intron 8. The full sequences were extracted using their accession number (32 recently published mammalian *ENAMs* [GQ352330 to GQ352361] and humans [NM_031889], mouse [NM_017468], rat [NM_001106001], and pig [NM_214241] sequences) and aligned using *Se-Al*. The alignment of these 36 mammalian ENAMs was recently published (Al-Hashimi et al. 2009) with the indication of lineage relationships following mammalian phylogeny (Springer and Murphy 2007). Readers can refer to this alignment for further information. Exons 8 and 9 sequences were identified and used to blast the mammalian genomes available in databases (NCBI and Ensembl). The nucleotide sequences of intron 8 were extracted and explored with UniDPlot, using the lizard and crocodile exon 8b sequence.

Previously, we showed that ENAM sequences were conserved within the six major mammalian lineages (Al-Hashimi et al. 2009). Therefore, in order to characterize the nonmammalian ENAM sequences obtained in this study, we chose representative mammalian ENAM sequences: *Homo sapiens* (of 11 full-length ENAM sequences available in the primate lineage), *Mus musculus* (8 ENAM in Glires), *Sus scrofa* (10 ENAM in leurasiatherians), *Loxodonta africana* (4 ENAM in afrotherians), *Monodelphis domestica* (2 ENAM in marsupials), and *Ornithorhynchus anatinus* (1 ENAM in monotremes).

Gallus gallus ENAM

In modern birds, tooth-specific *EMPs* have been invalidated since approximately 100 Ma, the estimated date from which the ancestor of modern birds lost the capability to develop teeth (Sire et al. 2008; Davit-Béal et al. 2009). As a consequence, chicken *EMPs*, although having accumulated numerous mutations, might still be present in the chicken genome as pseudogenes, as recently demonstrated for chicken *ψAMEL* (Sire et al. 2008).

Because it was not possible to find long-lasting invalidated gene sequences in the chicken genome using BLAST, we searched for the chicken *ENAM* in the genomic region where the genes syntenic to *ENAM* were found in the lizard genomic sequence. Once the *ENAM* sequence was localized in the lizard genome, we explored the regions on both sides of this gene to find genes that could be also annotated in the chicken genome (last genome assembly: build 2.1 at http://www.ensembl.org/Gallus_gallus/index.html). Then, we extracted the target region from the chicken genome and searched for chicken *ψENAM* with UniDPlot using the conserved regions of lizard, crocodile, and mammalian *ENAM* sequences.

Molecular Analyses

Sequence Alignment

The protein-coding regions of nonmammalian *ENAM* sequences were translated into putative amino acid sequences, aligned to the published mammalian *ENAM* sequences using Clustal X 2.0.12 (Higgins et al. 1996), and manually corrected using Se-Al v2.0.

Signal Peptide Analysis, Cleavage Site, Remarkable Residues, and Amino Acid Composition

The putative signal peptides (SPs) were analyzed using SignalP 3.0 server (http://www.cbs.dtu.dk/services/SignalP). This software predicts the location of the three characteristic regions (n, h, and c regions) in a SP, the putative cleavage site of the SP, and calculates the probability of each predicted SP to be functional. The sequences were scanned for remarkable domains using Prosite database (http://www.expasy.ch/prosite). The amino acid composition of frog, lizard, crocodile, opossum, and human ENAM sequences was calculated using THGS database (Transmembrane Helices in Genome Sequences: http://144.16.71.10/thgs/index.html). The residue proportion was determined for the entire sequence, for the P/Q-rich and the putative 32-kDa regions, and for exon 10 encoded sequences.

Substitution Rate Analysis

In order to estimate the substitution rates among the different lineages, we performed a phylogenetic analysis using HyPhy (for Hypothesis testing using Phylogenies) software (http://hyphy.org; Kosakovsky Pond et al. 2005) based on Maximum Likelyhood. In order to calculate the substitution rate, we used the JTT model (Jones et al. 1992) based on SWISSPROT version 22 data. The topology was fixed as follows: (frog, (lizard, crocodile), (platypus, (opossum, (elephant, (pig, (mouse, human)))))).

Molecular Clock Analysis

In order to determine the origin of rate variation among the different lineages, a molecular clock test was performed using HyPhy method on JTT model (Jones et al. 1992) based on SWISSPROT version 22 data. Both the local and the global molecular clocks were tested. The local molecular clock was tested using the "molecular clock" module and the "local molecular clock" module, the latter being especially developed for such an analysis. For each test, the topology was fixed as described above. A *P* value derived from a two-tailed extended binomial distribution was used to assess significance.

PCR Amplification

mRNAs were converted to cDNA by a reverse transcriptase using an oligo$(dT)_{18}$ primer (First Strand cDNA; MBI Fermentas). *ENAM* transcripts were recovered from cDNA using normal, then RACE-PCRs. In the frog and lizard, the primers were defined from the gDNA sequence. In the crocodile, the primers were constructed for phylogenetically conserved regions identified from the alignment of mammalian and lizard *ENAMs*. All primers were designed using Primer3 (v.0.4.0) software (http://frodo.wi.mit.edu/).

Normal PCR

Each PCR was performed in a total volume of 50 µl containing 500 ng of cDNA, 0.2 µM of sense and antisense primers, 1× GoTaq reaction buffer, 0.2 mM dNTPs, and 1.25 U of GoTaq DNA Polymerase (Promega). Amplification was

performed in a thermal cycler (G-Storm GS1; GRI, UK) for 30 cycles, each cycle consisting of 1 min of denaturation at 94 °C, 1 min of annealing at 50–60 °C (depending on the primers), and 1 min of extension at 72 °C. The final extension was for 20 min at 72 °C. Expected fragments were amplified and sent to GATC Biotech SARL (http://www.gatc-biotech.com/fr/) for sequencing.

+ Primers used for lizard *ENAM*: Ano 1 (sense: 5′-AATCCCTATTTTGGACCTGGC-3′) was designed to hybridize the 5′ region of exon 10, Ano 2 (antisense: 5′-GTCTGGTGATGAGTTGGATTGTAT-3′) for the central region of exon 10, Ano 3 (antisense: 5′-TGCTGGAGATTGGCTCTGG-3′) for the end of the coding region of exon 10, Ano 4 (sense: 5′-TTTGGAAGTAAGAGTGAAGAA-3′) for exon 5, and Ano 5 (antisense: 5′-CATCTCTTCAGAATAATATGGAGG-3′) for the 5′ region of exon 10.

+ Primers used for crocodile *ENAM*: Croc 1 (sense: 5′-GGATTTGGAAGTAAGAGTG-3′) for the 3′ region of exon 5 and Croc 2 (antisense: 5′-TATTATTCTGAAGAAATGTTTG-3′) for the 5′ region of exon 10.

3′ and 5′ RACE-PCR

The reverse transcriptase-PCR method of RACE was used to complete the mRNA sequences of lizard and crocodile *ENAM* upstream and dowstream, the regions obtained with normal PCR, and to find the mRNA sequences upstream and downstream, the 5′ and 3′ regions of exon 10 of frog *ENAM* gDNA. Most of the large gDNA sequence of exon 10 was conserved for our analysis. The RACEs allowed us to identify the transcription and termination site at the 5′ and 3′ end of the mRNAs, respectively.

PCR master mix was used for the 3′ and 5′ RACE reactions. For each PCR the mixture (50 μl) was composed of 34.5 μl PCR-grade water, 5 μl 10X Advantage 2 PCR buffer, 1 μl dNTP mix (10 mM), 1 μl 50X Advantage 2 polymerase mix, 1 μl 3′ or 5′ RACE primers (GSP1 or GSP2), 5 μl universal mix primer, and 2.5 μl RACE cDNA. We use a specific touch down thermal cycling program for the RACE reaction as follows: 5 cycles (94 °C for 30 s and 72 °C for 3 min); 5 cycles (94 °C for 30 s, 70 °C for 30 s, and 72 °C for 3 min); and 20 cycles (94 °C for 30 s, 68 °C for 30 s, and 72 °C for 3 min). The first run was always followed by a Nested PCR. Sequencing was performed by GATC.

+ Primers used for the RACEs: 5′-RACE: Xenopus-GSP1 (antisense: 5′-TTCAGCCTTTGCAGGTTCCTCATC-3′), then (nested): Xenopus-NGSP1 (antisense: 5′-CATTGTTAGTTGTGGCGTTTCCTT-3′), Anolis-GSP1 (antisense: 5′-GCTTAAGTCGTGGCCTGCTGTTTGGTTT-3′), then Anolis-NGSP1 (antisense: 5′-AACTGGCATCTGTTGTGGCCAGAGGTAA-3′) were designed for the 5′ region of exon 10 to amplify the 5′-untranslated region (UTR); Crocodile-GSP1 (antisense: 5′-GCAGGGGGTTGTACTGGTTTCTGTTGC-3′), then Crocodile-NGSP1 (antisense: 5′-CATACTGGCTGCTGCTGGAAGACCTGT-3′) were designed for exon 7 to amplify the 5′ UTR.

3′-RACE: Xenopus-GSP2 (sense: 5′-AACCAGGCCTACTGCATCTTTGTT-3′), then (nested) Xenopus-NGSP2 (sense: 5′-AACTCAGTGCAGATGCAATACCAG-3′), Anolis-GSP2 (sense: 5′-CCAAGAGGATCCCGTGTTTTGGAAGC-3′), then Anolis-NGSP2 (sense: 5′-CCAGAGCCAATCTCCAGCAGCTTTC-3′) were designed for the end of exon 10 coding region to amplify the 3′ UTR.

Crocodile-GSP2 (sense: 5′-GCCTTGGCACATCCCACAGATTTACAA-3′), then Crocodile-NGSP2 (sense: 5′-AACCCACAACACAGACAAATGCCTCCA-3′) were the first designed primers for exons 8a/8b and 8b/9 to amplify exon 10 and the 3′ UTR.

Results

Frog ENAM

We found part of *ENAM* in the frog genome, available in Ensembl. Then, we amplified *ENAM* cDNA using PCR primers designed on the genomic sequence from a frog studied in our laboratory and determined the cDNA sequence, with the exception of the middle portion of the large exon 10. A comparison of the cDNA and gDNA sequences allowed us to define all the exon–intron boundaries, intron length, the transcription start site (TSS), and the polyadenylation signal of frog *ENAM* (supplementary material S1, Supplementary Material online). A single PCR product was always observed, which indicates that frog *ENAM* is transcribed as a single isoform (no alternative splicing), at least in the jaws. Frog *ENAM* occupies 17.3 kb in scaffold 392 (*vs.*, e.g., 18.0 kb in humans [chr. 4] and 24.1 kb in the opossum [chr. 5]). The full length of the transcript consists of 3,420 nucleotides distributed into eight exons, with a coding sequence of 3,216 bp (fig. 1).

Three ATGs, that is, putative translation initiation site (TIS), are located in the 5′ region of the *ENAM* transcript: one in exon 1 and two, adjacent, in the second exon (supplementary material S1, Supplementary Material online). The ATG in exon 1 has pyrimidines at both the −3 and the +4 positions. This weak Kozak consensus suggests that this ATG is not really a potential TIS (Kozak 1981). Both ATGs in the second exon have purines at both the −3 and the +4 positions. They meet the requirements of valid ATGs. In the second exon, the gDNA and cDNA sequences differ in that the latter has 1) four additional nucleotides located before the ATGs and 2) a substitution G/T in the 3′ coding region that changes the residue from Ala to Ser. In order to identify coding exons by means of sequence similarity, these sequences were translated into putative amino acid sequences and then aligned with several mammalian ENAMs. In both sequences, the only two putative TIS located in the second exon led to a correct reading frame and did not generate a stop codon downstream. Using SignalP 3.0, these two TIS were predicted as valid in both the gDNA and the cDNA sequences ($P = 0.999$ and 0.998, respectively), and the cleavage site was predicted to occur between Ala/Ser[17] and Val[18] with a probability of 0.865 and 0.880, respectively (fig. 2). Therefore, the most 5′ ATG in the second exon is chosen as the very probable correct TIS. The SP of frog ENAM is composed of 17 aa, and the first two residues of the mature protein are encoded by the six nucleotides coded by the end of the second exon.

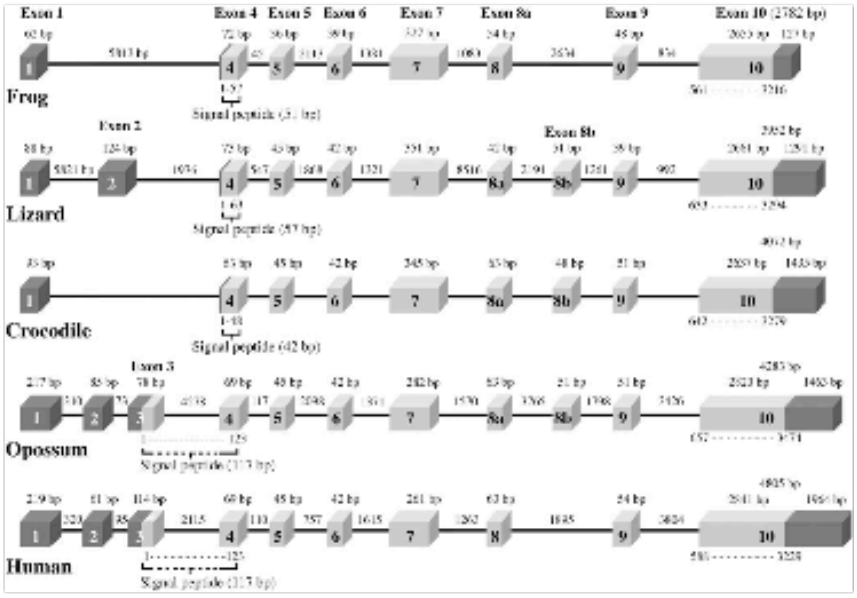


FIG. 1. Structure of the *ENAM* gene for frog (*Xenopus tropicalis*), lizard (*Anolis carolinensis*), crocodile (*Crocodylus niloticus*), opossum (*Monodelphis domestica*), and human (*Homo sapiens*). Frog and crocodile ENAM possess a single noncoding exon, whereas there are two in lizard and mammals. Exon 3, which houses a translation initiation site in mammals, is absent in frog, lizard, and crocodile *ENAM*. In frog, lizard, and crocodile *ENAM*, the SP is encoded by the only exon 4. Lizard, crocodile, and opossum *ENAM*s possess an additional coding exon 8b. The size (base pairs) of the exons (blocks) and introns (lines) are indicated (not to scale). The exons encoding the protein are in light gray. The 5′ and 3′ UTR are in dark gray.

The second exon of frog *ENAM* displays sequence similarities with mammalian *ENAM* exon 4, including a methionine codon at a similar position. Therefore, we considered both exons are orthologous and referred the second exon of frog ENAM as exon 4 (fig. 1). Exons 2 and 3, which are present in mammalian *ENAM*s, are absent in the frog.

Frog ENAM is encoded by seven exons (fig. 1). The first protein-coding exon, exon 4, is composed of 72 bp, and it starts with 15 untranslated nucleotides in our cDNA. The other exons are distributed into four small and two large exons. The 5′ UTR is composed of 77 nucleotides distributed in exon 1 and the beginning of exon 4 (15 bp). The 3′ UTR consists of 127 nucleotides located at the end of exon 10. All coding exons are in phase 0, that is, introns do not split codons. The seven exons encode a protein of 1,072 amino acids (figs. 1 and 2). Compared with various other proteins (McCaldon and Argos 1988), frog ENAM is particularly rich in proline, asparagine, serine, and glutamine, whereas poorer in leucine, lysine, alanine, valine, lysine, and aspartic acid (supplementary material S2, Supplementary Material online). The proline/glutamine-rich domain (aa 61–181) is encoded by exons 7, 8, and 9. In addition to its high number of prolines and glutamines, this region is also characterized by a large percentage of leucine and is particularly poor in acidic residues. The putative 32-kDa region of frog ENAM (deduced from the alignment with that of pig) is particularly rich in alanine, glycine, serine, glutamic acid, threonine, and asparagine, together representing more than 60% of its content. The large sequence encoded by the rest of exon 10 (representing nearly 80% of the entire sequence) is roughly similar to that of the full-length sequence (supplementary material S2, Supplementary Material online).

The comparison of the *ENAM*-coding sequences of the two frogs (ours and that available in Ensembl) revealed the presence of six single-nucleotide polymorphisms (SNPs), with the exception of exon 10, for which the cDNA was only sequenced in the 5′ and 3′ regions (supplementary material S1, Supplementary Material online). Of the six nucleotide differences, two are synonymous (not changing amino acid) and four are nonsynonymous (changing amino acid). Interestingly, one of these variable residues (Ala/Ser) is located at the SP-cleavage site.

A number of functionally important amino acids that were identified in porcine ENAM by Hu et al. (2005) are present in frog ENAM (fig. 2). They are the three putatively phosphorylated serines (SXE motifs) in the region encoded by exon 5 (SEE), exon 9 (SNE), and the beginning of exon 10 (SEE); the three putatively N-glycosylated asparagines in the region encoded by the beginning of exon 10 (NTT, NST, and NAT); and seven cysteines in the C-terminal region. An RGD motif (aa 756-758) is located in the C-terminal region (fig. 2).

Lizard ENAM

We identified the ENAM gene in the lizard genome sequence, available in Ensembl. Then, using primers designed from this sequence, we isolated this gene and determined

FIG. 2. Amino acid (aa) sequences of frog (1,072 aa), lizard (1,097 aa), and crocodile (1,092 aa) ENAM deduced from the transcripts. Remarkable residues known in mammalian ENAMs (three phosphorylated serines [S], three N-glycosylated asparagines [N], and six cysteines [C]) are present in the three ENAM sequences and boxed in gray background. The SP is boxed and the arrow indicates the cleavage site of the protein. An RGD motif is boxed in gray background. The proline/glutamine-rich domain is underlined. The asterisk indicates the end of the translation.

the full-length sequence of *ENAM* cDNA using the specimen studied in our laboratory. Exon–intron boundaries, intron length, and 5′ and 3′ UTR were defined (supplementary material S3, Supplementary Material online).

Lizard *ENAM* occupies 29.4 kb in scaffold 132 and is transcribed as a single isoform. The transcript consists of 4,807 nucleotides distributed into ten exons, and the protein-coding region is 3,294 bp in length (fig. 1). Six putative TIS are located in the 5′ region of the transcript: four in the second and two in the third exon (supplementary material S3, Supplementary Material online). Only the two TIS located in the third exon led to a correct reading frame, and they were both predicted as valid (SignalP 3.0, $P = 0.997$ and 1.0, respectively). In both cases, the cleavage site was predicted to occur between Ala^{19} and Val^{20} with a probability of 0.998 (fig. 2). By assuming that the first TIS should be the right one, the SP of lizard ENAM is composed of 19 aa, and the two residues of the protein are encoded by the last six nucleotides of the third exon. This exon shows sequence similarities with mammalian *ENAM* exon 4 and was therefore named exon 4 (fig. 1). In contrast to mammals, there was no functional TIS identified in one of the two exons located upstream lizard *ENAM* exon 4. Any of these two non–protein-coding exons showed sequence similarities with mammalian *ENAM* exon 3: We considered exon 3 being absent in lizard *ENAM*, and the two noncoding exons located at the 5′ end of the lizard *ENAM* transcript were called exon 1 and exon 2. However, they display no sequence similarity with exons 1 and 2 of mammalian *ENAM*s.

Sequence alignment of frog, lizard, and mammalian *ENAM* transcripts revealed that lizard *ENAM* possessed an additional coding exon located between exons 8 and 9 (fig. 1; supplementary material S3, Supplementary Material online). In order to conserve the current nomenclature of ENAM exons, we named this additional exon, exon 8b, and the former exon 8 was named exon 8a. Lizard ENAM is therefore encoded by eight exons. The 5′ extremity of the first coding exon, exon 4, includes ten non–protein-coding nucleotides. The following exons are distributed into five small and two large exons. The 5′ UTR is composed of 222 nucleotides distributed in exon 1, exon 2, and the beginning of exon 4. The 3′ UTR consists of 1,246 nucleotides located at the end of exon 10. All coding exons are in phase 0. The eight exons encode a protein of 1,098 amino acids (figs 1 and 2). Compared with the average overall amino acid composition of other proteins, lizard ENAM is richer in proline, glutamine, serine, asparagine, glutamic acid, arginine, and tyrosine, whereas it is poorer in leucine, alanine, lysine, and valine (supplementary material S2, Supplementary Material online). The proline/glutamine-rich domain, encoded by exons 7, 8a, 8b, 9, and beginning of exon 10 (aa 62–227) possesses a high number of prolines and glutamines and is also characterized by a large percentage of glycine and phenylalanine. In contrast, it is particularly poor in serine and acidic residues. The putative 32-kDa region of lizard ENAM deduced from the alignment is particularly rich in glycine, serine, glutamic and aspartic acids, threonine, proline, phenylalanine, and asparagine. Altogether, these amino acids represent more than 60% of the residues of this region. The amino acid composition of the large sequence encoded by the rest of exon 10 is roughly similar to that of the full-length sequence with a large number of serines, glutamic acids, arginines, and asparagines (supplementary material S2, Supplementary Material online).

Comparison of the *ENAM*-coding sequences in the two specimens reveals the presence of 43 SNPs (supplementary material S3, Supplementary Material online). There are 32 synonymous and 11 nonsynonymous differences. We also identified an insertion of 46 bp in the 3′ UTR of our transcript compared with the sequence available in GenBank.

The important amino acids identified in mammalian ENAMs are present in lizard ENAM (fig. 2): three putatively phosphorylated serines, three putatively N-glycosylated asparagines, and six cysteines in the C-terminal region. An RGD motif (aa 740–742) is present in the C-terminal region (fig. 2).

Crocodile ENAM

Our PCR using crocodile cDNA yielded a product of expected size (approximately 600 bp). The product was sequenced, translated into an amino acid sequence, validated by means of alignment with lizard ENAM, and identified as a partial sequence of crocodile *ENAM*, that is, from the end of exon 5 to the beginning of exon 10. Using 5′ and 3′ RACE-PCRs, transcript sequences were obtained from the end of exon 5 toward the 5′ extremity and from the beginning of exon 10 toward the 3′ extremity. After translation into the amino acid sequence, the organization of crocodile ENAM was validated by means of alignment with lizard and mammalian sequences. The complete coding sequence of crocodile *ENAM* was obtained from exon 4 to exon 10 (fig. 2), along with the entire 5′ and 3′ UTRs (supplementary material S4, Supplementary Material online). Alignment of the 5′ UTR of crocodile and lizard *ENAM* showed sequence similarity between lizard *ENAM* exon 1 and most of the 5′ UTR of crocodile *ENAM* (data not shown). This suggests the existence of a single noncoding exon, exon 1, in crocodile *ENAM* in contrast to two exons in the lizard. Two ATGs are located in the 5′ UTR of crocodile *ENAM*. Curiously, the ATG in exon 1 meets the requirements of a valid ATG, but it is located in an AG-rich region close to the 5′ end of exon 1. In fact, the only ATG in the second exon leads to a correct reading frame. Using SignalP 3.0, this ATG was confirmed as the probable TIS ($P = 0.897$), and the cleavage site was predicted to occur between Ala^{14} and Val^{15} ($P = 0.763$). We deduced that crocodile ENAM possesses a short SP encoded by exon 4. This SP is composed of 14 aa and the two possible first residues of the mature protein are encoded by the last six nucleotides of exon 4 (fig. 2).

The coding sequence of crocodile *ENAM* consists of 3,279 bp distributed into eight exons, including an additional exon 8b, as identified in lizard *ENAM* (fig. 1). The 5′ UTR is putatively composed of exon 1 and the beginning

FIG. 3. (*A*) Identification of exon 8b within the large intron 8 of opossum *ENAM* (not shown entirely). Intron splice sites are underlined. (*B*) Validation of exon 8b sequence of opossum and wallaby *ENAM* by means of alignment with lizard and crocodile sequences. The percentage of nucleotide identity is indicated on the right margin.

of exon 4 (14 bp). The 3′ UTR is composed of 1,435 bp (supplementary material S4, Supplementary Material online). The encoded protein, including SP, is composed of 1,093 aa, and it exhibits a roughly similar frequency of amino acid residues as described in lizard ENAM (supplementary material S2, Supplementary Material online).

The functionally important amino acids identified in mammalian ENAMs are also present in crocodile ENAM (fig. 2): three phosphorylated serines, two N-glycosylated asparagines of the three identified in porcine ENAM (Hu et al. 2005), and six cysteines located in the C-terminal region. An RGD motif (aa 729–731) is found in the C-terminal region (fig. 2).

The Additional Coding Exon 8b Discovered in Reptilian ENAM Is Present in Marsupials

Given the discovery of an exon 8b in lizard and crocodile *ENAM*, we tested the hypothesis that this exon either appeared in the sauropsid (birds and reptiles) lineage or was present earlier in *ENAM* evolution.

In the frog, the cDNA sequence did not contain this additional exon 8b. In order to look for a pseudoexon 8b, indicative of an earlier presence of this exon in tetrapod history, we blasted the frog *ENAM* intron 8 (2.6 kb) using either the lizard or the crocodile exon 8b; no valuable hit was obtained (less than 50% of nucleotide similarity).

In the two marsupials (opossum and wallaby), in large *ENAM* intron 8 (5 kb), a sequence of 51 bp was identified as possessing a high nucleotide identity with the two reptilian exon 8b; moreover, it exhibited correct splice sites (fig. 3A). This finding, which was already strongly suggested by the presence of coding exon 8b in marsupial *ENAM*, was furthermore supported by amino acid identity (>60%) with lizard and crocodile ENAM region encoded by exon 8b (fig. 3B).

In 33 placental species and a monotreme (platypus), *ENAM* intron 8 was blasted using opossum, lizard, and crocodile exon 8b. Lineage relationships and genomic mammalian sequences were indicated in our previous paper (Al-Hashimi et al. 2009). We found weak similarity (51% identity max.) with these sequences in some of these mammals, and for these species, no correct splice sites were identified. For example, in primates, leurasiatherians and afrotherians, remains of exon 8b were still identifiable in *ENAM* intron 8 as a pseudoexon sequence. This suggests that this no longer transcribed exon was probably invalidated in the common ancestor of these lineages, more than 100 My (Hedges 2002). However, such a "ghost" sequence of exon 8b was not found in intron 8 of platypus and rodent *ENAM*. It appears that exon 8b was lost in the platypus lineage independently. The reason why the remnant of exon 8b was not found in rodents may be due to high substitution rates in this lineage.

The presence of exon 8b in both the sauropsid and the mammalian lineages indicates that its origin is to be found before the divergence of these lineages. Although the size of exon 8b is close to that of exon 8a and exon 9, the comparison of the exon 8b sequence with these two exons did not show evidence that exon 8b could have originated from a duplication of either exon 8a or exon 9. However, we cannot exclude this hypothesis because its loss in most mammalian lineages could indicate that functional constraints operating on exon 8b are not strong, and hence, this exon could accumulate numerous mutations.

Comparison of ENAM Sequences in Amniotes

The organization of the coding sequences of frog, lizard, and crocodile *ENAM* is similar to that of the human and opossum *ENAMs* (fig. 1). However, there are two differences. First, 5′ UTR is distributed in three exons in mammalian *ENAM*, whereas it consists of two exons in the lizard and only one exon in the frog and, probably, in the crocodile *ENAM*. Interestingly, in mammals both exon 3 and exon 4 contain a putative functional TIS, an organization that is not present in nonmammalian *ENAM*. The second difference is the presence of an additional exon, exon 8b, in reptiles and opossum, in comparison to the frog and other mammalian *ENAM*. In the two reptilian *ENAMs*, exon 5 (45 bp) and exon 6 (42 bp)

encode the same number of amino acids as in mammals, whereas they are both shorter (36 and 39 bp, respectively) in the frog. Similarly, the sizes of the other exons of nonmammalian *ENAM* are different from those of the mammalian *ENAMs*. In particular, the size of exon 7 is much larger, whereas that of exon 10 is considerably smaller in mammals.

The amino acid sequences of ENAM were compared across nonmammals and six representative mammals, that is, human, mouse, pig, elephant, opossum, and platypus. The alignment resulted in total of 1498 positions including insertions and deletions (fig. 4). In the following, if not mentioned, the amino acid positions refer to those in this alignment. The alignment of all available amino acid sequences of mammalian ENAM was published elsewhere (Al-Hashimi et al. 2009).

The estimation of dN/dS is problematic because dS is highly likely saturated for the sequences shown in figure 4 of this study. Indeed, the divergence of the major groups of placental mammals occurred around 100 Ma, and it is known that the synonymous substitution is likely to be saturated within such a long period (Gojobori 1983).

The sequence variations of *ENAM* among the different lineages can be an indication of different functional constraints. Indeed, it is generally admitted that the mutation rate on synonymous sites is quite constant in different lineages (molecular clock). Our analysis indicates that the molecular clock is rejected in all cases (high *P* values) (supplementary material S5, Supplementary Material online). Although we cannot exclude the possibility of a change in mutation rate in the various lineages, it is rather unlikely, and the differences in amino acid sequences can be an indication of change of functional constraints among the different lineages.

The phylogenetic tree built using *ENAM* sequences in figure 4 highlights the presence of longer branches in platypus and mouse compared with the other tetrapods (supplementary material S6, Supplementary Material online). Concerning the mouse such a long branch is generally interpreted as the consequence of the combined effects of short generation times (driving a higher mutation rate) and large population size (resulting in more effective selection against mildly deleterious mutations). In contrast, the long branch obtained with platypus *ENAM* is probably the result of change of functional constraints in this lineage. Indeed, only milk teeth are present in juvenile platypus. When the primary teeth are lost, they are replaced with keratinized pads, which means that tooth proteins are no longer useful. We have previously hypothesized that the sequence differences in platypus compared with the other mammalian *ENAM* sequences could be related to relaxed selective pressures on enamel protein genes (Al-Hashimi et al. 2009).

Most variable positions and some short indels that generally concern only a few residues are located in exons 7 and 10. This may mean that functional constraints on each position are weak but not that there is no selective pressure on the amino acid regions encoded by these exons. Most of exon 7, for instance, encodes for a large part of the proline/glutamine-rich domain. In mammalian ENAMs, this region is composed of 100–130 amino acids and is characterized by a large percentage of prolines (from 27% in humans to 33.6% in opossum) and glutamines (18.7–11.7, respectively), with numerous lysines in humans and glycines and phenylalanines in opossum (supplementary material S2, Supplementary Material online). This P/Q-rich region is conserved in frog (121 aa), lizard (162 aa), and crocodile (151 aa) ENAM. The percentage of prolines in these nonmammalian ENAM is roughly similar to that observed in mammals, whereas the number of glutamines is slightly higher in nonmammals. These residues are accompanied with a high number of leucines in the frog, glycines and phenylalanines in the lizard, and alanines in the crocodile.

Besides these variable positions, 49 positions are found unchanged (80 when excluding the frog sequence), suggesting their biological significance (fig. 4). In our recent evolutionary analysis of mammalian ENAM, 25 of these unchanged positions were recognized as being important positions (Al-Hashimi et al. 2009). Of the 49 unchanged positions, 39 are located in the N-terminal region (aa 24–340). This is indicative of high sequence conservation in this region, and particularly in the regions containing the three putative phosphorylated serines identified in porcine ENAM (S^{54}, S^{251}, and S^{276}). This confirms the presence of strong functional constraints acting in these ENAM regions. Two of the three putative glycosylated asparagines (N^{309} and N^{332}) are conserved in sauropsids, whereas the third, N^{316}, is present in the lizard and frog ENAM but absent in the crocodile (fig. 4).

These important serines and asparagines are located in the region that corresponds to the so-called 32-kDa ENAM fragment characterized in porcine ENAM (aa 215–350). This short peptide is the most stable ENAM fragment that remains after MMP20 proteolysis, and it appears as a good candidate region for controlling crystal nucleation or growth as it possesses high affinity to bind apatite crystals (Tanabe et al. 1990). In the two reptiles and the opossum, the putative sequence corresponding to porcine 32 kDa should include, if it was similarly sized, the residues encoded by exon 8b (fig. 4). Of these 137 residues, 24 were unchanged during tetrapod evolution, that is, approximately 350 My (Hedges 2002). Most of these conserved positions compose two conserved motifs: GRPPXSNEEGGNPY and GXGGRPPYYSEEMFE. In nonmammalian as well as mammalian ENAMs, the 32-kDa region is characterized by a high proportion of proline and six other amino acids, glycine, serine, threonine, glutamic acid, asparagine, and phenylalanine (see the proportions of amino acids in the nonmammalian ENAMs in supplementary material S2, Supplementary Material online). Therefore, the conservation of these residues in this ENAM region during hundreds of millions of years suggests that a functional constraint keeps this region largely hydrophilic.

For most of the large protein sequence encoded by exon 10 (aa 359–1,108, not shown in fig. 4), numerous substitutions and indels hamper correct alignment (low percentage of unchanged residues). These highly variable sequences may

indicate that each position of this ENAM region is not under important functional constraints and evolved differently in the lineages leading to the species analyzed here.

In mammalian ENAMs, an RGD motif (aa 740–742) corresponding to a cell attachment sequence is present in several species (e.g., human, elephant, platypus) but absent in a few other species (e.g., mouse, pig, opossum). In reptilian ENAMs, an RGD motif is absent in this region (fig. 4). However, in the amphibian, the two reptilian, and the platypus ENAM sequences, additional RGD motif was identified (aa 1,334–1,336). This motif is absent in the other mammalian sequences. It is worthy to note that platypus ENAM houses the third RGD motif (aa 1,379–1,381, fig. 4).

The six cysteines that are involved in three disulfide bridges (C^{1147}–C^{1149}, C^{1301}, C^{1328}, C^{1404}, and C^{1492}) are phylogenetically well conserved, with the exception of crocodile ENAM in which C^{1328} is substituted by a tyrosine (Y). However, in this species, the sixth cysteine does exist at position C^{1272}, which probably forms the third disulfide bridge. In the frog ENAM, the seventh cysteine is present at position C^{1126}. The only six or seven amino acids located at the C-terminal extremity are well conserved and two remained unchanged during the evolution of these tetrapods (fig. 4).

Chicken *ψENAM*

In mammals, gene synteny is well conserved on both sides of the *EMP* gene cluster, with *IGJ* (immunoglobulin J) and *SULT1E1* (sulfotransferase 1) chosen here as dowstream and upstream boundaries, respectively. In humans (genome build 37.1), *IGJ*, *ENAM*, and *SULTIE1* are located on chromosome 4 (fig. 5A). In the chicken (genome build 2.1), *IGJ* and *SULT1E1* are annotated on chr. 4 (fig. 5B). Therefore, if gene synteny was conserved in sauropsids as in mammals, *ENAM* should be located in this region of chicken chr. 4. In lizard, *IGJ* is located in scaffold 209 and *SULT1E1* in scaffold 431, whereas the *EMP* cluster including *ENAM* is located in scaffold 132, downstream *LPL* (lipoprotein lipase) and upstream *NRG1* (neuregulin 1) (fig. 5C). In humans, *LPL* is not located on chr. 4 but is found on chr. 8, along with *NRG1* and *FUT10* (fucosyltransferase 10) (fig. 5D). In the chicken, *LPL*, *NRG1*, and *FUT10* are annotated on chr. Z (fig. 5E). Finally, in the frog, *LPL*, *NRG1*, and *FUT10* are found in scaffold 79, whereas *ENAM* and *AMBN* are located in scaffold 392 (fig. 5F). It is worthy to note that frog *ENAM* is found in a region which contains several genes (*RCHY1*: ring finger and CHY zinc finger domain containing 1; *CDKL2*: cyclin-dependent kinase-like 2; *G3BP2*: GTPase-activating protein [SH3 domain] binding protein 2; and *USO1*: USO1 homolog, vesicle-docking protein [yeast]) located on human chr. 4 (fig. 5A and F). Taken together, these findings strongly suggest that the gene synteny known in mammals around the *EMP* cluster could not be conserved in sauropsid genomes. In addition, on chicken chr. 4, several genes are not similarly oriented as in mammals, which suggests occurrence of chromosomal rearrangements (fig. 5B). Therefore, in the chicken genome, three target regions were identified as putative housers of *ψENAM*: two regions on chr. 4, one close to *IGJ* and the other close to *SULT1E1*, and, more probably, one region on chr. Z, between *LPL* and *NRG1* (fig. 5B and E).

In our previous study, in order to find chicken *ψENAM*, these regions of chr. 4 were explored with UniDPlot using the nucleotide sequences of exons encoding well-conserved regions of the putative ancestral mammalian *ENAM*, but no hits were obtained (Sire et al. 2008). In the present study, we looked for *ψENAM* in the same regions using the lizard and crocodile *ENAM* sequences that are more similar to chicken *ENAM* than mammalian ones. Again, no hits were obtained, which strongly suggested that *ENAM* was not present in these regions.

Using the same approach, we explored the region downstream of *LPL* located in chicken chr. Z. Using the well-conserved 5′ sequence of reptilian exon 10, the first hit was obtained in the target region, approximately 43 kb from *LPL* (fig. 5D). Although exhibiting numerous substitutions (as expected when considering the 100 My–long period of gene invalidation), there was no doubt that this sequence belonged to chicken *ENAM* exon 10. Therefore, we explored carefully the region close to this sequence and obtained the putative sequence of the pseudoexons of chicken *ENAM*, including exon 8b (supplementary material S7, Supplementary Material online). The chicken *ψENAM* sequence was compared with crocodile *ENAM*, its closest relative in sauropsid lineage (fig. 6). The percentage of nucleotide identity between the two sequences varies for each exon (50–70%, see fig. 6).

Discussion

We answered positively our initial question by showing the presence of *ENAM* in nonmammalian tetrapod lineages, amphibians, and sauropsids. However, these objectives would have not been reached without the availability of the sequenced frog and lizard genomes in databases. Indeed, the only *ENAM* exons that are accessible by PCR are the two large ones (exons 7 and 10), but these exons are highly variable, as demonstrated in our study. Because of sequence variations in this gene, primer design was not easy, which explains the difficulties previously encountered to identify this gene using PCR. In silico exploration of target regions, inferred from gene synteny in the sequenced genomes of frog and lizard, proved to be a successful method. Using conserved sequences of mammalian *ENAM* that were previously identified (Al-Hashimi et al. 2009), we obtained full-length sequences of *ENAM* cDNA in the frog, lizard, and crocodile, and identified the complete gDNA region of frog and lizard *ENAM* in the sequenced genomes. We also found the chicken *ψENAM*, although 1) the split of the archosaurian lineages leading to the crocodile and the chicken occurred approximately 250 Ma (Hedges 2002) and 2) the common ancestor of modern birds lost the capability to develop teeth 100 Ma (Sire et al. 2008). Therefore, all tetrapods possessing teeth covered with enamel probably have *ENAM* in their genome.

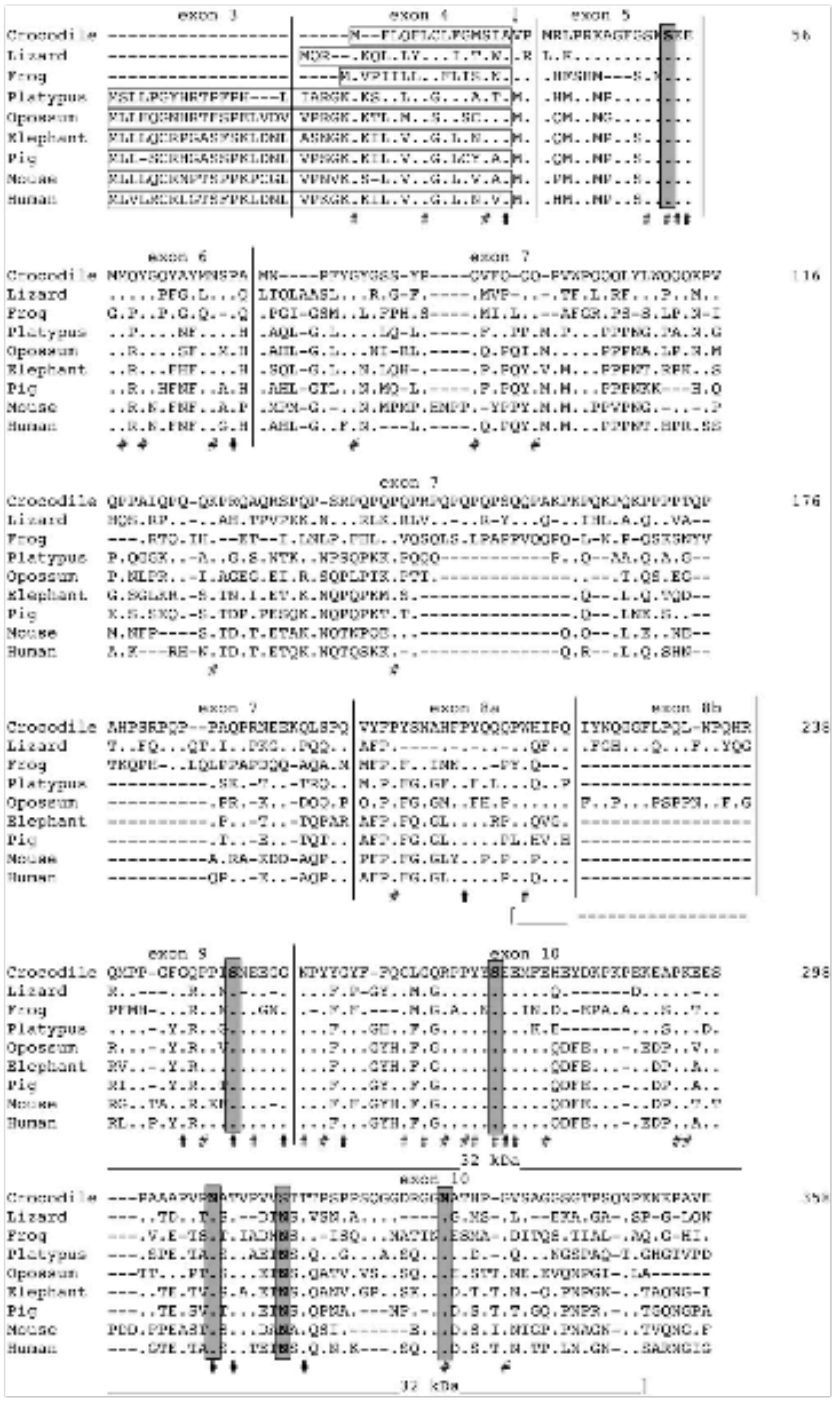


FIG. 4. Alignment of the amino acid sequence of crocodile (*Crocodylus niloticus*), lizard (*Anolis carolinensis*), and frog (*Xenopus tropicalis*) ENAM with the sequences of six species representative of the main mammalian lineages, that is, monotremes (platypus, *Ornithorhynchus anatinus*, accession no = GQ352352), marsupials (opossum, *Monodelphis domestica*, accession no = GQ352349), afrotherians (elephant, *Loxodonta africana*, accession no = GQ352337), leurasiatherians (pig, *Sus scrofa*, accession no = NM_214241), glires (mouse, *Mus musculus*, accession no = NM_017468), and primates (human, *Homo sapiens*, accession no = NM_031889). In crocodile, lizard, and frog ENAM exon 3 is absent. Exon 8b found in the lizard and crocodile is also present in the marsupial ENAM but is absent in the other mammalian species. The residues 359–718,

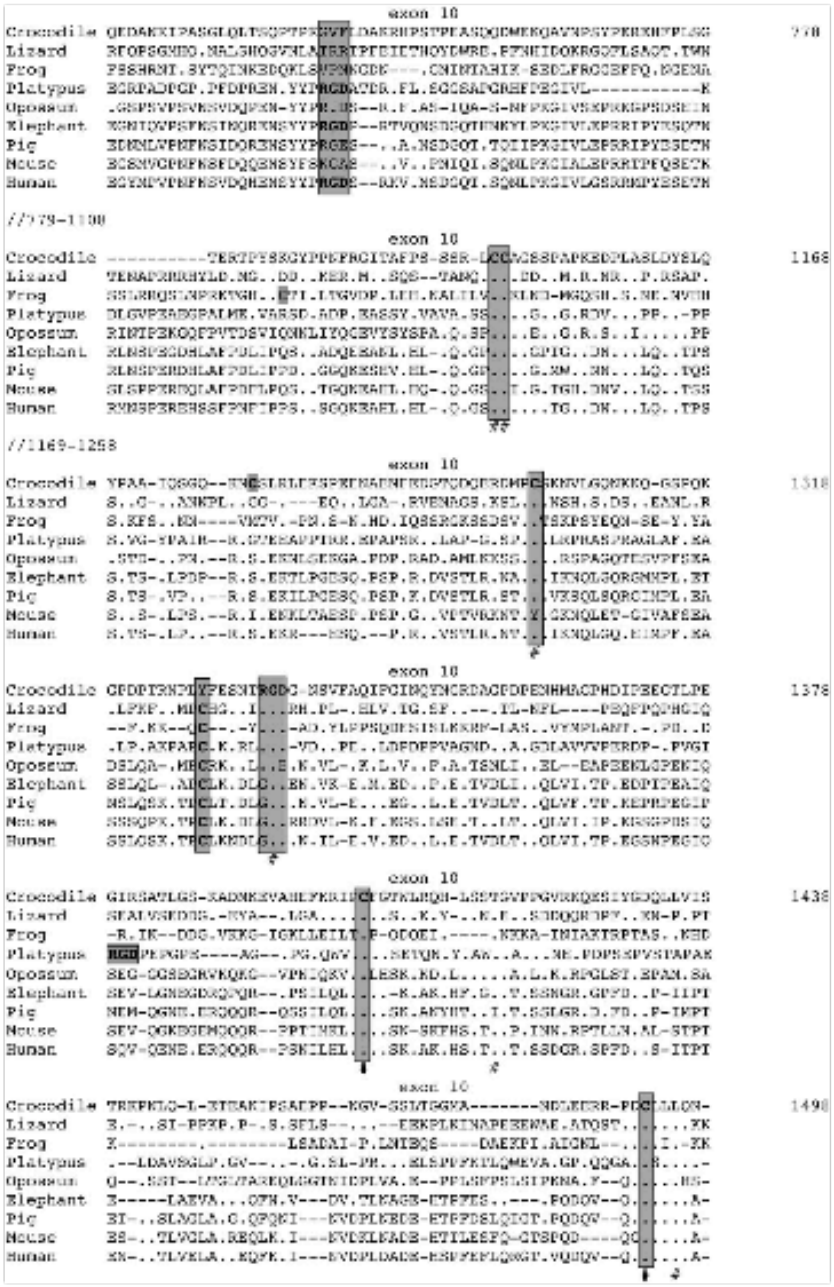


FIG. 4. (Continued).

Tetrapod ENAMs Push the Origin of ENAM Deep in Vertebrate Evolution

This is the first comparative sequence study of ENAM in representatives of various nonmammalian vertebrates. Concerning EMP evolution in vertebrates, Sire et al. have previously predicted not only that ENAM was the oldest EMP but also that this gene probably arose more than 500 Ma (Sire et al. 2005, 2006, 2007). Until now, this

←
779–1,108, and 1,169–1,258 encoded by exon 10 are not shown in the figure (//) because this highly variable region could not be aligned. SPs are boxed. Important residues and motifs known in mammals are boxed in gray background. Several RGD motifs are boxed in gray background. The 32-kDa region as known in porcine ENAM is indicated.][: limits of exons; (.): residue identical to the crocodile ENAM residue; (-): indel; (#): unchanged residue. (#): unchanged residue both in this study and when adding 36 mammalian sequences (data not shown, see Al-Hashimi et al. 2009).

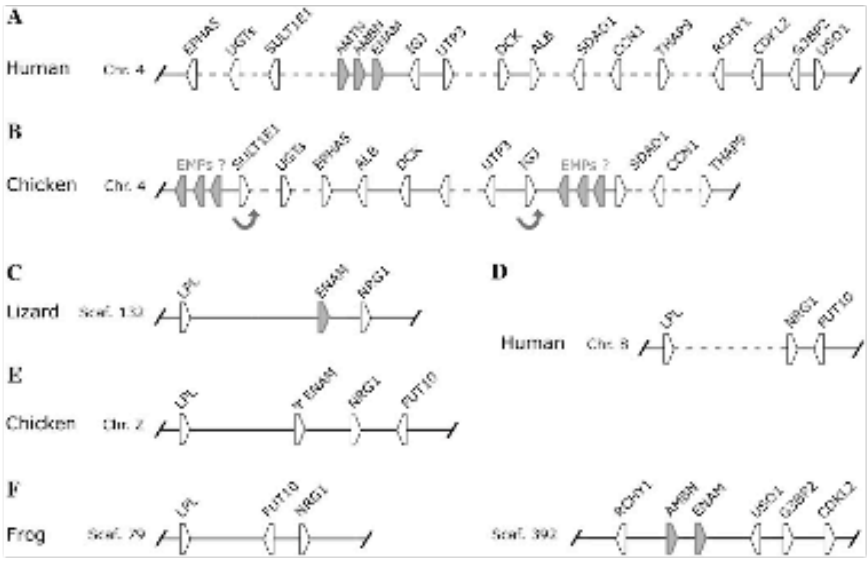


FIG. 5. Search for *ENAM* in the chicken genome by means of gene synteny. Comparison of *ENAM* location in human chromosome 4 (*A*), in lizard scaffold 132 (*C*), and in frog scaffold 392 (*F*). In the lizard, *ENAM* resides between *LPL* (lipoprotein lipase) and *NRG1* (neuregulin 1), while *IGJ* (immunoglobulin J) and *SULT1E1* (sulfotransferase 1) are located in other scaffolds (data not shown). In the frog, *ENAM* is found in a region homologous to human chr. 4, between *RCHY1* and *USO1-G3BP2-CDKL2*, while *LPL* and *NRG1* are located elsewhere (*F*). In the chicken, two chromosomes (chr. 4 and chr. Z) were targeted in order to look for *ENAM* (*B* and *E*). In chr. 4, rearrangements (curved arrows) have occurred either in the chicken or in mammals (*A* and *B*). There were, therefore, three possible locations of these genes in the chick: two on chr. 4 (*B*) and one on chr. Z (*E*). *ψENAM* was found in the latter, downstream of *LPL* and upstream of both *NRG1* and *FUT10*. In humans, *LPL*, *NRG1*, and *FUT10* are located on chr. 8 (*D*). The genes are depicted by oriented pentagons.

hypothesis was contradicted by the lack of data in non-mammalian tetrapods. Here, we clearly demonstrate that *ENAM* was present at least in the last common tetrapod ancestor of amphibians and amniotes, more than 350 Ma, that is, a jump of circa 150 My back. However, a large gap remains, which separates this date from the probable origin of *ENAM* deep in osteichthyan origins, 450–500 Ma. One support for the early origin of *ENAM* in the common ancestor of tetrapods is the presence of the three EMPs in all tetrapod genomes studied. Indeed, *AMEL* and then

```
Crocodile ex4   ATGTTCCTCCAGTTCCTGTGCCTCTTTGGCATGTCTATAGCAGTGCCG
Chicken ex4     .AAAAA....TC.............G..CA ....CTGG.......TA  (66.66%)

Crocodile ex5   ATGCGCTTGCCCCGTAAGGCCGGATTTGGAAGCAAGAGTGAAGAG
Chicken ex5     C.CTA.CAAA..TACG.A..T.T.....C...G..A.T.......   (57.78%)

Crocodile ex6   ATGATGC-AGTATGGCCAATATGCCTACATGAACTCGCCCGCC
Chicken ex6     C....C.T.T..GAATAC.CT...-..GC.AG.AAT..T...T  (48.84%)

Crocodile ex7   ATGAATCCCTTTTATGGC-------TATGGCTCCAGCTACCCACAGGTCTTCCAGCAGCAGCCAGTA-TGGC
Chicken ex7     ..T--...TAG..--...TTCCATC.......-G..GATTAA..T---..CC.....TCTT-.A..AGC..TG  (45.83%)

Crocodile ex7   CTCA-GCAGCAGCTCTACTTATGGCAGCAACAG-AAA-CCAGTACAACCCCCTGCAATTCAGCCCCAGCA //
Chicken ex7     A.G.CA..AG..TAAAG.AACCA.T..T.....C...AG..A.....G....A.G.-.A........A.. //  (60.00%)

Crocodile ex8a  GTGTATTTTCCCTATAGTAATGCACATTTTCCTTATCAGCAGCAGCCTTGGCACATCCCACAG
Chicken ex8a    A.....G....A...G.C........AA....A.T.......T....A..AT.GT.T.TGT..  (69.84%)

Crocodile ex8b  ATTTACAACCAGGGAGGCTTCTTACCACAGCTAAACCCACAACACAGA
Chicken ex8b    .........T.A-.........GG.T......G...ATGTTC..AG.T (66.67%)

Crocodile ex9   CAAA-TGCCT-CCAGGGTTTGGACAG-CCACCTATCAGCAACGAAGAAGGAGGG
Chicken ex9     ..TCT..TT.TA..T...G.TC.G.TA....T..AT.T..T.CT..G..C.AAA  (51.85%)

Crocodile ex10  AATCCTTACTATGGATACTTCTTTCAGGGGCTTGGACAGAGACCCCCCTATTACTCAGAGGA-AATGTT //
Chicken ex10    ...G....T.T.....CTA.......AA.AT....TGCC.TGT....AC.C..T.G....TTG..... //  (63.77%)
```

FIG. 6. Alignment of the nucleotide sequence of the exons encoding crocodile ENAM and the putative exon sequences of chicken *ψENAM* found on chromosome Z. The percentage of nucleotide identity is indicated in parentheses on the right. The entire sequence of exon 7 and exon 10 are not shown as they are variable and cannot be aligned.

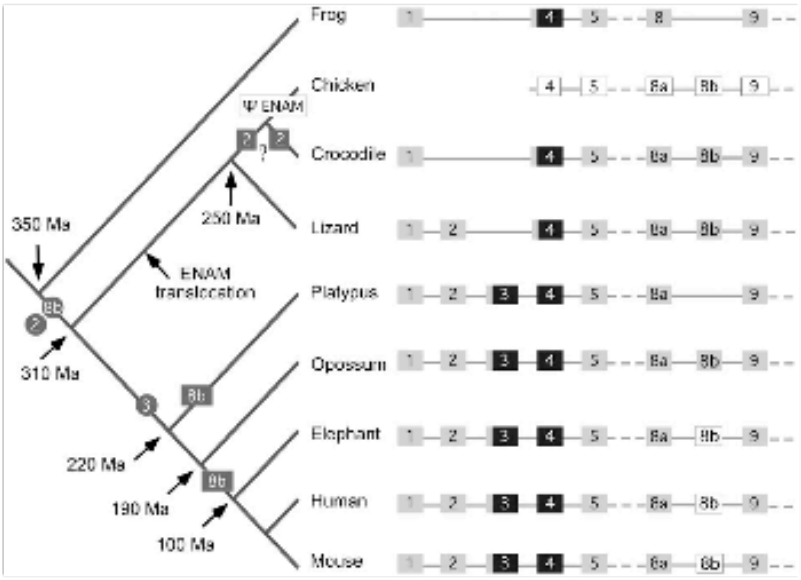


FIG. 7. Schematic localization of the events that occurred for *ENAM* during tetrapod lineage evolution. *ENAM* translocation occurred in an ancestral sauropsid. *ψENAM* indicates pseudogeneization of *ENAM* that occurred in the modern bird lineage. Valid exons are shown as gray squares, whereas "ghost exons" are shown as white squares. Numbers in gray circles and gray squares indicate the numbers of gain and loss of exons, respectively. If *ENAM* gene contains valid exons 3 and 4 (mammals), the regions shown in black squares encode large SP, whereas if there is only exon 4 and no exon 3 (frog, crocodile, and lizard), the region in black square encodes short SP. Estimation dates for lineage divergence are from Hedges (2002) and van Rheede et al. (2006).

AMBN have been already identified in amphibians and both genes show also well-conserved positions in reptiles and mammals (Toyosawa et al. 1998; Shintani et al. 2003). This means that the three EMPs were already well differentiated when the tetrapod lineages split. In addition, these EMPs arose by gene duplication from a common ancestor, and *AMEL* and *AMBN* might be derived from *ENAM* by gene duplication. Such a differentiation process may take a few tens of millions years though it is generally thought that the substitution rate is higher right after gene duplication (Hurles 2004). This would push the probable origin of the duplication toward the vertebrate origin.

Further investigations are therefore needed, for instance, in basal sarcopterygians (lungfish and coelacanths), in actinopterygians (polypteriforms, lepisosteiforms, and teleosts), and in chondrichthyans (sharks and rays). Unfortunately, as indicated above, such data are difficult to obtain without the availability of sequenced genomes in representatives of these lineages, and a correct annotation of these genomes. In the currently available teleost genomes, so far we were not able to find EMP orthologues using, for example, gene synteny. However, all teleost species possess enameloid, a well-mineralized tissue resembling enamel and evolutionarily related to the enameloid present in chondrichthyans and larval caudates (Sire et al. 2009). Either EMP genes were translocated on other chromosomes and are now too much changed to be identified in teleost genomes by searching sequence homology or they have disappeared and their role is now played by other members of the SCPP family. Indeed, several other members of the SCPP family have been also identified in teleosts as involved in bone and tooth mineralization. These genes are probably paralogs of the SIBLING (for Small Integrin-Binding Ligand *N*-linked Glycoprotein) genes (Fisher and Fedarko 2003), a subfamily of the SCPPs (Kawasaki and Weiss 2003; Kawasaki et al. 2004, 2005; Kawasaki 2009).

Alternatively, the tetrapod EMP genes all arose in the sarcopterygian lineage initially from the odontogenic ameloblast associated (ODAM) gene, as recently suggested by Kawasaki (2009). The ODAM gene is expressed during the maturation process of both tetrapod enamel and teleost enameloid.

Tetrapod ENAMs Exhibit Different Gene Organization

A Variable, Complex 5′ UTR

Until now, in mammals, the 5′ UTR of *ENAM* was classically described as being composed of either four exons, exons 1–4, or three exons, as exon 2 is absent in some species. The current nomenclature retained the presence of four exons, and *ENAM* is classically described as being composed of ten exons (e.g., Hu and Yamakoshi 2003). This gene structure is uncommon compared with that of the other EMP genes and, more generally, of all SCPP genes, in which the 5′ UTR is composed of two exons (fig. 7). In addition, mammalian *ENAMs* possess two translation initiation site (TIS) (Al-Hashimi et al. 2009). The first one is located in exon 3, which is unusual for an SCPP gene because it generates a large SP. The second TIS is found in exon 4, which

has a similar sequence in all SCPP genes. This particular organization, which probably can lead to two isoforms through alternative splicing of exon 3, was previously discussed in detail (Al-Hashimi et al. 2009). The presence of the TIS in exon 3 could be explained through exon shuffling (Gilbert 1978), but such an event is rare in vertebrate genomes. Alternatively, exon 3 could be the result of a duplication of the ancestral exon 2 (i.e., the ortholog of the current mammalian exon 4) along with the intronic acceptor and donor splice sites. Further mutations in the copy located upstream changed the environment of the SP and of the cleavage site. In frog and reptilian *ENAM*, the knowledge of the 5′ UTR brings some light to the evolution of this region, although rather complex. First, frog and reptilian *ENAM*s have a single TIS located in the second exon as expected for an SCPP gene. Therefore, our hypothesis of the recruitment of exon 3 in mammalian *ENAM* through exon shuffling could be correct. Second, in frog and, probably, in crocodile, the 5′ UTR is composed of two exons only: the first noncoding exon (exon 1) and the second exon, named exon 4 because homologue to mammalian exon 4, in which the correct TIS is located (fig. 7). Such a feature corresponds to the organization encountered in most SCPPs and was highlighted as one of the major SCPP characteristics (Kawasaki and Weiss 2003). Such an organization could possibly be the ancestral organization of *ENAM*. Third, in lizard, the second noncoding exon (exon 2) is present, whereas the TIS is located in the third exon, named exon 4 as homologous to mammalian exon 4.

The situation in the 5′ UTR is therefore complex, but worthy of interest for understanding *ENAM* evolution (fig. 7). It seems correct to propose that the ancestral tetrapod *ENAM* possessed a single noncoding exon 1 followed by an exon, in which the TIS was located, as shown in the frog. In addition, such an organization is similar to that of all SCPPs, which adds support to our hypothesis. Then, the second noncoding exon (*ENAM* exon 2) was recruited in the amniote lineage, prior to the divergence of sauropsids and mammals. The reason of the presence of the second noncoding exon is still obscure (Al-Hashimi et al. 2009). This exon was conserved in lepidosaurs (lizard) and mammals, whereas lost in crocodiles. This hypothesis seems more parsimonious than recruitment of this exon independently in both the lepidosaurian and the mammalian lineages. However, it is difficult to confirm homology of lizard and mammalian exon 2 by sequence similarities because these two noncoding exons have evolved separately for 310 My (Hedges 2002). Finally, in an ancestral mammal, the third exon (exon 3) housing a TIS was recruited probably from a duplication of the ancestral exon 2 of this gene, as suggested by the number of amino acids from the methionine codon to the last codon encoded in the exon (18 aa) and the following phase 0 intron.

The Additional Coding Exon 8b

When compared with mammals, the two reptilian *ENAM*s exhibit an additional coding exon, exon 8b. We showed that this exon is probably present in marsupial *ENAM*, whereas absent in monotremes and placentals. In the latter, however, a pseudoexon 8b, that is, a no longer functional exon, is still detectable in many species, except in rodents. In the frog, the only representative of the amphibian lineage in this study, exon 8b, was not identified. In tetrapods, the following evolutionary scenario for exon 8b could be proposed (fig. 7): 1) in the last common tetrapod ancestor of amphibians and amniotes, the coding sequence of *ENAM* was composed by seven exons, and this organization was conserved in the amphibian lineage; 2) a duplication of either exon 8 or exon 9 occurred in the amniote lineage leading to the creation of exon 8b. This exon was still present in the *ENAM* sequence when the two amniote lineages, sauropsids and mammals, diverged. Exon 8b was conserved in sauropsid *ENAM*, and even in the last toothed common ancestor of modern birds; and 3) in mammals, this exon was invalidated early in the monotreme lineage that separated from the therian lineage 220 Ma. A long period from the first invalidation event may explain why no pseudoexon 8b was recognized in platypus *ENAM*. In the therian lineage, exon 8b was conserved in the marsupial lineage, but invalidated, probably later, in the common ancestor of the extant placental lineages that diverged 100 Ma (Murphy et al. 2001; Hedges 2002) (fig. 7).

The presence of exon 8b in these modern species indicates that this exon is of some biological importance in both sauropsids and marsupials, otherwise it would have accumulated mutations and invalidated during the 310 My of sauropsid evolution (Hedges 2002) or the 190 My of marsupial evolution (van Rheede et al. 2006). Five residues are unchanged when comparing the four available sequences in the crocodile, the lizard, and the two marsupials.

Frog and Reptilian ENAMs Support the Putatively Important Function of Some Residues

The comparison of frog and reptilian ENAM sequences with sequences of representative mammalian ENAMs revealed that 47 amino acids have been unchanged for 350 My of tetrapod evolution. Some of these important residues are regrouped into motifs. When compared with our data set of ENAM sequences in mammals (36 species; Al-Hashimi et al. 2009), we found that 25 of these positions are unchanged in all ENAMs studied so far (fig. 4). These data add more weight to our recent findings, suggesting that such conserved positions are of high biological importance in mammals (Al-Hashimi et al. 2009). In addition to elucidate the putative ancestral condition of tetrapod ENAM, these new data allow us to predict that amino acid substitutions in these unchanged positions would lead to an ENAM-associated genetic disease (type 2 amelogenesis imperfecta: AIH2). Many single amino acid substitutions, which either reduce the efficiency of the protein or lead to important disorder, have already been observed in many proteins including amelogenin (Delgado et al. 2005, 2008) and alcohol dehydrogenase (Chen et al. 2009).

Most of the 25 well-conserved amino acids identified in tetrapod ENAM belong to the 32-kDa fragment, a keystone of this protein (Al-Hashimi et al. 2009). The main role of

this peptide is probably to initiate enamel mineralization (Tanabe et al. 1990; Uchida et al. 1991; Yamakoshi 1995; Hu and Yamakoshi 2003). In mammals, the 32-kDa fragment contains two phosphorylated serines and three glycosylated asparagines that do have important functions, including, among others, adsorption onto apatite crystals (phosphorylated Ser) and protection against precocious degradation by MMP20 (glycosylated Asn) (Hu and Yamakoshi 2003). Our study supports such important biological functions in showing that the corresponding sequence of porcine 32 kDa in the frog and the two reptiles possess these two serines and at least two asparagines at the right place. It is worth to note that the replacement of one of these phosphorylated serines by a leucine (p.S216L) was recently reported to lead to amelogenesis imperfecta (Chan et al. forthcoming).

Chicken *ψENAM* and Lizard ENAM Location Support Translocation of EMPs

In the chicken, *ψENAM* is present on chromosome Z. This location was unexpected from gene synteny observed for the surrounding region of mammalian *ENAM*. This explains why our previous search for *ENAM* in the target region of chr. 4 was unsuccessful (Sire et al. 2008). Chicken *ψENAM* would not be discovered without the knowledge of the location of *ENAM* in scaffold 132 of the currently assembled lizard genome, between *LPL* and *NRG1*. Although confirmation of this location is needed through both the annotation of these genes on a lizard chromosome and the location of the *ENAM* gene in the crocodile genome, one could suspect that the EMP gene cluster was translocated in sauropsids from a chromosome homologous to, for example human chr. 4, to the current location on chr. Z in the chicken. Indeed, in the frog, EMP genes reside in a region homologous to a region of human chr. 4, a finding that strongly supports translocation in sauropsids. In order to confirm this hypothesis, further investigations are necessary; for instance, finding the EMP genes in another amphibian lineage, for example in caudates, the sister group to the frogs, and/or better annotating lizard and frog chromosomes.

Chicken *ψENAM* sequence exhibits many substitutions and indels that occurred randomly for about 100 My. Random occurrence of mutations in this invalidated gene explains why nucleotide sequence similarity of each exons is low when compared with crocodile *ENAM*, the closest living relative of birds. Both the chicken and the crocodile are terminal taxa of lineages that separated 250 My (Hedges 2002). However, presence of pseudoexon 8b sequence in chicken suggests that this exon was functionally important in the lineage leading to modern birds until their tooth loss, for about 150 My after the separation of the bird and crocodile lineages.

Data Deposition

GenBank accession no. for *Xenopus (Silurana) tropicalis* enamelin mRNA = EU642606; *Anolis carolinensis* enamelin mRNA = GU198361; *Crocodylus niloticus* enamelin mRNA = GU344683; *Gallus gallus* enamelin pseudogene = GU198360

Supplementary Material

Supplementary materials S1-S7 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

We are grateful to S. Martin, director of "La ferme aux crocodiles," in Pierrelatte, France, for the generous gift of a juvenile Nile crocodile and to K. Daouès, director of "La ferme tropicale," Paris, who gave us green anole lizards.

References

Al-Hashimi N, Sire J-Y, Delgado S. 2009. Evolutionary analysis of mammalian enamelin, the largest enamel protein, supports a crucial role for the 32 kDa peptide and reveals selective adaptation in rodents and primates. *J Mol Evol*. 69:635–656.

Chan H-C, Mai L, Oikonomopoulou A, Chan HL, Richardson AS, Wang S-K, Simmer JP, Hu JC-C. Forthcoming. Altered enamelin phosphorylation site causes amelogenesis imperfecta. *J Dent Res*.

Chen Y-C, Peng G-S, Wang M-F, Tsao T-P, Yin S-J. 2009. Polymorphism of ethanol-metabolism genes and alcoholism: correlation of allelic variations with the pharmacokinetic and pharmacodynamic consequences. *Chem Biol Interact*. 178:2–7.

Davit-Béal T, Tucker T, Sire J-Y. 2009. Loss of teeth and enamel in tetrapods: fossil record, genetic data and morphological adaptations. *J Anat*. 214:277–501.

Delgado S, Casane D, Bonnaud L, Laurin M, Sire J-Y, Girondot M. 2001. Molecular evidence for Precambrian origin of amelogenin, the major protein of vertebrate enamel. *Mol Biol Evol*. 18(12):2146–2153.

Delgado S, Girondot M, Sire J-Y. 2005. Molecular evolution of amelogenin in mammals. *J Mol Evol*. 60(1):12–30.

Delgado S, Vidal N, Véron G, Sire J-Y. 2008. Amelogenin, the major protein of tooth enamel: a new phylogenetic marker for ordinal mammal relationships. *Mol Phylogenet Evol*. 47:865–869.

Deméré TA, McGowen MR, Berta A, Gatesy J. 2008. Morphological and molecular evidence for a stepwise evolutionary transition from teeth to baleen in mysticete whales. *Syst Biol*. 57:15–37.

Donoghue PCJ, Sansom IJ. 2002. Origin and early evolution of vertebrate skeletonization. *Microsc Res Tech*. 59:185–218.

Donoghue PCJ, Sansom IJ, Downs JP. 2006. Early evolution of vertebrate skeletal tissues and cellular interactions, and the canalization of skeletal development. *J Exp Zool B Mol Dev Evol*. 306B:278–294.

Fisher LW, Fedarko NS. 2003. Six genes expressed in bones and teeth encode the current members of the SIBLING family of proteins. *Connect Tissue Res*. 44(Suppl 1):33–40.

Fukumoto S, Kiba T, Hall B, Iehara N, Nakamura T, Longenecker G, Krebsbach PH, Nanci A, Kulkarni AB, Yamada Y. 2004. Ameloblastin is a cell adhesion molecule required for maintaining the differentiation state of ameloblasts. *J Cell Biol*. 167(5):973–983.

Gibson CW, Yuan ZA, Hall B, et al. (12 co-authors). 2001. Amelogenin-deficient mice display an amelogenesis imperfecta phenotype. *J Biol Chem*. 276(34):31871–31875.

Gilbert W. 1978. Why genes in pieces? *Nature* 271:501.

Gojobori T. 1983. Codon substitution in evolution and the "saturation" of synonymous changes. *Genetics* 105:1011–1027.

Hart PS, Michalec MD, Seow WK, Hart TC, Wright JT. 2003. Identification of the enamelin (g.8344delG) mutation in a new kindred and presentation of a standardized ENAM nomenclature. *Arch Oral Biol*. 48:589–596.

Hedges SB. 2002. The origin and evolution of model organisms. *Nat Rev Genet*. 3:838–849.

Higgins DG, Thomson JD, Gibson TJ. 1996. Using CLUSTAL for multiple sequence alignments. *Meth Enzymol.* 266:383–402.

Hu JC, Hu Y, Smith CE, et al. (11 co-authors). 2008. Enamel defects and ameloblast-specific expression in Enam knock-out/lacz knock-in mice. *J Biol Chem.* 283(16):10858–10871.

Hu JC, Yamakoshi Y. 2003. Enamelin and autosomal-dominant amelogenesis imperfecta. *Crit Rev Oral Biol Med.* 14:387–398.

Hu JCC, Yamakoshi Y, Yamakoshi F, Krebsbach PH, Simmer JP. 2005. Proteomics and genetics of dental enamel. *Cell Tissues Organs.* 181:219–231.

Hurles M. 2004. Gene duplication: the genomic trade in spare parts. *PLoS Biol.* 2(7):e206.

Huysseune A, Takle H, Soenens M, Taerwe K, Witten PE. 2008. Unique and shared gene expression patterns in Atlantic salmon (*Salmo salar*) tooth development. *Dev Genes Evol.* 218:427–437.

Janvier P. 1996. Early vertebrates. Oxford: Clarendon Press.

Jones DT, Taylor WR, Thornton JM. 1992. The rapid generation of mutation data matrices from protein sequences. *Comput Appl Biosci.* 8:275–282.

Kang HY, Seymen F, Lee SK, Yildirim M, Tuna EB, Patir A, Lee KE, Kim JW. 2009. Candidate gene strategy reveals ENAM mutations. *J Dent Res.* 88:266–269.

Kawasaki K. 2009. The SCPP gene repertoire in bony vertebrates and graded differences in mineralized tissues. *Dev Genes Evol.* 219:147–157.

Kawasaki K, Suzuki T, Weiss KM. 2004. Genetic basis for the evolution of vertebrate mineralized tissue. *Proc Natl Acad Sci USA.* 101:11356–11361.

Kawasaki K, Suzuki T, Weiss KM. 2005. Phenogenetic drift in evolution: the changing genetic basis of vertebrate teeth. *Proc Natl Acad Sci USA.* 102:18063–18068.

Kawasaki K, Weiss KM. 2003. Mineralized tissue and vertebrate evolution: the secretory calcium-binding phosphoprotein gene cluster. *Proc Natl Acad Sci USA.* 100:4060–4065.

Kawasaki K, Weiss KM. 2006. Evolutionary genetics of vertebrate tissue mineralization: the origin and evolution of the secretory calcium-binding phosphoprotein family. *J Exp Zool B Mol Dev Evol.* 306:295–316.

Kim JW, Seymen F, Lin BP, Kiziltan B, Gencay K, Simmer JP, Hu JC. 2005. ENAM mutations in autosomal-dominant amelogenesis imperfecta. *J Dent Res.* 84:278–282.

Kosakovsky Pond SL, Frost SD, Muse SV. 2005. HyPhy: hypothesis testing using phylogenies. *Bioinformatics* 21:676–679.

Kozak M. 1981. Possible role of flanking nucleotides in recognition of the AUG initiator codon by eukaryotic ribosomes. *Nucleic Acids Res.* 9:5233–5262.

McCaldon P, Argos P. 1988. Oligopeptide biases in protein sequences and their use in predicting protein coding regions in nucleotide sequences. *Proteins* 4:99–122.

Meredith RW, Gatesy J, Murphy WJ, Ryder OA, Springer MS. 2009. Molecular decay of the tooth gene enamelin (ENAM) mirrors the loss of fenamel in the fossil record of placental mammals. *PLoS Genet.* 5(9):e1000634.

Murphy WJ, Elzirik E, Johnson WE, Zhang YP, Ryder OA, O'Brien SJ. 2001. Molecular phylogenetics and the origin of placental mammals. *Nature* 409:614–618.

Rambaut A. 1996. Se-Al: Sequence alignment editor. Available from: http://tree.bio.ed.ac.uk/software/seal/. Oxford: University of Oxford.

Shintani S, Kobata M, Toyosawa S, Fujiwara T, Sato A, Ooshima T. 2002. Identification and characterization of ameloblastin gene in a reptile. *Gene* 283:245–254.

Shintani S, Kobata M, Toyosawa S, Ooshima T. 2003. Identification and characterization of ameloblastin gene in an amphibian, *Xenopus laevis*. *Gene* 318:125–136.

Sire J-Y, Davit-Béal T, Delgado S, Gu X. 2007. The origin and evolution of enamel mineralization genes. *Cells Tissues Organs.* 186(1):25–48.

Sire J-Y, Delgado S, Fromentin D, Girondot M. 2005. Amelogenin: lessons from evolution. *Arch Oral Biol.* 50:205–212.

Sire J-Y, Delgado S, Girondot M. 2006. Amelogenin story: origin and evolution. *Eur J Oral Sci.* 114(Suppl 1):64–77.

Sire J-Y, Delgado S, Girondot M. 2008. Hen's teeth with enamel cap: from dream to impossibility. *BMC Evol Biol.* 8:e246.

Sire J-Y, Donoghue PCJ, Vickaryous MK. 2009. Origin and evolution of the integumentary skeleton in non-tetrapod vertebrates. *J Anat.* 214:409–440.

Smith CE, Wazen R, Hu Y, Zalzal SF, Nanci A, Simmer JP, Hu JC-C. 2009. Consequences for enamel development and mineralisation resulting from loss of function of ameloblastin and enamelin. *Eur J Oral Sci.* 117:485–497.

Springer MS, Murphy WJ. 2007. Mammalian evolution and biomedicine: new views from phylogeny. *Biol Rev Camb Philos Soc.* 82:375–392.

Tanabe T, Aoba T, Moreno EC, Fukae M, Shimuzu M. 1990. Properties of phosphorylated 32 kd nonamelogenin proteins isolated from porcine secretory enamel. *Calcif Tissue Int.* 46:205–215.

Termine JD, Belcourt AB, Christner PJ, Conn KM, Nylen MU. 1980. Properties of dissociatively extracted fetal tooth matrix proteins. I. Principal molecular species in developing bovine enamel. *J Biol Chem.* 255:9760–9768.

Toyosawa S, O'HUigin C, Figueroa F, Tichy H, Klein J. 1998. Identification and characterization of amelogenin genes in monotremes, reptiles, and amphibians. *Proc Natl Acad Sci USA.* 95:13056–13061.

Uchida T, Tanabe T, Fukae M, Shimizu M. 1991. Immunocytochemical and immunochemical detection of a 32 kDa nonamelogenin and related proteins in porcine tooth germs. *Arch Histol Cytol.* 54:527–538.

van Rheede T, Bastiaans T, Boone DN, Hedges SB, de Jong WW, Madsen O. 2006. The platypus is in its place: nuclear genes and indels confirm the sister group relation of monotremes and therians. *Mol Biol Evol.* 23:587–597.

Yamakoshi Y. 1995. Carbohydrate moieties of porcine 32 kDa enamelin. *Calcif Tissue Int.* 56:323–330.

J Mol Evol (2009) 69:635–656
DOI 10.1007/s00239-009-9302-x

Evolutionary Analysis of Mammalian Enamelin, The Largest Enamel Protein, Supports a Crucial Role for the 32-kDa Peptide and Reveals Selective Adaptation in Rodents and Primates

Nawfal Al-Hashimi · Jean-Yves Sire · Sidney Delgado

Received: 3 July 2009 / Accepted: 6 November 2009 / Published online: 30 November 2009

Abstract Enamelin (ENAM) plays an important role in the mineralization of the forming enamel matrix. We have performed an evolutionary analysis of mammalian ENAM to identify highly conserved residues or regions that could have important function (selective pressure), to predict mutations that could be associated with amelogenesis imperfecta in humans, and to identify possible adaptive evolution of ENAM during 200 million years ago of mammalian evolution. In order to fulfil these objectives, we obtained 36-ENAM sequences that are representative of the mammalian lineages. Our results show a remarkably high conservation pattern in the region of the 32-kDa fragment of ENAM, especially its phosphorylation, glycosylation, and proteolytic sites. In primates and rodents we also identified several sites under positive selection, which could indicate recent evolutionary changes in ENAM function. Furthermore, the analysis of the unusual signal peptide provided new insights on the possible regulation of ENAM secretion, a hypothesis that should be tested in the near future. Taken together, these findings improve our understanding of ENAM evolution and provide new information that would be useful for further investigation of ENAM function as well as for the validation of mutations leading to amelogenesis imperfecta.

Keywords Enamelin · Evolution · Teeth · Mammals · Purifying selection · Positive selection

Introduction

Teeth are constructed with remarkably tough materials making them the most resistant part of the body toward destruction. They are crucial for animal survival and make us realize the importance of natural selection preserving the integrity of its components throughout evolution. Their loss or malfunction would lead to animal death (Davit-Béal et al. 2009). Enamel, which protects teeth, is the hardest mineralized tissue in mammals. Its uniqueness is reflected in the tissue specificity of its main matrix constituents: amelogenin (AMEL), ameloblastin (AMBN), and enamelin (ENAM) (Termine et al. 1980; Deutsch 1989). These three proteins constitute the Enamel Matrix Protein (EMP) family. Evolutionary analyses showed that EMPs are ancient and evolutionarily related (Delgado et al. 2001; Sire et al. 2006), and they belong to a large family of proteins that link calcium, the Secretory Calcium binding PhosPhoproteins (SCPPs) (Kawasaki and Weiss 2003). AMEL is the major protein of the forming enamel (about 90% of the total organic matrix in bovine; Termine et al. 1980) and the two other enamel proteins (AMBN and ENAM) combine with various minor components to account for the remaining matrix. In the previous years, using a large dataset of mammalian sequences, our group brought information on the evolution and relationships of AMEL (Delgado et al. 2001; Sire et al. 2005, 2006, 2007). Two additional studies dealing with AMELX evolution in mammals, showed how an evolutionary analysis is particularly useful to (i) identify conserved and variable positions (which indicates either strong or relaxed functional

Electronic supplementary material The online version of this article (doi:10.1007/s00239-009-9302-x) contains supplementary material, which is available to authorized users.

N. Al-Hashimi · J.-Y. Sire (✉) · S. Delgado
Université Pierre et Marie Curie, UMR 7138—Systématique, Adaptation, Evolution, Case 5, 7 Quai Saint-Bernard, Bâtiment A, 4e étage, 75005 Paris, France
e-mail: jean-yves.sire@upmc.fr

constraints, respectively), (ii) understand the mode of evolution, and (iii) find evolutionary relationships with the other EMPs (Sire et al. 2006, 2007). The first point is of importance as it allowed to validate AMEL mutations responsible for X-linked amelogenesis imperfecta (AIH1) and also to predict AIH1-associated mutations (Delgado et al. 2007; see also Subramanian and Kumar 2006; Springer and Murphy 2007).

This study is devoted to ENAM, a phosphorylated, enamel-specific glycoprotein that plays a key role in enamel formation. Although the precise functions of ENAM are not clearly known, it is clear that this EMP is necessary for proper enamel formation as demonstrated with ENAM-null mice. Their teeth have virtually no enamel as the deposited matrix does not mineralize (Hu et al. 2008). In humans, the important role of ENAM is manifested indirectly by mutations that lead to type-2 autosomal dominant amelogenesis imperfecta (AIH2) (Dong et al. 2000; Hu et al. 2000; Rajpar et al. 2001; Hart et al. 2003a, b; Kim et al. 2005a, b).

Only four *ENAM* mRNA sequences were published so far: pig (Hu et al. 1997a), mouse (Hu et al. 1998), human (Hu et al. 2001), and rat (Hu and Yamakoshi 2003). These sequences were not representative of the mammalian ENAM diversity, and therefore did not allow an evolutionary analysis; indeed, such an analysis needs a large dataset of sequences obtained in species representative of the main lineages in the current mammalian phylogeny. ENAM is supposed to be the ancestor of EMPs, which were probably present at the onset of vertebrates, 500 million years ago (Ma) (Delgado et al. 2001; Sire et al. 2006). Therefore, ENAM could be the keystone for enamel evolution, especially when considering the appearance of the first dental tissues, approximately 450 Ma (Sansom et al. 1992). Moreover, ENAM was identified as a target of recent positive selection among human populations (Kelley and Swanson 2008).

The objectives of our evolutionary analysis of mammalian ENAMs were (i) to identify highly conserved amino acids and regions, as such residues or peptides could play an important role for ENAM function, and hence to use these results to predict AIH2-associated mutations in humans; and (ii) to identify possible adaptive evolution of ENAM during approximately 220 Ma of mammalian evolution (van Rheede et al. 2006). To reach these objectives we obtained and analyzed ENAM sequences from 36 species, representative of all main mammalian lineages. ENAM sequences of toothless and enamel-less species (e.g., pangolins, xenarthrans, and aardvark; Davit-Béal et al. 2009) were excluded from this study as possible bias could be introduced given the relaxed functional pressure on this protein (see Deméré et al. 2008; Sire et al. 2008).

Material and Methods

Database Search

The four published mammalian *ENAM* nucleotide sequences (human, pig, mouse, and rat) were extracted from databases, including the untranslated regions (UTR). We then searched *Ensembl* (www.ensembl.org) and found 19 other ENAM sequences. They were computer-predicted sequences from the automatic analysis of sequenced mammalian genomes. Several sequences were not complete and some contained errors. In order to work with the most complete dataset, we have used the published sequences to search the sequenced genomes (*Ensembl*) with BLAST. We completed some sequences and found three other *ENAMs*. The final step was to search *NCBI Trace archives* (www.ncbi.nlm.nih.gov/BLAST) with the aim of filling in some gaps in the sequences and also to find other sequences in the genomes that were in course of sequencing. Ten new *ENAMs* were added to our dataset. Our last access to the databases was on January 2009. A total of 36 *ENAMs* were used for the evolutionary analysis (Table 1). Among them, 24 nucleotide sequences were complete.

The positions with missing data were included in our analysis and treated "unknown data". There were 2,048 unknown nucleotides only versus a total of 43,068, which means that less than 5% of the data were missing in our analysis. In addition, no bias resulted from these missing data because there was always at least one complete sequence from a species representative of the mammalian lineage in which the data were missing.

Sequence Alignment

The coding region of the published *ENAM* sequences were translated into putative amino acid sequences, aligned to the human sequence using *Clustal X 2.0* (Higgins et al. 1996), and checked by hand using *Se-Al v.2.0a11* software (http://tree.bio.ed.ac.uk/software/seal) (Rambaut and Bromham 1998). The new nucleotide sequences found in *Ensembl* and *NCBI* databases were aligned to the published sequences used as templates. Unchanged residues during mammalian evolution were identified by hand in the final alignment.

Purifying Selection Analysis Using Selecton Method

The identification of site-specific purifying selection (i.e., biologically significant amino acids) in ENAMs was carried out using the *Selecton Server 2.2* (http://selecton.tau.ac.il), in which ML estimates d_N/d_S for each site as described above (Doron-Faigenboim et al. 2005; Stern et al. 2007). In this method, the analysis is performed by means of a comparison between a null model assuming no

Table 1 Names of the 36 species used in our evolutionary analysis of mammalian enamelins (ENAM) and references

Common name	Genus and species	Order	Source
Baboon	*Papio hamadryas*	Primates	GQ352330
Bushbaby	*Otolemur garnettii*	Primates	GQ352331
Cat	*Felis catus*	Carnivora	GQ352332
Chimpanzee	*Pan troglodytes*	Primates	GQ352333
Cow	*Bos taurus*	Cetartiodactyla	GQ352334
Dog	*Canis familiaris*	Carnivora	GQ352335
Dolphin	*Tursiops truncatus*	Cetartiodactyla	GQ352336
Elephant	*Loxodonta africana*	Afrotheria	GQ352337
Fruit bat	*Pteropus vampyrus*	Chiroptera	GQ352338
Gibbon	*Nomascus leucogenys*	Primates	GQ352339
Gorilla	*Gorilla gorilla*	Primates	GQ352340 and EU482103
Guinea pig	*Cavia porcellus*	Rodentia	GQ352341
Hedgehog	*Erinaceus europaeus*	Insectivora	GQ352342
Horse	*Equus caballus*	Perissodactyla	GQ352343
Human	*Homo sapiens*	Primates	NM_031889
Hyrax	*Procavia capensis*	Afrotheria	GQ352344
Kangaroo rat	*Dipodomys ordii*	Rodentia	GQ352345
Marmoset	*Callithrix jacchus*	Primates	GQ352346
Microbat	*Myotis lucifugus*	Chiroptera	GQ352347
Mouse	*Mus musculus*	Rodentia	NM_017468
Mouse lemur	*Microcebus murinus*	Primates	GQ352348
Opossum	*Monodelphis domestica*	Didelphimorpha	GQ352349
Orangutan	*Pongo pygmaeus*	Primates	GQ352350
Pig	*Sus scrofa*	Cetartiodactyla	NM_214241
Pika	*Ochotona princeps*	Lagomorpha	GQ352351
Platypus	*Ornithorhynchus anatinus*	Monotremata	GQ352352
Rabbit	*Oryctolagus cuniculus*	Lagomorpha	GQ352353
Rat	*Rattus norvegicus*	Rodentia	NM_001106001
Rhesus monkey	*Macaca mulatta*	Primates	GQ352354
Shrew	*Sorex araneus*	Insectivora	GQ352355
Squirrel	*Spermophilus tridecemlineatus*	Rodentia	GQ352356
Tarsier	*Tarsius syrichta*	Primates	GQ352357
Tenrec	*Echinops telfairi*	Afrotheria	GQ352358
Tree shrew	*Tupaia belangeri*	Scandentia	GQ352359
Vicugna	*Vicugna vicugna*	Cetartiodactyla	GQ352360
Wallaby	*Macropus eugenii*	Diprotodontia	GQ352361

Out of these, 32 new ENAM sequences were obtained in this study. The species are arranged in alphabetical order of common names

positive selection and a model that allows positive selection. The results were then displayed on the human sequence. The *Selecton Server* program uses a color-coding scheme to represent the different levels of selection (purifying selection, positive selection, or lack of selection). For convenience, we have chosen the two highest levels of purifying selection as they allow a clear interpretation of the results.

Selective Pressure Analysis Using the Hyphy Method

In the analyses described below, four values were computed for every variable site: observed and normalized expected numbers of synonymous (N_S and E_S) and non-synonymous (N_N and E_N) substitutions. HyPothesis Testing Using Phylogenies (for *Hyphy*) software (http://hyphy.org; Pond et al. 2005a, b, c), or its improved online version, Single Likelihood Ancestor Counting (for *SLAC*; http://www.datamonkey.org), estimates $d_N = N_N/E_N$ and $d_S = N_S/E_S$. When $d_N > d_S$ a codon is considered positively selected. A P value derived from a two-tailed extended binomial distribution was used to assess significance. It is worthy to note that the extended binomial distribution is an approximation to the true distribution of non-synonymous and synonymous under the hypothesis of neutrality (Pond et al. 2005a, b, c). The model assumes that under neutrality a random

substitution will be synonymous with probability $P = E_S/(E_S + E_N)$, and computes how likely P is when N_S, out of $N_N + N_S$ substitutions, are synonymous. SLAC uses a P value of the extended binomial distribution that is different from the P values that derive from a simulation of the null distribution (i.e., $d_N = d_S$).

The parameter chosen for significance level was 0.1. Indeed, given our dataset of 37 sequences, such a P value is considered appropriate to detect true positives in datasets containing more than 30 sequences, and *SLAC* is one of the only methods that allow a high P value (Kosakovski et al. 2005b).

Sliding Window Analysis

In order to identify strong functional constraints a sliding window analysis of nucleotide sequence variability was conducted on *ENAM* alignment using *HyPhy*. The mean substitution rate was calculated using the maximum likelihood (ML) method based on HKY 85 model (Hasegawa, Kishino, Yano 1985). In contrast to other methods that use a sliding window analysis with large window sizes (e.g., Endo et al. 1996; Tsunoyama and Gojobori 1998; Schmid and Yang 2008), *Hyphy* utilizes the Ln likelihood to measure the selective pressure. At each position, the probability for the observed data is calculated by the likelihood algorithm taking into account the phylogenetic relationships. Then, the logarithm (Ln) of the product of the probabilities is calculated for a window of 15 bp with an overlap of 5 bp between windows. Indeed, when applying the *Hyphy* method it is not necessary to use large sliding windows and it is even recommended to avoid a "smoothing" effect, i.e., a loss of evolutionary information. In addition, when using the Ln likelihood, the evolution rate in a given sequence is not represented by a rate of change but by a probability. This value is more interesting as it does not need to take into account numerous parameters and does not need to identify non-synonymous and synonymous mutations.

We performed the analysis with and without the divergent platypus sequence.

Substitution Rate Analysis

The codon-based ML *SLAC* method was used to identify accurate regions with high selective constraints (low amino acid substitution rate) (Kosakovski et al. 2005a, b). Non-synonymous substitutions (d_N) were estimated at every site of the alignment and compared to normalized expected number of non-synonymous substitutions.

Positive Selection

SLAC was used to determine site-specific positive Darwinian selection by estimation of the normalized d_N/d_S ratio at every site of the alignment. A positively selected codon is identified when $d_N > d_S$. These evolutionary events were localized on the mammalian phylogeny under *MacClade* environment (see below) using the "trace character" option. We retained mutations that only occurred in a mammalian lineage and that were conserved afterwards. Mutations that appeared in terminal branches were considered as not being informative enough.

Distance Matrix

The evolutionary distance of the *ENAM* nucleotide sequences taken two by two was calculated using *HyPhy* with Tamura and Nei (1993) distance algorithm. This was particularly useful for the evolutionary analysis in determining the relevance of highly variable sequences.

Phylogenetic Tree

Gaps were eliminated from our alignment (11 residues removed out of 1,164 positions in the alignment). Then these sequences were transferred to *MacClade 4.08* (http://macclade.org) in order to place ENAM sequences into recent mammalian phylogeny (Springer and Murphy 2007). The phylogenetic tree was computed using the neighbor-joining method and the distances were estimated using pairwise ML parameter estimation, under Dayhoff model for amino acid substitutions (Dayhoff et al. 1978).

Deduction of the Putative Ancestral Sequence

The alignment (nucleotide sequences) of the 36 ENAM sequences (i.e., including platypus) was transferred into *HyPhy* from a nexus file, along with mammalian phylogeny (Springer and Murphy 2007). The putative ancestral characters were calculated at each node using the following parameters: optimality criterion = ML, substitution model = Dayhoff with local parameters estimated from dataset, frequencies estimated via ML, among-site rate variation assumed, proportion of invariable sites estimated by ML, and gamma distribution of rates at variable sites (discrete approximation). The sequence at the base of the tree nodes was retained as the putative ancestral mammalian sequence, i.e., at the time the monotreme and therian lineages diverged (210–220 Ma; van Rheede et al. 2006).

Analysis of the 5′-UTR

The putative 5′-UTRs, i.e., exon 1, exon 2 and the 5′-region of exon 3 upstream the translation initiation site (TIS) of ENAM, were aligned against the four published sequences using *Clustal X 2.0* software and checked by hand using

Se-Al v.2.0a11 software. The goal was to check whether or not these putative transcripts could include exon 2.

Analysis of Translation Initiation Sites

Because most 5′-UTRs contained several TIS, we used *DNA functional site miner* (http://dnafsminer.bic.nus.edu.sg/) to predict which TIS in each ENAM sequence were functional, i.e., possessing the highest Kozak (1984) consensus score.

Signal Peptide (SP) Analysis and Cleavage Site

For each sequence, the putative SP identified from the TIS analyses were analyzed using *SignalP 3.0 server* (www.cbs.dtu.dk/services/SignalP). This software predicts the location of the three characteristic regions (n-, h-, and c- regions) of a SP, the putative cleavage site of the protein, and calculates the probability for each SP to be functional.

Amino Acid Ratio

The percentage of amino acids [with particular focus on proline (P) and glutamine (Q)], and their mean values (P, Q, and P + Q) were calculated using *Microsoft Excel*.

Results

Brief Overview of ENAM

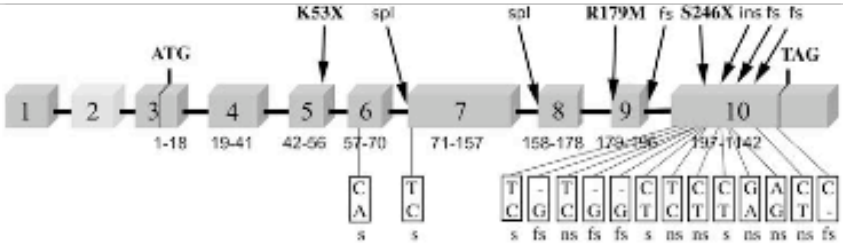
Extensive sequence analysis of pig ENAM was first reported by Fukae et al. (1996). This is the largest EMP, but it is less abundant than the other structural proteins in forming enamel. Murine *ENAM* transcripts display ten exons, including two non-coding exons in the 5′-UTR (Hu et al. 2001). In human DNA, the *ENAM* gene spans 18 kb on the chromosome 4 at 4q11–21. The transcripts show nine exons with a single non-coding exon 1. However, a region corresponding to the non-coding murine exon 2 (i.e., with a similar sequence and appropriate splice sites) is present in human intron 1. As exon 2 is absent from the few human *ENAM* transcripts sequenced so far, it was hypothesized that this exon could be subjected to splicing in most transcripts (Hu et al. 2001). Therefore, unlike the other SCPP members, *ENAM* is the only SCPP gene that possesses (at least in rodents) two non-coding exons (exons 1–2) at the 5′-UTR (Kawasaki and Weiss 2003).

Although exon 2 was not reported in *ENAM* transcripts of non-rodent species, the nomenclature took this exon into consideration. *ENAM* exons were numbered from 1 to 10. In humans and pigs, the eight exons following exon 2 encode a protein with 1,142 amino acids (aa) including the leader peptide. Most of these residues (946 aa) are encoded by the large exon 10 (Fig. 1). Recently, the *ENAM* promoter region and potential transcription factor-binding sites (Runx2, Dlx, and Msx) were found in the mouse (Hu et al. 2008). A 5.2-kb region located upstream on the translation site appears necessary for appropriate ENAM expression.

Another uncommon feature of ENAM is the SP, which consists of 39 aa (38 aa in the pig), encoded by exons 3 and 4 (Fukae et al. 1996). Such a large SP is different from all mammalian SCPPs, in which it is composed of 16 aa (Kawasaki and Weiss 2003; Sire et al. 2005, 2006).

Like in all *SCPPs*, *ENAM*-coding exons have type-0 splice junctions, meaning that none of the introns splits a codon, allowing exons to function as modules (Kawasaki and Weiss 2003). Therefore, skipping exons by alternative splicing would not shift the reading frame. Despite this feature, no alternative splicing has ever been identified in the few ENAM cDNAs analyzed so far.

ENAM is characterized, as the other EMPs, by a proline-rich domain. It is located near the N-terminal region (residues 86–189) and is encoded by exons 7, 8, and 9 (Fig. 2). Full-length ENAM has an apparent molecular mass of 186 kDa, but it is rapidly degraded or reabsorbed once secreted (Hu et al. 1997b). In the forming porcine enamel, several ENAM fragments resulting from the proteolysis were isolated as 155-, 142-, 89-, 34-, 32- and 25-kDa peptides (Fukae and Tanabe 1985, 1987; Uchida et al. 1991a; Tanabe et al. 1994; Fukae et al. 1996) (Fig. 2). ENAM processing during enamel formation is probably performed by enamelysin (MMP 20), as this is the most predominant protease at this stage (Ryu et al. 1999). The large molecular-weight cleavage products are found in the superficial layer of enamel matrix, at the mineralization front, near Tomes' processes. This suggests a possible role in nucleation and extension of mineral crystals (Hu et al. 1997b; Masuya et al. 2005). The 89-kDa fragment consists of a 627 amino acid peptide. It is supposed to control the growth of apatite crystals and to inhibit nucleation of new crystallites (Fukae et al. 1996). Numerous hydrophobic amino acids are present in this fragment, with a highly hydrophobic region composed of 21–62 aa (Fukae et al. 1996). In porcine ENAM, three phosphorylation sites (serines), and three glycosylation sites (asparagines) were identified in the N-terminal region, and the presence of six cysteines in the C-terminal region suggests possible disulfide bridges (Fig. 2). Five other asparagines are supposed to be glycosylated (Hu et al. 2005). Two fragments of 25 and 34-kDa were isolated from the outer, thin layer of porcine enamel (Fukae et al. 1996), and a 32-kDa fragment accumulates in the entire thickness of the enamel matrix at advanced stage of mineralization (Tanabe et al. 1990; Uchida et al. 1991b; Yamakoshi 1995; Hu and Yamakoshi 2003). The three glycosylated asparagines and two phosphorylated serines are located in this 32-kDa fragment that



Gene	cDNA	Protein	References
g.2382A>T	c.157A>T	p.K53X	Mardh et al., 2002
g.4806A>C		p.M71-Q157del	Kim et al., 2005
g.6331G>A		p.A158-Q178del	Rajpar et al., 2001
g.8291G>T	c.536G>T	p.R179M	Gutierrez et al., 2007
g.8344delG		p.N197fsX277	Kida et al., 2002
g.12663C>A	c.737C>A	p.S246X	Ozdemir et al., 2005
g.12946-12947 insAGTC AGTACCAGTACTGTGTC	c.1020-1021 insAGTCA GTACCAGTACTGTGTC	p.V340-M341 ins SQYQYCV	Ozdemir et al., 2006
g.13185-13186 ins AG	c.1258-1259 ins AG	p.P422fsX448	Hart et al., 2003
g.14917delT	c.2991delT	p.L998fsX1062	Kang et al., 2009

Fig. 1 *Top ENAM* gene structure with indication of the nine reported mutations leading to type-2 autosomal-dominant amelogenesis imperfecta (*arrows*). *ENAM* exons are numbered from 1 to 10. Exon 2, in *light gray*, was not found in published human and pig *ENAM* transcripts. The number of encoded amino acids is indicated below each exon, and the start and stop codons are shown. SNPs registred in databases for the coding *ENAM* regions are also figured in *small boxes*. *Bottom* nomenclature of the nine *ENAM* mutations leading to amelogeneis imperfecta type-2 in human (modified after Kim et al. 2005a, b). The nomenclature was modified to fit with the following official reference sequences: *ENAM* gene, NC_000004, 1–18076 bp; *ENAM* transcript, NM_031889, ATG = +1; ENAM protein, NP_114095, Met[1] = +1. *fs* frameshift, *ins* insertion, *ns* non-synonymous substitution, *s* synonymous substitution, *spl* splicing site

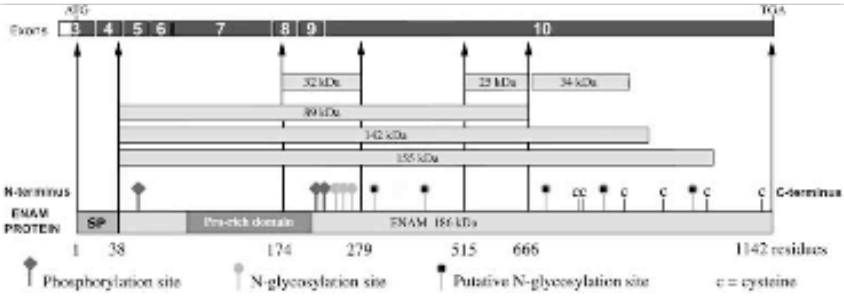


Fig. 2 Diagram showing the cleavage products of pig ENAM and their corresponding location on the protein sequence (encoded by exons 3–10). The three phosphorylated residues, the three *N*-glycosylated residues, the five putative *N*-glycosylated residues, and the six cysteines are indicated. The first 38 aa constitute the signal peptide. The secreted protein has an apparent molecular weight of 186 kDa; partially characterized ENAM cleavage products are the 155-, 142-, 89-, 34-, 32- and 25-kDa fragments (after Fukae et al. 1996). The proline-rich domain is indicated

consists of 106 aa (L^{174}–R^{279}) in the pig, the only species in which such data are available so far (Yamakoshi 1995; Fukae et al. 1996; Yamakoshi et al. 1998). This peptide is encoded by exons 8, 9, and the beginning of exon 10, which represents less than one-tenth of the full-length protein. However, this is the most stable ENAM fragment that remains after MMP20 proteolysis and it appears as a good candidate for controlling crystal nucleation or growth as it possesses high affinity to bind apatite crystals (Tanabe et al. 1990). We will see that our evolutionary analysis allowed us to predict which positions are important for the correct function of the protein.

Although the role of the 32-kDa fragment remains poorly understood, it may potentially bind AMEL (Ravindranath et al. 1999) or its small, 20-kDa AMEL, cleavage product (Yamakoshi et al. 2003). This supports the hypothesis by Weiner (1986) who suggested that hydrophobic molecules (like AMEL) could provide a skeletal or space-filling structure, while (acidic) hydrophilic molecules (like ENAM and AMBN) could be involved in the regulation and growth of crystal nucleation. AMEL fragments show low adsorption affinity onto the crystallites (Aoba and Moreno 1987), while the 32-kDa ENAM, which accumulates at the same time, adsorbs strongly onto apatite crystals (Tanabe et al. 1990; Brookes et al. 2002). This suggests a dual role between AMEL and ENAM in the regulation of enamel mineralization. Recently, a study of the secondary structure of the 32 kDa fragment revealed that it has a high content of alpha-helix and that it undergoes conformational changes with structural preference of beta-sheets when calcium concentration increases, suggesting that this structure improves interactions with the apatite crystal surfaces (Fan et al. 2008). Finally, during enamel maturation stage, MMP20 is not active against the 32-kDa fragment and it is assumed that the terminal proteolysis of this peptide is performed by kallikrein-related peptidase 4 (KLK4) with five cleavages sites identified (Hu and Yamakoshi 2003).

To date, nine different ENAM mutations (either indels or substitutions) leading to AIH2 have been reported in the literature (Fig. 1). They result in several phenotypes, ranging from relatively minor localized enamel pitting to severely hypoplastic enamel (Hart et al. 2003a, b; Kim et al. 2005a, b). Among these mutations, only one is a single amino acid substitution (p.R179M). Experimentally induced mutations in the mouse *ENAM* (using *N*-ethyl-*N*-nitrosourea, a mutagen) resulted in the identification of four mutations with a single-base substitution (Masuya et al. 2005; Seedorf et al. 2007). Two mutations in exon 5 led to non-synonymous substitutions (=p.S54I and p.E56G in human ENAM) with local hypoplastic enamel in the heterozygotes. One mutation in exon 8 led to a stop codon and a mutation in the splicing donor site in intron 4 gave rise to a reading frameshift with premature stop codon. The enamel was hypomature in the heterozygote mice and was lacking in the homozygous mice. These missense mutations are of prime importance in helping the identification of possible crucial function played by specific amino acids and/or regions of ENAM. In identifying conserved residues (i.e., positions subjected to selective pressure) the evolutionary analysis allows to predict AIH2-associated mutations.

Beside these nine reported mutations in humans, 15 single nucleotide polymorphisms (SNPs) are registered in *Ensembl* and *NCBI* databases for the coding regions of *ENAM* (Fig. 1). Out of them, six are non-synonymous substitutions (i.e., changing the residue) and surprisingly, four are indels leading to reading frame shifts, which might result in AIH2. However, the SNPs described in dbSNP without a validation status could also result from sequencing errors.

Comparison and Analysis of our ENAM Dataset

Our dataset of 36 sequences (32 new), among which 24 are complete, is well representative of the mammalian diversity (16 orders and 24 families; Table 1). The amino acids in our alignment were numbered from the initiation start site, i.e., the N-terminal methionine in exon 3 to the last residue preceding the stop codon in exon 10. This alignment resulted in a total of 1,550 positions when including insertions (Supplementary material 1). In the following, if not mentioned otherwise the amino acid positions refer to this alignment.

The comparison of the 24 complete sequences indicates that the *ENAM* structure is well conserved in mammals, with eight coding exons (numbered 3–10). However, although two non-coding exons 1 and 2 are reported in the mouse, exon 2 is probably not transcribed in several species (see below, section "5′-UTR").

The analysis of the intron–exon boundaries in representatives of the main lineages indicates that they are well conserved in all coding sequences (Supplementary material 2, "Prediction and validation of AIH2-associated mutations in human ENAM"). These data are useful to validate AIH2-associated mutations. Indeed, mutations of nucleotides of splice sites account for numerous cases of AIH2 (Fig. 1) and of numerous genetic diseases. Here, we show a high conservation of the consensus sequences of the donor [GT(A/G)AG] and acceptor [(C/T)AG] sites.

Variable Positions

Our alignment indicates numerous variable positions and reveals the presence of a large number of indels (from 1 to 9 residues). Most of these variations are located in exons 7 and 10. In addition, six ENAM sequences possess large insertions (Supplementary material 1). In the pika, 10 residues (mostly prolines) are inserted in the region encoded by exon 9 (positions 200–209). The other insertions are located in exon 10. In the mouse and the rat, 33 nucleotides encoding the motif [VGANPASNKPF] were duplicated 13 and 15 times, resulting in the insertion of 143 and 165 aa, respectively (positions 499–664). In the kangaroo rat, the closest relative of rat and mouse in our analysis, such an insertion was not found. Instead, there was an insertion of four repeats of [ARPGNPT] leading to 28 additional residues (positions 436–460). In the tenrec, the motif

[KEYLTYTLENPSKPR] was duplicated three times with 41 aa inserted (positions 985–1026). In the platypus, we found three repeats of [RPVG] (12 residues inserted; positions 378–389), and an insertion of 60 aa (positions 1271–1331). In contrast to these large insertions, extended deletions (e.g., ≥10 residues) were not found in the sequences analyzed. Taken altogether, the high number of amino acid substitutions, the numerous indels and the large insertions indicate that ENAM regions encoded by exon 10, and to a lesser degree by exon 7, are highly variable, and that exon 10 can include large insertions that have apparently no negative consequence on protein function and enamel structure.

Among ENAMs, the two marsupial (opossum and wallaby) and the monotreme (platypus) sequences are more variable than placental sequences. This could result either from large evolutionary distances between these lineages or from high substitution rates within each lineage. This led us to wonder whether or not these sequences were relevant for our evolutionary analysis of ENAM. In order to answer this question, we quantified the substitution rates in mammalian *ENAMs* using a distance matrix, which allowed us to calculate the pairwise distance within our dataset of sequences taken two by two (Supplementary material 3); the differences between substitution rate values are illustrated by variations in branch lengths in the mammalian phylogenetic tree (Fig. 3). For instance, in primates the substitution rate was low (from 0.009 changes when comparing chimpanzee and human ENAM to 0.141 for tarsier and marmoset ENAM) as illustrated by short branches in the tree. Conversely, in rodents long branches result from substitution rate values comprised between 0.2 and 0.3, reflecting rapidly diverging sequences; the highest values were found with the guinea pig ENAM (Fig. 3). In all placental *ENAMs*, the substitution rate values did not exceed 0.353. When including the two marsupial *ENAMs*, the mean substitution rates increased to 0.479, and with the platypus it was 0.779, a value which falls largely outside the range of the other values (Supplementary material 3), while also causing a long branch in the tree (Fig. 3). Given this difference, we decided to carry out the substitution rate analysis both with and without platypus *ENAM* to minimize the effects of a divergent sequence on the results (see "Selective pressure").

Conserved Positions

Beside the numerous variable positions described above, our ENAM alignment reveals several conserved positions (i.e., unchanged or with a low level of substitution), among which 77 residues were never changed during evolution (Supplementary material 1, see also Fig. 5). We will deal with these conserved residues in the following sections as they indicate high functional constraints on ENAM.

Functional Constraints on the ENAM Regions

The selective forces acting on ENAM were inferred (with and without the platypus sequence) using a sliding window analysis (d_N/d_S ratio) (Fig. 4a). Four regions subjected to high functional constraint (low Ln likelihood values) were

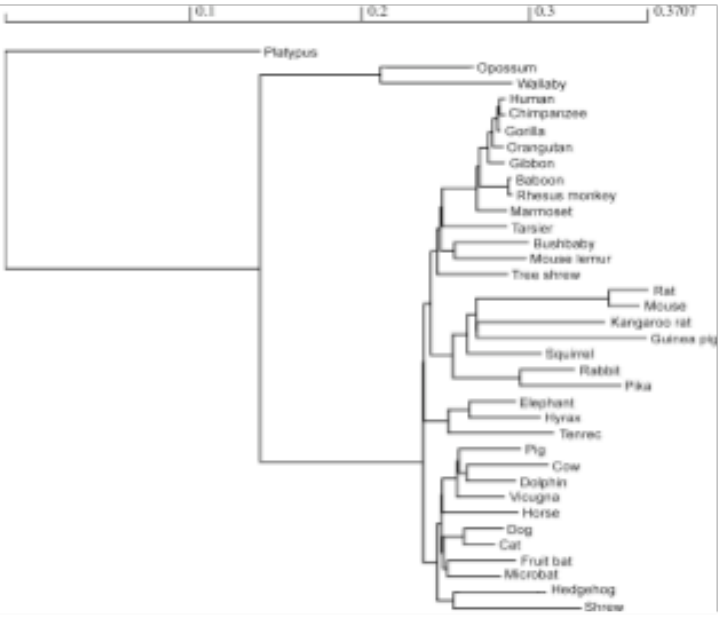


Fig. 3 Maximum likelihood tree obtained under the HKY85 model using our nucleotide dataset to evaluate the evolutionary rate of the 36 mammalian *ENAMs*. The longer the branches, the higher the evolutionary rates were. Branch lengths are indicated at the *top*

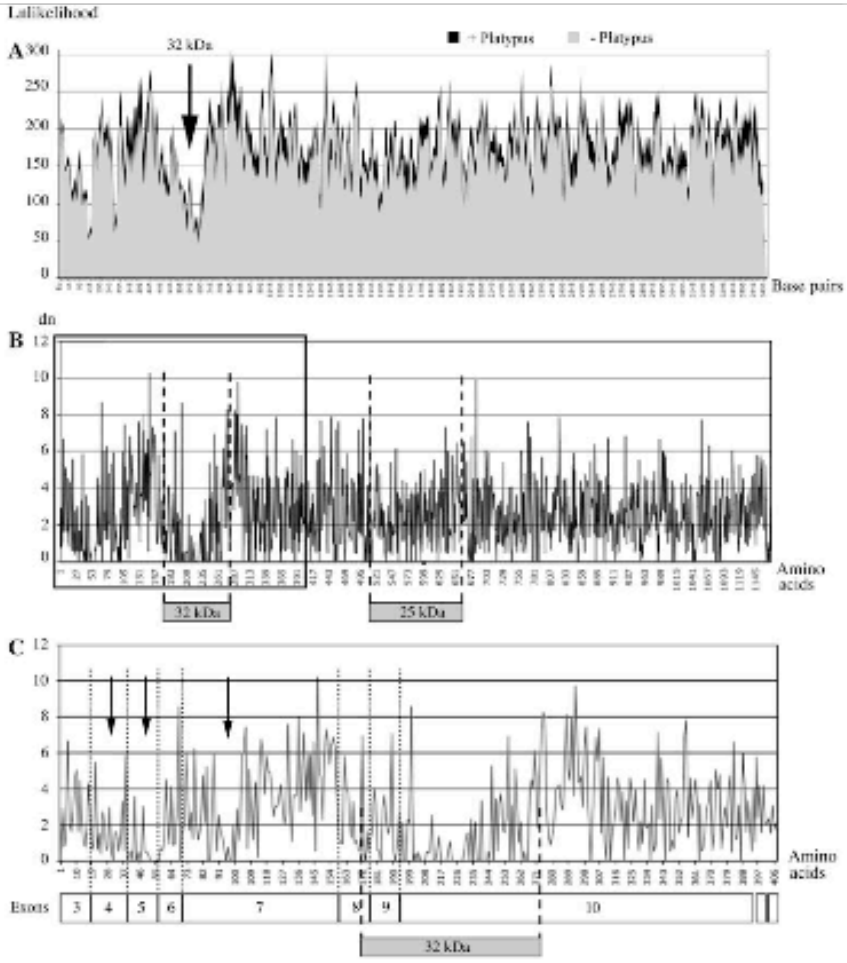


Fig. 4 Substitution rate analysis of the 36 mammalian *ENAMs*. **a** Evolutionary divergence in the full-length nucleotide coding sequences using a Sliding Window analysis. The plot shows the Ln of substitution rate per site along each branch of the mammalian phylogeny estimated for 15-bp windows for every 5 bp. Substitutions were estimated from the best-fit maximum likelihood model under Hasegawa et al. (1985) model. The analysis was performed with (+ platypus) and without platypus (− platypus) *ENAM*. The divergent platypus sequence does not change much the results. Low substitution rates (unchanged base pairs) are identified in the N-terminal region, including the region encoding the 32-kDa fragment. **b** Observed non-synonymous changes versus normalized expected number of non-synonymous changes in the full-length amino acid sequences. The regions of the 25- and 32-kDa fragments are indicated. **c** Enlargement of the *boxed region* in (**b**) to highlight the three regions with low substitution rate encoded by exons 4, 5, and 7 (*arrows*) and the 32-kDa fragment encoded by exons 8, 9, and the beginning of exon 10. These regions are rich in functionally constrained residues and they reflect the loci of high selective pressures

identified in the first quarter of *ENAM* and at the 3′ extremity. In contrast, most of exon 10 sequence, from nucleotide 750 onwards, was characterized by an alternation of high and low selective pressures. The profile of this analysis remained largely unchanged when including the platypus *ENAM*, even if some regions were characterized by a lower selective pressure than when therians were analyzed alone, in particular in exon 10 (Fig. 4a).

The non-synonymous substitution rate analysis allowed to identify accurately the selective constraint acting on regions encoded by exons 4, 5, 7, 8, 9, and the beginning of exon 10 (Fig. 4b, c). The lowest substitution rates (=highest functional constraints) were found in the region encoding the 32-kDa fragment (aa 174–279) (Fig. 4c). This analysis revealed also that the 25-kDa fragment was not under a high selective pressure (Fig. 4b).

Purifying Selection

In order to detect site-specific negative selection we have chosen the *Server Selecton* program because it uses a more powerful method than *Hyphy*. The analysis identified 423 (out of 1,142 aa in human ENAM) positions under purifying selection, i.e., having biological significance (Fig. 5). These include 77 unchanged residues and 346 conservative residues (i.e., positions that can be substituted with an amino acid possessing the same characteristics). As indicated by the selective pressure analysis, most of these conserved positions are located in the regions encoded by

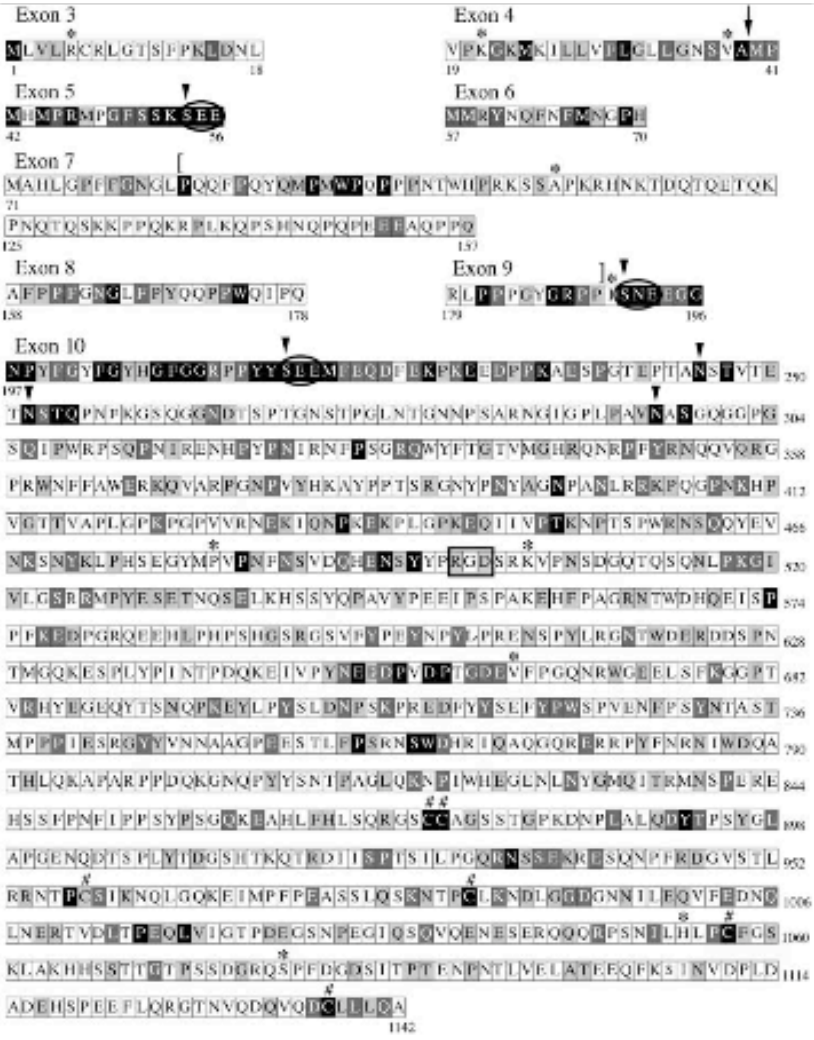


Fig. 5 Amino acid sequence of human ENAM. It is composed of 1,142 residues including the putative leader peptide encoded by exons 3 and 4. The *arrow points* to the cleavage site of the SP. The conserved positions, i.e., subjected to purifying selection during therian (placentals + marsupials) evolution (approximately 160 million years of evolution), are indicated on *black* and *dark gray* (level 1) and on *light gray* (level 2) background. Unchanged (=fixed) residues (77) are on *black background*. By place, well-conserved motifs are composed of several amino acids under selective pressure. These residues are encoded by exons 4–6, part of exon 7, exons 8, 9, and the beginning of exon 10. The three putative phosphorylation sites (SXE) are surrounded by an *oval*. The RGD motif (*boxed*) in exon 10 is not subjected to functional constraint. The regions encoded by most of exons 7 and 10 are variable. The proline-rich domain is localized between ([) in exon 7 and (]) in exon 9. All unchanged positions are predicted to lead to enamel disorders when substituted, as probably most of the other conserved positions when substituted by amino acids with different characteristics. * 10 site-specific positive selections in the lineage leading to human ENAM. # Six cysteines that could be involved in disulfide bridges; all of them are conserved. *Arrowheads* point to phosphorylated and *N*-glycosylated residues known in the pig ENAM

exon 4 (15 conserved aa, representing 65% of the residues of this region), exon 5 (13: 87%), exon 6 (11: 79%), part of exon 7 (20), exon 8 (12: 57%), exon 9 (13: 72%), and by the beginning of exon 10 (42). This confirmed the presence of strong functional constraints in these regions. In contrast, in the other regions encoded by exons 7 and 10, conserved residues represented a low percentage. The other amino acids did not appear constrained, i.e., they were under neutral selection. The three phosphorylated serines (S^{54}, S^{191}, S^{216}) and the three *N*-glycosylated asparagines (N^{245}, N^{252}, N^{296}) known in porcine ENAM were identified as unchanged positions (Fig. 5). In addition, in the region encoded by exon 10 several serines and asparagines were found subjected to purifying selection, which is suggestive of a functional constraint. Conversely, the RGD motif (cell attachment sequence) located in exon 10 is not under selective pressure (Supplementary material 1: positions 756–758). Finally, in the pig, there is a hydroxylated proline (context: egiPspak), which is an unusual post-translational modification. This proline (P^{820}) was unchanged during mammalian evolution except in the platypus, in which it is substituted by a glutamic acid (Supplementary material 1).

Out of the 106 residues composing the 32-kDa peptide, 67 (63%) were found under purifying selection, and among them 28 were unchanged during mammalian evolution (Fig. 6a). In this region, we identified two motifs (GRPPISNEEGG and GFGGRPPYYSEEMFEQD) that are remarkably conserved. Our search in protein databases (*Prosite*) revealed that these motifs are not shared with other known proteins. However, the former motif could contain a Golgi casein kinase phosphorylation site (SneE). In addition, out of the five cleavage sites of this fragment by KLK4, three were found subjected to purifying selection (Fig. 6a).

In comparison, the 25-kDa fragment of porcine ENAM (152 aa) was not found highly constrained; 45 aa (30%) were under purifying selection, and only 5 were unchanged during mammalian evolution. However, at the end of this fragment we have identified a conserved motif composed of 12 residues (NEEDPIDPTGDE) (Fig. 6b). This motif is not shared with other known proteins (*Prosite*), but it includes a probable casein kinase II phosphorylation site (TgdE).

Surprisingly, the only amino acid substitution reported so far in the literature as leading to AIH2 (p.R179M, Fig. 1), i.e., the first residue encoded by exon 9, arginine (R^{195} in our alignment), is not conserved during mammalian evolution and our analysis did not indicate this position as a site-specific positive selection (Supplementary material 1, Fig. 5).

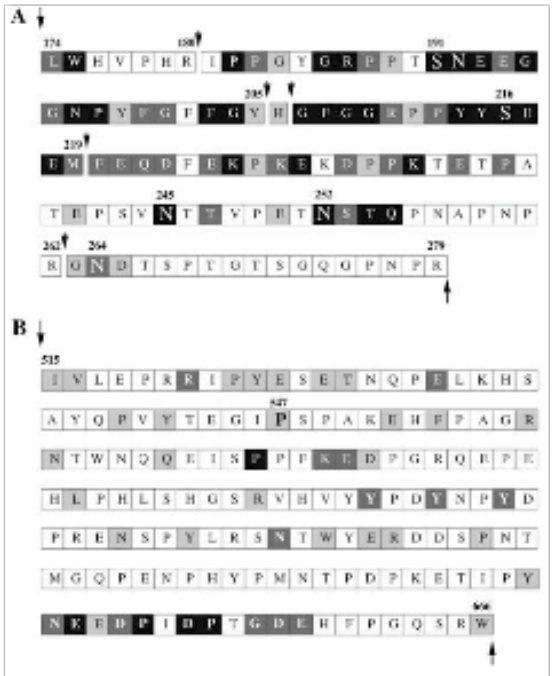


Fig. 6 **a** 32-kDa fragment of pig ENAM (L^{174}–R^{279}). The numerous residues under purifying selection are identified on *dark* and *light gray* background, and unchanged residues are on *black* background. The *long arrows* point to the initial cleavage sites by MMP20. The *short arrows* indicate the 5-cleavage sites by KLK4 reported in the literature. The two phosphorylated serines, S^{191} and S^{216}, and the three N-linked glycosylated asparagines N^{245}, N^{252}, and N^{264} are also under purifying selection. **b** 25-kDa fragment of pig ENAM showing a smaller number of residues under purifying selection compared to the 32-kDa fragment. At the end of the fragment the motif (NEEDPIDPTGDE) is well conserved. The *long arrows* point to the initial cleavage sites by proteases, probably MMP20. No positive selection was identified in these two fragments of pig ENAM

Positive Selection

A total of 19 positions were found subjected to positive selection during mammalian ENAM evolution ($P < 0.1$; Table 2, Fig. 7). It is worth noting that similar positions (e.g., 121, 216, 739, …) were positively selected in various lineages. These site-specific positive selections are located in the regions encoded by exon 3 (1 site), exon 4 (2), exon 7 (1), exon 9 (1), and exon 10 (14) (Fig. 5). Most of these positions are concentrated in Primates (10) and Rodentia (8) (Fig. 7). The 10 positions that are suggested as positively selected in the Primate lineage are not under purifying selection when considering all mammalian sequences (i.e., they are substituted in various, non-primate species: see Supplementary material 1). From them, one occurred after Primates–Scandentia (tree shrew) divergence, five after Tarsiiformes (tarsier) divergence and four in Simiiformes after separation of Platyrrhini (marmoset) (Fig. 7). Out of the eight positions found in rodents, four became unchanged early after the Lagomorpha–Rodentia divergence (rabbit and pika); the four others positive selections

occurred after Caviidae–Rodentia (guinea pig) divergence. In comparison with Rodentia and Primates, the other mammalian lineages possessed a fewer number of residues under positive selection in ENAM (Fig. 7): two occurred early in the Afrotheria and in the Laurasiatheria lineages, two in the common ancestor of Carnivora, Perissodactyla, Chiroptera and Cetartiodactyla, and only one in the Cetartiodactyla. No positive selection was detected in Marsupialia and Monotremata.

Table 2 Results of the codon model fit for positive selection during mammalian evolution using the SLAC method

Codon	$d_N - d_S$	Normalized $d_N - d_S$	P value
5	5.39	1.37	0.08
21	5.84	1.48	0.04
38	5.83	1.48	0.05
121	5.64	1.43	0.07
216	5.85	1.49	0.06
390	5.55	1.41	0.08
739	6.17	1.57	0.02
763	3.68	0.94	0.10
830	3.91	0.99	0.09
871	5.47	1.39	0.07
949	6.95	1.77	0.04
1051	3.96	1.01	0.07
1088	5.18	1.32	0.02
1135	5.42	1.38	0.04
1159	6.40	1.63	0.03
1364	7.20	1.83	0.04
1460	5.56	1.41	0.04
1486	6.02	1.53	0.02
1487	3.85	0.98	0.07

A total of 19 positions were identified within an appropriate significance level of $P < 0.1$. The number of the position refers to our alignment (Supplementary material 1)

5′-UTR

The 5′-UTR of murine *ENAM* consists of 299 bp distributed into exon 1 (214 bp), exon 2 (61 bp) and UTR of exon 3 (24 bp). However, the presence of exon 2 is not reported in human and pig *ENAM* transcripts. In humans, the 5′-UTR is composed of 279 bp (exon 1 = 219 bp; exon 3 = 60 bp) while only 243 bp in pigs (exon 1 = 221 bp; exon 3 = 24 bp). The alignment of 1 kb of the DNA region upstream the TIS in exon 3 with the published sequences allowed to identify the putative UTR of all *ENAMs* analyzed (not shown): exon 1, putative exon 2, and UTR of exon 3. In all *ENAMs*, appropriate splice sites exist for exon 1 (donor site) and exon 3 (acceptor site). Concerning the putative exon 2, correct boundaries were found in Rodentia (rat, kangaroo rat, guinea pig, and squirrel), Primates (from humans to lemurs), Scandentia (tree shrew), Cetartiodactyla (cow, dolphin), and Carnivora (dog, cat). The presence of correct splice sites and the large similarity of nucleotide sequence suggest that exon 2 is probably transcribed in these *ENAMs*. For instance, in humans, putative exon 2 (61 bp) is similar to murine exon 2, with 73% nucleotide identity. In contrast, in some species although a putative exon 2 sequence is easily distinguishable in the intronic region between exons 1 and 3, defaults

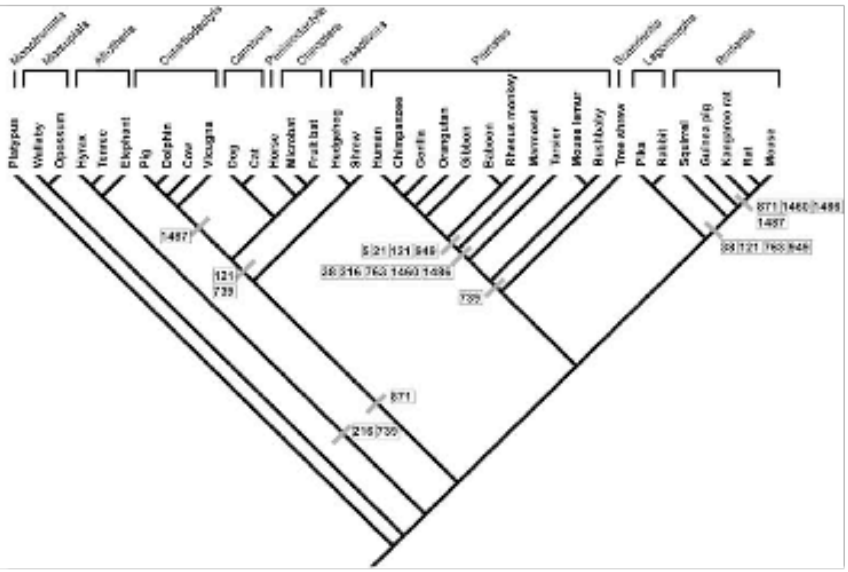


Fig. 7 Location of 12 site-specific positive selections (out of 19; Table 2) during mammalian ENAM evolution. Several similar positions were subjected to positive selection independently in several lineages. The 12 numbers refer to the only informative positions, i.e., residues that were changed then conserved early in the various lineages. Seven sites are not indicated in the figure as they are located in terminal branches and, therefore, not informative enough. Mammalian phylogeny after Springer and Murphy 2007

in putative splice sites along with numerous variations in the sequence when compared to rodents indicate that this DNA sequence is deleted during splicing along with introns 1 and 2. This occurs in Lagomorpha (rabbit and pika), Chiroptera (bats), Afrotheria (elephant, tenrec, hyrax), and Marsupialia (opossum, wallaby). No putative exon 2 sequence was identified in the platypus (Monotremata) *ENAM*.

Signal Peptide Analysis

So far, the only amino acid sequence of ENAM SP obtained through peptide sequencing was in the mouse. The other SP sequences were deduced from mRNA sequences. In mouse ENAM, the SP was composed of 39 residues, i.e., encoded by exons 3 and 4, and this is different from the SPs described in the other SCPPs, i.e., consisting in 16 aa encoded by a single exon. In fact, the sequence of the murine ENAM SP contained a second methionine (M^{24}) in the region encoded by exon 4 (Supplementary material 1). If this start codon was the true one, it would lead to a 16-aa-long SP similar to those in other SCPPs. Therefore, we wonder whether the 39-aa-long SP is not only specific for rodent ENAMs. Another surprising feature in the 5′ region of murine *ENAM* transcripts was the presence of several ATGs that could act each as a TIS, upstream the valid start codon in exon 3.

The alignment of the 5′-UTR of the 36 *ENAM* nucleotide sequences (not shown) indicates that (i) the ATGs located in the putative exon 1 (excepting the rat and the mouse) are not valid TIS; either they are not in the correct reading frame (i.e., that leads to the translation of a functional ENAM protein) or, when they are, in-frame stop codons are present dowstream; (ii) in the rat and the mouse, the ATG at positions 75–77 in *ENAM* exon 1 is in the correct reading frame but this feature is not conserved in evolution, it is absent in the other rodent *ENAM* analyzed (e.g., kangaroo rat, guinea pig, and squirrel), and the putatively encoded SP was not found in the protein sequenced by Fukae et al. (1996); and (iii) the two ATGs located in exons 3 and 4 are in the correct reading frame in all sequences analyzed, and the encoded methionines were both unchanged during mammalian evolution.

The analysis of these two ATGs (using *dnafsminer*) in all the sequences always provided a high probability score for both (Fig. 8c). These ATGs are in agreement with "Kozak's consensus", i.e., having a purine at the 3′ position (Kozak 1984) (Fig. 8a, b).

In the rat and the mouse ENAMs the analysis of the possible SPs (using *SignalP 3.0*) resulting from the predicted TIS showed a low probability for the SP to be starting in exon 1. In contrast, high probability scores were obtained for the SPs starting in exon 3 (39 aa) or in exon 4 (16 aa), with always a slight advantage for the latter (Fig. 8c). As illustrated with the analysis of human ENAM SP, the three typical regions characterizing a SP were found for both SPs: the positively charged n-region, the hydrophobic h-region, and the polar c-region (Fig. 8a, b). In the large SP (39 aa), the 23 residues of the n-region are not under purifying selection, except the start codon (Met, M^1) that was unchanged during mammalian evolution (Supplementary material 1, Fig. 5). The residues comprising the h- and c-regions are subjected to purifying selection. In the short SP, the three regions are subjected to purifying selection and the methionine (M^{23}) was also fixed.

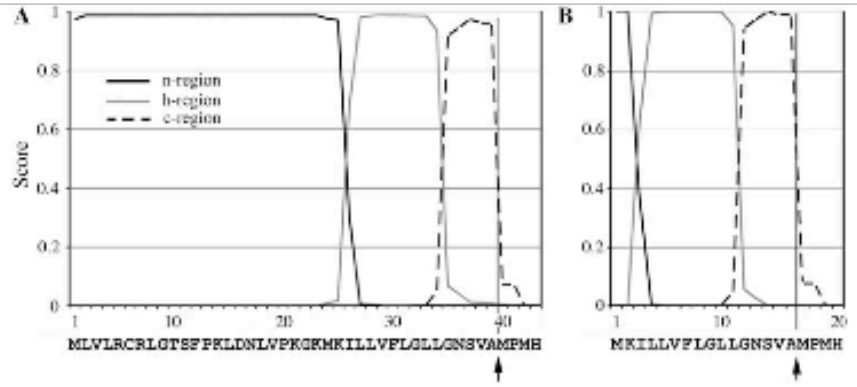


Fig. 8 Prediction of signal peptides encoded either by exon 3 (**a**) or by exon 4 (**b**) and of the cleavage site (*arrows*), taking the human ENAM sequence as an example. For both SP sequences the score is high (0.973 and 1.000, respectively). In both, the maximum cleavage site probability is the same for (**a**) (between positions 39 and 40) and (**b**) (between 16 and 17). The three characteristical regions of SP (n-, h- and c-regions) are indicated. Scores obtained for the long and short SP of 30 mammalian ENAMs. In dolphin (*bold*) ENAM only, the low score for the long SP does not support this SP as functional

The maximal cleavage site probability was always located between A^{39} and M^{40}. This site is highly conserved in all ENAMs analyzed (Supplementary material 1, Fig. 5).

Amino Acid Ratio and Richness in Proline and Glutamine

ENAM belongs to a P/Q-rich SCPP cluster, the members of which are characterized by a high content in proline (P) and glutamine (Q). Indeed, proline is by far the most represented amino acid (14% in average) in all ENAM analyzed (Table 3; Fig. 9). A value that is nearly three times higher than the 5.1% of P in normalized values of proteins. The relative percentage of glutamine (Q) (7% in average) is also high when compared to the 3.8% of normalized value. Serine (S), asparagine (N), glycine (G), and glutamic acid (E) are the other frequent residues. Taken together these six residues account for more than 50% of the residues in ENAMs. Compared to normalized values, ENAM differs in having a larger percentage of P, Q, N, and E, while having a smaller percentage of L (leucine), K (lysine), V (valine), and A (alanine): these differences are related to the proline-rich (P, Q) and acidic region (N, E) (Table 3).

P and Q are concentrated in exons 7–9, in which a proline-rich domain is identified from P^{83} to P^{189} (Figs. 2, 5; Supplementary material 1). The numerous S, N, G, and E are distributed apparently at random in the large exon 10.

We have especially analyzed P and Q ratios, as they represent possible EMP relationships (P/Q-rich proteins). In all the ENAM analyzed, the two residues represent an average of 21% of the total residues (P = 14%; Q = 7%) (Fig. 9). The highest percentage of P (21%) is reached in platypus ENAM. In human ENAM, exon 7 possesess 28% of P (32% in platypus) and 17% of Q (23% in platypus). The putative ancestral ENAM sequence reflects this high percentage of P and Q, as they both represent 24% of the total residues (Fig. 9).

Ancestral Mammalian Sequence

The putative ancestral sequence of mammalian ENAM (210–220 Ma) was calculated from our dataset of 36 sequences (Fig. 10). The coding sequence consisted of eight exons and its primary structure was composed of 1,159 residues. All the important positions were found in this sequence. Among them are the 77 unchanged residues during mammalian evolution, which include the three

Table 3 Amino acid counts and their percentage in representative mammalian ENAM sequences (human, mouse, cow, elephant, opossum, and platypus) and in the putative sequence of ancestral mammalian ENAM, when compared to normalized percentage in most proteins

Amino acid	Human 1,142 aa	Mouse 1,274 aa	Cow 1,143 aa	Elephant 1,141 aa	Opossum 1,140 aa	Platypus 1,209 aa	Ancestral 1,159 aa	Normalized percentage
Pro (P)	156 (**13.7**)	194 (**15.2**)	165 (**14.4**)	161 (**14.1**)	162 (**14.2**)	248 (**20.5**)	189 (**16.3**)	5.1
Ser (S)	93 (**8.1**)	110 (**8.6**)	88 (**7.7**)	87 (**7.6**)	109 (**9.6**)	89 (**7.4**)	93 (**8.0**)	7.0
Asn (N)	91 (**8.0**)	116 (**9.1**)	87 (**7.6**)	87 (**7.6**)	77 (**6.7**)	40 (3.3)	83 (**7.2**)	4.0
Gly (G)	90 (7.9)	93 (7.3)	89 (7.8)	92 (8.1)	93 (8.2)	137 (**11.3**)	103 (8.9)	8.9
Gln (Q)	86 (**7.5**)	81 (**6.4**)	79 (**6.9**)	80 (**7.0**)	70 (**6.1**)	64 (**5.3**)	88 (**7.6**)	3.8
Glu (E)	80 (**7.0**)	73 (**5.7**)	74 (**6.5**)	77 (**6.7**)	81 (**7.1**)	72 (**5.9**)	85 (**7.3**)	5.0
Arg (R)	62 (**5.4**)	50 (3.9)	60 (**5.2**)	61 (**5.3**)	60 (**5.3**)	81 (**6.7**)	62 (**5.3**)	4.1
Leu (L)	60 (5.2)	57 (4.5)	66 (5.8)	67 (5.9)	65 (5.7)	61 (5.0)	59 (5.1)	8.5
Thr (T)	59 (5.2)	70 (5.5)	72 (**6.3**)	60 (5.3)	55 (4.8)	42 (3.5)	50 (4.3)	5.8
Lys (K)	53 (4.6)	73 (5.7)	46 (4.0)	49 (4.3)	51 (4.5)	37 (3.1)	44 (3.8)	8.1
Tyr (Y)	48 (**4.2**)	50 (**3.9**)	50 (**4.4**)	45 (**3.9**)	52 (**4.6**)	46 (**3.8**)	46 (**4.0**)	3.0
Asp (D)	43 (3.8)	38 (3.0)	46 (4.0)	47 (4.1)	34 (3.0)	59 (**4.9**)	41 (3.5)	4.7
Phe (F)	43 (3.8)	53 (**4.2**)	43 (3.8)	40 (3.5)	38 (3.3)	35 (2.9)	40 (3.4)	4.0
Ala (A)	38 (3.3)	65 (5.1)	38 (3.3)	49 (4.3)	49 (4.3)	84 (6.9)	57 (4.9)	8.7
Val (V)	36 (3.1)	46 (3.6)	31 (2.7)	36 (3.1)	41 (3.6)	41 (3.4)	31 (2.7)	6.5
Iso (I)	32 (2.8)	34 (2.7)	39 (3.4)	29 (2.5)	35 (3.1)	10 (0.8)	28 (2.4)	3.7
His (H)	31 (2.7)	20 (1.6)	30 (2.6)	32 (2.8)	26 (2.3)	23 (1.9)	24 (2.1)	3.4
Met (M)	21 (**1.8**)	28 (**2.2**)	19 (**1.7**)	19 (**1.7**)	23 (**2.0**)	18 (1.5)	16 (1.4)	1.5
Try (W)	15 (**1.3**)	16 (**1.3**)	14 (**1.2**)	16 (**1.4**)	12 (1.0)	15 (**1.2**)	14 (**1.2**)	1.0
Cys (C)	7 (0.6)	7 (0.5)	7 (0.6)	7 (0.6)	7 (0.6)	7 (0.6)	6 (0.5)	3.3

The percentages are larger than the normalized values are in bold characters

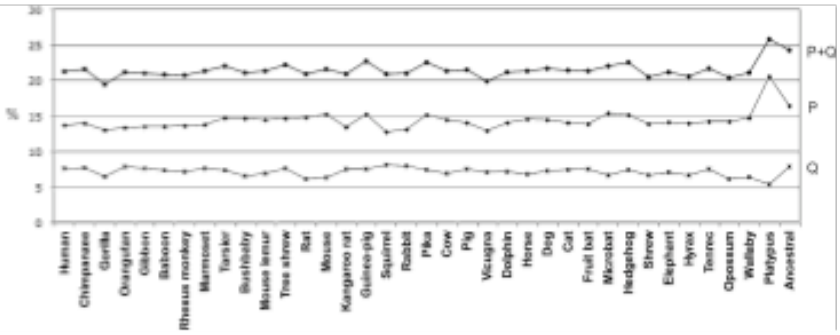


Fig. 9 Percentage of proline (P), glutamine (Q), and both (P + Q) in the 36 mammalian ENAMs analyzed and in the deduced putative ancestral sequence

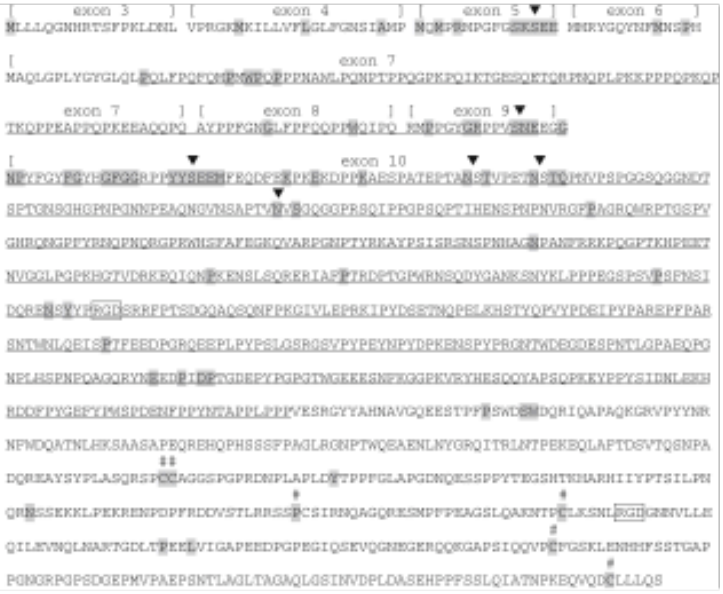


Fig. 10 Ancestral amino acid sequence of mammalian ENAM calculated from the entire dataset. The exons are indicated and delimited by brackets []. The residues unchanged during mammalian evolution are on *gray* background. The residues known as being phosphorylated or *N*-glycosylated are indicated (*arrowheads*). Two RGD motifs are *boxed* and the Pro-rich domain is *underlined*

phosphorylated serines, the three glycosylated asparagines, and the six cysteines (Fig. 10). It is worthy to note that two RGD motifs (cell attachment sequences) are located in this region; although, these amino acids were not subjected to selective pressure. The proline-rich domain was also identified but it extends larger into the region encoded by exon 10 compared to, e.g., human ENAM. No repeats were detected in exons 7 and 10. The two putative TIS (M^1 and M^{23}) in exons 3 and 4 were present.

SignalP 3.0 analysis (not shown) provided a score of 0.887 for the functional probability of the large SP starting with the first methionine (39 aa) (with signal anchor probability = 0.029) and a score of 0.999 for the short SP starting with the second methionine (16 aa) (with signal anchor probability = 0.0). In both cases the maximum cleavage site probability was between A^{39} and M^{40}.

Discussion

The objectives of our evolutionary analysis of mammalian *ENAM* were fulfilled. Using a large set of *ENAM* sequences (36 mammals) that are representative of the main mammalian lineages, we (i) highlighted numerous residues (some of them forming motifs) which were conserved during approximately 220 Ma that is strongly indicative of potential biological or structural functions, and (ii) identified several positions that were recently positively selected in rodents and primates (positive selection). Such a large dataset is not only useful to predict mutations that could lead to ENAM-associated genetic disease, type-2 amelogenesis imperfecta (AIH2), but will also be of importance in helping validate *ENAM* mutations that are identified in patients suffering from AIH2. In addition, we revealed that

the proline-rich domain is conserved, analyzed the large 5′-UTR and the unusually large SP, and also calculated the putative ancestral ENAM sequence in mammals. The latter proved to be useful in recent studies investigating *ENAM* in reptiles (Al-Hashimi et al., unpublished results).

Selective Constraints and Potential Functions for ENAM Residues

Structurally or biologically important residues or motifs, ligand-binding sites or regions involved in protein–protein interaction of ENAM were found subjected to selective pressure as revealed by the analysis of purifying selection. Indeed, as observed in many proteins (e.g., alcohol dehydrogenase or aldehyde dehydrogenase; Chen et al. 2009) the substitution of a single amino acid may reduce the efficiency of ENAM, or even may lead to important disorders such as AIH2. Our analysis indicates that either positions must not be changed (77 unchanged positions), or can only change when replaced with a residue possessing similar biochemical characteristics (more than 300 conservative positions). In addition, 19 positions were identified (principally in rodents and primates) as having high evolution rates, but became unchanged in some lineages. This positive selection is well known (Stern et al. 2007) and is of great interest for evolutionary analyses as it could reflect a recent adaptation of ENAM (improvement of previous functions, new properties, etc.).

The 32-kDa Fragment is the Keystone of ENAM

The 32-kDa fragment, which comprises about 1% of total enamel protein, accumulates in the entire thickness of the enamel. It has been long suggested that this peptide plays a crucial role in the initiation of enamel mineralization (Tanabe et al. 1990; Uchida et al. 1991b; Yamakoshi 1995; Hu and Yamakoshi 2003). Here, we show that 67 out of the 106 residues of this peptide are subjected to purifying selection. Such a high selective pressure can only be explained as an important function of this region. In particular, we identified two remarkably conserved motifs possibly containing a phosphorylated site. This indicates that these motifs are probably the main actors for the function of this peptide.

The 32-kDa peptide is phosphorylated and glycosylated, and it was shown to adsorb strongly onto apatite crystals (Tanabe et al. 1990). Here, we show that the two serines described as being phosphorylated in porcine ENAM were unchanged during mammalian evolution. This suggests that these serines are phosphorylated in the other species as well and that their phosphorylation plays a crucial role for the right function of this peptide. The glycosylation was shown to protect the 32-kDa peptide from degradation by MMP20 (Yamakoshi et al. 2006). Here, we show that the three N-linked glycosylation sites described in porcine ENAM were unchanged during mammalian evolution. This means that selective forces are acting in order to prevent precocious degradation of this peptide during the early phases of enamel mineralization. In contrast, later, during the enamel maturation stage, degradation of the 32-kDa fragment by KLK4 is required as this proteolysis provides space for apatite crystal growth. An extended retention of this peptide in the forming enamel matrix would disturb the mineralization process and may result in enamel hypomaturation. We also show that three out of the five KLK4 cleavage sites identified in the 32-kDa peptide of porcine ENAM are under high purifying selection in mammals. This means that these three KLK4-specific sites were kept unchanged because they provide an optimal substrate for the function of this enzyme. The two other variable cleavage sites are either no longer useful because of alternative sites being present in other locations of the other mammals, or the variable cleavage sites are less important for KLK4 action as, for instance, the N-terminal cleavage site described in porcine ENAM. However, in humans this site appears to be highly important for ENAM function, because the substitution of a single residue was reported to lead to AIH2 (Gutierrez et al. 2007; see below). The results of our evolutionary analysis strongly suggest that further functional studies should target this 32-kDa peptide.

The 25-kDa ENAM Fragment Includes ja Well-Conserved Motif

The 25-kDa ENAM peptide, isolated from the outer thin layer of porcine secretory enamel, is the result of the final process of the proteolytic cleavage of the large 89-kDa fragment (Fukae et al. 1996). Here, we show that most residues of this fragment are not under selective pressure. This could lead to the conclusion that this peptide does not play an important role in enamel mineralization, and was only identified as a degradation product of ENAM. However, in this peptide, we identified a conserved motif (NEEDPIDPTGDE in the pig sequence) with a probable phosphorylation site (Thr). This raises the question of a possible role for this short region in early phases of mineralization, before being entirely degraded by MMP20.

Identification of Functional Motifs

Several motifs (SxE) containing a putative phosphorylation site are well conserved in regions encoded by exons 4, 9, and 10. Such an evolutionary constraint indicates that these residues are essential for ENAM function, in particular by increasing protein affinity to apatite crystals (Tanabe et al. 1990; Hu et al. 1997b). Moreover, similar motifs are

present in all SCPPs and they are characteristic features of these proteins (Kawasaki and Weiss 2003).

Although the proline-rich domain encoded by exon 7 appears variable, we identified a conserved motif in its middle region. It is worthy to note that we did not find a similar motif in protein databases, including proline-rich proteins. This suggests that this motif could have an important function for enamel formation, in particularly protein–protein interactions when considering its location in the core of the proline-rich domain. This hypothesis should be tested in the near future.

The presence of an RGD motif in the ENAM region encoded by exon 10, first led Fisher and Fedarko (2003) to consider ENAM as a member of the SIBLING (Small Integrin-Binding Ligand N-linked Glycoprotein) family. However, they dropped this idea when they found out that a RGD was absent in murine ENAM. Here, we show that the RGD is not present in several mammalian ENAM and is not subjected to purifying selection. This could mean that the RGD motif was inherited from an ancestral SCPP early in ENAM evolution, but was not functionally/structurally important for the protein (e.g., no interaction with the cell membrane in this ENAM region) and, therefore, was not evolutionary constrained. Another possibility is that the RGD motif was generated randomly in several species. Indeed, this ENAM region is rich in Arg, Gly, and Asp residues. Our analysis of positive selection did not reveal that this motif was selected recently in some lineages, such as primates. We conclude that the RGD motif is not necessary for the correct function of ENAM.

Positive Selection and ENAM Adaptation

Nineteen sites subjected to positive selection were identified in ENAM as having a greater evolution rate than observed for neutrally evolving sites. Generally, this phenomenon results from a mutation with a strong selective value, i.e., increasing the mean fitness of an organism. Some of these site-specific residues are close to motifs subjected to purifying selection. This indicates that they could improve the specific function of these motifs.

It is worthy to note that positive selection was detected principally in primates (10 sites) and rodents ENAM (8 sites), while only two occurred in afrotherians and in laurasiatherians. The presence of recently selected residues in primates and rodents could be the result of adaptive evolution. This brings up to the question on which changes could these positively selected positions be related to? Recently, using *PAML* program, Kelley and Swanson (2008) have identified three positions subjected to positive selection in the 32-kDa peptide of primate ENAM. These authors found a correlation between these newly selected residues, an increase in enamel thickness and a shift in diet from folivorous to frugivorous. However, this study did not take into consideration the non-primate ENAMs. The *PAML* program does not take into account variations in synonymous substitution rate, and therefore may lead to false positives (Kosakovski et al. 2005a). In our study, we tried to overcome these limitations by comparing numerous mammalian ENAMs and by also using the *HyPhy* program that displays some advantages over other methods in detecting amino acid positions subjected to positive selection (Kosakovski et al. 2005b). Finally, it is supposed that the 32-kDa peptide plays a crucial role in the initiation of mineralization, not in enamel microstructure and thickness. This function could be ensured by the proline-rich region of the three EMPs (Delgado et al. 2005; Sire et al. 2007). Positively selected residues in the 32-kDa peptide probably reinforce the function of the 32-kDa fragment as, for instance, a better mineral initiation.

In the absence of more precise data on the role played by enamel proteins in the precise structure of this tissue, it was not possible to find a correlation between positively selected amino acids identified in ENAM and dietary change. Indeed, we cannot explain that several positions were positively selected in primates and rodents, and nearly absent in leurasiatherians and afrotherians. Further studies are needed to understand which advantages, if any, are related to these positively selected residues.

Proline-Rich Region of ENAM Witnesses for EMP Relationships

The presence of a proline-rich region is not a specific feature of ENAM; the two other EMPs, AMEL, and AMBN, also possess this region (Kawasaki and Weiss 2008). The proline-rich ENAM region encoded by exon 7 is variable and, as such, it appears to be under relaxed evolutionary pressure. In a previous evolutionary analysis of AMEL, we showed that the proline-rich domain was not less evolutionary constrained than the other regions, but that the selective pressure was focused on the conservation of a high percentage of prolines (and glutamines) (Delgado et al. 2005). We proposed also that the proper function of AMEL in enamel formation (enamel microstructure and thickness) resided in its proline-rich region. Proline-rich regions are also supposed to favor protein–protein interactions (Dunker et al. 2002) and there is some evidence that the proline-rich region of AMEL interacts with AMBN (Ravindranath et al. 2004).

In AMEL, we showed that the high percentage of proline (P) and glutamine (Q) was acquired through repeated insertions of PxQ triplets. We suggested the hypothesis that most of the region encoded by exon 6 was generated through this mechanism early in AMEL evolution (Delgado et al. 2005, 2006; Sire et al. 2007). We also

demonstrated that the three EMPs are phylogenetically related (Sire et al. 2006, 2007). The proline-rich regions encoded by *ENAM* exon 7, *AMEL* exon 6, and *AMBN* exon 6 are probably homologous, and this condition was inherited from a common ancestral *EMP* gene. However, in the putative ancestral ENAM sequence we did not identify PxQ repeats in the region encoded by exon 7. We can conclude that the ancestral condition of ENAM was to possess a proline-rich region, and that the triplet repeats have appeared later, when ENAM gave rise to AMBN and AMEL (Sire et al. 2007). Further analyses of lissamphibian and reptilian sequences could help to better understand the evolution of this proline-rich domain.

Prediction and Validation of AIH2-Associated Mutations in Human ENAM

Our evolutionary analysis of ENAM revealed 77 unchanged positions, i.e., residues that were not modified during approximately 220 Ma of mammalian evolution, when taking into account the divergence time between therians and monotremes (Warren et al. 2008). As discussed above, amino acid conservation during long geological times certainly means that they play an important role for the correct function of the protein. This correlation is strongly supported by studies demonstrating that more than 95% of the amino acid substitutions leading to a genetic disease occur on residues unchanged during hundreds million years of evolution, i.e., those under strong selective pressure (Subramanian and Kumar 2006). As a consequence, a disease-associated mutation (DAM) can be predicted when such an unchanged position is changed. It is therefore important, in this context, to perform an evolutionary analysis of proteins to identify accurately all unchanged ENAM positions. Such a dataset will be of great help to validate human mutations (amino acid substitutions) suspected to be responsible for AIH2. Such a prediction/validation of DAM was recently performed on AMEL, the major enamel protein. The study revealed that numerous unchanged residues, among which eight positions, already reported in the literature to be responsible for X-linked AI when substituted, were validated (Delgado et al. 2007). Here, we predict that the substitution of one of these 77 conserved positions in human ENAM will lead to AIH2.

In addition, a large number of positions were identified as being conservative (i.e., they slowly evolve and can be replaced by amino acids possessing the same biochemical characteristics). We postulate that changing the properties of these positions could lead either to a severe enamel disorder or to discrete enamel abnormalities, as for instance higher susceptibility to dental caries.

However, it is worthy to note that, out of the eight mutations reported in the literature as leading to AIH2, only one concerns a single amino acid substitution: the first residue, arginine (positively charged), encoded by exon 9 is replaced by a methionine (hydrophobic) (p.R179M; Gutierrez et al. 2007). Surprisingly, the evolutionary analysis indicates that this position is not conserved during mammalian evolution: Arg is substituted either in Thr (rat), Lys (cow), His (wallaby), or Gln (platypus). Moreover, this position is not under positive selection. Four hypotheses can be suggested to explain this contradiction: (1) this position is conservative and Arg can be replaced with a polar (Thr, Gln) or charged (Lys, His) residue, but not with a non-polar residue (Met); (2) as discussed before, the importance of this position is related to its putative function as the KLK4 N-terminal cleavage site of the 32-kDa fragment. Another cleavage site could exist in species in which Arg is substituted, but we failed to find it; (3) the mutation is in the first codon of exon 9 and may also represent a splice-junction mutation, even though it is in the exon. There are a number of such mutations in the 5′-end of DSPP that cause inherited dentin defects (Kim et al. 2005a, b); and (4) the reported mutation is not valid, i.e., another mutation may be responsible for the AI in the patient in which Arg is substituted, but the probability is very low. The first hypothesis seems the most plausible as none of the substitutions observed in the dataset are involving a methionine. According to various authors kallikreins are normally targeted to Arg, Lys, or Gln (Gomis-Ruth et al. 2002; Yoon et al. 2007).

Nevertheless, we failed to demonstrate that KLK4 is able to process this site when substituted by Thr or His.

Is the Role of 5′-UTR of ENAM Different from that of the Other SCPPs?

In all *SCPPs*, the 5′-UTR is composed of 100–110 nucleotides distributed into a short exon 1 and the 5′-end of exon 2 (Kawasaki and Weiss 2003). This is not the case in most *ENAMs*, in which the 5′-UTR possesses a large exon 1 (>200 bp), an additional exon (approximately 60 bp), and the 5′-end of exon 3 (60 bp). Such a large 5′-UTR, and the presence of a large SP (see below), are interesting features of ENAM.

The length of the 5′-UTR seems to play a role on the fidelity of the translation initiation signal (Kozak 1991). Indeed, the minimal length necessary to allow the recognition of the right ATG by the transcription machinery is about 20 nucleotides, but the fidelity increases with the 5′-UTR length. This observation is ascribable only to the length of 5′-UTR, not to a particular sequence. The only known characteristic of 5′-UTR sequences consists in their poorness in guanine, which is not favorable to the formation of secondary structures (Kozak 1991; Reuter et al. 2008). The capacity for a long 5′-UTR to increase the effectiveness of the translation

initiation can be explained partly by the accumulation of many ribosome 40s sub-units on this mRNA. Moreover, Reuter et al. (2008) suggested recently that the length of 5′-UTRs is selectively neutral and evolves under a process of stochastic destruction and recruitment of core promoter elements. This phenomenon is combined with a selective pressure against a premature translation initiation. This means that most of 5′-UTR sequence is not subjected to selective constraints to conserve DNA sequences, but only to keep their poorness in guanine (Reuter et al. 2008). Therefore, we can deduce that the presence of a supplementary exon in ENAM 5′-UTR is either in relation to the initiation of the second ATG (see below) or a random recruitment process of promoter elements.

An Unusual SP in ENAM

All members (22 in humans) of the SCPP family possess a SP ranging from 17 to 26 aa, except ENAM, in which the SP is unusually large (39 residues in humans, Hu et al. 2000; 38 aa in the pig, Hu et al. 1997a) (Kawasaki and Weiss 2003). Such a large SP is rarely encountered in eukaryotes; comparative analysis of a large number of SP indicated a length of 20–25 aa (von Heijne 1985; Martoglio and Dobberstein 1998). This large SP is encoded by exon 3, in which the TIS is well conserved during mammalian evolution, and exon 4 that encodes well-conserved residues (leucine-rich region) composing the hydrophobic core (h-) region required for protein targeting and membrane insertion (von Heijne 1985). Conversely, the large n-region (of unknown function) encoded by exon 3 (the N-terminal region of this SP) is not subjected to strong selective pressure, although present in all mammalian ENAMs. The h-region is homologous to the SP of the other SCPPs, while the large n-region is proper to ENAM. In the other SCPPs the SP possesses a short n-region.

The existence of well-conserved second TIS encoded by exon 4, and homologous to the SP of the other SCPPs, raises the question of the need/function of the first one. It appears clearly that the second TIS was inherited from the ancestral gene of this family, and we showed that it could be functional if it was not preceded by the first TIS of the large SP. The latter being located in a different exon (exon 3) than the second TIS (exon 4), it is probable that ENAM exon 3 results from the insertion of an exon from another gene (the so-called exon shuffling process; Patthy 1999). This event occurred at least in the common ancestor to mammals or earlier. However, using Psi Blast search we did not find similarity of this exon 3 with other proteins in the database (NCBI). The conservation of this new SP during ENAM evolution means that the presence of a large SP (with a large n-region) was a positive event. The strong conservation of the two TIS can only be explained by the existence of two isoforms that result from alternative splicing of exon 3: one with a short SP (homologous to that of others SCPPs) and another with a long SP, the function of which is still to be found.

Our analysis revealed a single exception of this rule: the large dolphin ENAM SP exhibits a low probability score for a secretory-related SP. In contrast, our SP analysis indicates a high probability score for an anchoring SP. Therefore, either this large SP is not functional or it could play another function. We can also wonder whether this difference when compared to the other ENAMs could be related to the homodont dentition in toothed cetaceans.

The presence of two (predicted) SP in mammalian ENAM is not unique and the use of different SP is a common feature of many proteins (Davis et al. 2006). For example, interleukin-15 presents both a short and a long SP revealing complex pathways of intracellular trafficking (Kurys et al. 2000). In general, complex SP organization means multiple functional properties: protein secretion, cytoplasm localization, mitochondrial targeting, membrane protein (Hiss et al. 2008; Davis et al. 2006). Unfortunately, the pathway of the predicted ENAM isoforms remains complex to determine because these isoforms are not reported so far and no statistical data can predict the exact role of a SP from its amino acid sequence. However, from our analyses we can conclude that the ENAM SPs are probably used only for protein secretion and not for other cellular pathways. We hypothesize a different extracellular export efficiency of the two isoforms (long and small SP) as reported from mutations in the SP of the protein Shrew-1 (Hiss et al. 2008). Indeed, the export efficiency appears to be correlated with the existence and integrity of an area separating the n- and c- regions of the Shrew-1-long SP. This region called "transition area" exists in many long SP and is characterized by 4–7 aa including a glycine (G) and some large polar residues (Hiss et al. 2008). These authors have shown that controlled mutations (progressive deletions) in this area decreased the secretion activity.

The ENAM SP seems to present such a transition area at the beginning of exon 4. Interestingly, this putative transition area is variable in mammals, except the glycine conserved at position 22. This suggests a complex regulation of ENAM secretion in mammals. The regulation of secretion could decrease the amount of ENAM in the forming enamel matrix and would facilitate the orderly replacement of organic matrix with mineral during the transition- and maturation-stages (Lu et al. 2008).

The Case of Platypus ENAM

The platypus (monotremes) is an interesting species with regard to EMPs, not only because the monotreme lineage diverged from the therian lineage approximately 200 Ma,

but also because the milk teeth are not replaced in juveniles. A keratinized beak and large keratinized pads are used instead of teeth for food processing (Davit-Béal et al. 2009). Therefore, the selective pressure on tooth enamel and on its constituents is rapidly relaxed during ontogeny, and we wonder whether or not this low functional pressure could be detected in the ENAM sequence. First, our analysis showed that platypus ENAM is functional: neither stop codon nor reading frame shift, or large deletions were detected. This means that ENAM is important for enamel formation, even in a species with a short enamel life. In addition, the 32-kDa peptide is well conserved as in the other mammalian ENAMs, despite the long lasting separate evolution. Therefore, this finding strongly supports the 32-kDa ENAM peptide as a crucial actor for enamel formation in mammals. The only difference identified in the platypus ENAM concerns its larger P + Q content when compared to the other ENAMs (see above). Another difference consists in a richer guanine content (48% on average) of the 5′-UTR region than in the other ENAMs (data not show). In eukaryotes, it is known that a 5′-UTR guanine-rich sequence decrease the capacity of transcription initiation (Kozak 1991), which could suggest a lower level of ENAM expression in platypus when compared to therian species.

Conclusion

The role that ENAM plays in enamel formation is still largely within speculation; although, it is assumed to work as a nucleator during the early phases of enamel mineralization and/or could be critical for enamel crystal elongation. Our evolutionary analysis reinforces some previous findings, such as the crucial function of the phosphorylated and glycosylated residues, cysteines, and the important role played by the 32-kDa peptide, because these positions were conserved for more than 220 Ma. We also shed light on numerous unchanged residues and motifs that were unknown and certainly are important for the correct function of the molecule. All of them should be targets for future studies aiming to clarify their function. These data can also be used to predict or validate AI-associated mutations in ENAM. Although ENAM represents less that 5% of the matrix during enamel formation, the important role in enamel formation of this protein exists for long geological times, as revealed by its presence and conservation in the platypus. In addition to the 32-kDa fragments, the proline-rich domain and the conserved motif in the 25-kDa fragment should have an interesting function, still to be elucidated.

Acknowledgments We thank Mehboob Chilwan (Erasmus, University of Keele, UK) for English corrections. This work was supported by CNRS and UPMC (UMR 7138) Grants.

References

Aoba T, Moreno EC (1987) The enamel fluid in the early secretory stage of porcine amelogenesis: chemical composition and saturation with respect to enamel mineral. Calcif Tissue Int 41:86–94

Brookes SJ, Lyngstadaas SP, Robinson C, Shore RC, Wood SR, Kirkham J (2002) Enamelin compartmentalization in developing porcine enamel. Connect Tissue Res 43:477–481

Chen Y-C, Peng G-S, Wang M-F, Tsao T-P, Yin S-J (2009) Polymorphism of ethanol-metabolism genes and alcoholism: correlation of allelic variations with the pharmacokinetic and pharmacodynamic consequences. Chem Biol Int 178:2–7

Davis MJ, Hanson KA, Clark F, Fink JL, Zhang F, Kasukawa T, Kai C, Kawai J, Carninci P, Hayashizaki Y, Teasdale RD (2006) Differential use of signal peptides and membrane domains is a common occurrence in the protein output of transcriptional units. PLoS Genet 2:e46

Davit-Béal T, Tucker T, Sire JY (2009) Loss of teeth and enamel in tetrapods: fossil record, genetic data and morphological adaptations. J Anat 214:277–501

Dayhoff MO, Schwartz R, Orcutt BC (1978) A model of evolutionary change in proteins, matrixes for detecting distant relationships. In: Dayhoff MO (ed) Atlas of protein sequence and structure, vol 5. National Biomedical Research Foundation, Washington, DC, pp 345–358

Delgado S, Casane D, Bonnaud L, Laurin M, Sire JY, Girondot M (2001) Molecular evidence for Precambrian origin of amelogenin, the major protein of vertebrate enamel. Mol Biol Evol 18:2146–2153

Delgado S, Girondot M, Sire JY (2005) Molecular evolution of amelogenin in mammals. J Mol Evol 60:12–30

Delgado S, Couble ML, Magloire H, Sire JY (2006) Cloning, sequencing, and expression of the amelogenin gene in two scincid lizards. J Dent Res 85:138–143

Delgado S, Ishiyama M, Sire JY (2007) Validation of amelogenesis imperfecta inferred from amelogenin evolution. J Dent Res 86:326–330

Deméré TA, McGowen MR, Berta A, Gatesy J (2008) Morphological and molecular evidence for a stepwise evolutionary transition from teeth to baleen in mysticete whales. Syst Biol 57:15–37

Deutsch D (1989) Structure and function of enamel gene products. Anat Rec 224:189–210

Dong J, Gu TT, Simmons D, MacDougall M (2000) Enamelin maps to human chromosome 4q21 within the autosomal dominant amelogenesis imperfecta locus. Eur J Oral Sci 108:353–358

Doron-Faigenboim A, Stern A, Mayrose I, Bacharach E, Pupko T (2005) Selecton: a server for detecting evolutionary forces at a single amino-acid site. Bioinformatics 21:2101–2103

Dunker AK, Brown CJ, Lawson JD, Iakoucheva LM, Obradovic Z (2002) Intrinsic disorder and protein function. Biochemistry 41:6573–6582

Endo T, Ikeo K, Gojobori T (1996) Large-scale search for genes on which positive selection may operate. Mol Biol Evol 13:685–690

Fan D, Lakshminarayanan R, Moradian-Oldak J (2008) The 32 kDa enamelin undergoes conformational transitions upon calcium binding. J Struct Biol 163:109–115

Fisher LW, Fedarko NS (2003) Six genes expressed in bones and teeth encode the current members of the SIBLING family of proteins. Connect Tissue Res 44(Suppl 1):33–40

Fukae M, Tanabe T (1985) Separation of non-amelogenin component from purified amelogenin preparation of immature porcine enamel. Jpn J Oral Biol 27:1249–1251

Fukae M, Tanabe T (1987) Nonamelogenin components of porcine enamel in the protein fraction free from the enamel crystals. Calcif Tissue Int 40:286–293

Available online at www.sciencedirect.com
ScienceDirect
Molecular Phylogenetics and Evolution 47 (2008) 865–869
MOLECULAR PHYLOGENETICS AND EVOLUTION
www.elsevier.com/locate/ympev

Short Communication

Amelogenin, the major protein of tooth enamel: A new phylogenetic marker for ordinal mammal relationships

Sidney Delgado [a], Nicolas Vidal [b], Géraldine Veron [c], Jean-Yves Sire [a,*]

[a] *UMR 7138, Equipe "Evolution et développement du squelette", Université Pierre & Marie Curie—Paris 6, 7 quai St-Bernard, Case 05, 75252 Paris, France*
[b] *UMR 7138, Equipe "Phylogénie", Muséum national d'Histoire naturelle, Paris, France*
[c] *UMR 5202, Unité "Origine, Structure et Evolution de la Biodiversité", Muséum national d'Histoire naturelle, Paris, France*

Received 3 May 2007; revised 14 January 2008; accepted 23 January 2008
Available online 2 February 2008

1. Introduction

Teeth and their tissues, dentin and enamel, have a long, well-defined history. Their origin was traced back to the extra-oral dermal skeleton of early jawless vertebrates, approximately 500 million years ago, mya (see reviews in Huysseune and Sire, 1998; Smith and Coates, 2000; Sire and Huysseune, 2003). Once recruited into the mouth in early gnathostomes, circa 450 mya, teeth were subjected to strong selective pressure due to their crucial function. This explains why teeth, and particularly their developmental processes, organization and structural components, were conserved nearly unchanged through geological times.

In mammals, as in most tetrapod taxa, teeth are covered by a thick and highly mineralized, protective tissue, enamel. The amelogenin gene (*AMEL*) encodes the major protein of enamel (90% of the organic matrix). Recent molecular analyses have brought insights into the evolutionary pattern of *AMEL* in mammals (Delgado et al. 2005) and have shown that the history of this protein could have started by the end of the Precambrian period (Sire et al., 2007). Comparative studies of *AMEL* in mammals, reptiles and amphibians have revealed highly conserved residues located at the C- and N-terminal regions and have indicated that a large part of the hydrophobic, central region of the molecule, encoded by the largest exon 6, was more variable (Ishiyama et al., 1998; Toyosawa et al., 1998; Delgado et al., 2005). Because *AMEL* is X-linked in many mammal lineages, the gene as a whole is predicted to be particularly strongly conserved (under Ohno's rule in general, Ohno, 1967, and because of the X's bias toward transmission through the slowly mutating female mammal germline, Li et al., 2002). In eutherians, *AMEL* was shown to span an ancient pseudoautosomal boundary on the X-chromosome, exon 6 being a formerly pseudoautosomal segment of the gene (Iwase et al., 2003). This additional stringency at this particular location may have reinforced the conservation of exon 6 sequence because recombination has been shown to have little effect on the rate of sequence divergence in this pseudoautosomal boundary among humans and great apes (Yi et al., 2004). This possibility has been also discussed in a recent article (Richard et al., 2007).

Both functional constraints and sequence variation indicate that *AMEL*, and particularly the variable region, could contain a useful phylogenetic signal for deep cladogenetic events, even if exon 6, the only exon easily retrieved using PCR, is rather short (approximately 400 bp). We have therefore tested the utility of this region of *AMEL* for inferring a mammalian phylogeny above the family level.

Comparative genomic data from mammals have accumulated rapidly in the recent past and have contributed significantly to resolving long-standing phylogenetic controversies. Mitochondrial then nuclear DNA sequence analyses revealed new interordinal mammalian relationships (e.g., Springer et al., 1997; Stanhope et al., 1998; Madsen et al., 2001; Murphy et al., 2001; Delsuc et al., 2002; Waddell and Shelley, 2003). Four superordinal eutherian clades are recognized: Laurasiatheria (six orders: Cetartiodactyla, Perissodactyla, Carnivora, Pholidota, Chiroptera and Eulipotyphla), Euarchontoglires (five orders: Primata, Dermoptera, Scandentia, Rodentia and

* Corresponding author. Fax: +33 1 44 27 35 72.
E-mail address: sire@ccr.jussieu.fr (J.-Y. Sire).

1055-7903/$ - see front matter
doi:10.1016/j.ympev.2008.01.025

Lagomorpha), Xenarthra, and Afrotheria (six orders: Macroscelidea, Afrosoricida, Tubulindentata, Sirenia, Hyracoidea, Proboscidea). Eighteen eutherian (placentals) orders were defined, to which are added the metatherian order (marsupials) and the prototherian order (monotremes) to give a total of 20 orders encompassing all extant mammalian species (Waddell and Shelley, 2003).

Most of the recent molecular phylogenies confirm these relationships (e.g., Amrine-Madsen et al., 2003; Springer et al., 2003; Hallström et al., 2007; Murphy et al., 2007). However, controversies still persist both among molecular phylogenies and when comparing these data to evolutionary relationships based on morphology. This is particular pertinent to several superordinal eutherian relationships such as between Afrotheria, Xenarthra and Boreoeutheria (Euarchontoglires + Laurasiatheria), although monophyly of Afrotheria was recently supported by morphological features (Sanchez-Villagra et al., 2005; Tabuce et al., 2007). It is clear, however, that more nuclear data are required as early placental divergences may have been compressed in time (Kriegs et al., 2006; but see Bininda-Emonds et al., 2007 for further discussion on the diversification of today's mammals).

In the present study, we have used 55 sequences of the amelogenin exon 6 from species representative of all main mammalian lineages. We show that *AMEL* exon 6 is an additional efficient marker for ordinal mammal relationships.

2. Material and methods

The species and accession numbers of *AMEL* sequences used in this study are listed in Table 1. Eighteen sequences were found in databases. The other sequences were obtained from genomic DNA extracted from either frozen or ethanol-preserved soft tissues (kidney, liver, spleen, skin) using the DNeasy Tissue System kit (Qiagen). The source of material is indicated in "Acknowledgments" section.

AMEL exon 6 was amplified using the following primers: *Mam1* (*sense*: 5′-TACGAACCATGGGTGGATGGC TGC-3′) or *Mam3* (*sense*: 5′-TACCCTTCCTATGGTTAC GAG-3′) to hybridize the 5′ region, and *Mam2* (*antisense*: 5′-CACTTCCTCCCGCTTGGTCTT-3′) or *Mam4* (*antisense*: 5′-GCCAAGCTTCCAGAGTCAGAT-3′) to hybridize the 3′ region.

Amplification was performed in 38 cycles, each cycle comprising: 1 min denaturation at 94 °C, 1 min annealing at 59 °C and 1 min extension at 72 °C. The final extension was for 30 min at 72 °C. Sequencing of PCR products was done by Genome Express S.A.

Sequences were aligned manually using the editor Se-Al software (Rambaut, 1996) and amino acid properties were used. Resulting gaps were treated as missing data in all analyses. The 5′ (36 first bp) and 3′ (21 last bp) regions of exon 6 are highly conserved and were deleted. This resulted in 567 sites for 55 taxa (322 variable sites, 203 of which are informative). The alignment is available upon request.

Table 1
Species studied (55 taxa)

Human	Hominidae	*Homo sapiens*
Orangutan	Hominidae	*Pongo pygmaeus*
Rhesus monkey	Cercopithecidae	*Macaca mulatta*
Squirrel monkey	Cebidae	*Saimiri boliviensis*
Marmoset	Cebidae	*Callithrix jacchus*
Ring-tailed lemur	Lemuridae	*Lemur catta*
Bushbaby	Galagidae	*Otolemur garnettii*
Tree shrew	Tupaiidae	*Tupaia belangeri*
Flying lemur	Cynocephalidae	*Cynocephalus variegatus*
Mouse	Muridae	*Mus musculus*
Rat	Muridae	*Rattus norvegicus*
Hamster	Muridae	*Mesocricetus auratus*
Guinea pig	Caviidae	*Cavia porcellus*
Squirrel	Sciuridae	*Spermophilus tridecemlineatus*
Cow	Bovidae	*Bos taurus*
Goat	Bovidae	*Capra hircus*
Japanese serow	Bovidae	*Capricornis crispus*
Pig	Suidae	*Sus scrofa*
Hippopotamus	Hippopotamidae	*Hexaprotodon liberiensis*
Dolphin	Delphinidae	*Tursiops truncatus*
Porpoise	Phocoenidae	*Phocoena phocoena*
Horse	Equidae	*Equus caballus*
Tapir	Tapiridae	*Tapirus terrestris*
Rhinoceros	Rhinocerotidae	*Ceratotherium simum*
Dog	Canidae	*Canis familiaris*
Black bear	Ursidae	*Ursus americanus*
Panda	Ursidae	*Ailuropoda melanoleuca*
Gray seal	Phocidae	*Halichoerus grypus*
Sea lion	Otariidae	*Otaria byronia*
Cat	Felidae	*Felis catus*
Tiger	Felidae	*Panthera tigris*
Cheetah	Felidae	*Acinonyx jubatus*
Pangolin	Manidae	*Manis javanica*
Fruit bat	Pteropodidae	*Cynopterus brachyotis*
Flying fox	Pteropodidae	*Pteropus vampyrus*
Roundleaf bat	Rhinolophidae	*Hipposideros ater*
Microbat	Vespertilionidae	*Myotis lucifugus*
Hedgehog	Erinaceidae	*Erinaceus europaeus*
Shrew	Soricidae	*Sorex araneus*
Armadillo	Dasypodidae	*Dasypus novemcinctus*
Tamandua	Myrmecophagidae	*Tamandua tetradactyla*
Three-toed sloth	Bradypodidae	*Bradypus infuscatus*
Two-toed sloth	Megalonychidae	*Choloepus hoffmanni*
African elephant	Elephantidae	*Loxodonta africana*
Tenrec	Tenrecidae	*Echinops telfairi*
Golden mole	Chrysochloridae	*Chrysochloris asiatica*
Aardvark	Orycteropidae	*Orycteropus afer*
Hyrax	Procaviidae	*Procavia capensis*
Elephant shrew	Macroscelididae	*Elephantulus edwardii*
Manatee	Trichechidae	*Trichechus manatus*
Opossum	Didelphidae	*Monodelphis domestica*
Aquatic opossum	Didelphidae	*Chironectes minimus*
Wallaby	Macropodidae	*Macropus eugenii*
Echidna	Tachyglossidae	*Tachyglossus aculeatus*
Platypus	Ornithorhynchidae	*Ornithorhynchus anatinus*

The entries in the table are ordered alphanumerically (by Accession No: EU168848–EU168899).

We built phylogenies using probabilistic approaches with Maximum Likelihood (ML) and Bayesian methods of inference. ML analyses were performed with PAUP*4 (Swofford, 1998). Bayesian analyses were performed with MrBayes 3.1 (Ronquist and Huelsenbeck, 2003). For both approaches an appropriate model of sequence evolution

was inferred from the data themselves using ModelTest (Posada and Crandall, 1998). The model selected (Akaike Information Criterion) was the TrN+G model with substitution parameters as A–C/A–T/C–G = 1, A–G = 4.1029, and C–T = 2.5407, base frequencies as A = 0.2398, C = 0.4136, G = 0.1682 and T = 0.1785, and a Γ parameter of 0.707. Bayesian analyses were run with model parameters estimated as part of the Bayesian analyses, and the best-fit model as inferred by Modeltest. ML results are presented under the form of a bootstrap consensus tree (1000 replicates, NJ starting tree with NNI branch swapping) which is considered to be a reliable estimate of phylogeny. Bayesian analyses were performed by running 2,000,000 generations in four chains, saving the current tree every 100 generations. The last 18,000 trees were used to construct a 50% majority-rule consensus tree.

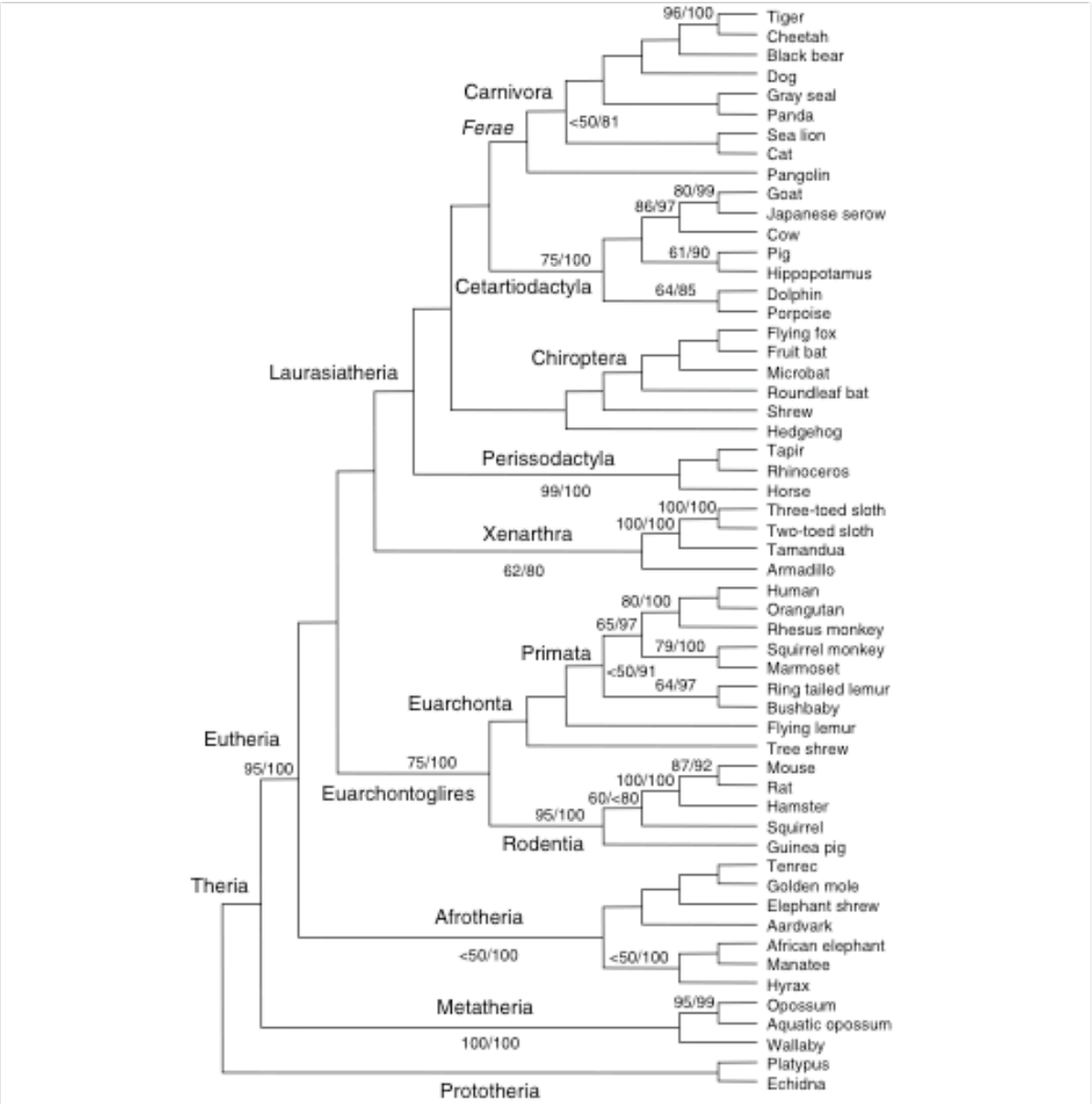


Fig. 1. Phylogenetic relationships of mammals based on *AMEL* sequences (bootstrap ML consensus tree). ML bootstrap values above 50% are shown, followed by Bayesian posterior probabilities above 80%.

3. Results and discussion

The 50% majority-rule ML bootstrap consensus tree is shown in Fig. 1. As expected based on analyses from one portion of gene only (less than 600 sites), most basal nodes show low robustness values. However, remarkable congruence with previously published molecular mammalian phylogenies is obtained.

Our analyses based on *AMEL* exon 6 from 55 species recovered the marsupial and eutherian clades with high support values.

The four superordinal clades of eutherian mammals, Laurasiatheria, Euarchontoglires, Xenarthra and Afrotheria, recognized by, e.g., Madsen et al. (2001), Murphy et al. (2001), Amrine-Madsen et al. (2003), and Nishihara et al. (2006) are all identified. Moreover, seven eutherian orders are identified as monophyletic: Carnivora, Cetartiodactyla, Chiroptera, Perissodactyla, Primata, Rodentia and Afrosoricida.

Several other higher-level mammalian relationships are found congruent with the most recent eutherian mammal phylogenies cited above: Ferae: Carnivora and Pholidota; Euarchonta: Primata, Dermoptera and Scandentia; Tethytheria: Sirenia and Proboscidea; Paenungulata: Hyracoidea and Tethytheria; Afroinsectivora: Afrosoricida and Macroscelidea; Afroinsectiphilia: Afroinsectivora and Tubulidentata.

This tree shows only a few discrepancies with other molecular phylogenies based on large concatenated DNA sequences: for instance, Cetartiodactyla is identified as Cetacea + Artiodactyla, but Hippopotamidae + Cetacea is not retrieved; also, a recent article by Hallström et al. (2007) strongly supports Xenarthra as the sister lineage to Afrotheria, reinforcing the clade 'Atlantogenata' already proposed by Delsuc et al. (2002).

These results indicate that *AMEL* exon 6, although composed of approximately 400 bp, is a very efficient phylogenetic marker for higher-level mammalian relationships that could be added to the current large data sets of DNA sequences.

Acknowledgments

We are greatly indebted to the following colleagues and laboratory collections for providing material for this work. Gray seal, tamandua, tapir, rhinoceros, manatee come from the Université de Montpellier 2, France (UMR 5554, Dr F. Catzeflis); tiger, cheetah, panda, pygmy hippopotamus, from the Zoo de Vincennes, Muséum national d'Histoire naturelle, France (Dr F. Ollivet and A. Lécu); dolphin and porpoise from the Muséum de la Rochelle, France (Dr W. Dabin); sea lion, black bear, flying lemur, tree shrew, fruit bat, roundleaf bat, shrew, three-toed sloth, tamandua, pangolin, hedgehog, tenrec, aquatic opossum, wallaby from the Laboratoire Mammifères et Oiseaux, Muséum national d'Histoire naturelle, France; African elephant, golden mole, elephant shrew, aardvark, hyrax, manatee from the University of Stellenbosch, South Africa (Dr. T.J. Robinson).

References

Amrine-Madsen, H., Koepfli, K.-P., Wayne, R.K., Springer, M.S., 2003. A new phylogenetic marker, apolipoprotein B, provides compelling evidence for eutherian relationships. Mol. Phylogent. Evol. 28, 225–240.

Bininda-Emonds, O.R.P., Cardillo, M., Jones, K.E., MacPhee, R.D.E., Beck, R.M.D., Grenyer, R., Price, S.A., Vos, R.A., Gittleman, J.L., Purvis, A., 2007. The delayed rise of present-day mammals. Nature 446, 507–512.

Delgado, S., Girondot, M., Sire, J.-Y., 2005. Molecular evolution of amelogenin in mammals. J. Mol. Evol. 60, 12–30.

Delsuc, F., Scally, M., Madsen, O., Stanhope, M.J., de Jong, W.W., 2002. Molecular phylogeny of living xenarthrans and the impact of character and taxon sampling on the placental tree rooting. Mol. Biol. Evol. 19, 1656–1671.

Hallström, B., Kullberg, M., Nilsson, M., Janke, A., 2007. Phylogenomic data analyses provide evidence that Xenarthra and Afrotheria are sistergroups. Mol. Biol. 24, 2059–2068.

Huysseune, A., Sire, J.-Y., 1998. Evolution of patterns and processes in teeth and tooth-related tissues in non-mammalian vertebrates. Eur. J. Oral Sci. 106 (suppl. 1), 437–481.

Ishiyama, M., Mikami, M., Shimokawa, H., Oida, S., 1998. Amelogenin protein in tooth germs of the snake *Elaphe quadrivirgata*, immunohistochemistry, cloning and cDNA sequence. Arch. Histol. Cytol. 61, 467–474.

Iwase, M., Satta, Y., Hirai, Y., Hirai, H., Imai, H., Takahata, N., 2003. The amelogenin loci span an ancient pseudoautosomal boundary in diverse mammalian species. Proc. Natl. Acad. Sci. USA 100, 5258–5263.

Kriegs, J.O., Churakov, G., Kiefmann, M., Jordan, U., Brosius, J., Schmitz, J., 2006. Retroposed elements as archives for the evolutionary history of placental mammals. PLoS Biol. 4, e91.

Li, W.-H., Yi, S., Makova, K.D., 2002. Male-driven evolution. Curr. Opin. Genet. Dev. 12, 650–656.

Madsen, O., Scally, M., Douady, C.J., Kao, D.J., Debry, R.W., Adkins, R., Amrine, H.M., Stanhope, M.J., de Jong, W.W., Springer, M.S., 2001. Parallel adaptive radiations in two major clades of placental mammals. Nature 409, 610–614.

Murphy, W.J., Elzirik, E., Johnson, W.E., Zhang, Y.P., Ryder, O.A., O'Brien, S.J., 2001. Molecular phylogenetics and the origin of placental mammals. Nature 409, 614–618.

Murphy, W.J., Pringle, T.H., Crider, T.A., Springer, M.S., Miller, W., 2007. Using genomic data to unravel the root of the placental mammal phylogeny. Genome Res. 17, 413–421.

Nishihara, H., Hasegawa, M., Okada, N., 2006. Pegasoferae, an unexpected mammalian clade revealed by tracking ancient retroposon insertions. Proc. Natl. Acad. Sci. USA 103, 9929–9934.

Ohno, S., 1967. Sex chromosomes and sex-linked genesMonographs on Endocrinology, vol. 1. Springer-Verlag, Berlin.

Posada, D., Crandall, K.A., 1998. MODELTEST: testing the model of DNA substitution. Bioinformatics 14, 818–819.

Rambaut, A., 1996. Se-Al. <http://evolve.zoo.ox.ac.uk/Se-Al/Se-Al.html>.

Richard, B., Delgado, S., Gorry, P., Sire, J.-Y., 2007. A study of polymorphism in human *AMELX*. Arch. Oral Biol. 52, 1026–1031.

Ronquist, F., Huelsenbeck, J.P., 2003. Mr Bayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19, 1572–1574.

Sanchez-Villagra, M.R., Narita, Y., Kuratani, S., 2005. Thoracolumbar vertebral number: the first skeletal synapomorphy for afrotherian mammals. Syst. Biodiversity, 1–7.

Sire, J.-Y., Davit-Béal, T., Delgado, S., Gu, X., 2007. The origin and evolution of enamel mineralization genes. Cells Tissues Organs 186, 25–48.

Sire, J.-Y., Huysseune, A., 2003. Formation of skeletal and dental tissues in fish: a comparative and evolutionary approach. Biol. Rev. 78, 219–249.

Smith, M.M., Coates, M.I., 2000. Evolutionary origins of teeth and jaws: developmental models and phylogenetic patterns. In: Teaford, M.F.,

Smith, Ferguson, M.W.J. (Eds.), Development, Function and Evolution of Teeth. Cambridge University Press, Cambridge, MA, pp. 133–151.

Springer, M.S., Cleven, G.C., Madsen, O., de Jong, W.W., Waddell, V.G., Amrine, H.M., Stanhope, M.J., 1997. Endemic African mammals shake the phylogenetic tree. Nature 388, 61–64.

Springer, M.S., Murphy, W.J., Eizirik, E., O'Brien, S.J., 2003. Placental mammal diversification and the Cretaceous–Tertiary boundary. Proc. Natl. Acad. Sci. USA 100, 1056–1061.

Stanhope, M.J., Waddell, V.G., Madsen, O., de Jong, W.W., Hedges, S.B., Cleven, G.C., Kao, D., Springer, M.S., 1998. Molecular support for multiple origins of insectivora and for a new order of endemic African insectivore mammals. Proc. Natl. Acad. Sci. USA 95, 9967–9972.

Swofford, D.L., 1998. PAUP*: phylogenetic analysis using parsimony (*and other methods), version 4. Sinauer Associates, Sunderland, Mass.

Tabuce, R., Marivaux, L., Adaci, M., Bensalah, M., Hartenberger, J.-L., Mahboubi, M., Mebrouk, F., Tafforeau, P., Jaeger, J.-J., 2007. Early Tertiary mammals from North Africa reinforce the molecular Afrotheria clade. Proc. R. Soc. Lond. B 274, 1159–1166.

Toyosawa, S., O'HUigin, C., Figueroa, F., Tichy, H., Klein, J., 1998. Identification and characterization of amelogenin genes in monotremes, reptiles, and amphibians. Proc. Natl. Acad. Sci. USA 95, 13056–13061.

Waddell, P.J., Shelley, S., 2003. Evaluating placental inter-ordinal phylogenies with novel sequences including RAG1, gamma-fibrinogen, ND6, and mt-tRNA, plus MCMC-driven nucleotide, amino acid, and codon models. Mol. Phylogenet. Evol. 28, 197–224.

Yi, S., Summers, T.J., Pearson, N.M., Li, W.-H., 2004. Recombination has little effect on the rate of sequence divergence in pseudoautosomal boundary 1 among humans and great apes. Genome Res. 14, 37–43.

Fukae M, Tanabe T, Murakami C, Dohi N, Uchida T, Shimizu M (1996) Primary structure of the porcine 89-kDa enamelin. Adv Dent Res 10:111–118

Gomis-Ruth FX, Bayes A, Sotiropoulou G, Pampalakis G, Tsetsenis T, Villegas V, Aviles FX, Coll M (2002) The structure of human prokallikrein 6 reveals a novel activation mechanism for the kallikrein family. J Biol Chem 277:27273–27281

Gutierrez SJ, Chaves M, Torres DM, Briceno I (2007) Identification of a novel mutation in the enamelin gene in a family with autosomal-dominant amelogenesis imperfecta. Arch Oral Biol 52:503–506

Hart PS, Michalec MD, Seow WK, Hart TC, Wright JT (2003a) Identification of the enamelin (g.8344delG) mutation in a new kindred and presentation of a standardized ENAM nomenclature. Arch Oral Biol 48:589–596

Hart TC, Hart PS, Gorry MC, Michalec MD, Ryu OH, Uygur C, Ozdemir D, Firatli S, Aren G, Firatli E (2003b) Novel ENAM mutation responsible for autosomal recessive amelogenesis imperfecta and localised enamel defects. J Med Genet 40:900–906

Hasegawa M, Kishino H, Yano T (1985) Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J Mol Evol 22:160–174

Higgins DG, Thompson JD, Gibson TJ (1996) Using CLUSTAL for multiple sequence alignments. Methods Enzymol 266:383–402

Hiss JA, Resch E, Schreiner A, Meissner M, Starzinski-Powitz A, Schneider G (2008) Domain organization of long signal peptides of single-pass integral membrane proteins reveals multiple functional capacity. PLoS ONE 3:e2767

Hu JC, Yamakoshi Y (2003) Enamelin and autosomal-dominant amelogenesis imperfecta. Crit Rev Oral Biol Med 14:387–398

Hu C-C, Fukae M, Uchida T, Qian Q, Zhang CH, Ryu OH, Tanabe T, Yamakoshi Y, Murakami C, Dohi N, Shimizu M, Simmer JP (1997a) Cloning and characterization of porcine enamelin mRNAs. J Dent Res 76:1720–1729

Hu CC, Fukae M, Uchida T, Qian Q, Zhang CH, Ryu OH, Tanabe T, Yamakoshi Y, Murakami C, Dohi N, Shimizu M, Simmer JP (1997b) Sheathlin: cloning, cDNA/polypeptide sequences, and immunolocalization of porcine enamel sheath proteins. J Dent Res 76:648–657

Hu CC, Simmer JP, Bartlett JD, Qian Q, Zhang C, Ryu OH, Xue J, Fukae M, Uchida T, MacDougall M (1998) Murine enamelin: cDNA and derived protein sequences. Connect Tissue Res 39:47–61

Hu CC, Hart TC, Dupont BR, Chen JJ, Sun X, Qian Q, Zhang CH, Jiang H, Mattern VL, Wright JT, Simmer JP (2000) Cloning human enamelin cDNA, chromosomal localization, and analysis of expression during tooth development. J Dent Res 79:912–919

Hu JC, Zhang CH, Yang Y, Karrman-Mardh C, Forsman-Semb K, Simmer JP (2001) Cloning and characterization of the mouse and human enamelin genes. J Dent Res 80:898–902

Hu JC, Yamakoshi Y, Yamakoshi F, Krebsbach PH, Simmer JP (2005) Proteomics and genetics of dental enamel. Cells Tissues Organs 181:219–231

Hu JC, Hu Y, Smith CE, McKee MD, Wright JT, Yamakoshi Y, Papagerakis P, Hunter GK, Feng JQ, Yamakoshi F, Simmer JP (2008) Enamel defects and ameloblast-specific expression in Enam knock-out/lacz knock-in mice. J Biol Chem 283:10858–10871

Kang HY, Seymen F, Lee SK, Yildirim M, Tuna EB, Patir A, Lee KE, Kim JW (2009) Candidate gene strategy reveals ENAM mutations. J Dent Res 88:266–269

Kawasaki K, Weiss KM (2003) Mineralized tissue and vertebrate evolution: the secretory calcium-binding phosphoprotein gene cluster. Proc Natl Acad Sci USA 100:4060–4065

Kawasaki K, Weiss KM (2008) SCPP gene evolution and the dental mineralization continuum. J Dent Res 87:520–531

Kelley JL, Swanson WJ (2008) Dietary change and adaptive evolution of enamelin in humans and among Primates. Genetics 178:1595–1603

Kim JW, Seymen F, Lin BP, Kiziltan B, Gencay K, Simmer JP, Hu JC (2005a) ENAM mutations in autosomal-dominant amelogenesis imperfecta. J Dent Res 84:278–282

Kim JW, Hu JCC, Lee JI, Moon SK, Kim YJ, Jang KT, Lee SH, Kim CC, Hahn SH, Simmer JP (2005b) Mutational hot spot in the DSPP gene causing dentinogenesis imperfecta type II. Hum Genet 116:186–191

Pond SLK, Frost SD (2005a) Not so different after all: a comparison of methods for detecting amino acid sites under selection. Mol Biol Evol 22:1208–1222

Pond SLK, Frost SD (2005b) Datamonkey: rapid detection of selective pressure on individual sites of codon alignments. Bioinformatics 21:2531–2533

Pond SLK, Frost SD, Muse SV (2005) HyPhy: hypothesis testing using phylogenies. Bioinformatics 21:676–679

Kozak M (1984) Compilation and analysis of sequences upstream from the translational start site in eukaryotic mRNAs. Nucleic Acids Res 12:857–872

Kozak M (1991) A short leader sequence impairs the fidelity of initiation by eukaryotic ribosomes. Gene Expr 1:111–115

Kurys G, Tagaya Y, Bamford R, Hanover JA, Waldmann TA (2000) The long signal peptide isoform and its alternative processing direct the intracellular trafficking of interleukin-15. J Biol Chem 275:30653–30659

Lu Y, Papagerakis P, Yamakoshi Y, Hu JC, Bartlett JD, Simmer JP (2008) Functions of KLK4 and MMP-20 in dental enamel formation. Biol Chem 389:695–700

Mårdh CK, Bäckman B, Holmgren G, Hu JC, Simmer JP, Forsman-Semb K (2002) A nonsense mutation in the enamelin gene causes local hypoplastic autosomal dominant amelogenesis imperfecta (AIH2). Hum Mol Genet 11(9):1069–1074

Martoglio B, Dobberstein B (1998) Signal sequences: more than just greasy peptides. Trends Cell Biol 8:410–415

Masuya H, Shimizu K, Sezutsu H, Sakuraba Y, Nagano J, Shimizu A, Fujimoto N, Kawai A, Miura I, Kaneda H, Kobayashi K, Ishijima J, Maeda T, Gondo Y, Noda T, Wakana S, Shiroishi T (2005) Enamelin (Enam) is essential for amelogenesis: ENU-induced mouse mutants as models for different clinical subtypes of human amelogenesis imperfecta (AI). Hum Mol Genet 14:575–583

Ozdemir D, Hart PS, Firatli E, Aren G, Ryu OH, Hart TC (2005) Phenotype of ENAM mutations is dosage-dependent. J Dent Res 84:1036–1041

Patthy L (1999) Genome evolution and the evolution of exon-shuffling—a review. Gene 238:103–114

Rajpar MH, Harley K, Laing C, Davies RM, Dixon MJ (2001) Mutation of the gene encoding the enamel-specific protein, enamelin, causes autosomal-dominant amelogenesis imperfecta. Hum Mol Genet 10:1673–1677

Rambaut A, Bromham L (1998) Estimating divergence dates from molecular sequences. Mol Biol Evol 15:442–448

Ravindranath RM, Moradian-Oldak J, Fincham AG (1999) Tyrosyl motif in amelogenins binds N-acetyl-D-glucosamine. J Biol Chem 274:2464–2471

Ravindranath HH, Chen LS, Zeichner-David M, Ishima R, Ravindranath RM (2004) Interaction between the enamel matrix proteins amelogenin and ameloblastin. Biochem Biophys Res Commun 323:1075–1083

Reuter M, Engelstadter J, Fontanillas P, Hurst LD (2008) A test of the null model for 5' UTR evolution based on GC content. Mol Biol Evol 25:801–804

Ryu OH, Fincham AG, Hu CC, Zhang C, Qian Q, Bartlett JD, Simmer JP (1999) Characterization of recombinant pig enamelysin

BMC Evolutionary Biology

Research article

Open Access

Hen's teeth with enamel cap: from dream to impossibility

Jean-Yves Sire*[1], Sidney C Delgado[1] and Marc Girondot[2]

Address: [1]Université Pierre & Marie Curie-Paris 6, UMR 7138 "Systématique, Adaptation, Evolution", 7 quai St-Bernard, 75005, Paris, France and [2]Université Paris-Sud, UMR 8079 "Ecologie, Systématique et Evolution", 91160 Orsay, & Département Systématique et Evolution, Muséum National d'Histoire Naturelle de Paris, 25 rue Cuvier, 75005, Paris, France

Email: Jean-Yves Sire* - jean-yves.sire@upmc.fr; Sidney C Delgado - sidney.delgado@upmc.fr; Marc Girondot - marc.girondot@ese.u-psud.fr

* Corresponding author

Published: 5 September 2008

BMC Evolutionary Biology 2008, **8**:246 doi:10.1186/1471-2148-8-246

Received: 18 January 2008
Accepted: 5 September 2008

This article is available from: http://www.biomedcentral.com/1471-2148/8/246

Abstract

Background: The ability to form teeth was lost in an ancestor of all modern birds, approximately 100-80 million years ago. However, experiments in chicken have revealed that the oral epithelium can respond to inductive signals from mouse mesenchyme, leading to reactivation of the odontogenic pathway. Recently, tooth germs similar to crocodile rudimentary teeth were found in a chicken mutant. These "chicken teeth" did not develop further, but the question remains whether functional teeth with enamel cap would have been obtained if the experiments had been carried out over a longer time period or if the chicken mutants had survived. The next odontogenetic step would have been tooth differentiation, involving deposition of dental proteins.

Results: Using bioinformatics, we assessed the fate of the four dental proteins thought to be specific to enamel (amelogenin, AMEL; ameloblastin, AMBN; enamelin, ENAM) and to dentin (dentin sialophosphoprotein, DSPP) in the chicken genome. Conservation of gene synteny in amniotes allowed definition of target DNA regions in which we searched for sequence similarity. We found the full-length chicken AMEL and the only N-terminal region of DSPP, and both are invalidated genes. AMBN and ENAM disappeared after chromosomal rearrangements occurred in the candidate region in a bird ancestor.

Conclusion: These findings not only imply that functional teeth with enamel covering, as present in ancestral Aves, will never be obtained in birds, but they also indicate that these four protein genes were dental specific, at least in the last toothed ancestor of modern birds, a specificity which has been questioned in recent years.

Background

Modern birds derive from theropod dinosaurs. The most ancient Avialae [1] is the well-known "dinobird" *Archaeopteryx lithographica*, which lived some 150 million years ago (mya) and possessed teeth. The most recent toothed Avialae in the fossil record, the ornithurine birds *Hesperornis regalis* and *Ichthyornis dispar*, are known from the late Cretaceous. To date, *Ichthyornis* is the closest Avialae to the common ancestor of modern birds (Aves) [2]*Ichthyornis* specimens trace from the late Cenomanian, 95 mya, to early Campanian, 80 mya, but we do not know whether fossil taxa closer than *Ichthyornis* to the most recent common ancestor of Aves have teeth. Therefore, we can estimate that tooth loss in crown Aves arose maximally on the stem lineage between *Ichthyornis* and Aves and minimally in the most recent common ancestor of Aves, the origin of

modern birds (Neornithes). Neornithine fossils are found near the end of the Cretaceous period (Campanian, 80 mya) [3], and the recent discovery of a close relative to ducks (Anseriformes) in the Maastrichtian of Antarctica (70 mya) indicates that Aves originated long before the Cretaceous/Tertiary boundary [4]; they probably arose even earlier than 80 mya, although they may have diversified later, during the early Cenozoic [5]. The deep Cretaceous origination inferred from molecular studies (120–130 mya) [6] is, however, still earlier, but establishing accurate calibration times for molecular phylogenies on the basis of fossil data is difficult [7].

Would birds be able to rebuild teeth with reactivation of the odontogenic pathway under appropriate conditions? In other words, are all genes required for complete odontogenesis still active 100-80 million years (at least) after tooth loss in a bird ancestor? A positive answer would mean that these genes serve functions other than building teeth [8]. Otherwise, no-longer-useful dental-specific genes might have been invalidated through random accumulation of mutations.

There are two justifications for asking this question: the first is the growing evidence in mammals that some dental proteins, believed to be specific to enamel or dentin matrix, are expressed in other organs and therefore are suspected of having other functions [9-12]; The second reason is that several recombination experiments and the observations made on a chicken mutant strongly suggest that resurrecting teeth in birds could be possible. In 1980, Kollar and Fischer [13] recombined chick dental epithelium with mouse mesenchyme and obtained teeth with an enamel cover, the famous "hen's teeth." However, a possible contamination of the mouse mesenchyme by mouse epithelium makes the interpretation uncertain. Chen et al. [14] have shown that the early odontogenic pathway remains inducible in chicken. They suggested that the loss of odontogenic *Bmp4* expression (i.e., inactivation of the genetic pathway leading to tooth formation) may be responsible for the early arrest of tooth development in birds. Performing transplantations of mouse neural crest cells into the chick embryo, Mitsiadis et al. [15] showed that avian dental epithelium can still induce a nonavian developmental program in mouse neural crest-derived mesenchyme, resulting in tooth germ formation. These last two experiments indicate that under appropriate conditions, the odontogenic capacity of chicken dental epithelium can be reactivated. However, if the re-activation of such an odontogenic pathway is a prerequisite to initiating tooth development and to reaching an advanced stage of tooth morphogenesis, it is insufficient for forming functional teeth with a dentin cone covered with enamel. At the end of the pathway, structural genes might have been activated, but it seems they have not. Unfortunately, the duration of these experiments was too short for determining whether or not tooth differentiation would have eventually occurred. Also interesting are recent observations made in *talpid*^{2} (*ta*^{2}), a mutant chicken in which the development of several organ systems is affected. *ta*^{2} was shown to develop rudimentary teeth reminiscent of first-generation teeth in crocodiles [16]. Unfortunately again, the oldest *ta*^{2} died at stages E16, before hatching, and further tooth development was not assessable.

An alternative approach for determining whether or not obtaining hen's teeth similar to crocodile and lepidosaurian teeth is not an impossible dream was to look for the fate of the dental protein genes, 100 million years (my) after tooth loss. Four structural proteins are considered specific to dental tissues: one dentin matrix protein, dentin sialophosphoprotein (DSPP), and three enamel matrix proteins (EMPs) – amelogenin (AMEL, the major protein of the enamel matrix), ameloblastin (AMBN), and enamelin (ENAM). AMEL and AMBN genes have been sequenced in reptiles and they were shown to share conserved regions with their mammalian orthologs [17,18]. In addition, during reptilian amelogenesis both genes are similarly expressed as described in mammals, and ameloblasts are similarly differentiated [19,20]. Therefore, there is no doubt that they played a similar function and were necessary for proper enamel formation not only in the ancestral theropod dinosaurs, but also in archeopteryx and in the last common toothed Aves ancestor to modern birds. For what concerns ENAM and DSPP, the two other tooth-specific genes, we recently found that they are also present in a lizard genome http://pre.ensembl.org/Anolis_carolinensis/index.html and expressed (Sire et al., unpublished data). All of this supports the idea that these four dental proteins were present and functional when the teeth were lost in the last common ancestor to modern birds.

Previous molecular attempts to localize AMEL in chicken DNA have been unsuccessful [21]. Even when the chicken genome sequence became available http://www.ensembl.org/Gallus_gallus/index.html, the genes encoding the four dental proteins were not found using either computer prediction or bioinformatics [22,23]. Here, using software designed to screen large DNA regions for weak sequence similarity (UniDPlot, Girondot and Sire, unpublished), we have found that AMEL and DSPP are invalidated genes and that ENAM and AMBN have probably disappeared from the chicken genome through chromosomal rearrangement.

Methods

Blast search

AMEL, AMBN, ENAM, DSPP were searched (BLASTN) in the most recent chicken assembly genome (WASHUC2)

using either full-length amniote sequences or various e-primers defined from conserved regions. In addition to various mammalian sequences available for these four genes in databanks (see NCBI and Ensembl websites), we used crocodile AMEL and AMBN sequences (GenBank accession: AF095568 and AY043290, respectively). For ENAM and DSPP, only mammalian sequences were available in the databanks.

Search of target genes usingUniDPlot

Gene synteny in mammals and chicken was established using the NCBI website (mapviewer).

We searched for sequence similarity with UniDPlot software (Girondot and Sire, unpublished), using crocodile AMEL exon 2 (54 bp), which is well conserved [17]. Basically, UniDPlot uses a projection of the maximum of the matrix of similarity from a 2D dot-plot along the largest axis.

Alignments were performed using Se-Al (v2.0a11 Carbon) and checked by hand.

Results and discussion

Search for dental protein genes in the chicken genome using BLASTN

Searching for the four genes (AMEL, AMBN, ENAM, DSPP) in the chicken genome failed to return any result. Blast searching for these genes proved to be unfruitful, even when low sensitivity (distant homology) was used. The crocodile-bird divergence is estimated to have occurred approximately 250 mya [24], and the mammal-reptile (birds) divergence is estimated to have occurred 310 mya [25]. If AMEL and AMBN were not dental specific in ancestral toothed birds and had other functions, they might still be present in the chicken genome as functional genes. We at least expected that conserved coding regions, which are subjected to strong constraints, would have been found. This negative result means that either the sequences have strongly derived over 250 my (acquisition of a new function or pseudogenization) or these genes have disappeared. For ENAM and DSPP, the lack of positive hits could be (in addition to the two hypotheses evoked above) the consequence of this evolutionary distance, which could have led to large differences between mammalian and chicken sequences.

Whatever their fate, the complete deletion of all four genes (e.g., as a consequence of chromosomal rearrangements) in the chicken genome was unlikely because they are not located in the same genomic regions in mammals. Because gene synteny has been shown to be largely conserved in comparisons of mammalian and chicken genomes, we decided to use a synteny-based approach to try to find the chicken dental protein genes.

Search of target genes using synteny

Amelogenin (AMEL)

In placental mammals, AMEL maps on the X chromosome (e.g., primates, rodents, cow, horse, and dog) and a copy is located on the Y chromosome in some species. In opossum (marsupials), AMEL is mapped on chromosome 7. In these species, AMEL is located close to the rhoGTPase activating protein 6 gene (ARHGAP6). For instance, in humans, AMELX is located at position Xp22.3, between ARHGAP6 and HCCS (holocytochrome C synthetase) gene. MID1 (midline 1) and MSL3L1 (male-specific lethal 3-like 1) mark out this region (Fig. 1A). AMELX codes in antisense within the 200 kb large intron 1 of ARHGAP6, and its 5' UTR is located at approximately 40 kb far from the 5' region of ARHGAP6 exon 2. In the opossum, AMEL is similarly located but 58 kb from ARHGAP6 exon 2.

In chicken, ARHGAP6 (LOC418642), MID1, and MSL3L1 (LOC418641) are found close one to another on chromosome 1 (Fig. 1B), but compared to their location in humans, chicken MID1 and MSL3L1 are inverted, while HCCS is located on chicken chromosome 8 (LOC424482). In the target region, i.e., between ARHGAP6 and MID1, the GenBank prediction program indicates neither the presence of a putative candidate gene locus nor of a pseudogene, which might have been Ψ-AMEL (Fig. 1B).

In the chicken, we localized exon 2 of ARHGAP6 and selected a 200-kb DNA strand, running from the 5' region of exon 2 to the 5' region of MID1, as the most probable region for housing chicken AMEL. Searching for sequence similarity using crocodile AMEL exon 2 led to a positive hit, approximately 38 kb upstream of chicken ARHGAP6 exon 2 (Fig. 2). Such a distance from ARHGAP6 was expected when considering the location of AMEL in mammals (e.g., 40 kb in human, 58 kb in opossum). We extracted and aligned this sequence with crocodile AMEL exon 2 (Fig. 3). With the exception of four inserted nucleotides, the chicken sequence was unequivocally identified as the ortholog of crocodile AMEL exon 2, with 68.8% nucleotide identity. When the four inserted nucleotides are removed, the deduced putative amino acid sequence encoded by chicken AMEL exon 2 is similar to known sequences. However, the insertion of four nucleotides would lead to a shift in the reading frame, changing the amino acid sequence and the chemical nature of chicken AMEL (Fig. 3). Therefore, we conclude that the chicken AMEL gene is invalidated and has become a pseudogene (Ψ-AMEL).

We proceeded similarly using crocodile AMEL exons 3, 5, and 6, focusing on the chicken DNA region adjacent to AMEL exon 2. The full-length sequence of chicken Ψ-AMEL was retrieved (Fig. 4); GenBank accession number;

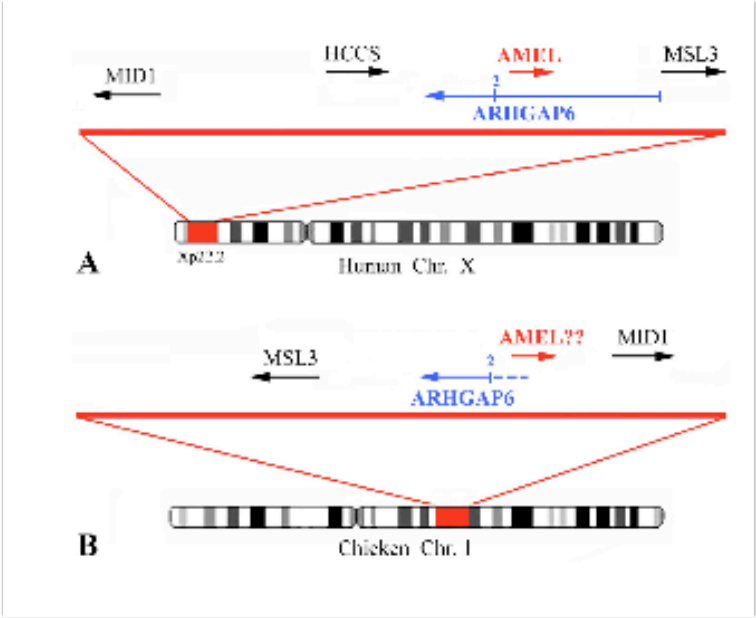


Figure 1 Location of amelogenin (AMEL) on human chromosome X.
(*A*) Location of amelogenin (AMEL) on human chromosome X. (*B*) Homologous region on chicken chromosome 1 and the putative location of AMEL. In chicken, HCCS is located on chromosome 8 (LOC424482). ARHGAP6 exon 2 is indicated by the numeral 2. Gene descriptions corresponding to the symbols can be found at the NCBI web site: http://www.ncbi.nlm.nih.gov/.

EU340348). In tetrapods, exons 2, 3, and 5 and the 5' and 3' regions of exon 6 encode the well-conserved N- and C-terminal AMEL regions, while most of exon 6 encodes the largest and variable region [26,27]. Chicken Ψ-AMEL exon 3 (indels), exon 5 (no indel), 5' exon 6 (indels), and 3' exon 6 (indels) show a high percentage of nucleotide identity with crocodile AMEL sequences (63.2, 73.3, 54.8, and 64.0%, respectively), while the central region of exon 6 shows less than 50% nucleotide identity (Fig. 5). Such a low percentage in this variable region is not surprising if we consider that mutations have accumulated in this region during the long period from the divergence of the crocodile-bird lineages to the last common ancestor of modern birds. In addition to point substitutions, Ψ-AMEL exon 6 shows numerous indels. Nevertheless, when included in a phylogenetic analysis (using PAUP 4.0) with currently available AMEL sequences in amniotes, chicken Ψ-AMEL locates, as expected, as the sister gene of crocodile AMEL (Fig. 6). In addition to confirm that chicken Ψ AMEL is really an AMEL gene, this finding indicates that the mutations that have occurred at random during approximately 100 my have not blurred the phylogenetic signal contained in the AMEL sequence [28,29].

Ameloblastin (AMBN) and enamelin (ENAM)

AMBN and ENAM are located adjacent one another on autosomal chromosomes: chr. 4 in human and chimpanzee, chr. 5 in rhesus macaque, mouse, and opossum, chr. 14 in rat, chr. 6 in cow, chr. 3 in horse, and chr. 13 in dog. Because gene synteny is conserved in these regions, we searched for AMBN and ENAM using the same approach as described for AMEL.

In humans and in the other mammals in which they have been mapped, AMBN and ENAM are flanked on the one side by the immunoglobulin J peptide gene (IGJ) and on the other side by the other members of the secretory calcium-binding phosphoprotein (SCPP) family, which comprises ameloblast-secreted protein genes (amelotin, or AMTN, and odontogenic ameloblast associated, or ODAM) and several salivary and milk protein genes [30,31]. The SCPPs are flanked by SULT1E1, a member of the sulfotransferase family 1E (Fig. 7A). In chicken, IGJ is located on chr. 4, but no members of the SCPPs (i.e., enamel, salivary, and milk protein genes) adjacent to it on mammalian chromosomes were predicted by computer analysis to reside in this region (Fig. 7B). Moreover, in a

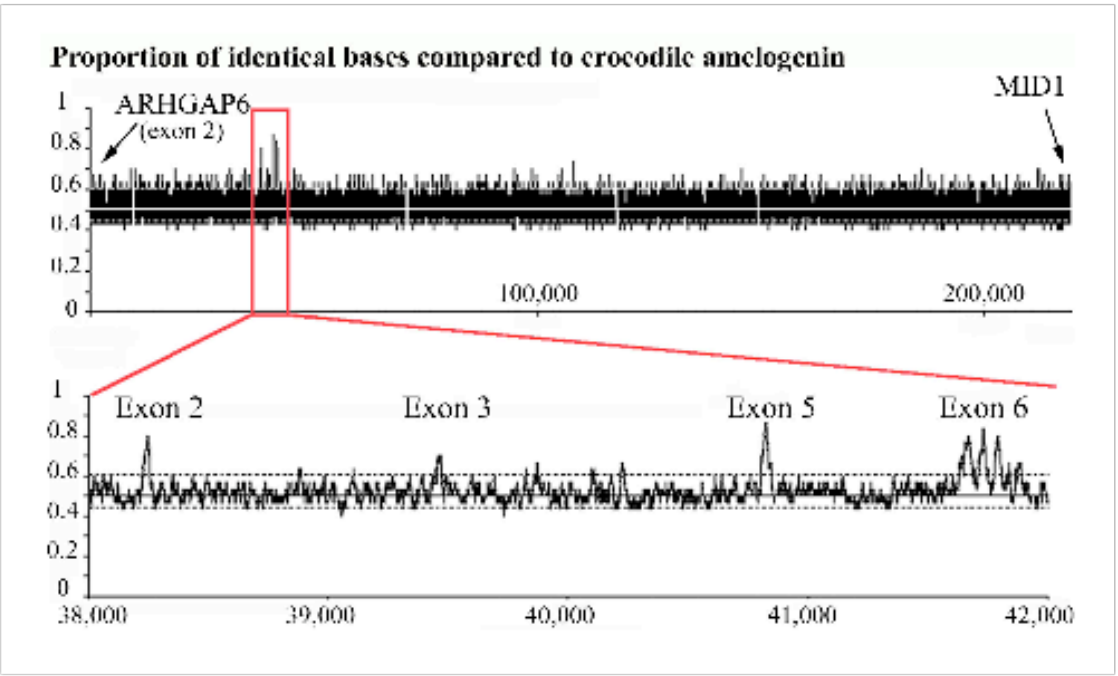


Figure 2
Result of sequence similarity search for AMEL in the target region of the chicken genome. This region is delimited by two flanking genes, ARHGAP6 (exon 2) and MID1. This region (200 kb) was extracted, and a similarity search was performed using crocodile AMEL exon 2, then exons 3 and 5 and the beginning of exon 6 (Figure 5). We used UniDPlot software (Girondot and Sire, unpublished), an extension of the dot-plot method, in which the maximum similarity index between both sequences is shown on the axis of the largest sequence. Significant identity was tested by calculating the distribution under H_0 limits obtained by random sampling of sequences. Top: Candidate region of chicken DNA showing the hits. Bottom: Detail of the chicken AMEL gene region found 38 kb from the 5' region of ARHGAP6 exon 2.

```
Crocodile   ATGGAGGGCTGG--ATG--TTGATCACTTGCCTACTAGGTGCAACATTTGCTATACCA
Ψ-Chicken   ATGGAGGACAGACTATTTATTGACTGCTTGCCTCCTAGGAGCACTGTTTGCTATGCCA
            ******* * *    ** *  ***    ******* ***** ***    ******** ***

Ψ-Chicken    MEDRLFIDCLPPRSTVCYA

Crocodile   MEGWMLITCLLGATFAIP
"Chicken"   MEDRILTACLLGALFAMP
            **    *  ***** ** *
```

Figure 3
Chicken Ψ-AMEL exon 2 analysis. Top: Alignment of chicken and crocodile AMEL exon 2 sequences. Four nucleotides are inserted (red) in chicken Ψ-AMEL exon 2 (signal peptide), leading to a shift in the reading frame. Middle: Putative deduced amino acid sequence from chicken Ψ-AMEL exon 2. Bottom: The four inserted codons were removed from the Ψ-AMEL sequence, which was translated and aligned to the crocodile sequence; both amino acid sequences are highly similar.

Chicken Ψ-AMEL mRNA

ex2 ATGGAGGACAGA**CT**ATT**TA**TTGACTGCTTGCCTCCTAGGAGCACTGTTTGCTATGCCA
ex3 CTACCCCTTCCCCCTCCTATCTAACACACCCTGGTTTCATCAACTTGAGTTGAG
ex5 GCACAAACACCTTTGAAAAGGCATCAGAGCATGATGACACCCCAG
ex6 TTCCCATT**TAA**TGGTTACAACTAGACAGAAGCTGACAAGAACACCAACCAGTTACAAGCAACATCTACAAA
TGGAGAGCTTACTATCACCCCAGCACACCCCCTTGGTGGCACTCCAGCACCAGCTGATGTAAATTCCCAGGCTAT
TTCCAGTTCTACCACTAGCGCAGCACCTACCAAGCCTGCCAATGCCAGCTCAAACCACACAGCTGCACACAACAA
AAGAGGCCTCAGCATCCTGCAAATCCCAACCCACCGTTGCACCCCAGTGGCTGGGGAGTCCCCATATGCACATGT
GCCCCCTGTCAGGGACTCCTCTGGAGCCAAGGCAGCCAGACAACAAAGCAAAGGAAAACA
ex7 TAT

Chicken Ψ-AMEL protein

MEDRLFIDCLPPRSTVCYAIPLLSNTPWFHQLELRCTNTFEKASEHDDTPVFLFSSHLMVTTRQKTSYKQHLQME
SLLSPQHTPLLTRTPVALQHQLM**Z**

Chicken Ψ-AMEL : exon-intron boundaries and intron size

gaagtactttctcctcttacttcag **ex2** gtgagtattacggtcatcttgcaac

intron 2 = 1163 bp

tatataccagttttgttttttctag **ex3** gtaaaatgttttgatctttttgaca

intron 3 = 1319

tttctccttttcttcccttttagaag **ex5** gtatcacacttcagttttcttcagc

intron 5 = 702 bp

tggatggctttctcttcctctttag **ex6** gtaagaaagctttggttcttccccc

intron 6 = 735 bp

ttattttctgagttaaatagaacag **ex7**

Figure 4 Pseudo-gene mRNA and deduced amino acid sequence.
Chicken Ψ-amelogenin mRNA and deduced amino acid sequence. Insertion of four nucleotides (in red) in exon 2 leads to a reading frameshift, which changes the amino acids in the N-terminal region and results in a premature stop codon in exon 6 (in red). The intron-exon boundary and the intron size are also indicated.

comparison of the chicken and human chromosomal regions adjacent to IGJ, it appears that intrachromosomal rearrangements have occurred. In the chicken chromosome, we identified two inversions in the candidate region adjacent to IGJ.

Two regions (14 and 700 kb) were designated as possibly housing AMBN and ENAM (Fig. 7B). We performed a sequence similarity search using the well-conserved exon 2 sequences (54 bp) of crocodile AMBN and human ENAM [32]. No positive hit was obtained in these regions.

Figure 5 Exon 2 of chicken and crocodile amelogenin mRNA, with percentage of nucleotide identity (in brackets)
Comparison of chicken and crocodile amelogenin mRNA, with percentage of nucleotide identity (in brackets). Start and stop of translation, and polyadenylation sites are underlined. Crocodile = *Paleosuchus palpebrosus* (accession no: AF095568).

These genes have been likely deleted from the chicken genome as a consequence of intrachromosomal rearrangements, which have probably occurred in the lineage that led to the last common ancestor of modern birds. The recently sequenced lizard genome (*Anolis carolinensis*) in which we found the enamel protein genes (Sire et al., unpublished data) will be useful for determining whether or not the synteny observed in this region in mammals was conserved until the divergence of the lepidosaurian and archosaurian lineages, 255 mya [24].

Dentin sialophosphoprotein (DSPP)

In human and other sequenced mammalian genomes, DSPP belongs to the so-called SIBLING cluster, which consists of five genes coding for dentin and bone proteins. It is located on the same autosomal chromosome as AMBN and ENAM (i.e., chr. 4 in humans), except in the dog in which the SIBLINGs are mapped on chr. 14 instead of chr. 13. These five genes are arranged side to side on the chromosome and flanked on the one side by SPARC-like 1 (SPARCL1) and on the other side by polycystic kidney

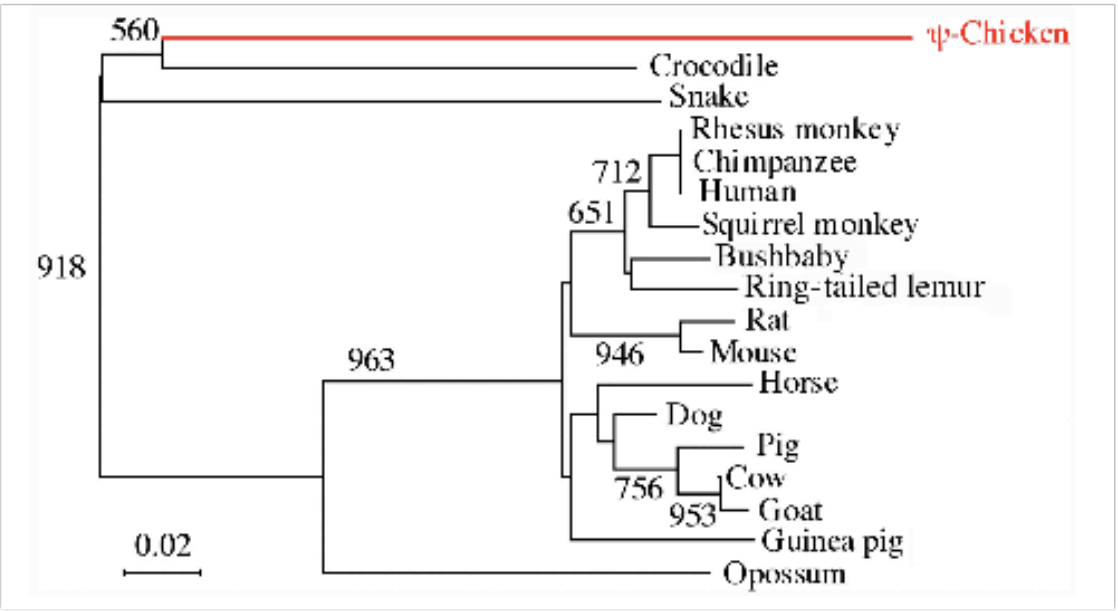


Figure 6
Phylogenetic analysis of chicken Ψ-AMEL. GenBank accession number: chicken Ψ-AMEL, *Gallus gallus*, EU340348; Crocodile (Caiman), *Paleosuchus palpebrosus*, AF095568; Snake, *Elaphe quadrivirgata*, AF118568; Rhesus monkey, *Macaca mulatta*, EF537871; Chimpanzee, *Pan troglodytes*, AB091781; Human, *Homo sapiens*, M86932; Squirrel monkey, *Saimiri sciureus*, AB091783; Bushbaby, *Otolemur garnettii*, AB091787; Ring-tailed lemur, *Lemur catta*, AB091785; Rat, *Rattus norvegicus*, U67130; Mouse, *Mus musculus*, D31769; Horse, *Equus caballus*, AB032193; Dog, *Canis familiaris*, XM_548858; Pig, *Sus scrofa*, U43405; Cow, *Bos taurus*, M63499; Goat, *Capra hircus*, AF215889; Guinea pig, *Cavia porcellus*, AJ012200; Opossum, *Monodelphis domestica*, U43407.

disease 2 (PKD2) (Fig. 8A). DSPP is located between SPARCL1 and DMP1 (dentin matrix protein 1).

In the chicken genome, the SIBLINGs are conserved in synteny and are mapped on chromosome 4 (Fig. 8B). The SIBLING cluster is more than 12 times denser in chicken than in human genome (40 kb versus 510 kb, respectively), with the genes oriented in the opposite direction from that in mammals. However, between SPARCL1 and DMP1, the GenBank computer prediction program indicates the presence of neither a putative candidate gene locus for DSPP, nor a pseudogene, although the 5' UTRs of these two genes are separated by a DNA region of 10.9 kb, strongly suggesting the possible presence of DSPP (Fig. 8B).

We extracted this candidate DNA region and performed a sequence similarity search using human DSPP exon 2 (51 bp), the best conserved exon in mammals (Sire, unpublished results). We obtained a positive hit, located in the middle region of the intergenic sequence, approximately 5,800 bp from DMP1 (Fig. 8C). This sequence (50 bp) was found to share 54% nucleotide identity with human DSPP exon 2, indicating that we have identified the putative chicken DSPP exon 2 (Fig. 8D). In addition to numerous substitutions of well-conserved residues in mammalian DSPP, one nucleotide has been deleted, leading to a reading frame shift were this sequence to be translated. Therefore, in chicken, DSPP was invalidated through pseudogenization. Using the other exons of human DSPP (exons 3, 4, and 5), we screened the DNA region located between Ψ-DSPP exon 2 and DMP1 but did not identify regions having more than 50% nucleotide identity. Nevertheless, on the one hand, these regions are more variable than exon 2 in mammals and, on the other hand, the evolutionary distance between chicken and human is 310 my [24]. Additional DSPP sequences in reptiles, and particularly in the lizard *Anolis carolinensis* (Sire et al., unpublished data), would allow a better detection of the other DSPP exons in this target region of chicken chromosome 4. It is noteworthy, however, that this region in the chicken genome is very short (10.9 kb), and we did not find the numerous and typical SDSSD repeats characterizing DSPP exon 5, which strongly suggests that this

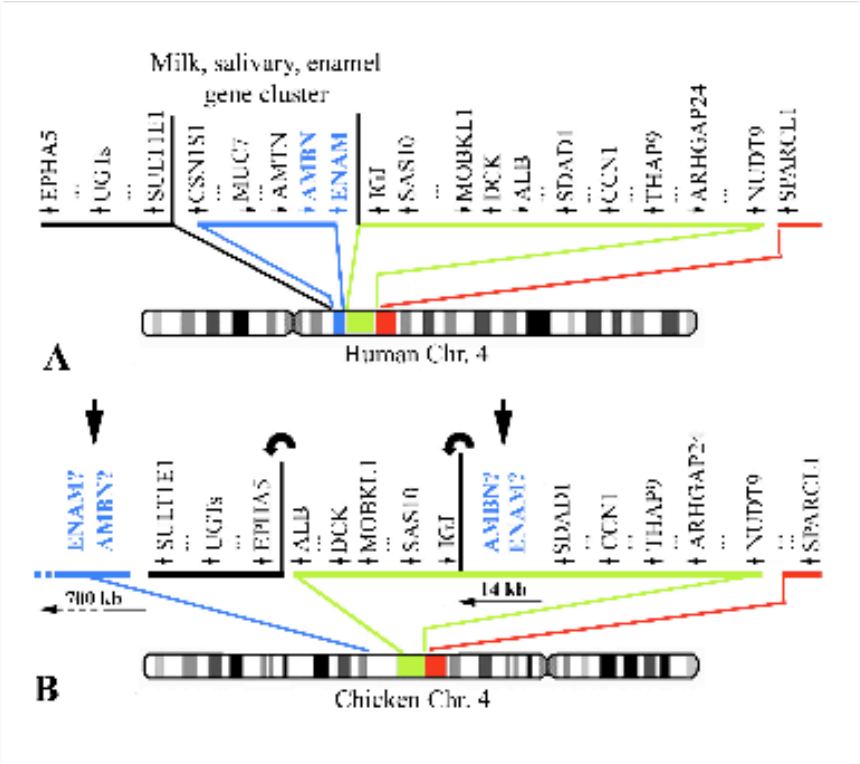


Figure 7 amelin (AMBN) and enamelin (ENAM) on human chromosome 4. (*B*) Homologous region on chicken **(*A*) Location of ameloblastin (AMBN) and enamelin (ENAM) on human chromosome 4. (*B*) Homologous region on chicken chromosome 4**. The position of several gene clusters is different in both chromosomes. Two gene inversions (curved arrows) have occurred in the candidate region putatively housing AMBN and ENAM leading to two likely locations for these genes on chicken chromosome: either adjacent to sulfotransferase 1E1 (SULT1E1) or to immunoglobulin J peptide. See the NCBI website for gene descriptions corresponding to the symbols.

exon has been deleted from the chicken genome. These numerous mutations in chicken Ψ-DSPP exon 2 and the disappearance of most of the sequence indicate that DSPP was invalidated for a long evolutionary period, which could correspond to the loss of teeth in the last ancestor of modern birds.

Conclusion

Eliciting well-developed, reptilian teeth (i.e. with enamel cap) in chicken will remain unachievable because all genes encoding the structural proteins crucial for enamel and dentine formation have been invalidated or have disappeared from the chicken genome. The odontogenic pathway remains inducible in chicken embryos because the genes required for tooth morphogenesis remain active in the chicken, involved in many developmental processes. We can speculate that the tooth germs that form with experimental reactivation of this pathway or in *ta*^{2} chicken mutants could develop until an advanced stage of predentin deposition because the process to this point requires mainly collagen matrix deposition. However, the next step of tooth development, during which enamel matrix proteins are deposited, either could never be activated or if it was (in the lack of data on the promoter sequence we cannot demonstrate that the AMEL gene is not translated) the protein would not be functional, and enamel will not form.

Another focus of this study is to demonstrate clearly that the four dental protein genes were tooth specific, at least in the last common toothed ancestor of modern birds. After the loss of teeth 100-80 mya, the four dental proteins became no longer useful; when the functional pressure relaxed on the coding genes, they started to accumulate

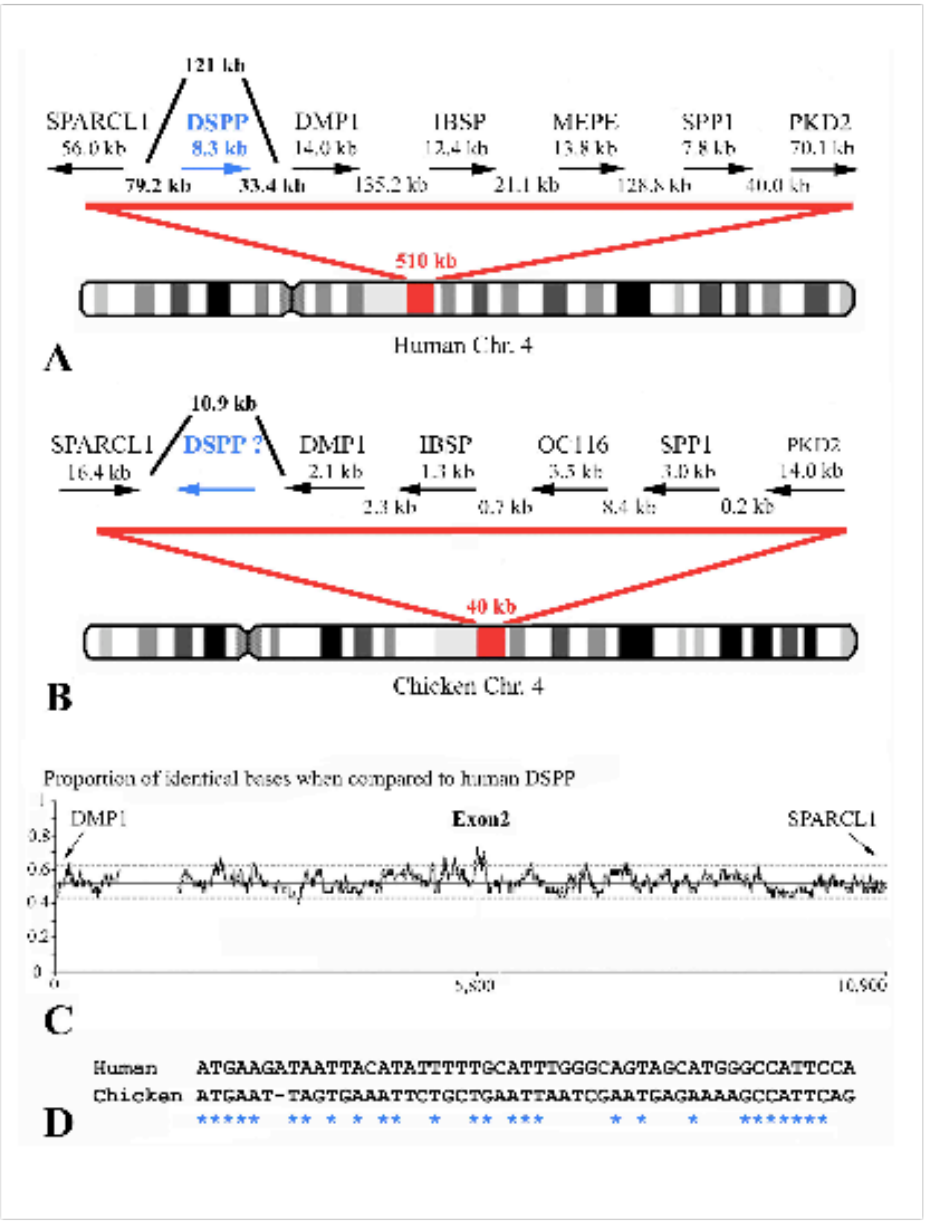


Figure 8 an of the dentin sialophosphoprotein (DSPP) and other SIBLING genes on human chromosome 4. **(*A*) Location of the dentin sialophosphoprotein (DSPP) and other SIBLING genes on human chromosome 4.(*B*)** Homologous region on chicken chromosome 4 and putative location of DSPP. Note that the SIBLING cluster is more compact in chicken than in human. OC116 and MEPE are orthologs. **(*C*)** Result of the similarity search in the candidate region between DMP1 and SPARCL1 using human DSPP exon 2. Chicken Ψ-DSPP exon 2 was found 5,800 bp from DMP1. **(*D*)** Alignment of chicken and crocodile DSPP exon 2 showing 54% nucleotide identity. See the NCBI website for gene descriptions corresponding to the symbols.

mutations at random. After a period of 100 my, it is not surprising that they are now pseudogenes or have disappeared after chromosomal rearrangement events. In the currently ongoing sequencing of the genome of the zebrafinch, a passeriform, we have found AMEL exon 2, with a deletion of 12 bases and a base substitution leading to a premature stop codon. The AMEL gene mutations in these two bird species indicate that this crucial gene for enamel formation has lost its functional constrainsts long before the split between Passeriformes and Galliformes (Sire et al, unpublished data).

Authors' contributions

JYS and MG designed the research and analyzed the data; JYS, MG, and SD performed the research; MG contributed analytic tools; and JYS wrote the paper. All authors read and approved the final manuscript.

Acknowledgements

SD and JYS were financially supported by the Centre National de la Recherche Scientifique (CNRS) and the Université Pierre & Marie Curie-Paris 6 via UMR 7138; MG supports came from the CNRS and the Université Paris Sud via UMR 8079.

References

1. Gauthier J: **Saurischian monophyly and the origin of birds.** *Mem Calif Acad Sci* 1986, **8**:185-197.
2. Clarke JA: **Morphology, phylogenetic taxonomy, and systematics of *Ichthyornis* and *Apatornis* (Avialae: Ornithurae).** *Bull Amer Mus Nat Hist* 2004, **286**:1-179.
3. Fountaine TMR, Benton MJ, Dyke GJ, Nudds RL: **The quality of the fossil record of Mesozoic birds.** *Proc R Soc London, B* 2004, **272**:289-294.
4. Clarke JA, Tambussi CP, Noriega JI, Erickson GM, Ketcham RA: **Definitive fossil evidence for the extant avian radiation in the Cretaceous.** *Nature* 2005, **433**:305-308.
5. Zhou Z: **The origin and early evolution of birds: discoveries, disputes, and perspectives from fossil evidence.** *Naturwissenchaften* 2004, **91**:455-471.
6. Smith AB, Peterson KJ: **Dating the time of origin of major clades: Molecular clocks and the fossil record.** *Ann Rev Earth Planet Sci* 2002, **30**:65-88.
7. Graur D, Martin W: **Reading the entrails of chickens: molecular timescales of evolution and the illusion of precision.** *Trends Genet* 2004, **20**:80-86.
8. Marshall CR, Raff EC, Raff RA: **Dollo's law and the death and resurrection of genes.** *Proc Natl Acad Sci USA* 1994, **91**:12283-12287.
9. Qin C, Brunn JC, Cadena E, Ridall A, Tsujigiwa H, Nagatsuka H, Nagai N, Butler WT: **The expression of dentin sialophosphoprotein gene in bone.** *J Dent Res* 2002, **81**:392-394.
10. Veis A: **Amelogenin gene splice products: potential signaling molecules.** *Cell Mol Life Sci* 2003, **60**:38-55.
11. Spahr A, Lyngstadaas SP, Slaby I, Pezeshki G: **Ameloblastin expression during craniofacial bone formation in rats.** *Eur J Oral Sci* 2006, **114**:504-511.
12. Haze A, Taylor AL, Blumenfeld A, Rosenfeld E, Leiser Y, Dafni L, Shay B, Gruenbaum-Cohen Y, Fermon E, Haegewald S, (and 2 co-authors): **Amelogenin expression in long bone and cartilage cells and in bone marrow progenitor cells.** *Anat Rec* 2007, **290**:455-460.
13. Kollar EJ, Fisher C: **Tooth induction in chick epithelium: expression of quiescent genes for enamel synthesis.** *Science* 1980, **207**:993-995.
14. Chen Y, Zhang Y, Jiang T-X, Barlow AJ, St Amand TR, Hu Y, Heaney S, Francis-West P, Chuong C-M, Maas R: **Conservation of early odontogenic signaling pathways in Aves.** *Proc Nat Acad Sci USA* 2000, **97**:10044-10049.
15. Mitsiadis TA, Chéraud Y, Sharpe P, Fontaine-Pérus J: **Development of teeth in chick embryos after mouse neural crest transplantations.** *Proc Nat Acad Sci USA* 2003, **100**:6541-6545.
16. Harris MP, Hasso SM, Ferguson MWJ, Fallon JF: **The development of archosaurian first-generation teeth in a chicken mutant.** *Curr Biol* 2006, **16**:371-377.
17. Toyosawa S, O'hUigin C, Figueroa F, Tichy H, Klein J: **Identification and characterization of amelogenin genes in monotremes, reptiles, and amphibians.** *Proc Natl Acad Sci USA* 1998, **95**:13056-13061.
18. Shintani S, Kobata M, Toyosawa S, Fujiwara T, Sato A, Ooshima T: **Identification and characterization of ameloblastin gene in a reptile.** *Gene* 2002, **283**:245-254.
19. Delgado S, Couble ML, Magloire H, Sire JY: **Cloning, sequencing, and expression of the amelogenin gene in two scincid lizards.** *J Dent Res* 2006, **85**:138-143.
20. Shintani S, Kobata M, Toyosawa S, Ooshima T: **Expression of ameloblastin during enamel formation in a crocodile.** *J Exp Zool B Mol Dev Evol* 2006, **306**:126-133.
21. Girondot M, Sire JY: **Evolution of the amelogenin gene in toothed and toothless vertebrates.** *Eur J Oral Sci* 1998, **106**:501-508.
22. Kawasaki K, Weiss KM: **Evolutionary genetics of vertebrate tissue mineralization: The origin and evolution of the secretory calcium-binding phosphoprotein family.** *J Exp Zool B (Mol Dev Evol)* 2006, **306B**:295-316.
23. Kawasaki K, Weiss KM: **Gene duplication and the evolution of vertebrate skeletal mineralization.** *Cell Tissue Organ* 2007, **186**:7-24.
24. Hedges SD: **The origin and evolution of model organisms.** *Nat Rev Genet* 2002, **3**:838-849.
25. Reisz RR, Muller J: **Molecular timescales and the fossil record: a paleontological perspective.** *Trends Genet* 2004, **20**:237-241.
26. Delgado S, Girondot M, Sire JY: **Molecular evolution of amelogenin in mammals.** *J Mol Evol* 2005, **60**:12-30.
27. Sire JY, Delgado S, Fromentin D, Girondot M: **Amelogenin: lessons from evolution.** *Arch Oral Biol* 2005, **50**:205-212.
28. Vidal N, Hedges SD: **The phylogeny of squamate reptiles (lizards, snakes and amphisbaenians) inferred from nine nuclear protein-coding genes.** *C R Biol* 2005, **328**:1000-1008.
29. Delgado S, Vidal N, Véron G, Sire JY: **Amelogenin, the major protein of tooth enamel: a new phylogenetic marker for ordinal mammal relationships.** *Mol Phylogenet Evol* 2008, **47(2)**:865-869.
30. Kawasaki K, Weiss KM: **Mineralized tissue and vertebrate evolution: the secretory calcium-binding phosphoprotein gene cluster.** *Proc Natl Acad Sci* 2003, **100**:4060-4065.
31. Kawasaki K, Weiss KM: **Genetic basis for the evolution of vertebrate mineralized tissue.** *Proc Natl Acad Sci* 2004, **101**:11356-11361.
32. Sire JY, Davit-Béal T, Delgado S, Gu X: **The origin and evolution of enamel mineralization genes.** *Cell Tissue Organ* 2007, **186**:25-48.

activity and cleavage of recombinant pig and mouse amelogenins. J Dent Res 78:743–750

Sansom IJ, Smith MP, Armstrong HA, Smith MM (1992) Presence of the earliest vertebrate hard tissue in conodonts. Science 256:1308–1311

Schmid K, Yang Z (2008) The trouble with sliding windows and the selective pressure in BRCA1. PLoS ONE 3(11):e3746

Seedorf H, Klaften M, Eke F, Fuchs H, Seedorf U, Hrabe de Angelis M (2007) A mutation in the enamelin gene in a mouse model. J Dent Res 86:764–768

Sire JY, Delgado S, Fromentin D, Girondot M (2005) Amelogenin: lessons from evolution. Arch Oral Biol 50:205–212

Sire JY, Delgado S, Girondot M (2006) The amelogenin story: origin and evolution. Eur J Oral Sci 114(Suppl 1):64–77

Sire JY, Davit-Béal T, Delgado S, Gu X (2007) The origin and evolution of enamel mineralization genes. Cells Tissues Organs 186:25–48

Sire JY, Delgado S, Girondot M (2008) Hen's teeth with enamel cap: from dream to impossibility. BMC Evol Biol 8:e246

Springer MS, Murphy WJ (2007) Mammalian evolution and biomedicine: new views from phylogeny. Biol Rev Camb Philos Soc 82:375–392

Stern A, Doron-Faigenboim A, Erez E, Martz E, Bacharach E, Pupko T (2007) Selecton 2007: advanced models for detecting positive and purifying selection using a Bayesian inference approach. Nucleic Acids Res 35:W506–W511

Subramanian S, Kumar S (2006) Evolutionary anatomies of positions and types of disease-associated and neutral amino acid mutations in the human genome. BMC Genomics 7:e306

Tamura K, Nei M (1993) Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. Mol Biol Evol 10:512–526

Tanabe T, Aoba T, Moreno EC, Fukae M, Shimuzu M (1990) Properties of phosphorylated 32 kd nonamelogenin proteins isolated from porcine secretory enamel. Calcif Tissue Int 46:205–215

Tanabe T, Fukae M, Shimizu M (1994) Degradation of enamelins by proteinases found in porcine secretory enamel in vitro. Arch Oral Biol 39:277–281

Termine JD, Belcourt AB, Christner PJ, Conn KM, Nylen MU (1980) Properties of dissociatively extracted fetal tooth matrix proteins. I. Principal molecular species in developing bovine enamel. J Biol Chem 255:9760–9768

Tsunoyama K, Gojobori T (1998) Evolution of nicotinic acetylcholine receptor subunits. Mol Biol Evol 15:518–527

Uchida T, Tanabe T, Fukae M, Shimizu M, Yamada M, Miake K, Kobayashi S (1991a) Immunochemical and immunohistochemical studies, using antisera against porcine 25 kDa amelogenin, 89 kDa enamelin and the 13–17 kDa nonamelogenins, on immature enamel of the pig and rat. Histochemistry 96:129–138

Uchida T, Tanabe T, Fukae M, Shimizu M (1991b) Immunocytochemical and immunochemical detection of a 32 kDa nonamelogenin and related proteins in porcine tooth germs. Arch Histol Cytol 54:527–538

van Rheede T, Bastiaans T, Boone DN, Hedges SB, de Jong WW, Madsen O (2006) The platypus is in its place: nuclear genes and indels confirm the sister group relation of monotremes and therians. Mol Biol Evol 23:587–597

von Heijne G (1985) Signal sequences: the limits of variation. J Mol Biol 184:99–105

Warren WC, Hillier LW, Marshall Graves JA, Birney E, Ponting CP, Grutzner F, Belov K, Miller W, Clarke L, Chinwalla AT, Yang SP, Heger A, Locke DP, Miethke P, Waters PD, Veyrunes F, Fulton L, Fulton B, Graves T, Wallis J, Puente XS, Lopez-Otin C, Ordonez GR, Eichler EE, Chen L, Cheng Z, Deakin JE, Alsop A, Thompson K, Kirby P, Papenfuss AT, Wakefield MJ, Olender T, Lancet D, Huttley GA, Smit AF, Pask A, Temple-Smith P, Batzer MA, Walker JA, Konkel MK, Harris RS, Whittington CM, Wong ES, Gemmell NJ, Buschiazzo E, Vargas Jentzsch IM, Merkel A, Schmitz J, Zemann A, Churakov G, Kriegs JO, Brosius J, Murchison EP, Sachidanandam R, Smith C, Hannon GJ, Tsend-Ayush E, McMillan D, Attenborough R, Rens W, Ferguson-Smith M, Lefevre CM, Sharp JA, Nicholas KR, Ray DA, Kube M, Reinhardt R, Pringle TH, Taylor J, Jones RC, Nixon B, Dacheux JL, Niwa H, Sekita Y, Huang X, Stark A, Kheradpour P, Kellis M, Flicek P, Chen Y, Webber C, Hardison R, Nelson J, Hallsworth-Pepin K, Delehaunty K, Markovic C, Minx P, Feng Y, Kremitzki C, Mitreva M, Glasscock J, Wylie T, Wohldmann P, Thiru P, Nhan MN, Pohl CS, Smith SM, Hou S, Renfree MB (2008) Genome analysis of the platypus reveals unique signatures of evolution. Nature 453:175–183

Weiner S (1986) Organization of extracellularly mineralized tissues: a comparative study of biological crystal growth. CRC Crit Rev Biochem 20:365–408

Yamakoshi Y (1995) Carbohydrate moieties of porcine 32 kDa enamelin. Calcif Tissue Int 56:323–330

Yamakoshi Y, Pinheiro FH, Tanabe T, Fukae M, Shimizu M (1998) Sites of asparagine-linked oligosaccharides in porcine 32 kDa enamelin. Connect Tissue Res 39:39–46

Yamakoshi Y, Hu JC, Liu S, Zhang C, Oida S, Fukae M, Simmer JP (2003) Characterization of porcine dentin sialoprotein (DSP) and dentin sialophosphoprotein (DSPP) cDNA clones. Eur J Oral Sci 111:60–67

Yamakoshi Y, Hu JC, Fukae M, Yamakoshi F, Simmer JP (2006) How do enamelysin and kallikrein 4 process the 32-kDa enamelin? Eur J Oral Sci 114(Suppl 1):45–51

Yoon H, Laxmikanthan G, Lee J, Blaber SI, Rodriguez A, Kogot JM, Scarisbrick IA, Blaber M (2007) Activation profiles anad regulatory cascades of the human kallikrein-related peptidases. J Biol Chem 282:31852–31864

Cells Tissues Organs

Cells Tissues Organs 2007;186:25–48
DOI: 10.1159/000102679

The Origin and Evolution of Enamel Mineralization Genes

Jean-Yves Sire[a] Tiphaine Davit-Béal[a] Sidney Delgado[a] Xun Gu[b]

[a]UMR 7138, Université Pierre et Marie Curie-Paris 6, Paris, France; [b]Department of Genetics, Development, and Cell Biology, Iowa State University, Ames, Iowa, USA

Key Words
Enamel · Evolution · Genomics · Mineralization · Tooth

Abstract
Background/Aims: Enamel and enameloid were identified in early jawless vertebrates, about 500 million years ago (MYA). This suggests that enamel matrix proteins (EMPs) have at least the same age. We review the current data on the origin, evolution and relationships of enamel mineralization genes. ***Methods and Results:*** Three EMPs are secreted by ameloblasts during enamel formation: amelogenin (AMEL), ameloblastin (AMBN) and enamelin (ENAM). Recently, two new genes, amelotin (AMTN) and odontogenic ameloblast associated (ODAM), were found to be expressed by ameloblasts during maturation, increasing the group of ameloblast-secreted proteins to five members. The evolutionary analysis of these five genes indicates that they are related: AMEL is derived from AMBN, AMTN and ODAM are sister genes, and all are derived from ENAM. Using molecular dating, we showed that AMBN/AMEL duplication occurred >600 MYA. The large sequence dataset available for mammals and reptiles was used to study AMEL evolution. In the N- and C-terminal regions, numerous residues were unchanged during >200 million years, suggesting that they are important for the proper function of the protein. ***Conclusion:*** The evolutionary analysis of AMEL led to propose a dataset that will be useful to validate AMEL mutations leading to X-linked AI.

Introduction

Living vertebrates possess a great diversity of mineralized elements, comprising not only endochondral and dermal bone (including osteoderms and scutes), mineralized cartilage, and teeth (dentin and enamel), but also scales, fin rays and otoliths, and egg shells [Huysseune and Sire, 1998; Sire and Huysseune, 2003]. The first mineralized elements, which have given rise to the current

Abbreviations used in this paper

AIH1	amelogenesis imperfecta 1, hypoplastic/hypomaturation, X-linked
AIH2	amelogenesis imperfecta 2, hypoplastic local, autosomal dominant
AMBN	ameloblastin
AMEL	amelogenin
AMTN	amelotin
EMP	enamel matrix protein
ENAM	enamelin gene
KLK4	kallikrein-4 (prostase, enamel matrix, prostate)
MMP20	matrix metallopeptidase 20 (enamelysin)
MYA	million years ago
ODAM	odontogenic, ameloblast associated (APIN, FLJ20513)
SCPP	secretory calcium-binding phosphoprotein
SIBLING	small integrin-binding ligand, N-linked glycoprotein
SPARC	secreted protein, acidic, cysteine-rich (osteonectin)
SPARCL1	SPARC-like 1 (mast9, HEVIN)
C4orf7	chromosome 4 open reading frame 7 (FDC-SP, MGC71894)

KARGER

Fax +41 61 306 12 34
E-Mail karger@karger.ch
www.karger.com

© 2007 S. Karger AG, Basel
1422–6405/07/1861–0025$23.50/0

Accessible online at:
www.karger.com/cto

Dr. Jean-Yves Sire
Equipe 'Evolution et Developpement du Squelette', UMR 7138
Université Pierre et Marie Curie-Paris 6, 7 quai St-Bernard, Case 5
FR–75252 Paris (France)
Tel./Fax +33 1 44 27 35 72, E-Mail sire@ccr.jussieu.fr

skeletal diversity in vertebrates, were identified as early as 500 million years ago (MYA) [Sansom et al., 1992, 1994; Janvier, 1996; Donoghue, 1998, 2001]. In fact, the occurrence of mineralized tissue in vertebrates was a major innovation, which was fundamental to the radiation of modern vertebrates in relation to the important roles of the skeletal elements in protection, predation and locomotion [Reif, 1982; Smith and Hall, 1990; Janvier, 1996; Donoghue, 2002; Donoghue and Sansom, 2002; Donoghue et al., 2006].

Our understanding of the mechanisms by which organisms form mineralized elements is still at a rudimentary stage, but we know that biomineralization is mediated by the organic matrix, either through its biological activity or in controlling nucleation, growth and microarchitecture of the mineral deposited [Carter, 1990]. It is assumed that the basic processes of biomineralization are common to all systems and that mineral formation by any individual biological system may diverge from this common pathway. This general definition applies to vertebrates in which the main skeletal elements derive from common ancestral elements [Huysseune and Sire, 1998; Sire and Huysseune, 2003] and there is growing evidence that most proteins currently involved in mineralization of skeletal tissues (bone, dentin, and enamel) also have diverged from a common ancestor [Kawasaki and Weiss, 2003; Kawasaki et al., 2004; Kawasaki and Weiss 2006]. The evolutionary analysis of genes coding for these 'mineralizing' proteins not only has the potential to provide insight into the debated question of the origin of mineralization in vertebrates and of its subsequent diversification, but could also bring important information for humans, as mutations of these proteins lead to genetic disorders (bone [Rowe, 2004]; dentin [Zhang et al., 2001], and enamel [Stephanopoulos et al., 2005]).

This review is devoted only to our current knowledge on the origin and evolution of the genes coding for enamel matrix proteins (EMPs). The reader is referred to the paper by Kawasaki et al., published in this issue [pp 7–24], regarding the history of the other mineralizing proteins in vertebrates.

In living and extinct vertebrates, teeth are protected by a hypermineralized tissue, either 'true' enamel (e.g. in tetrapods) or 'enamel-like' tissue, enameloid (e.g. in cartilaginous and ray-finned fish). These hard dental tissues are identified early in the history of the mineralized integument. They were present in the dermal skeleton of various lineages of jawless vertebrates [Ørvig, 1967, 1977; Reif, 1982; Smith and Hall, 1990; Janvier, 1996; Donoghue and Sansom, 2002; Donoghue et al., 2006]. Enamel and enameloid are homologous tissues that correspond to different aspects of the same hypermineralization process [Donoghue et al., 2006]. Enamel has replaced enameloid in the lineage leading to tetrapods, probably by a process of heterochrony[1] [Smith, 1995], but enameloid was conserved in two important lineages, chondrichthyans[2] and actinopterygians[3]. The close evolutionary relationships, the similar features of the ameloblasts during their formation, and the same maturation process strongly indicate that both enamel and enameloid matrices could contain similar mineralizing proteins, and that some of them (if not all) were already present in tooth-related elements of early vertebrates, 500 MYA. Unfortunately, our knowledge on EMP genes is restricted to the tetrapods, and the road is still long before we will be able to test the hypothesis of an early origin of EMPs.

In mammals, the enamel matrix is composed of three specific proteins secreted by ameloblasts: amelogenin (AMEL), which represents 90% of the matrix deposited, and enamelin (ENAM) and ameloblastin (AMBN), which are components of the remaining 10% organic matrix. Evolutionary analyses have indicated that these three EMPs constitute a family, which, itself, is included into a larger family, the secretory calcium-binding phosphoprotein (SCPP) family. This SCPP family comprises other Ca-binding proteins: some saliva proteins, milk caseins and small integrin-binding ligand, N-linked glycoproteins (SIBLINGs) [Fisher and Fedarko, 2003], which contain five dentin and bone proteins [Kawasaki and Weiss, 2003]. Interestingly, with the exception of AMEL that is located elsewhere (on sex chromosomes X and Y in placental mammals), all SCPP genes are located in two clusters on the same autosomal chromosome. This supports the idea that SCPP genes originated by tandem duplication followed by neofunctionalization.

In humans, several types of amelogenesis imperfecta (AI), leading to enamel hyploplasia or hypomineralization, are related to mutations in AMELX (14 X-linked AI, AIH1, identified to date [Hart et al., 2002; Kim et al., 2004; Stephanopoulos et al., 2005]) or in ENAM (5 autosomal-dominant AI, AIH2 [Hart et al., 2003; Hu and Yamakoshi, 2003; Kim et al., 2005]) genes [review in Stephanopoulos et al., 2005]. In contrast, although being con-

[1] Heterochrony: developmental changes in the timing of events, leading to changes in size and shape from an ancestral state.
[2] Chondrichthyans: the cartilaginous fishes, including sharks, rays and chimaeras.
[3] Actinopterygians: the ray-finned fish, which are the dominant group of vertebrates.

sidered as a candidate gene, AMBN was excluded from a causative role within the families studied [Mardh et al., 2001].

Since a few years, we focus our attention on EMP gene relationships (AMEL, AMBN, and ENAM), and more precisely on the origin and evolution of AMEL, the best known member of the family [Delgado et al., 2001; 2005; Sire et al., 2005; 2006; Delgado et al., in press]. Here, (i) we summarize these previous data, (ii) we provide new information on two newly identified genes, amelotin and APIN protein, that are expressed by the ameloblasts, (iii) we provide a date for EMP gene origin and discuss this result in the light of our knowledge of enamel and/or enameloid appearance in vertebrate evolution, and (iv) we show how evolutionary analysis of AMEL can help to identify structural features that might be important for the protein function, and to validate mutations responsible for genetic diseases.

Ameloblast Products: EMPs, and Amelotin and APIN Proteins

In mammals, the synthetic activity of ameloblasts is divided in two successive phases corresponding to two stages of enamel formation: secretion and maturation, separated by a transition stage. To our knowledge, during the former step ameloblasts deposit four proteins in the extracellular matrix: three EMPs (AMEL, ENAM, and AMBN) and a tooth-specific, calcium-dependent peptidase, MMP20 (= enamelysin) [Bartlett et al., 1998; Bartlett, 2004]. During the transition and maturation stages, ameloblasts have been shown to produce a fifth protein, kallikrein 4 (KLK4), a pleiotropic, calcium-independent protease, which is involved in the final proteolysis of the remaining organic matrix [Simmer et al., 1998; Hu et al., 2002; Simmer and Hu, 2002].

Recently, two novel genes were found to be also expressed by ameloblasts during tooth formation: amelotin (AMTN, but annotated UNQ689 in human genome build 36.2) [Iwasaki et al., 2005; Moffat et al., 2006b] and ODAM ('odontogenic, ameloblast associated', previously named APIN or FLJ20513) [Moffat et al., 2006a]. Can the proteins encoded by these two genes be considered EMPs? In other words, although being produced by ameloblasts, are AMTN and ODAM structural proteins playing a role in enamel matrix formation and/or mineralization? In rats, AMTN was localized to the basal lamina of maturation stage ameloblasts [Moffatt et al., 2006b]. This location seems to indicate a possible role of AMTN in cell adhesion, and it also demonstrates the absence of AMTN participation in enamel matrix formation. In humans, ODAM was first identified from extracts of amyloid deposits obtained from calcifying epithelial odontogenic tumors [Solomon et al., 2003]. Transcripts of this gene were also found at a high level in gastric cancer [Aung et al., 2006]. In rats, ODAM is specifically expressed in ameloblasts during maturation stage [Moffatt et al., 2006a], but the location of the protein in the extracellular matrix remains to be shown. However, late expression during tooth formation does not mean that the secreted ODAM protein is not incorporated in the enamel matrix at the end of the mineralization process. Such a location would not be surprising if one considers that the protein was first isolated from calcifying tissues of odontogenic tumors [Solomon et al., 2003]. Therefore, if AMTN cannot be considered an EMP, the few data available to date do not permit to exclude ODAM from this family.

Interestingly, these two genes are located in the same cluster as EMPs, and they share structural similarities with the members of this family (see below). This indicates that AMTN and ODAM were probably created after duplication of an ancestral EMP gene and, therefore, that they belong to the SCPP family [Kawasaki and Weiss, 2006].

In the following, we provide some data on these two newly identified ameloblast-secreted proteins, although concentrating on the evolutionary relationships of EMPs (AMEL, AMBN, and ENAM).

Evolutionary Relationships of AMTN, ODAM, and EMP Genes

EMPs are evolutionarily related, forming a gene family that belongs to a super-gene family called SCPP [Kawasaki and Weiss, 2003]. All SCPP genes probably derive from a common ancestor by gene duplications. The key gene could be SPARC-like1 (SPARCL1, also called HEVIN or SC1), which was created after a duplication of SPARC (osteonectin) [Kawasaki et al., 2004]. Four lines of evidence have permitted to establish SCPP relationships, and SPARCL1 may resemble the ancestral form of SCPP: (i) common gene structure and similar protein characteristics in the N-terminal region, (ii) in most SCPPs, presence of an SXY phosphorylation site encoded in the 3′ region of the second coding exon, suggesting Ca-binding properties, (iii) location on the same chromosome, and (iv) presence of SPARCL1 on this chromosome, adjacent to the dentin-bone protein gene cluster [reviewed by Kawasaki and Weiss, 2006].

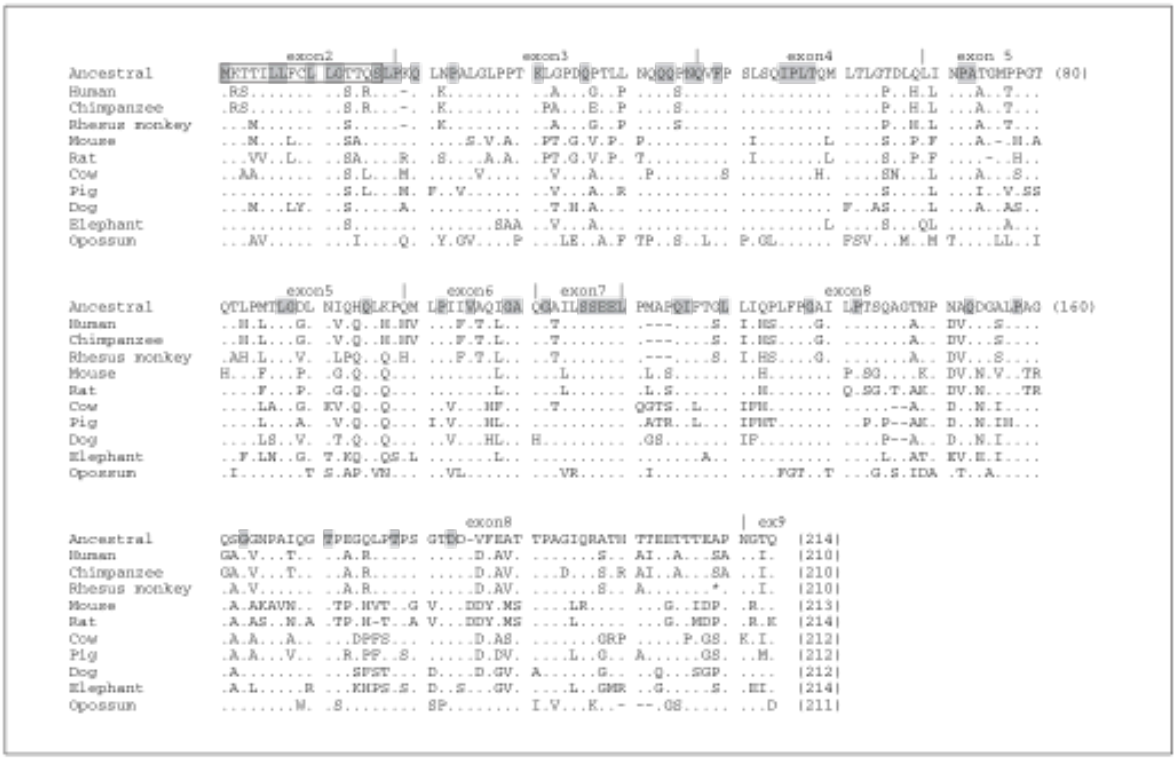


Fig. 1. Amelotin (AMTN): alignment of 10 complete mammalian amino acid sequences and of the putative ancestral sequence (shown at the top). Six sequences were inferred from DNA sequences retrieved in databases (blast search against sequenced genomes). These sequences were checked against three published complete coding sequences: human, accession number AY358528; mouse, AK017352, and rat, DQ198381. The pig sequence was obtained from the literature [Moffatt et al., 2006]. The putative ancestral sequence was calculated using PAUP 4.0 and MacCLADE 3.06. Vertical bars indicate the limits between exons. The signal peptide is in a box. The total number of residues in each protein is indicated at the end of each sequence. Unchanged residues are shown on a gray background. · · · · · = Identical residue; – – – = indel.

Recent studies on the origin and evolution of AMEL in tetrapods have extended our knowledge on EMP relationships [Sire et al., 2005, 2006]. A phylogenetic analysis using a large set of sequences demonstrated that AMEL and AMBN are sister genes, and that AMEL was created from a duplication of AMBN. In addition, it was shown that both genes are related to ENAM, which was recognized as a more ancient member of the EMP family. The calculation of putative ancestral sequences of EMP genes and the use of SPARCL1 as an outgroup were helpful for this phylogenetic analysis. Putative ancestral sequences permit to go back to the gene origin, while the whole dataset of sequences is less informative to reveal possible relationships. Indeed, although they are phylogenetically related, EMP genes show large sequence variations when comparing evolutionary distant lineages. Moreover, since their creation, hundreds of million years ago, AMEL, AMBN, and ENAM have acquired specific functions and their sequences diverged rapidly. However, currently available sequences permit to calculate putative ancestral sequences of EMPs at the origins of the mammals only, i.e. when monotremes[4] and therians[5] diverged, 225 MYA [van Rheede et al., 2006]. Using amniote or tetrapod sequences was not possible because of the few sequences

[4] Monotremes: egg-laying mammals: extant members are the echidnas and the duck-billed platypus.

[5] Therians: marsupials and placental (eutherian) mammals.

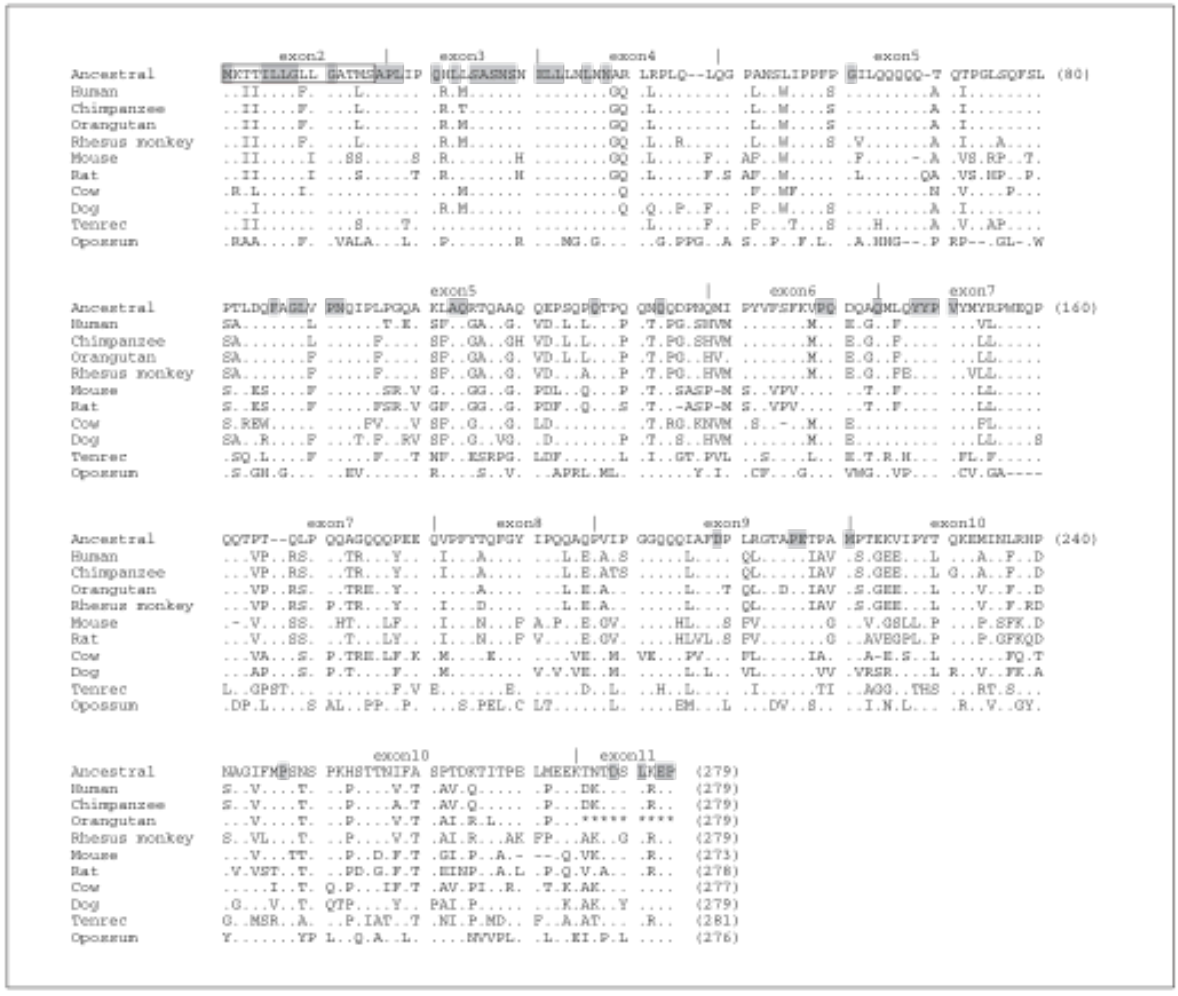


Fig. 2. ODAM (APIN protein): alignment of 10 complete mammalian amino acid sequences and of the putative ancestral sequence (shown at the top). Seven sequences were inferred from DNA sequences retrieved in databases (blast search against sequenced genomes and trace archive-Whole Genome Shotgun). The sequences were checked against three published complete coding sequences: human, NM17855; rat, DQ198380, and mouse, NM27128. For further information, see legend to figure 1. * = Unknown residue.

available in reptiles, and amphibians are not representative enough of EMP evolution in these lineages (see below).

Here, we use the same approach to try to identify the origins of the two newly identified genes, AMTN and ODAM, with regard to the EMPs. Ten complete coding therian [a metatherian (opossum) + nine eutherian species] sequences of both genes were retrieved from databases and the literature. The inferred protein sequences were aligned using CLUSTALX and hand-checked using the sequence alignment editor Se-Al 2.0 (fig. 1, 2). The putative ancestral sequences of therian AMTN and ODAM (i.e. 190 million years old [van Rheede et al., 2006]) were calculated with PAUP 4.0 (Sinauer, Sunderland, Mass., USA), taking into account the current mammalian phylogeny [Madsen et al., 2001; Murphy et al.,

2001; Delsuc et al., 2002; van Rheede, 2006] using MacCLADE 3.06 (Sinauer; fig. 1, 2). Given the small number of sequences available and the lack of sequences of representative species in some important mammalian lineages (e.g. Perissodactyla, Insectivora, Xenarthra, and prototherians: platypus or echidna), it was not possible to perform an evolutionary analysis. However, some findings from these alignments reveal some interesting points.

AMTN Analysis in Mammals

The amino acid sequences (ranging from 210 to 214 residues) were easily aligned without including numerous gaps (fig. 1). The presence of large conserved regions when comparing eutherian and metatherian AMTN, and the large differences with the other members of the family suggest that this gene was created long before mammalian lineage divergence, which occurred 310 MYA [Murphy et al., 2001; Hedges, 2002]. As a consequence, a functional AMTN gene might be found in reptile genomes. The putative mammalian ancestral sequence comprised 214 residues. Four residues were lost during primate evolution. Only a few residues (22%) remained unchanged during mammalian evolution (47 out of 214, fig. 1). Such relaxed selective constraints on AMTN suggest that some polymorphism could be encountered in humans. This idea is supported by the comparison of chimpanzee and human sequences: four amino acids (1.9%) were substituted within a time period of 5–7 million years, which separates the two lineages [Kumar et al., 2005]. In addition, most of the unchanged residues are dispersed through the sequence. This means that the number of conserved positions will almost certainly drop when sequences from other mammalian and reptilian species become available [Sire, unpubl. res.]. However, three important features emerge from this alignment.

(i) In the N-terminal region encoded by exon 2, 55% of the residues (10 out of 18) are unchanged. This region is similarly organized as in the other SCPPs, and it is mainly composed of the signal peptide, which plays an important role in the extracellular secretion of the proteins.

(ii) In positions 55–58 (exon 4), an IPLT motif is conserved, which means that this predicted O-glycosylated site could be important for the function of AMTN. Two other predicted O-glycosylated sites (threonines) are also conserved, but isolated, in exon 8.

(iii) In positions 116–120 (exon 6), a SSEEL motif is well conserved. This is a putative CK2 serine phosphorylation site [Moffatt et al., 2006b]. Surprisingly, in contrast to the condition observed in EMPs, there are no conserved residues in the C-terminal region of AMTN. It is clear that a further study, including new mammalian and reptilian sequences, is necessary to reveal further details on gene ancestry and to perform an accurate evolutionary analysis.

ODAM Analysis in Mammals

The amino acid sequences, which contain 273–281 residues depending on the species, were easily aligned without including numerous gaps (fig. 2). The absence of large sequence variations and the large differences compared with the other members of the family indicate that ODAM, like AMTN and the EMPs, arose before the mammalian/reptilian split. The putative ancestral mammalian sequence comprised 279 amino acids. Regarding AMTN, only a few residues (16.8%) are unchanged (47 out of 279, fig. 2), and this low selective pressure suggests that some polymorphism could occur in human ODAM (seven amino acid variations, 2.5%), are found between human and chimpanzee). Most of the conserved residues are dispersed along the sequence, but four features emerged from this alignment:

(i) the N-terminal region (signal peptide) in which 47% of residues (8 out of 17) are unchanged (exon 2);

(ii) in positions 25–33 (exon 3 and beginning of exon 4), a SASNSxELL motif is well conserved; this is a probable phosphorylation site;

(iii) in positions 147–150 (exon 7), a YYPV motif is kept unchanged, but its function remains to be discovered, and

(iv) in contrast to AMTN, four residues are conserved in the C-terminal region (exon 11). Here too, further sequences from species representative of other tetrapod lineages are needed to perform an accurate evolutionary analysis.

Relationships of Ameloblast-Expressed SCPP Genes

The structure and organization of the two newly identified ameloblast-expressed genes, AMTN and ODAM, were compared to the putative ancestral sequences previously calculated for the three EMP genes and SPARCL1 (fig. 3). A previous analysis of the putative ancestral sequences of EMPs had shown that:

(i) AMEL exon 4 was created during eutherian evolution (it is present in some eutherian lineages only), and two additional exons 8 and 9, that are unique to the mouse and rat, were created by duplication of exons 4 and 5 [Bartlett et al., 2006a];

(ii) AMBN exons 8 and 9 have appeared in primates only, as duplications of exon 7;

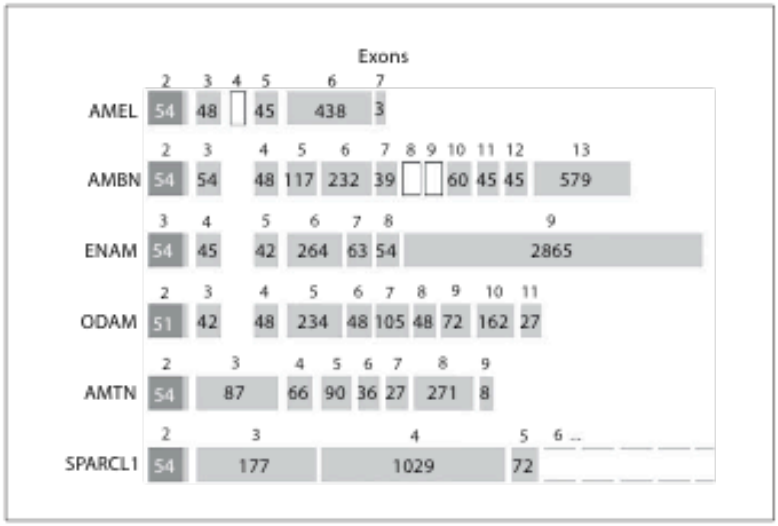


Fig. 3. Gene structure of the putative ancestral coding sequences calculated for the EMP genes (AMEL, AMBN, and ENAM), the two other ameloblast-expressed SCPP genes (ODAM and AMTN) and SPARCL1, the supposed SCPP ancestor. The reference to exon number on top of the boxes is that of the human sequences. Empty boxes indicate exons lacking in the basal mammalian taxa. The nucleotide number of each exon is indicated within the boxes (not to scale). Dark gray = Signal peptide.

(iii) ENAM exon 3 can be considered as homologous to exon 2 of the other genes, and

(iv) although considered the probable ancestor of SCPPs, the N-terminal organization of SPARCL1 is different from that of the EMPs, except for the first coding exon, exon 2 [Sire et al., 2005; 2006].

The structural comparison of the six putative ancestral genes, i.e. EMPs, AMTN and ODAM, and SPARCL1, confirms the previous findings that only the first three coding exons share similarities (fig. 3). As already shown for human genes, the strongest similarity of the ancestral sequences concerns exon 2 (exon 3 in ENAM), which encodes a well-conserved signal peptide and the first two residues of the protein [Kawasaki and Weiss, 2003; Kawasaki et al., 2004]. The two following exons in EMPs and ODAM are small and of roughly the same size (42–54 and 42–48 bp, respectively), with the third exon (exon 4 for ENAM) ending with an SXE phosphorylation motif. In mammals, such an organization is not observed in AMTN and in SPARCL1, which exhibit a larger third exon (87 and 177 bp, respectively), and which lack an SXE motif. The sizes of exon 3 in chicken and teleost fish SPARCL1 are small (54–57 bp), similar to the size of SPARC exon 3. This suggests that SPARCL1 originally had a small exon 3. However, in the absence of data for SPARCL1 in amphibians, crocodiles and squamates (lizards and snakes) we cannot claim that a small exon 3 was the condition when actinopterygian and sarcopterygian lineages separated. Our alignment (not shown) indicates that the third exon in AMTN could correspond to the two short exons 3 and 4 in the other genes. The phylogenetic position of AMTN suggests that this exon could have been created by a fusion of these two short exons (see below). The mere comparison of gene organization already suggests that these genes belong to a single family [Kawasaki and Weiss, 2003]. With the exception of AMTN, the structure of which is somewhat different from the four other genes, the N-terminal region of EMPs and ODAM is similar. In addition, the organization of ODAM is more similar to that of ENAM, which suggests closer relationships of ODAM with ENAM than with the other genes (fig. 3).

Since 2002, the study of EMP (and SCPP) relationships has highly benefited from gene mapping in humans, and new data have progressively accumulated in other tetrapod species (but unfortunately mainly in mammals) [http://www.ncbi.nlm.nih.gov/]. In humans, SCPP genes are located on chromosome 4, on which they form two clusters, separated by 15 Mb: the dentin and bone protein cluster (4q22, approximately 375 kb), to which SPARCL1 is adjacent, and the saliva, milk and ameloblast-secreted protein cluster (4q13, approximately 770 kb; fig. 4). The only exception is AMEL, two copies of which are found on the sex chromosomes. The most important copy, which encodes 90% of the transcripts, resides on chromosome X (fig. 4). In humans AMELX is located in antisense in the intron 1 of the ARHGAP6 gene. As AMEL belongs to the EMP family, it is clear that it was translo-

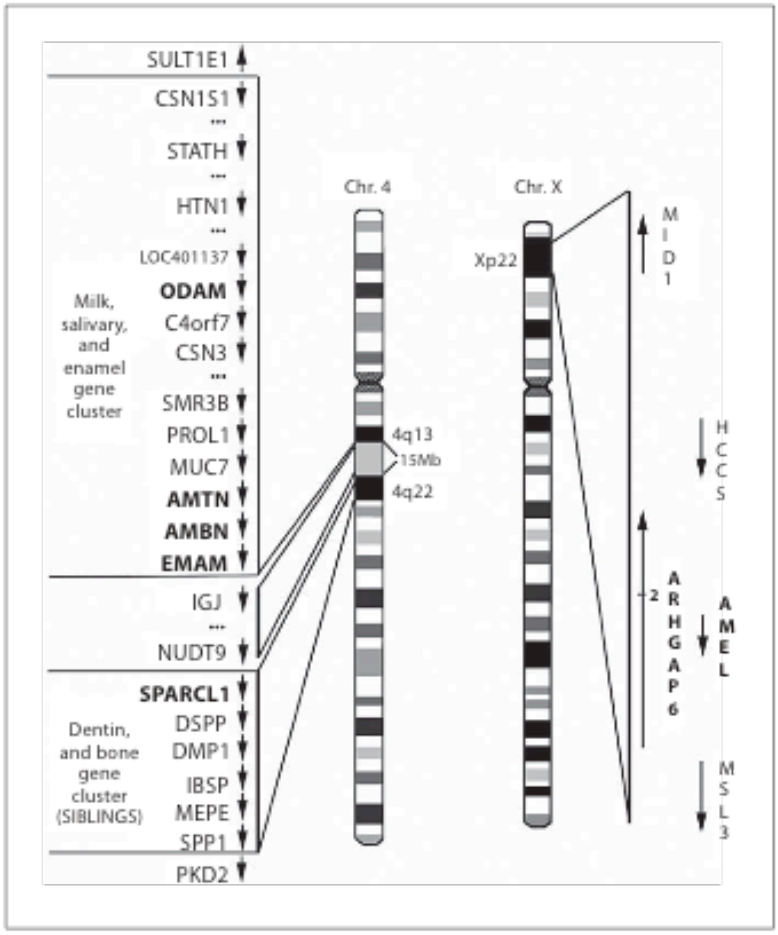


Fig. 4. Location of the ameloblast-expressed SCPP genes and of SPARCL1 on human chromosomes. ENAM, AMBN, AMTN, ODAM, and SPARCL1 are located on chromosome 4, in two clusters separated by 15 Mb. AMEL is the only SCPP found elsewhere, on the sex chromosomes. The most important AMEL copy is on chromosome X, located in antisense within ARHGAP6 intron 2. SCPP genes are identically oriented on chromosome 4.

cated from the 'EMP family' chromosome to another chromosome (ARHGAP6 gene intron), either immediately after its duplication, or during a particular event, which occurred some time after a tandem duplication. ENAM, AMBN, and AMTN are adjacent genes on chromosome 4, while ODAM is located between C4orf7 (follicular dendritic cell secreted peptide) and LOC401137 (a hypothetical protein), at some distance from the three ameloblast-expressed genes, and separated from them by some salivary protein and milk casein genes (fig. 4). This syntheny is conserved in the few mammalian species for which genes are mapped [http://www.ncbi.nlm.nih.gov/map view/]. In birds, which lost teeth approximately 80–100 MYA [Huysseune and Sire, 1998], the SIBLING genes are found in syntheny, while the enamel, saliva, and milk protein gene cluster is lacking [Kawasaki and Weiss, 2006]. In amphibians *(Xenopus)* the syntheny is roughly conserved, but some mineralizing protein genes, known to be important in mammals, are apparently lacking. However, this absence could be related to the currently incomplete assembly of this frog genome [Kawasaki and Weiss, 2006].

The five ameloblast-expressed genes (ENAM, AMBN, AMTN, ODAM, and AMEL) were created by tandem duplication from a common ancestor [Kawasaki and Weiss, 2003, Kawasaki et al., 2004; Kawasaki and Weiss, 2006]. These duplications were probably asymmetric, i.e. after each duplication one copy kept the former function of the protein and did not diverge much from the ancestral sequence, while the other copy differentiated rapidly and

acquired new functions (neofunctionalization) [Chung et al., 2006; Steinke et al., 2006]. These functions were positively selected, but they are still to be uncovered for most of these genes. This finding is deduced from comparison of the gene structure (fig. 3) and from sequence analysis (fig. 1, 2 and Sire et al. [2006]). Indeed, the roughly conserved features of the N-terminal region suggest not only a common origin but also some functional similarities (they are all ameloblast-expressed proteins). In contrast, the rest of the sequence (the largest part) houses the specificities of each protein (i.e. its proper functions) and, therefore, is strongly divergent. The specific function of each protein could reside either in this whole sequence, as for instance for most part of the region coded by AMEL exon 6 [Sire et al., 2006] (see below), or in some particular important loci, as for instance the conserved motifs that emerge from the alignment of AMTN and ODAM mammalian sequences (fig. 1, 2).

The next questions now are: how are these ameloblast-expressed genes related and which evolutionary scenario can be proposed for their origins in vertebrates?

AMEL and the Evolutionary Origin of EMP Genes

The current knowledge on the relationships and evolutionary origin of EMPs was acquired in several steps, and this study represents the last (but not least) one. This story can be briefly reconstructed as follows.

In 2001, Delgado et al. showed a high sequence similarity of the 5′ region (exon 2, which mainly encodes the signal peptide) of AMEL, SPARC, and SPARCL1, suggestive of a common origin of this region after duplication. Using a molecular-clock method to estimate SPARC/SPARCL1 divergence, these authors proposed that AMEL exon 2 was created >600 MYA (i.e. at the end of the Precambrian). This meant that AMEL could have been present before the origin of vertebrates, 530 MYA [Shu et al., 1999, 2003], and of the first evidence of mineralized elements in euconodonts, 500 MYA [Sansom et al., 1992; 1994; Janvier, 1996].

Two years later, taking advantage of the availability of the sequenced human genome and gene mapping, Kawasaki and Weiss [2003] convincingly demonstrated that (i) EMPs comprise a subfamily, (ii) EMP, milk casein, and salivary protein families together are regrouped into a cluster on chromosome 4, forming a larger family, and (iii) this family also contains the SIBLING gene cluster, which is located in another locus on the same chromosome. The SCPP family was now a fact.

Another chapter was added to the story when SPARCL1 was proposed to be the common ancestor of SCPP genes on the basis of its location, adjacent to the SIBLING cluster on chromosome 4, and of the structure of its N-terminal region [Kawasaki et al., 2004]. Therefore, although SPARC still remains at the origin of the mineralizing protein gene story, it was SPARCL1 that gave rise to the SCPP gene ancestor. SPARC is present in both protostomes and deuterostomes[6], where it influences cell behavior and interactions with the extracellular matrix, rather than being involved in the generation of mineralized tissues. Several runs of duplications, and subsequent sub- and/or neofunctionalization have occurred and led to the current diversity of this family. Using a molecular-clock method, the divergence date between SPARC and SPARCL1 was found to be inferior or equal to the current divergence date of cartilaginous fishes (estimated at 528 ± 56 MYA using molecular dating [Kumar and Hedges, 1998]). This led to the conclusion that the SCPP genes probably emerged after this date [Kawasaki et al., 2004]. This dating is more recent than the >600 MYA previously calculated by Delgado et al. [2001].

Taken together, these findings suggest that AMEL is more distantly related to SPARC and/or SPARCL1 than hitherto believed before, and that at least five duplication events took place from SPARC to AMEL [Sire et al., 2006]:

SPARC → SPARCL1 → SCPP ancestor →
ENAM → AMBN → AMEL

Below, we briefly review the current scenario for EMP gene relationships, which was established in the course of studies dealing with AMEL origins [Sire et al., 2005, 2006]. The previously published dataset is completed by additional information on AMTN and ODAM (fig. 1, 2), with the aim to clarify the relationships of all ameloblast-secreted SCPP proteins.

The Evolutionary Origin of AMEL

This study was performed in three steps:

Step 1: Evolutionary Analysis of AMEL Sequences in Tetrapods

A total of 80 AMEL sequences (including mammals, reptiles, and amphibians) were compiled (published se-

[6] Protostomes and deuterostomes: the two main divisions of bilateria mostly comprising animals with bilateral symmetry and three germ layers (endoderm, mesoderm, and ectoderm).

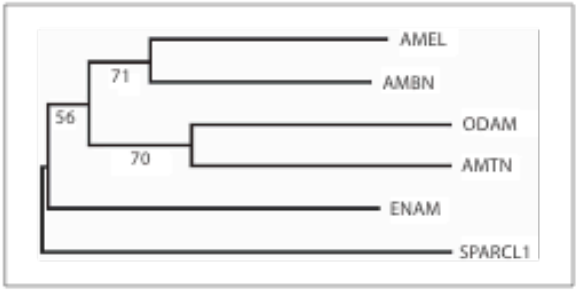


Fig. 5. Phylogenetic analysis (distance analysis with maximum likelihood using neighbor-joining method) of the five ameloblast-expressed SCPP genes (AMEL, AMBN, AMTN, ODAM, and ENAM) based on the 5′ region (288 bp) of their putative ancestral sequences. The ancestral sequence of SPARCL1, the probable ancestor of SCPP genes, was used to root the tree. Bootstrap values are indicated (1,000 replicates).

quences, sequences retrieved in the databases, and new sequences; see Sire et al. [2006] for the species list). The sequences were aligned as described above for AMTN and ODAM, and a putative AMEL ancestral sequence was calculated using PAUP 4.0. The conserved versus variable regions were determined and used for the next step.

Step 2: Search for Sequence Similarity in Databases

A PSI-blast search (National Center for Biotechnological Information) of statistically significant similar peptides was performed in GenBank [Sire et al., 2006]. The well-conserved regions of the putative ancestral AMEL were used, i.e. the N-terminal region: exon 2 (signal peptide), exon 3, exon 5, and beginning of exon 6. Sequence similarities were detected with AMBN, then with ENAM and, finally, with SPARCL1. It is noteworthy that the first non-AMEL sequence to be found using PSI-blast was crocodile AMBN, indicating that the latter is closer to ancestral AMEL than mammalian AMBN. This would mean that crocodile AMBN is more conservative of an ancestral state, and could have been subjected to a slower rate of evolution than mammalian AMBN after reptile/mammal divergence. At this time (July 2004), neither AMTN nor ODAM sequences were available in databases [Sire et al., 2005].

Step 3: Sequence Analysis

The putative ancestral sequences of AMEL, AMBN, ENAM, and SPARCL1 were calculated as described above for AMTN and ODAM. The dataset comprised AMEL sequences, 30 AMBN, 28 ENAM, and 20 SPARCL1 (entire and partial sequences), and those obtained here from 10 AMTN and 10 ODAM (fig 1, 2). The N-terminal region of SPARCL1 was only used because EMPs and the other SCPPs are supposed to be derived from this region [Kawasaki et al., 2004]. The N-terminal regions of these putative ancestral sequences were aligned to the same region of AMEL (i.e. the first 62 residues, from exon 2 to the TRAP proteolytic site at the beginning of exon 6) with CLUSTALX and hand-checked using Se-Al 2.0. The phylogenetic analysis was performed using maximum likelihood (neighbor-joining method) in PAUP 4.0 and the tree was rooted on SPARCL1, since this is the probable ancestor of the SCPPs. This analysis confirms with a good statistical support the previous finding that AMEL and AMBN are sister genes [Sire et al., 2006] (fig. 5). The two newly identified ameloblast-expressed genes, ODAM and AMTN, appear as two sister genes (this is well supported statistically), and their group is the sister group of the AMEL/AMBN group. ENAM is the sister gene of the two groups AMEL/AMBN + ODAM/AMTN, and SPARCL1 is the sister gene of the three. However, the relationships of ENAM and SPARCL1 are not strongly supported by our bootstrap analysis. This phylogenetic analysis means that AMEL/AMBN and ODAM/AMTN have a common ancestor, which was probably issued from a duplication of the ENAM ancestor, itself deriving from a copy of the SPARCL1 ancestor.

This phylogeny corresponds to our relatively weak knowledge of ameloblast-expressed genes and must be interpreted with caution. Indeed, even though a large number of sequences were used, most of them are from mammals, and even from eutherians only. Only a few AMEL and AMBN sequences are available in reptiles and amphibians, and no ENAM, AMTN, and ODAM sequences are known in these lineages. This lack of data in non-mammalian lineages does not allow to obtain representative putative ancestral sequences at the amniote and tetrapod levels. This means that the phylogenetic signal (i.e. gene relationships) is probably reduced by (i) the long evolutionary period (hundreds of million years) that separates each gene from its closest relative, (ii) the different evolution rate for each gene in each lineage, and (iii) the rapid divergence of some gene regions in relation to their proper functions. This phylogeny will become more accurate in the near future, when more ameloblast-expressed SCPP gene sequences will be known in reptiles and amphibians. Nevertheless, the present analysis supports AMBN/AMEL relationships and the hypothesis that both genes derive from ENAM. It furthermore indi-

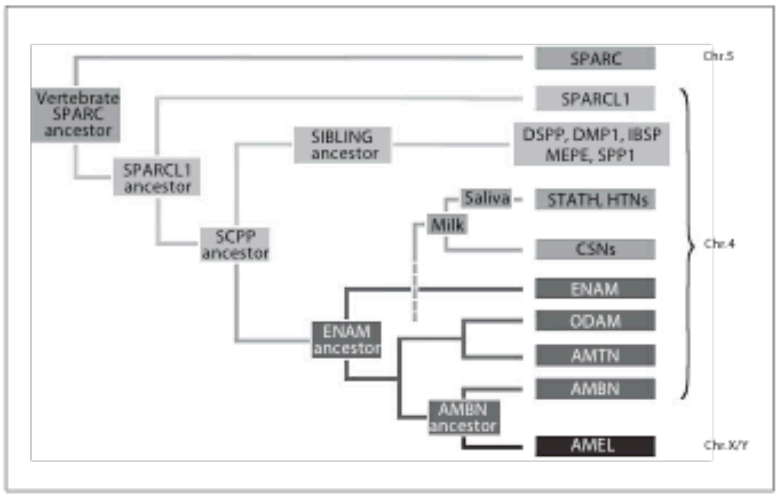


Fig. 6. Current probable scenario for the origin and evolution of SCPP genes and, in particular, of ameloblast-expressed genes (AMEL/AMBN, AMTN/ODAM, and ENAM). Early in deuterian evolution, SPARC duplicated into SPARCL1. During successive rounds of genome and gene duplication, SPARCL1 and its descendants were copied several times on the same chromosome, giving rise to two clusters: the ameloblast-expressed/milk/saliva protein gene cluster and the bone/dentin protein gene cluster (SIBLINGs). The ENAM ancestor duplicated from an SCPP ancestor and one ENAM copy was duplicated again, giving rise to the ancestors of AMBN/AMEL and of AMTN/ODAM. After its duplication from AMBN, AMEL was translocated to another chromosome.

cates that ODAM and AMTN could also be derived from ENAM. This implies that an additional duplication event has occurred between ENAM and the other ameloblast-expressed SCPP genes (fig. 6).

A preliminary, schematic scenario for SCPP evolution and for the place of the ameloblast-secreted actors (to which AMTN and ODAM are now added) can be drawn, but the story is far from complete (fig. 6). In particular, the relationships between SPARCL1 and the two gene clusters (SIBLINGs and enamel-milk-saliva protein genes), and among the SIBLINGs are not established. In contrast, within the salivary SCPPs, histatins 1 and 3 derive from statherin duplication, and the latter was created from a copy of a milk casein ancestor (CSN1S2) [Kawasaki and Weiss, 2003]. The evolutionary story of salivary SCPPs is relatively recent (they are known in some eutherians only), while the origin of milk caseins is more ancient in mammalian evolution. Indeed, α-, β- and κ-caseins are identified in the milk of metatherians (marsupials) [Ginger et al., 1999; Stasiuk et al., 2000]. Milk casein family members are also evolutionarily related and, given their structural similarity with EMP genes, the ancestral Ca-sensitive casein gene was probably derived from the duplication of an EMP [Kawasaki and Weiss, 2003], which remains to be found (fig. 6).

In summary, depending on the branches of the tree, SCPP relationships are either strongly or weakly supported. Strong relationships are: SPARC/SPARCL1; STATH/HTHs; CSN/STATH/HTHs; AMEL/AMBN, and AMEL/AMBN/ENAM. In contrast, there are (i) no clear relationships established within the SIBLING cluster, and between this cluster and SPARCL1; (ii) no clearly identified connection between CSNs and EMPs; (iii) weak (lack of non-mammalian sequences) relationships between ODAM/AMTN, and ENAM/ODAM/AMTN, and (iv) no clear relationship between the ameloblast-expressed genes (AMEL/AMBN, ODAM/AMTN, and ENAM) and SPARCL1.

Sequencing these SCPP genes in non-mammalian species [reptiles (crocodiles, lizards, and snakes) and amphibians (salamanders, caecilians, and frogs)] will help to improve our knowledge on the relationships in the family.

Dating of AMBN/AMEL Duplication

Now that AMEL and AMBN are clearly established sister genes, the last questions are: was the ancestral gene AMBN or AMEL and is it possible to date this duplication event? The stronger support to AMBN ancestry is indirectly suggested by the location of AMEL on sex chromosomes. Indeed, it is difficult to imagine that an AMEL copy (that would have become AMBN) was translocated by mere chance, on the chromosome housing the other SCPP genes, and close to ENAM, their close relative. In contrast, the close location of AMBN and ENAM on the same autosomal chromosome (fig. 4) strongly supports that AMBN was created from a copy of ENAM, and, as a consequence, that AMEL originated after a duplication of the ancestral AMBN, and then translocated to another chromosome. One could argue that AMEL translocation

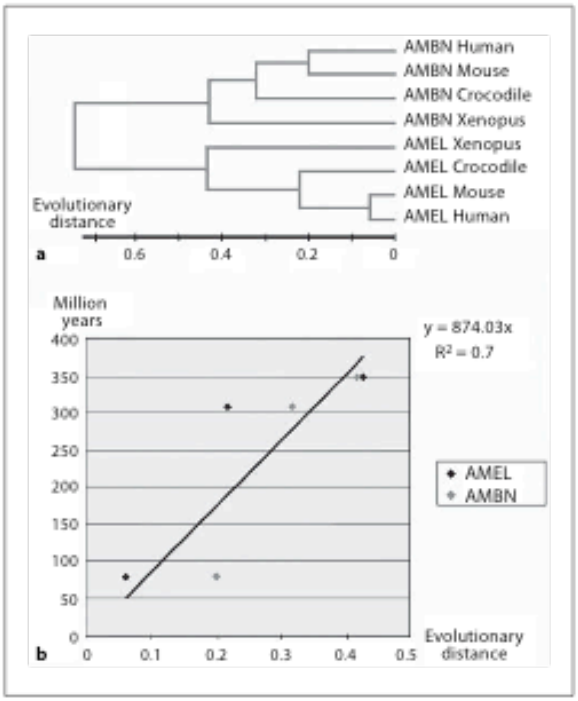


Fig. 7. a Linearized tree obtained from the phylogenetic analysis of AMBN and AMEL sequences in human, mouse, crocodile, and *Xenopus*. The calibration time used is: human/mouse: 90 MYA; human/crocodile: 310 MYA; human/*Xenopus:* 360 MYA [Hedges, 2002]. **b** Linear regression of time versus distance (y-x). Each point has two evolutionary distances of AMBN and AMEL. The duplication time of AMBN/AMEL can be estimated when we add the evolutionary distance of duplication to this linear equation, i.e. it occurred >600 MYA.

occurred after its duplication from the ENAM ancestor and that the copy remained close to ENAM and differentiated into AMBN. This scenario cannot be maintained since the similarities found in gene organization (fig. 3) and in amino acid pattern indicate that AMBN is closer to ENAM than AMEL is. Therefore, AMBN is the 'mother' of AMEL and not the opposite.

In summary, all ameloblast-expressed genes are phylogenetically related, and ENAM could be the ancestor of all of them. AMEL, which codes for the major protein of the forming enamel matrix in mammals (90% of the protein content) is the youngest EMP gene. This strongly suggests that AMEL divergence after AMBN duplication was an important innovation for enamel, at least in mammals. To date, the relationships of EMP genes with SPARCL1 are difficult to establish and more data are needed to test the hypothesis of SPARCL1 ancestry.

The availability of AMEL and AMBN sequences in various mammalian species, in a crocodile and in an amphibian *(Xenopus)* allowed to envisage a molecular dating of AMBN/AMEL duplication. A phylogenetic tree was inferred from the amino acid sequences using the neighbor-joining method (fig. 7a). From the phylogeny, it is apparent that the duplication event was much earlier than the speciation events such as the mammal/amphibian split, or the mammal/reptile split, and roughly two times of these events. To give an approximate estimate of when this duplication event occurred, we utilized the molecular dating technique developed by Gu et al. [2002], calibrated by the fossil record: primate/rodent split (around 90 MYA), mammal/reptile split (310 MYA), and amniote/amphibian split (360 MYA) [Hedges, 2002]. Our results are as follows.

1 If the amniote/amphibian split is used alone, the date of duplication (T) = 627 MYA.
2 If the mammal/reptile split is used alone: T = 896 MYA.
3 If the primate/rodent split is used alone: T = 480 MYA.
4 If all three calibrations are used: T = 682 MYA.

This is a molecular dating of gene duplication, so it should be compared to other molecular date profiling [Gu et al., 2002]. Here, (2) and (3) are unreliable because the distance between human-mouse or human-crocodile differs considerably in AMBN/AMEL genes. In contrast (1) is mostly reliable and (4) takes the average, but both give similar results, i.e. AMBN/AMEL duplication occurred >600 MYA (fig. 7b). This result confirms the previous dating of AMEL origins during the Precambrian period [Delgado et al., 2001]. A major peak of genome and gene duplication occurred around 700–500 MYA [Gu et al., 2002]. Therefore, like many developmental genes, EMPs were duplicated during this period, which preceded vertebrate diversification and skeletal mineralization.

In summary, two unrelated molecular dating methods of EMP origins (SPARC/SPARCL1 divergence date: Delgado et al. [2001] and AMBN/AMEL duplication date: this study) indicate that the genes encoding them were created from several duplication rounds that have occurred before the currently accepted dates of the appearance of the first vertebrates in the fossil record (>600 MYA). In contrast, the molecular dating of SPARC/SPARCL1 divergence proposed by Kawasaki et al. [2004] supports an emergence of EMPs after the di-

vergence of cartilaginous fish (approximately 500 MYA Kumar and Hedges [1998]). The knowledge of the divergence date of SPARC/SPARCL1 is of importance as SPARCL1 is considered the probable ancestor of SCPPs. However, the apparent different evolutionary rates of SPARC and SPARCL1 in various taxa, together with the fact that various gene regions were compared within each species or each clade, does not allow an accurate prediction of the divergence date. Indeed, these two paralogs share a well-conserved C-terminal region which is not easy to differentiate from one gene to the next in the vertebrate species examined. In contrast, their N-terminal region is not only extremely different but also, when comparing this region in various species, difficult to align due to a large number of sequence variations. Nevertheless, the N-terminal region of SPARCL1 is considered the probable ancestor of SCPPs. The divergence date of AMBN/AMEL seems to be more reliable because the relationships of these two genes are now well established. Also, the presence of enamel-like tissues in early vertebrates indicates that the divergence of SCPP genes might have preceded the origin of vertebrate tissue mineralization.

It is important to realize the following.

(i) The molecular dating of AMBN/AMEL duplication does not indicate the presence of these molecules in forming enamel, 600 MYA. After the duplication, several dozens of millions of years were probably necessary before one copy acquired its new function (new gene structure and new expression). This divergence could have occurred before, during or after the vertebrate diversification, reported to be in the Cambrian as demonstrated in the fossil record. Moreover, genetic evidence suggests that most animal phyla evolved dozens of millions of years before they started to leave behind fossil evidence, although this is debated by paleontologists. Given the lack of a temporal association between the birth of a gene (e.g. AMEL 600 MYA) and the advent of mineralized 'teeth' >50–100 millions of years later, the confidence in the assigned dating should be softened.

(ii) Tissue mineralization could not have occurred if the necessary tools were not already present. This implies that EMPs could have had other functions before the first enamel/enameloid tissues mineralized and before EMPs were recruited for mineralization later in vertebrate evolution. This novel trait (mineralization) therefore probably evolved by employing already existing materials.

Enamel/Enameloid and the Origin of EMPs

Morphological studies of enamel and enameloid in living taxa have shown that they are different in their mode of formation. The enamel organic matrix is secreted by the ameloblasts, and contains enamel-specific proteins. In contrast, enameloid organic matrix is mostly deposited by odontoblasts and contains a large amount of collagen, but the ameloblasts contribute to its formation, too [Prostak and Skobe, 1984; Sasagawa, 1984; Prostak and Skobe, 1988; Prostak et al., 1993; Sasagawa, 1995, 2002]. However, in functional teeth, the structure of both tissues is similar, i.e. highly mineralized with only a little organic matrix left (<5%). Given the same location, the same final structure and the same evolutionary origin, most authors have considered enamel and enameloid as homologous tissues. Enamel and enameloid matrices are only partially mineralized when laid down, and their final hardness is acquired during a second stage, maturation, during which the matrix is lost through the activity of proteolytic enzymes. This process creates space, allowing mineral crystal growth to eventually achieve a highly mineralized structure. Because they are highly mineralized, enamel and enameloid are easily recognizable in the fossil record and their relationships can be traced back deep in vertebrate evolution.

The question of which tissue appeared first, enamel or enameloid, has been long debated and it is not clearly answered yet. It is, however, accepted that enamel progressively replaced enameloid during evolution in various lineages (e.g. in tetrapods) [Smith, 1995; Donoghue, 2002; Donoghue and Sansom, 2002; Donoghue et al., 2006]. Odontoblasts progressively reduced their production of loose collagenous matrix, which characterizes forming enameloid, while ameloblast activity increased with the secretion of large amounts of enamel-specific products at the dentin surface. This evolutionary 'transition' between enameloid and enamel was, in fact, probably an enameloid-dentin transition, as recently demonstrated in the ontogeny of caudate amphibians [Davit-Béal et al., 2007]. However, enamel did not replace enameloid in all vertebrate lineages. A particular type of enameloid is present in chondrichthyans (cartilaginous fish [Prostak et al., 1993; Sasagawa, 2002]), and this supports an ancient origin for this tissue, at least for the gnathostome lineage. Enamel and enameloid were certainly present in basal actinopterygians (ray-finned fish), as in polypterids and lepisosteids [Sire et al., 1987; Sire, 1990, 1994, 1995]. This supports the idea that enamel was already present in early osteichthyans, which also indicates an ancient origin.

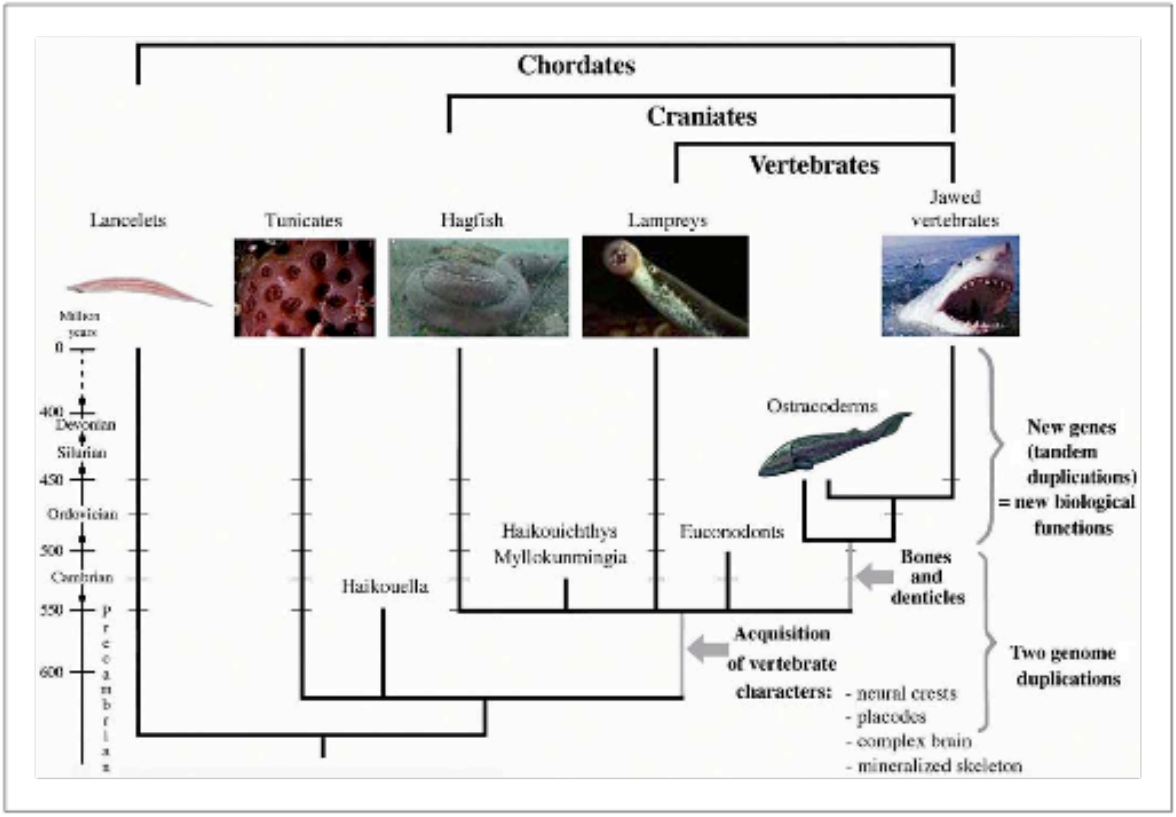


Fig. 8. Chordate relationships and the origin of the mineralized skeletal elements in vertebrates (adapted from Shimeld and Holland [2000]). Chordates are deeply anchored in the Precambrian era (>700 MYA). The acquisition of a mineralized skeleton, a major event for vertebrate radiation, occurred 600–500 MYA, a period which post-dates the two genome duplications [Gu et al., 2002]. Bone and dental tissues are clearly recognized in early, jawless vertebrates, 450 MYA. Skeletal diversification in jawed vertebrates was next favored by the appearance of new genes after tandem duplication.

Enamel is absent in more derived actinopterygian taxa (teleost fish), which possess enameloid only [Sasagawa, 1984; Prostak and Skobe, 1984; Sasagawa, 1995]. The large evolutionary distance between all living representatives of these chondrichthyan and actinopterygian lineages (430–420 MYA, respectively, in the fossil record: Janvier [1996]) explains why the current structure of these enameloids is so different.

Enamel and enameloid appear, therefore, to be merely grades of a hypermineralized tissue that has evolved independently in a number of vertebrate lineages [Donoghue, 2001]. The origin of these tissues can be traced back in early vertebrates, along with the appearance of a bony mineralized skeleton, one of the four main vertebrate character acquisitions, together with neural crest cells and their derivatives, neurogenic placodes, and an elaborate segmented brain (fig. 8). These vertebrate innovations appeared after the divergence between tunicates[7] *(Ciona)* and craniates[8] (recent genetic evidence indicates that tunicates could be closer to vertebrates than cephalochordates [Graham, 2004]), and probably after the divergence between craniates and vertebrates as witnessed by the fossil record. The absence of mineralized tissues in living hagfish and lampreys is probably primitive [Jan-

[7] Tunicates: subphylum of chordates that feed by siphoning plankton through a filter.
[8] Craniates: animals with skull.

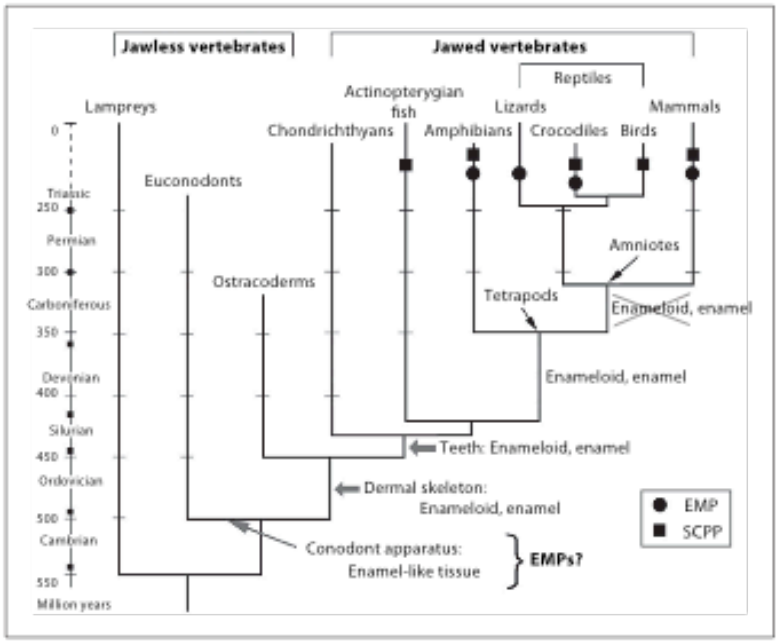


Fig. 9. Enamel/enameloid tissues during vertebrate evolution (as reported in the fossil record), and current knowledge of the presence of EMP and SCPP genes in vertebrate lineages. Enamel-like tissues are identified in early vertebrates, the euconodonts, and they display a different evolutionary history in the various lineages. Enameloid was conserved in chondrichthyan and actinopterygian lineages, but disappeared in amniotes. The early presence of enamel/enameloid tissues in vertebrate evolution strongly suggests that EMP divergence predates this time (>500 MYA). However, there is a large gap between this theoretical EMP presence in early vertebrate lineages and the current knowledge of the genes coding for these proteins, which is restricted to the tetrapod level (350 MYA). SCPPs are known, however, from actinopterygian fish.

vier, 1996]. Indeed, the most ancient vertebrates discovered in the Lower Cambrian of China (530 MYA), *Haikouichthys* (which looks like a hagfish) and *Myllokunmingia* (which looks like a lamprey), possessed a skeleton composed of unmineralized cartilage only [Shu et al., 1999, 2003].

The first mineralized elements encountered in vertebrates are the tooth-like organs (conodont apparatus) composed of enamel-like and dentine tissue found in euconodonts, fossil marine vertebrates known from the Middle Cambrian (500 MYA) to the Late Triassic (230 MYA) [Sansom et al., 1992, 1994; Janvier, 1996; Donoghue, 1998, 2001] (fig. 9). These minute comb-shaped denticles are located at the entrance of the pharynx (viscerocranium). Bone appears to be absent from these elements [Donoghue, 1998].

Enamel, or enameloid, is clearly identified in the skeleton of early jawless vertebrates (e.g. pteraspidomorphs, heterostracans, thelodonts, and 'ostracoderms') from the Early Ordovician (480 MYA) to Late Devonian (380 MYA) periods and of jawed vertebrates (early chondrichthyans and osteichthyans) [Janvier, 1996; Donoghue et al., 2006] (fig. 8). The earliest skeleton was a dermal skeleton comprising odontodes (tooth-like elements consisting of enameloid and dentine), ornamenting dermal plates composed of acellular bone [Sansom et al., 2005]. It is noteworthy that our current knowledge of early vertebrates reveals a gap of 30 million years between the appearance of the first vertebrates (530 MYA) and the first evidence of vertebrate mineralized elements (500 MYA).

It is clear that numerous gene families expanded by gene duplication in the vertebrate stem lineage (in particular gene families encoding transcription factors and signaling molecules) [Shimeld and Holland, 2000]. The acquisition of the mineralized skeleton followed the increased genetic complexity (two genome duplications and several gene duplications) which occurred early in vertebrate evolution (during the Precambrian and Cambrian periods) [Dehal and Boore, 2005; Panopoulou and Pouska, 2005]. These large scale genomic events facilitated the evolutionary success of the vertebrate lineage and, probably, led to the diversification of several members of the SCPP family. Additional tandem duplications certainly occurred during the long period of vertebrate evolution and resulted in new gene differentiation and in a further diversification of SCPPs into new biological functions (fig. 8).

The presence of enamel and enameloid tissues in early vertebrates strongly suggests that EMPs (and some other SCPPs) were present in these tissues at least 500 MYA

(fig. 9). This would mean that SCPPs diversified earlier. The hypothetical date of this diversification could be not so distant from the molecular dating of EMP origins (>600 MYA) if we consider that the duplication could have occurred long before the divergence of function/expression of the copies, and that vertebrates possessing a mineralized skeleton could have lived dozens of millions of years before any evidence of them in the fossil record. However, although structurally well-identified enameloid and enamel tissues are present in the teeth of chondrichthyans, actinopterygians, and basal sarcopterygians, EMP genes are known in tetrapods only (fig. 9). However, this statement relates to genes only; there is evidence from immunohistochemical studies or Southern hybridization that AMEL and/or ENAM proteins could be present in sharks [Slavkin et al., 1983; Herold et al., 1989], teleost fish [Lyngstadaas et al., 1990], polypterids [Zylberberg et al., 1997] and lungfish [Satchell et al., 2000].

Whilst the data on EMP genes (mainly in model mammals) slowly accumulated over a period of approximately 15 years, the last years witnessed a rapid increase in our knowledge, mainly because of genome sequencing in numerous species, and in particular in mammals. To date eight well-covered mammalian genomes are available and seven additional genomes are provided at a low coverage level (see http:/www.ensembl.org/). The current mammalian genome project aims to add 11 mammalian species to this list in a phylogenetic perspective (http:/www.broad.mit.edu/mammals). Therefore, within the next few months, we will have access to at least 26 mammalian genomes and, potentially, will be able to perform evolutionary analyses of any gene in the mammalian lineage. Opposite to this large covering of mammalian phylogeny, our knowledge of non-mammalian EMPs is, unfortunately, much less advanced (fig. 10). We can see two reasons: (1) the lack of sequenced genomes and (2) the divergence of EMP sequences.

The Lack of Sequenced Genomes

In toothed reptiles (crocodiles, snakes, and lizards), there is still no sequenced genome available, although the reptilian (sauropsid) lineage is the lineage closest to mammals (fig. 10). However, AMEL sequences are available in a crocodile [Toyosawa et al., 1998], in a snake [Ishiyama et al., 1998], and in two lizards [Delgado et al., 2006; Wang et al., 2006], and AMBN has been sequenced in a crocodile [Shintani et al., 2002]. At present, there are no data on reptilian ENAM but, fortunately, we will soon have access to a lizard genome (*Anolis carolinensis* genome is being sequenced). However, sequencing a crocodile genome (a representative of the lineage closest to birds) would be also extremely interesting for evolutionary analyses.

In amphibians, AMEL [Toyosawa et al., 1998] and AMBN [Shintani et al., 2003] have been sequenced in the pipid frog *Xenopus laevis*, and an AMEL sequence is available in another frog (*Rana pipiens* [Wang et al., 2005a]). Moreover, sequencing of a pipid genome *(Silurana tropicalis)* is well advanced (fig. 10). Surprisingly, although as expected AMEL and AMBN are present in this genome, to our knowledge ENAM has not been found yet [Kawasaki and Weiss, 2006]. It is questionable whether this EMP is really absent from this frog genome. Indeed, on the one hand our evolutionary analysis indicates that ENAM is the oldest representative of the EMP family and, on the other hand, ENAM plays important roles in enamel structure and organization as illustrated by AIH2 resulting from ENAM mutations. It is also clear that pipids have a well-formed enamel [Sato et al., 1986]. Therefore, this 'lack' is probably related to the fact that the pipid genome is still not entirely (or correctly) assembled. One should also take into consideration that pipids are highly derived anurans and, as a consequence, EMPs could be divergent compared to more basal amphibian species. Sequencing another frog, salamander/newt or caecilian genome would be, therefore, highly informative for evolutionary analysis.

No EMP is known in basal sarcopterygians, i.e. lungfish and coelacanth, nor in basal actinopterygians (polypterids and lepisosteids), and there is no sequenced genome available nor sequencing project running. However, these taxa possess enamel and they belong to lineages that are crucial to improve our understanding of EMP relationships and evolution. In contrast to this lack of data, the genome has been sequenced in four teleost species, and several SCPPs were identified. However, teleosts are derived actinopterygian lineages, and the long evolutionary distance (>420 million years) between actinopterygians and tetrapods explains the difficulty encountered when trying to identify homology between teleost and tetrapod SCPP genes [Kawasaki et al., 2005]. For instance, no EMP gene can be related to these SCPPs.

No SCPP is known in chondrichthyans (sharks and rays). Here too, the long evolutionary distance (>430 million years) between cartilaginous fish and tetrapods could lead to problems when trying to identify homologous genes, but the syntheny conservation of SCPP genes could help [Kawasaki et al., 2005; Kawasaki and Weiss, 2006].

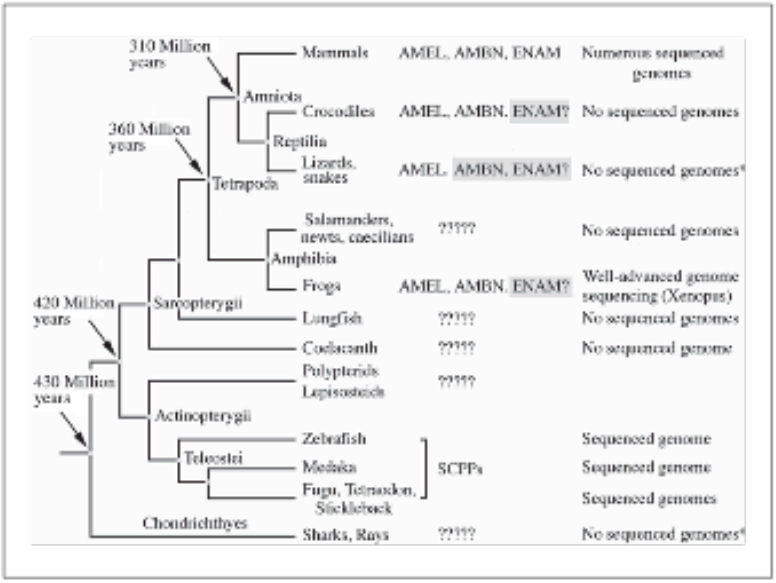


Fig. 10. Current knowledge of EMP genes in vertebrates. To date only two EMPs are characterized at the tetrapod level (AMBN and AMEL). ENAM is only known in mammals. The lack of data in non-mammalian lineages is clearly related to the absence of sequenced genomes. SCPP genes are identified in teleost fish, but the large evolutionary distance makes their relationships to EMPs uncertain. EMP genes on gray background are potentially accessible to sequencing. Question marks indicate lineages in which sequencing of EMP genes might be a priority to improve our understanding on their origin and evolution. * = Large DNA regions (Whole Genome Shotgun) have been sequenced in a lizard *(A. carolinensis)*.

The Divergence of EMP Sequences

The difficulty to find EMP (and other SCPP) genes using PCR or RT-PCR resides in their variability. Indeed, except for the short N-terminal region that is relatively well conserved in each member of the family, the largest part of the sequence is variable. For instance, although they probably conserve their main function, most of the mammalian AMEL exon 6 sequences (the largest part of AMEL) cannot be accurately aligned with the homologous region in reptiles and amphibians due to numerous substitutions and indels [Sire et al., 2006]. These highly variable sequences indicate that SCPPs are intrinsically disordered proteins [Dunker et al., 2001; Kawasaki et al., 2005] and there are only a few conserved residues. Therefore, the only means to find EMPs in evolutionary distant species, such as basal sarcopterygians or actinopterygians, is to study sequenced genomes or sequences of large DNA regions suspected to house these genes. For example, in a teleost fish (fugu), several SCPP genes were identified in a DNA region corresponding to the SIBLING cluster in mammals, meaning that the syntheny of the SIBLING cluster is conserved between fish and tetrapods [Kawasaki et al., 2005; Kawasaki and Weiss, 2006]. These SCPP genes were found not based on their similarity with known SCPP sequences but because they are located adjacent to SPARCL1, and because they share some structural features with tetrapod SCPPs. Fish SCPP genes are so different from tetrapod SIBLINGs that no homology could be recognized. Fish SCPP genes are expressed during tooth formation [Kawasaki et al., 2005] but one can wonder whether they play the same function as EMPs. Moreover, SIBLINGs (DSPP, DMP1, IBSP, and SPP1) are known to be expressed during tooth matrix formation in tetrapods [Fisher and Fedarko, 2003; Qin et al., 2004]. EMP genes could also be conserved in other regions of the teleost fish genome, but they remain to be discovered. Indeed, morphological studies strongly support that EMPs are present in the enamel-like tissue (ganoine) of basal actinopterygian lineages, polypterids and lepisosteids [Sire et al., 1987; Sire, 1994; 1995].

To date the information available for the three EMP genes largely relates to mammals and the few sequences available (or planned to be so) in other tetrapods are not sufficient to perform an evolutionary analysis at this level (fig. 10).

What Can the Evolutionary Analysis of EMP Genes Tell Us? The Case of AMEL

AMEL Evolution

AMEL is the main component of forming enamel and it plays crucial roles in enamel structure and mineralization [Diekwisch et al., 1993; reviews in Bartlett et al.,

2006b; Margolis et al., 2006]. Mutations of the encoding gene lead to AIH1 [Hart et al., 2002; Kim et al., 2004]. Given this importance it is not surprising that AMEL is the best-known EMP. Over the past years, AMEL studies on model animals have provided information on the gene structure and supposed functions of the various regions of the protein [Fincham et al., 1991; Fincham and Moradian-Oldak, 1995; Greene et al., 2002]. AMEL is subject to posttranslational modifications [Fincham and Moradian-Oldak, 1993] and it self-assembles to form nanospheres that are involved in enamel mineralization [Wen et al., 2001; Snead, 2003; Du et al., 2005; Veis, 2005]. The N- and C-terminal regions interact with mineral [Aoba et al., 1989; Aoba, 1996; Hoang et al., 2002; Paine et al., 2003; Snead, 2003] and are involved in adhesion with the ameloblast surface through membrane proteins (e.g. Cd63, annexin A2, and Lamp1 [Wang et al., 2005b; Tompkins et al., 2006]). AMEL interacts also with some keratins in ameloblasts through ligand-binding properties located in the N-terminal region [Ravindranath et al., 1999, 2000, 2001, 2003]. Some splice products have been proposed to be signaling molecules [Veis et al., 2000; Veis, 2003].

From these studies, increasing evidence accumulates to support the idea that the N-terminal, and to a lesser degree the C-terminal, regions are the most important regions for proper AMEL function. This importance is also revealed by several AIH1, caused by mutations modifying the functioning of these regions. The question of a possible role for the central variable region (encoded by most of exon 6) is completely ignored. Is it useless? Certainly not. Evolutionary analyses indicate that this core region of the protein, although intrinsically disordered, could be responsible for the well-ordered microstructure of enamel [Delgado et al., 2005; Sire et al., 2005; 2006]. More data are still needed to understand the relationships between structure and function of this region and, more generally, to reveal the amino acid positions and regions that could play an essential role.

As an alternative to biochemical and in vitro approaches, an evolutionary analysis of mammalian AMEL was performed using 56 sequences constituting a dataset representative of mammalian diversity [Delgado et al., 2005]. Here, we summarize and complete these results in proposing two alignments (fig. 11): one, illustrated with 20 sequences of the N- and C-terminal regions only, reveals the numerous well-conserved residues that are important for the proper function of the protein (interactions with the cell membrane and/or with mineral crystals). The other alignment, comprising 51 sequences, is centered in the variable central region of exon 6, which houses, in mammals, a hot spot of mutation. The putative ancestral sequence has been calculated for both alignments. Briefly, this evolutionary analysis reveals the following points.

(i) A total of 56 residues (out of 74 in the full-length sequence) have remained unchanged in the N- and C-terminal regions of AMEL during mammalian evolution, i.e. during 225 million years [van Rheede et al., 2006] (fig. 11a). This indicates that strong functional constraints act on these amino acids, meaning that they certainly play, either alone or with other conserved residues, an important role. Most variants are found in the C-terminal region of exon 5.

(ii) The hot spot of mutation (large insertions/deletions of residues) has appeared recently in mammals, and independently in several lineages (fig. 11b). Insertions are found in basal primates (lemurs), in tree shrews, in basal rodents (squirrel and guinea pig), in bovids (cow and goat) and cervids (deer), in only one family of carnivores (ursids), in bats (Macrochiroptera), in insectivores (hedgehog), in afrotherians (elephant shrew), and in marsupials (opossums). The perissodactyls (e.g. horse) and prototherians (platypus and echidna) are the only important lineages in which such large insertions are absent. These insertions contain a variable number of three amino acid (triplet) repeats (e.g. PIQ-PMQ-PLQ). These triplet repeats range from two (in the tree shrew) to 12 (in a fruit bat), in which a total of 36 residues (108 bp) are inserted. Within some lineages, e.g. bovids, the number of repeats can vary in closely related species (8 repeats in the African buffalo, 7 in cattle, and 5 in the other members of the family). It is noteworthy that AMELY, that is expressed at a low level in forming enamel (less than 10% [Salido et al., 1992]), does not show insertions in this region. This illustrates the separate evolution of the two AMEL copies on sex chromosomes [Girondot and Sire, 1998], AMELY being subjected to the particular mode of evolution of the Y chromosome [Iwase et al., 2001; Lahn et al., 2001; Iwase et al., 2003]. The lack of triplet insertions in AMELY versus AMELX exon 6 allows to easily discriminate males from females in lineages possessing the hot spot of mutation, e.g. bovids [Weikard et al., 2006] and ursids [Yamamoto et al., 2002]. Large deletions ($\geq$9 residues) are found in dolphin, Weddell seal, panda and roundleaf bat (Microchiroptera). However, we do not know whether these indels have a consequence on enamel microstructure in these species [Delgado et al., 2005]. It is clear, however, that the conservation of such large indels during evolution has no negative results on enamel function as protective tissue.

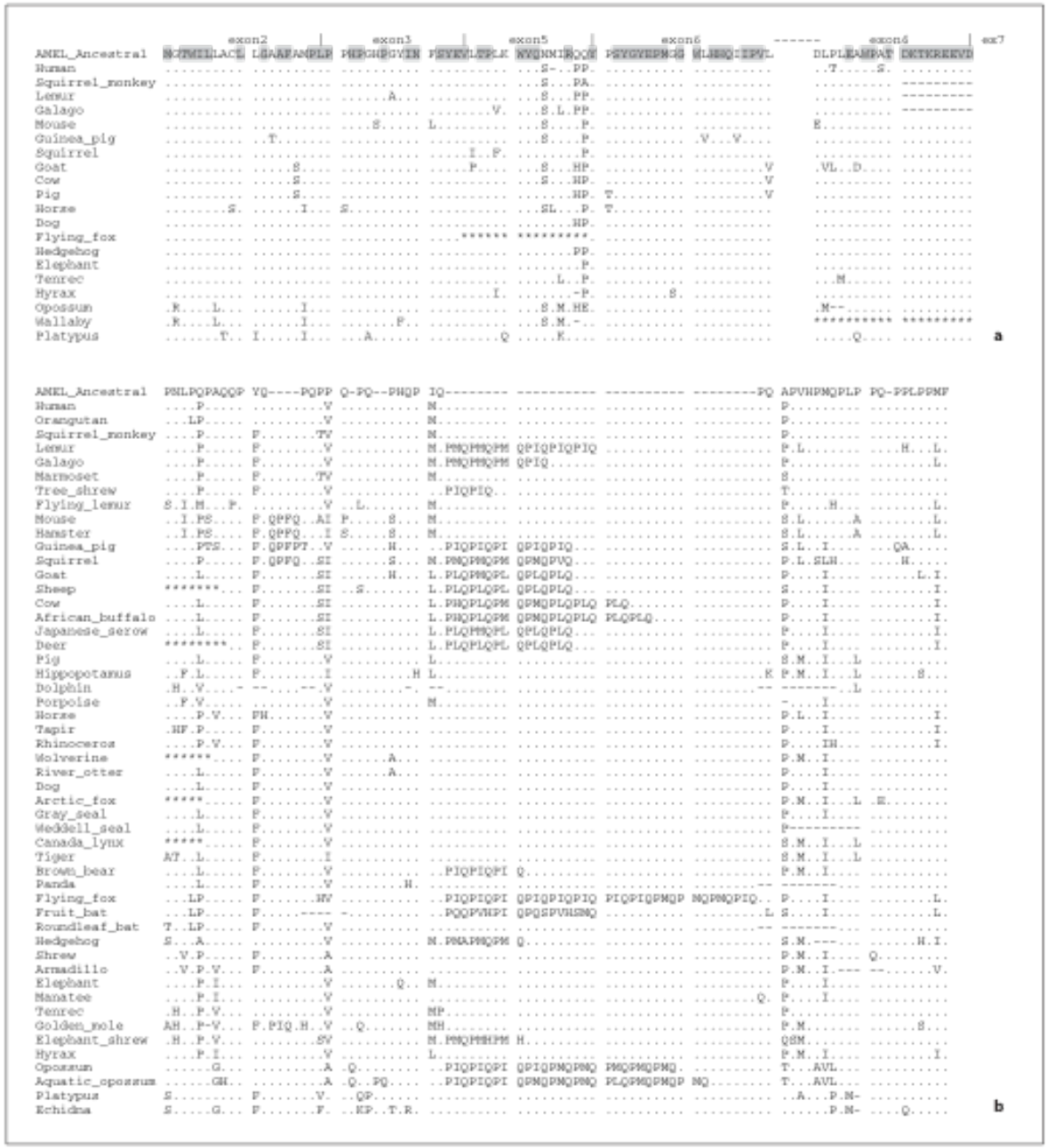


Fig. 11. Alignment of AMEL amino acid sequences in representative mammals. **a** Well-conserved N- and C-terminal regions in 20 species. Exon 4 is not represented because it is lacking in several species. Partial sequences were removed from this alignment. Vertical bars indicate the limits between exons. Unchanged residues are shown on a gray background. **b** Central region of exon 6 in 51 species emphasizing the region considered a hot spot of mutation. This region is characterized by amino acid triplet insertions or deletions. Identical sequences were not included in these alignments: e.g. human = chimpanzee and rhesus monkey; mouse = rat; cow = European bison. · · · · · = Identical residue; – – – = indel; * = unknown residue.

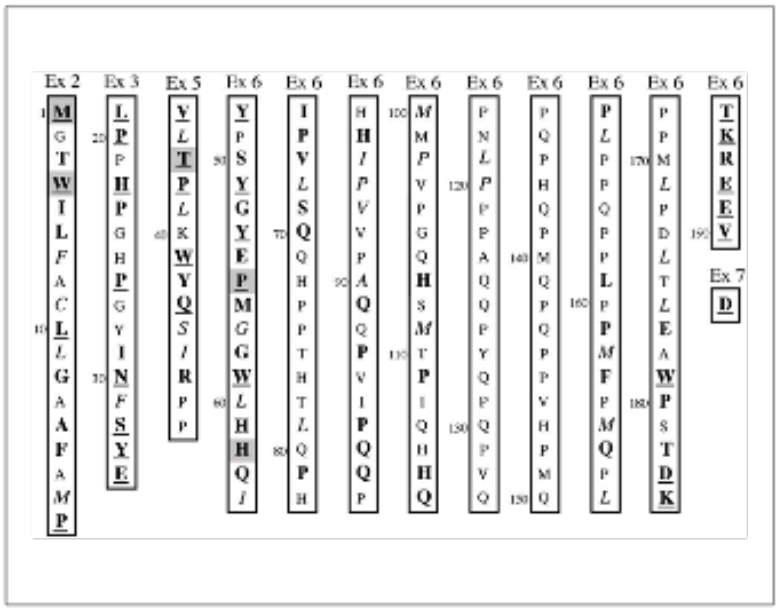


Fig. 12. Amino acid sequence of human amelogenin highlighting the residues which remained unchanged during the 225 million years of mammalian diversification. The importance of amino acids is inferred from the alignment of 60 mammalian sequences representative of the main lineages, as partially shown in figure 11. Exon 4 (14 residues) was not included because it is missing in several species studied. Signal peptide is on gray background. The protein sequence (191 amino acids) is numbered from methionine (1). Bold characters (n = 75) indicate residues unchanged in mammals, italics (n = 35) residues that can be substituted by an amino a{tb}cid from the same group only, small roman characters residues that can be substituted, characters on gray background (n = 5) residues that are known so far to lead to amelogenesis imperfecta when substituted, and underlined characters indicate (n = 31) residues that are unchanged in amniotes (mammals and reptiles) [Delgado et al., in press].

(iii) Although this central region of AMEL exon 6 is variable, it maintains its richness in proline (30%) and glutamine (20%) in all sequences studied. This means that this region is also subject to a functional constraint but that this selective pressure probably acts on the general conservation of the P and Q richness rather than on specific amino acid positions. This strongly suggests that this region could be subject to polymorphism in humans.

(iv) The origin of the largest of AMEL exon 6 has to be found in the repeats of nine nucleotides coding for three residues (triplets) PXQ or PXX [Delgado et al., 2005]. These repeats have not been blurred by substitutions during at least 310 million years of amniote evolution, because such triplet repeats have been identified in crocodile AMEL [Sire et al., 2006]. The triplet insertions found in the hot spot mutation in mammals are probably reminiscent of this mechanism. These repeats are to be found, probably, in the origin of AMEL after AMBN duplication, and also constitute the originality of AMEL compared to the other EMPs and to ameloblast-secreted SCPPs in general. This leads to the hypothesis that AMEL divergence consisted of the loss of most of the C-terminal region of the AMBN ancestor and of the development of exon 6 (probably from AMBN exon 5) through several runs of PXQ triplet repeats. This new protein was positively selected during enamel evolution in vertebrates because this hydrophobic region, rich in P and Q, improved the resistance of enamel to wear and microbreaks. This could explain why today AMEL represents 90% of the forming enamel matrix in mammals.

Validation of Mutations and Important Residues

The evolutionary analysis of AMEL in mammals reveals >70 residues (out of 191) that are certainly important for a correct function of AMEL because they have remained unchanged during 225 million years of evolution (fig. 12). The number of conserved residues is reduced to 34 when reptilian AMELs are added to this analysis [Delgado et al., in press]. These 34 positions conserved during 310 million years of amniote evolution are considered crucial residues for enamel formation. All of them are located in the N- and C-terminal regions of AMEL, known to play an important role in relation with the environment (interactions with the ameloblast surface and/or with the mineral crystals). The residues conserved only in mammals could indicate that they play new, important roles for enamel formation in this lineage.

As a consequence of their long-lasting conservation, substitution of the important amino acids revealed in this study could result in enamel defects (AIH1) when substi-

tuted in humans (fig. 12). The five substitutions leading to AIH1 are validated when using the mammalian, and four of them when using the amniote dataset. Therefore, this list of conserved residues in the human AMEL sequence (fig. 12) can be useful for the clinical diagnosis of AIH1 since it helps to validate any human AMEL mutation, which could be suspected for AIH1.

Conclusion

Although the origin of enamel can be traced back to early vertebrates, at least 500 MYA in the fossil record, our knowledge of enamel mineralization genes is still restricted to the tetrapod level (350 MYA) for AMEL and AMBN, and to the mammalian level (225 MYA) for ENAM. The difficulty encountered when looking for EMP genes in the vertebrate lineages that diverged earlier in evolution (i.e. chondrichthyans, 430 MYA, and actinopterygians, 420 MYA) resides in their high sequence variations (intrinsically disordered proteins) and in the lack of sequenced genomes in basal lineages such as lungfish, polypterids and sharks, which do not allow looking for EMP genes using syntheny. Our approach using putative ancestral sequences could help to obtain data in closely related but not in evolutionary distant lineages. Molecular dating of AMBN/AMEL duplication indicates that EMP genes probably appeared at the end of the Precambrian era (>600 MYA) after several rounds of genome/gene duplications that took place in this period. ENAM was created first, then AMBN and AMEL. After AMBN duplication, one copy lost a large part of the ancestral 3′ region and accumulated PXQ repeats. These events gave rise to a new protein: AMEL. AMEL was then positively selected (and constrained), probably because it improved enamel microstructure and thickness: it is now the major protein forming enamel in amniotes. The AMEL story is relatively well established now, but some details will be undoubtedly added when the evolutionary analyses in amphibians and reptiles will be achieved. Such a study will probably open the door to access the AMEL sequence in lungfish, the sister group to tetrapods. In contrast to our knowledge on AMEL, the other ameloblast-secreted SCPP proteins (AMBN, ENAM and the newly identified AMTN and ODAM) are poorly known. Efforts have to be made towards better knowledge of the relationships and evolution of these proteins, and the current genome sequencing programs will certainly be of great value in this quest. It is clear that evolutionary analyses are necessary not only for thorough knowledge of each protein (i.e. its origin, relationships, and mode of evolution) but also because they provide insights into residues that play important roles for the correct function of the protein. In addition, as illustrated with AMEL, sequence datasets obtained in a phylogenetic perspective will be helpful to validate mutations responsible for genetic diseases in humans.

Acknowledgments

We are grateful to Ann Huysseune (Ghent University, Belgium), J. Hu and J.P. Simmer (University of Michigan School of Dentistry, Ann Arbor, Mich., USA), N. Takahata (Graduate University for Advanced Studies, Kanagawa, Japan), and K. Kawasaki (Pennsylvania State University, University Park, Pa., USA) for helpful remarks and suggestions. We thank J.P. Simmer for his kind invitation for J.Y.S. to present this review to the 2006 Symposium of the International Association for Dental Research in Brisbane, Australia.

References

Aoba, T. (1996) Recent observations on enamel crystal formation during mammalian amelogenesis. Anat Rec *245*: 208–218.

Aoba, T., E.C. Moreno, M. Kresak, T. Tanabe (1989) Possible roles of partial sequences at N- and C-termini of amelogenin in protein-enamel mineral interaction. J Dent Res *68*: 1331–1336.

Aung, P.P., N. Oue, Y. Mitani, H. Nakayama, K. Yoshida, T. Noguchi, A.K. Bosserhoff, W. Yasui (2006) Systematic search for gastric cancer-specific genes based on SAGE data: melanoma inhibitory activity and matrix metalloproteinase-10 are novel prognostic factors in patients with gastric cancer. Oncogene *25*: 2546–2557.

Bartlett, J.D., O.H. Ryu, J. Xue, J.P. Simmer, H.C. Margolis (1998) Enamelysin mRNA displays a developmentally defined pattern of expression and encodes a protein which degrades amelogenin. Connect Tissue Res *39*: 101–109.

Bartlett, J.D., E. Beniash, D.H. Lee, C.E. Smith (2004) Decreased mineral content in MMP-20 null mouse enamel is prominent during the maturation stage. J Dent Res *83*: 909–913.

Bartlett, J.D., R.L. Ball, T. Kawai, C.E. Tye, M. Tsuchiya, J.P. Simmer (2006a) Origin, splicing, and expression of rodent amelogenin exon 8. J Dent Res *85*: 894–899.

Bartlett, J.D., B. Ganss, M. Goldberg, J. Moradian-Oldak, M.L. Paine, M.L. Snead, X. Wen, S.N. White, Y.L. Zhou (2006b) Protein-protein interactions of the developing enamel matrix. Curr Top Dev Biol *74*: 57–115.

Carter, J.G. (1990) Skeletal Biomineralization: Patterns, Processes and Evolutionary Trends. New York, Van Nostrand Reinhold, p 832.
Chung, W.Y., R. Albert, I. Albert, A. Nekrutenko, K.D. Makova (2006) Rapid and asymmetric divergence of duplicate genes in the human gene coexpression network. BMC Bioinformatics *7*: 46.
Davit-Béal, T., F. Allizard, J.-Y. Sire (2007) Enameloid/enamel transition through successive tooth replacements in *Pleurodeles waltl* (Lissamphibia, Caudata). Cell Tissue Res *328*: 167–183.
Dehal, P., J.L. Boore (2005) Two rounds of genome duplication in the ancestral vertebrate. PLoS Biol *3*: e314.
Delgado, S., D. Casane, L. Bonnaud, M. Laurin, J.-Y. Sire, M. Girondot (2001) Molecular evidence for Precambrian origin of amelogenin, the major protein of vertebrate enamel. Mol Biol Evol *18*: 2146–2153.
Delgado, S., M. Girondot, J.-Y. Sire (2005) Molecular evolution of amelogenin in mammals. J Mol Evol *60*: 12–30.
Delgado, S., M.L. Couble, H. Magloire, J.-Y. Sire (2006) Cloning, sequencing and expression of the amelogenin gene in two scincid lizards. J Dent Res *85*: 138–143.
Delgado, S., M. Ishiyama, J.-Y. Sire (in press) Validation tools for AIH1 inferred from amelogenin evolution. J Dent Res.
Delsuc, F., M. Scally, O. Madsen, M.J. Stanhope, W.W. de Jong (2002) Molecular phylogeny of living xenarthrans and the impact of character and taxon sampling on the placental tree rooting. Mol Biol Evol *19*: 1656–1671.
Diekwisch, T., S. David, P. Bringas, Jr., V. Santos, H.C. Slavkin (1993) Antisense inhibition of AMEL translation demonstrates supramolecular controls for enamel HAP crystal growth during embryonic mouse molar development. Development *117*: 471–482.
Donoghue, P.C.J. (1998) Growth and patterning in the conodont skeleton. Philos Trans R Soc Lond Ser B *353*: 633–666.
Donoghue, P.C.J. (2001) Microstructural variation in conodont enamel is a functional adaptation. Proc R Soc Lond Ser B *268*: 1691–1698.
Donoghue, P.C.J. (2002) Evolution and development of the vertebrate dermal and oral skeletons: unraveling concepts, regulatory theories, and homologies. Paleobiology *28*: 474–507.
Donoghue, P.C.J., I.J. Sansom (2002) Origin and early evolution of vertebrate skeletonization. Microsc Res Techn *59*: 352–372.
Donoghue, P.C.J., I.V. Sansom, J.P. Downs (2006) Early evolution of vertebrate skeletal tissues and cellular interactions, and the canalization of skeletal development. J Exp Zoolog B Mol Dev Evol *306*: 278–294.
Du, C., G. Falini, S. Fermani, C. Abbott, J. Moradian-Oldak (2005) Supramolecular assembly of amelogenin nanospheres into birefringent microribbons. Science *307*: 1450–1454.
Dunker, A.K., J.D. Lawson, C.J. Brown, R.M. Williams, P. Romero, J.S. Oh, C.J. Oldfield, A.M. Campen, C.M. Ratliff, K.W. Hipps, J. Ausio, M.S. Nissen, R. Reeves, C.H. Kang, C.R. Kissinger, R.W. Bailey, M.D. Griswold, W. Chiu, E.C. Garner, Z. Obradovic (2001) Intrinsically disordered protein. J Mol Graph Model *19*: 26–59.
Fincham, A.G., Y. Hu, E.C. Lau, H.C. Slavkin, M.L. Snead (1991) Amelogenin post-secretory processing during biomineralization in the postnatal mouse molar tooth. Arch Oral Biol *36*: 305–317.
Fincham, A.G., J. Moradian-Oldak (1993) Amelogenin post-translational modifications: carboxy-terminal processing and the phosphorylation of bovine and porcine 'TRAP' and 'LRAP' amelogenins. Biochem Biophys Res Commun *197*: 248–255.
Fincham, A.G., J. Moradian-Oldak (1995) Recent advances in amelogenin biochemistry. Connect Tissue Res *32*: 119–124.
Fisher, L.W., N.S. Fedarko (2003) Six genes expressed in bones and teeth encode the current members of the SIBLING family of proteins. Connect Tissue Res *44(suppl 1)*: 33–40.
Ginger, M.R., C.P. Piotte, D.E. Otter, M.R. Grigor (1999) Identification, characterisation and cDNA cloning of two caseins from the common brushtail possum (*Trichosurus vulpecula*). Biochim Biophys Acta *1427*: 92–104.
Girondot, M., J.-Y. Sire (1998) Evolution of the amelogenin gene in toothed and tooth-less vertebrates. Eur J Oral Sci *106*: 501–508.
Graham, A. (2004) Rise of the little squirts. Curr Biol *14*: R956–R958.
Greene, S.R., Z.A. Yuan, J.T. Wright, H. Amjad, W.R. Abram, J.A. Buchanan, D.I. Trachtenberg, C.W. Gibson (2002) A new frameshift mutation encoding a truncated amelogenin leads to X-linked amelogenesis imperfecta. Arch Oral Biol *47*: 211–217.
Gu, X., Y. Wang, J. Gu (2002) Age distribution of human gene families shows significant roles of both large- and small-scale duplications in vertebrate evolution. Nat Genet *31*: 205–209.
Hart, P.S., M. Aldred, P. Crawford, N. Wright, T. Hart, J.T. Wright (2002a) Amelogenesis imperfecta phenotype-genotype correlations with two amelogenin gene mutations. Arch Oral Biol *47*: 261–265.
Hart, P.S., T.C. Hart, J.P. Simmer, J.T. Wright (2002b) A nomenclature for X-linked amelogenesis imperfecta. Arch Oral Biol *47*: 255–260.
Hart, P.S., M.D. Michalec, W.K. Seow, T.C. Hart, J.T. Wright (2003) Identification of the enamelin (g.8344delG) mutation in a new kindred and presentation of a standardized ENAM nomenclature. Arch Oral Biol *48*: 589–596.
Herold, R.C., J. Rosenbloom, M. Granovsky (1989) Phylogenetic distribution of enamel proteins: immunolocalization with monoclonal antibodies indicates the evolutionary appearance of enamelins prior to amelogenins. Calcif Tissue Int *45*: 88–94.
Hedges, S.B. (2002). The origin and evolution of model organisms. Nat Rev Genet *3*: 838–849.
Hoang, A.M., R.J. Klebe, B. Steffensen, O.H. Ryu, J.P. Simmer, D.L. Cochran (2002) Amelogenin is a cell adhesion protein. J Dent Res *81*: 497–500.
Hu, J.C., Y. Yamakoshi (2003) Enamelin and autosomal-dominant amelogenesis imperfecta. Crit Rev Oral Biol Med *14*: 387–398.
Hu, J.C., X. Sun, C. Zhang, S. Liu, J.D. Bartlett, J.P. Simmer (2002) Enamelysin and kallikrein-4 mRNA expression in developing mouse molars. Eur J Oral Sci *110*: 307–315.
Huysseune, A., J.-Y. Sire (1998) Evolution of patterns and processes in teeth and tooth-related tissues in non-mammalian vertebrates. Eur J Oral Sci *106(suppl 1)*: 437–481.
Ishiyama, M., M. Mikami, H. Shimokawa, S. Oida (1998) Amelogenin protein in tooth germs of the snake *Elaphe quadrivirgata*, immunohistochemistry, cloning and cDNA sequence. Arch Histol Cytol *61*: 467–474.
Iwasaki, K., E. Bajenova, E. Somogyi-Ganss, M. Miller, V. Nguyen, H. Nourkeyhani, Y. Gao, M. Wendel, B. Ganss (2005) Amelotin – a novel secreted, ameloblast-specific protein. J Dent Res *84*: 1127–1132.
Iwase, M., Y. Satta, N. Takahata (2001) Sex-chromosomal differentiation and amelogenin genes in mammals. Mol Biol Evol *18*: 1601–1603.
Iwase, M., Y. Satta, Y. Hirai, H. Hirai, H. Imai, N. Takahata (2003) The amelogenin loci span an ancient pseudoautosomal boundary in diverse mammalian species. Proc Natl Acad Sci USA *100*: 5258–5263.
Janvier, P. (1996) Early Vertebrates. Oxford Monographs on Geology and Geophysics. New York, Oxford University Press, vol 33, p 393.
Kawasaki, K., K.M. Weiss (2003) Mineralized tissue and vertebrate evolution: the secretory calcium-binding phosphoprotein gene cluster. Proc Natl Acad Sci USA *100*: 4060–4065.
Kawasaki, K., K.M. Weiss (2006) Evolutionary genetics of vertebrate tissue mineralization: the origin and evolution of the secretory calcium-binding phosphoprotein family. J Exp Zoolog B Mol Dev Evol *306*: 295–316.
Kawasaki, K., T. Suzuki, K.M. Weiss (2004) Genetic basis for the evolution of vertebrate mineralized tissue. Proc Natl Acad Sci USA *101*: 11356–11361.
Kawasaki, K., T. Suzuki, K.M. Weiss (2005) Phenogenetic drift in evolution: the changing genetic basis of vertebrate teeth. Proc Natl Acad Sci USA *102*: 18063–18068.

Kim, J.-W., J.P. Simmer, Y.Y. Hu, B.P.-L. Lin, C. Boyd, J.T. Wright, C.J.M. Yamada, S.K. Rayes, R.J. Feigal, J.C.-C. Hu (2004) Amelogenin p.M1T and p.W4S mutations underlying hypoplastic X-linked amelogenesis imperfecta. J Dent Res *83*: 378–383.

Kim, J.-W., F. Seymen, B.P.-L. Lin, B. Kiziltan, K. Gencay, J.P. Simmer, J.C.-C. Hu (2005) ENAM mutations in autosomal-dominant amelogenesis imperfecta. J Dent Res *84*: 278–282.

Kumar, S., S.B. Hedges (1998) A molecular timescale for vertebrate evolution. Nature *392*: 917–920.

Kumar, S., A. Filipski, V. Swarna, A. Walker, S.B. Hedges (2005) Placing confidence limits on the molecular age of the human-chimpanzee divergence. Proc Natl Acad Sci USA *102*: 18842–18847.

Lahn B.T., N.M. Pearson, K. Jegalian (2001) The human Y chromosome, in the light of evolution. Nat Rev Genet *2*: 207–216.

Lyngstadaas, S.P., S. Risnes, H. Nordbo, A.G. Flones (1990) Amelogenin gene similarity in vertebrates: DNA sequences encoding amelogenin seem to be conserved during evolution. J Comp Physiol *160*: 469–472.

Madsen, O., M. Scally, C.J. Douady, D.J. Kao, R.W. DeBry, R. Adkins, H.M. Amrine, M.J. Stanhope, W.W. de Jong, M.S. Springer (2001) Parallel adaptive radiations in two major clades of placental mammals. Nature *409*: 610–614.

Mardh, C.K., B. Backman, D. Simmons, I. Golovleva, T.T. Gu, G. Holmgren, M. MacDougall, K. Forsman-Semb (2001) Human ameloblastin gene: genomic organization and mutation analysis in amelogenesis imperfecta patients. Eur J Oral Sci *109*: 8–13.

Margolis, H.C., E. Beniash, C.E. Fowler (2006) Role of macromolecular assembly of enamel matrix proteins in enamel formation. J Dent Res *85*: 775–793.

Moffatt, P., C.E. Smith, R. Sooknanan, R. St-Arnaud, A. Nanci (2006a) Identification of secreted and membrane proteins in the rat incisor enamel organ using a signal-trap screening approach. Eur J Oral Sci *114(suppl 1)*: 139–146.

Moffatt, P., C.E. Smith, R. St-Arnaud, D. Simmons, T. Wright, A. Nanci (2006b) Cloning of rat amelotin and localization of the protein to the basal lamina of maturation stage ameloblasts and junctional epithelium. Biochem J *399*: 37–46.

Murphy, W.J., E. Elzirik, W.E. Johnson, Y.P. Zhang, O.A. Ryder, S.J. O'Brien (2001) Molecular phylogenetics and the origin of placental mammals. Nature *409*: 614–618.

Ørvig, T. (1967) Phylogeny of tooth tissues: evolution of some calcified tissues in early vertebrates; in Miles, A.E.W. (ed): Structural and Chemical Organization of Teeth. New York, Academic Press, vol 1, pp 45–105.

Ørvig, T. (1977) A survey of odontodes ("dermal teeth") from developmental, structural, functional, and phyletic points of view; in Andrews, S.M., R.S. Miles, A.D. Walker (eds): Problems in Vertebrate Evolution. New York, Academic Press, pp 53–75.

Paine, M.L., W. Luo, D.-H. Zhu, P. Bringas, Jr., M.L. Snead (2003a) Functional domains for amelogenin revealed by compound genetic defects. J Bone Miner Res *18*: 466–472.

Paine, M.L., H.J. Wang, M.L. Snead (2003b) Amelogenin self-assembly and the role of the proline located within the carboxyl-teleopeptide. Connect Tissue Res *44*: 52–57.

Panopoulou, G., A.J. Pouska (2005) Timing and mechanism of ancient vertebrate genome duplications – the adventure of an hypothesis. Trends Genet *21*: 559–567.

Prostak, K., Z. Skobe (1984) Effects of colchicines on fish enameloid matrix formation; in Fearnhead, R.W., S. Suga (eds): Tooth Enamel IV. Amsterdam, Elsevier, pp 525–529.

Prostak, K., Z. Skobe (1988) Ultrastructure of odontogenic cells during enameloid matrix synthesis in tooth buds from an elasmobranch, *Raja erinacea*. Am J Anat *182*: 59–72.

Prostak, K., P. Sieffert, Z. Skobe (1993) Enameloid formation in two tetraodontiform fish species with high and low fluoride contents in enameloid. Arch Oral Biol *38*: 1031–1044.

Qin, C., O. Baba, W.T. Butler (2004) Post-translational modifications of SIBLING proteins and their roles in osteogenesis and dentinogenesis. Crit Rev Oral Biol Med *15*: 126–136.

Ravindranath, R.M, J. Moradian-Oldak, A.G. Fincham (1999) Tyrosyl motif in amelogenins binds N-acetyl-D-glucosamine. J Biol Chem *274*: 2464–2471.

Ravindranath, R.M.H., W. Tam, P. Nguyen, A.G. Fincham (2000) The enamel protein amelogenin binds to the N-acetyl-D-glucosamine-mimicking peptide motif of cytokeratins. J Biol Chem *275*: 39654–39661.

Ravindranath, R.M.H., W. Tam, P. Bringas, V. Santos, A.G. Fincham (2001) Amelogenin-cytokeratin 14 interaction in ameloblasts during enamel growth. J Biol Chem *276*: 36586–36597.

Ravindranath, R.M.H, R.M. Basilrose, N.H. Ravindranath, B. Vaitheesvaran (2003) Amelogenin interacts with cytokeratin-5 in ameloblasts during enamel growth. J Biol Chem *278*: 20293–20302.

Reif, W.E. (1982) Evolution of dermal skeleton and dentition in vertebrates. Evol Biol *15*: 287–368.

Rowe, P.S. (2004) The wrickkened pathways of FGF23, MEPE and PHEX. Crit Rev Oral Biol Med *15*: 264–281.

Salido, E., P. Yen, K. Koprivnikar, L.C. Yu, L. Shapiro (1992) The human enamel protein gene amelogenin is expressed from both the X- and Y-chromosomes. Am J Hum Genet *50*: 303–316.

Sansom, I.V., M.P. Smith, H.A. Armstrong, M.M. Smith (1992) Presence of earliest vertebrate hard tissues in conodonts. Science *256*: 1308–1311.

Sansom, I.J., M.P. Smith, M.M. Smith (1994) Dentine in conodonts. Nature *368*: 391.

Sansom, I.J., P.C.J. Donoghue, G.L. Albanesi (2005) Histology and affinity of the earliest armoured vertebrate. Biol Lett *2*: 446–449.

Sasagawa, I. (1984) Formation of cap enameloid in the jaw teeth of dog salmon, *Oncorhynchus keta*. Jpn J Oral Biol *26*: 477–495.

Sasagawa, I. (1995) Fine structure of the tooth germs during formation of the enameloid matrix in *Tilapia nilotica*, a teleost fish. Arch Oral Biol *40*: 801–814.

Sasagawa, I. (2002a) Fine structural and cytochemical observations of dental epithelial cells during the enameloid formation stages in red stingrays *Dasyatis akajei*. J Morphol *252*: 170–182.

Sasagawa, I. (2002b) Mineralization patterns in elasmobranch fish. Microsc Res Techn *59*: 396–407.

Satchell, P.G., C.F. Shuler, T.G.H. Diekwisch (2000) True enamel covering in teeth of the Australian lungfish *Neoceratodus forsteri*. Cell Tissue Res *299*: 27–37.

Sato, I., M. Kobayashi, R. Ueno, T. Sato (1986) The ultrastructure of the teeth in the Amphibia: differentiation of the enamel. Shigaku *73*: 1815–1820.

Shimeld, S.M., P.W.H. Holland (2000) Vertebrate innovations. Proc Natl Acad Sci USA *97*: 4449–4452.

Shintani, S., M. Kobata, S. Toyosawa, T. Fujiwara, A. Sato, T. Ooshima (2002) Identification and characterization of ameloblastin gene in a reptile. Gene *283*: 245–254.

Shintani, S., M. Kobata, S. Toyosawa, T. Ooshima (2003) Identification and characterization of ameloblastin gene in an amphibian, *Xenopus laevis*. Gene *318*: 125–136.

Shu, D.-G., H.-L. Luo, S. Conway Morris, X.-L. Zhang, S.-X. Hu, L. Chen, J. Han, M. Zhu, Y. Li, L.-Z. Chen (1999). Lower Cambrian vertebrates from South China. Nature *402*: 42–46.

Shu, D.-G., S.C. Morris, J. Han, Z.-F. Zhang, K. Yasui, P. Janvier, L. Chen, X.-L. Zhang, J.-N. Liu, Y. Li, H.-Q. Liu (2003) Head and backbone of the Early Cambrian vertebrate *Haikouichthys*. Nature *421*: 526–529.

Simmer, J.P., M. Fukae, T. Tanabe, Y. Yamakoshi, T. Uchida, J. Xue, H.C. Margolis, M. Shimizu, B.C. DeHart, C.-C. Hu, J.D. Bartlett (1998) Purification, characterization and cloning of enamel matrix serine proteinase 1. J Dent Res *77*: 377–386.

Simmer, J.P., J.C. Hu (2002) Expression, structure, and function of enamel proteinases. Connect Tissue Res *43*: 441–449.

Sire, J.-Y. (1990) From ganoid to elasmoid scales in the actinopterygian fishes. Neth J Zool *40*: 75–92.

Sire, J.-Y. (1994) A light and TEM study of non-regenerated and experimentally regenerated scales of *Lepisosteus oculatus* (Holostei) with particular attention to ganoine formation. Anat Rec *240*: 189–207.

Sire, J.-Y. (1995) Ganoine formation in the scales of primitive actinopterygian fishes, lepisosteids and polypterids. Connect Tissue Res *33*: 213–222.

Sire, J.-Y., A. Huysseune (2003) Formation of skeletal and dental tissues in fish: a comparative and evolutionary approach. Biol Rev *78*: 219–249.

Sire, J.-Y., S. Delgado, D. Fromentin, M. Girondot (2005) Amelogenin: lessons from evolution. Arch Oral Biol 50: 205–212.

Sire, J.-Y., S. Delgado, M. Girondot (2006) The amelogenin story: origin and evolution. Eur J Oral Sci *114*: 64–77.

Sire, J.-Y., J. Géraudie, F.J. Meunier, L. Zylberberg (1987) On the origin of ganoine: histological and ultrastructural data on the experimental regeneration of the scales of *Calamoichthys calabaricus* (Osteichthyes, Brachiopterygii, Polypteridae). Am J Anat *180*: 391–402.

Slavkin, H.C., N. Samuel, P. Bringas, Jr., A. Nanci (1983) Selachian tooth development: II. Immunolocalization of amelogenin polypeptides in epithelium during secretory amelogenesis in *Squalus acanthias*. J Craniofac Genet Dev Biol *3*: 43–52.

Smith, M.M. (1995) Heterochrony in the evolution of enamel in vertebrates; in McNamara, K.J. (ed): Evolutionary Change and Heterochrony. New York, Wiley, pp 125–150.

Smith, M.M., B.K. Hall (1990) Development and evolutionary origins of vertebrate skeletogenic and odontogenic tissues. Biol Rev *65*: 277–373.

Snead, M. (2003). Amelogenin protein exhibits a modular design: implications for form and function. Connect Tissue Res *44(suppl 1)*: 47–51.

Solomon, A., C.L. Murphy, K. Weaver, D.T. Weiss, R. Hrncic, M. Eulitz, R.L. Donnell, K. Sletten, G. Westermark, P. Westermark (2003) Calcifying epithelial odontogenic (Pindborg) tumor-associated amyloid consists of a novel human protein. J Lab Clin Med *142*: 348–355.

Stasiuk, S.X., E.L. Summers, J. Demmer (2000) Cloning of a marsupial kappa-casein cDNA from the brushtail possum (*Trichosurus vulpecula*). Reprod Fertil Dev *12*: 215–222.

Steinke, D., W. Salzburger, I. Braasch, A. Meyer (2006) Many genes in fish have species-specific asymmetry rates of molecular evolution. BMC Genom *8*: 7–20.

Stephanopoulos, G., M.E. Garefalaki, K. Lyroudia (2005) Genes and related proteins involved in amelogenesis imperfecta. J Dent Res *84*: 1117–1126.

Tompkins, K., A. George, A. Veis (2006) Characterization of a mouse amelogenin [A-4]'M59 cell surface receptor. Bone *38*: 172–180.

Toyosawa, S., C. O'hUigin, F. Figueroa, H. Tichy, J. Klein (1998) Identification and characterization of amelogenin genes in monotremes, reptiles, and amphibians. Proc Natl Acad Sci USA *95*: 13056–13061.

van Rheede, T., T. Bastiaans, D.N. Boone, S.B. Hedges, W.W. de Jong, O. Madsen (2006) The platypus is in its place: nuclear genes and indels confirm the sister group relation of monotremes and Therians. Mol Biol Evol *23*: 587–597.

Veis, A., K. Tompkins, K. Alvares, K. Wei, L. Wang, X.S. Wang, A.G. Brownell, S.-M. Jengh, K.E. Healy (2000) Specific amelogenin gene splicing products have signaling effects on cells in culture and in implants in vivo. J. Biol Chem *275*: 41263–41272.

Veis, A. (2003) Amelogenin gene splice products: potential signaling molecules. Cell Mol Life Sci *60*: 38–55.

Veis, A. (2005) A window on biomineralization. Science *307*: 1419–1420.

Wang, X., Y. Ito, X. Luan, A. Yamane, T.G.H. Diekwisch (2005a) Amelogenin sequence and enamel biomineralization in *Rana pipiens*. J Exp Zoolog B Mol Dev Evol *304*: 1–10.

Wang, H.J., S. Tannukit, D.H. Zhu, M.L. Snead, M.L. Paine (2005b) Enamel matrix protein interactions. J Bone Miner Res *20*: 1032–1040.

Wang, X., J.L. Fan, Y. Ito, X. Luan, T.G.H. Diekwisch (2006) Identification and characterization of a squamate reptilian amelogenin gene: *Iguana iguana*. J Exp Zool B Mol Dev Evol *306*: 393–406.

Weikard, R., C. Pitra, C. Kuhn (2006) Amelogenin cross-amplification in the family Bovidae and its application for sex determination. Mol Reprod Dev *73*: 1333–1337.

Wen, H.B., A.G. Fincham, J. Moradian-Oldak (2001) Progressive accretion of amelogenin molecules during nanospheres assembly revealed by atomic force microscopy. Matrix Biol *20*: 387–395.

Yamamoto, K., T. Tsubota, T. Komatsu, A. Katayama, T. Murase, I. Kita, T. Kudo (2002) Sex identification of Japanese black bear, *Ursus thibetanus japonicus*, by PCR based on amelogenin gene. J Vet Med Sci *64*: 505–508.

Zhang, X., J. Zhao, C. Li, S. Gao, C. Qiu, P. Liu, G. Wu, B. Qiang, W.H.Y. Lo, Y. Shen (2001) DSPP mutation in dentinogenesis imperfecta Shields type II. Nat Genet *27*: 151–152.

Zylberberg, L., J.-Y. Sire, A. Nanci (1997) Immunodetection of amelogenin-like proteins in the ganoine of experimentally regenerating scales of *Calamoichthys calabaricus*, a primitive actinopterygian fish. Anat Rec *249*: 86–95.

RESEARCH REPORTS
Biological

S. Delgado[1], M. Ishiyama[2], and J.-Y. Sire[1]*

[1]UMR 7138, Equipe "Evolution & Développement du Squelette", Université Paris 6, Case 05, 7 quai St-Bernard, 75005 Paris, France; and [2]Department of Histology, The Nippon Dental University, School of Dentistry, Niigata, Japan; *corresponding author, sire@ccr.jussieu.fr

J Dent Res 86(4):326-330, 2007

Validation of Amelogenesis Imperfecta Inferred from Amelogenin Evolution

ABSTRACT

We used the evolutionary analysis of amelogenin (AMEL) in 80 amniotes (52 mammalian and 28 reptilian sequences) to aid in the genetic diagnosis of X-linked amelogenesis imperfecta (AIH1). Out of 191 residues, 77 were found to be unchanged in mammals, and only 34 in amniotes. The latter are considered crucial residues for enamel formation, while the 43 residues conserved only in mammals could indicate that they play new, important roles for enamel formation in this lineage. The 5 substitutions leading to AIH1 were validated when the mammalian dataset was used, and 4 of them with the amniote dataset. These 2 sequence datasets will facilitate the validation of any human AMEL mutation suspected of involvement in AIH1. This evolutionary analysis also revealed numerous residues that appeared to be important for correct AMEL function, but their role remains to be elucidated.

KEY WORDS: amelogenin, amelogenesis imperfecta, molecular evolution, enamel, teeth, mammals, reptiles.

Received July 11, 2006; Last revision November 23, 2006; Accepted November 29, 2006

A supplemental appendix to this article is published electronically only at http://www.dentalresearch.org.

INTRODUCTION

Amelogenin (AMEL) is the major protein of forming enamel. In humans, the amelogenin genes (*AMEL*) are located on the X and Y chromosomes, but in males, 90% of the transcripts are expressed from *AMELX* (Salido *et al.* 1992). AMEL plays a crucial role in enamel formation, but its exact functions are not totally understood (Paine *et al.*, 2003). Its importance is well-illustrated, however, by the occurrence of a genetic disease, X-linked amelogenesis imperfecta (AIH1), resulting from *AMEL* mutations leading to various hypoplastic and hypomature enamel phenotypes. To date, 14 mutations leading to AIH1 are known (Hart *et al.*, 2002a; Kim *et al.*, 2004). The characterization of these mutations helps in identifying particular regions, or specific residues, that play a crucial role in AMEL function (Collier *et al.*, 1997; Ravindranath *et al.*, 1999). However, the few *AMEL* mutations known so far are insufficient to target all important residues.

Mutational analyses are time-consuming and expensive: analysis of the individual's pedigree, mapping the mutation on a chromosome to identify a candidate gene, sequencing, and sequence analysis to validate the mutation. Moreover, *AMEL* polymorphism could lead to diagnostic errors in the clinical context, and this possibility is largely underestimated. Indeed, if a person has an enamel defect, and there is a pedigree consistent with an X-linked mutation, then a polymorphism in *AMELX* is unlikely to be the cause of the defect.

Evolutionary analysis is an alternative for validating the *AMEL* mutations responsible for AIH1, and for highlighting all the residues that are important for the protein to function correctly (Delgado *et al.*, 2005; Sire *et al.*, 2005, 2006). Such an analysis is based on the following postulates: (i) Important residues must remain unchanged, because their change or loss could lead to severe enamel defaults; (ii) conversely, less important residues can be substituted without damage to enamel structure and organization, and must therefore be considered polymorphisms; and (iii) given the slow rate of mutations in most lineages, the studied sample must cover a large evolutionary period and must be representative of the various lineages in which the protein has similar functions. This is the case in mammals and reptiles (amniotes), in which enamel structure is roughly similar (Sander, 2000), although both lineages separated approximately 310 million years (my) ago (Hedges, 2002). Nevertheless, in reptiles, teeth are continuously replaced during life (polyphyodonty), and the constraints acting on enamel structure could be less important than in mammals, which are diphyodont or monophyodont.

In the present study, we compiled 52 mammalian and 28 reptilian AMEL sequences, with the aim of obtaining datasets that could be useful for a rapid and accurate validation of the mutations responsible for AIH1.

MATERIALS & METHODS

In humans, *AMELX* is composed of 7 exons. Exon 1 is not translated; exon 4 is subjected to alternative splicing (Hu *et al.*, 1996; Yuan *et al.*, 1996, 2001) and is absent in some mammals and in all reptiles (Ishiyama *et al.*, 1998; Delgado *et al.*, 2006); and exon 7 codes for a single amino acid. Nine exons have been identified in rat and mouse *AMEL* (Li *et al.*, 1998), but exons 8 and 9 are absent in all other species studied so far. Therefore, only *AMEL* exons 2, 3, 5, and 6 were included in the present study. Because the sequences of the small exons 2, 3, and 5 (fewer than 60 bp each) are well-conserved, we have concentrated our efforts primarily on exon 6 (> 400 bp), which is more variable. *AMELY* has evolved separately in various mammalian lineages (Girondot and Sire, 1998), in relation to the particular pattern of Y chromosome evolution (Iwase *et al.*, 2001, 2003). Therefore, *AMELY* was not included in our study.

Materials

Several *AMEL* sequences were found in GenBank, and one sequence was obtained from the literature (Yamamoto *et al.*, 2002). We completed this dataset by blasting sequenced genomes and by sequencing *AMEL* in representative species of most amniote lineages (Fig. 1). A dataset of 80 sequences (52 mammals and 28 reptiles) was obtained. References to species and sequences are found in APPENDIX 1. Taxa which have either no teeth [*e.g.*, baleen whales (*Mysticeti*), anteaters (*Xenarthra*), pangolins (*Pholidota*)] or no enamel [*e.g.*, armadillos (*Xenarthra*), aardvarks (*Tubulidentata*)] were not included in this study.

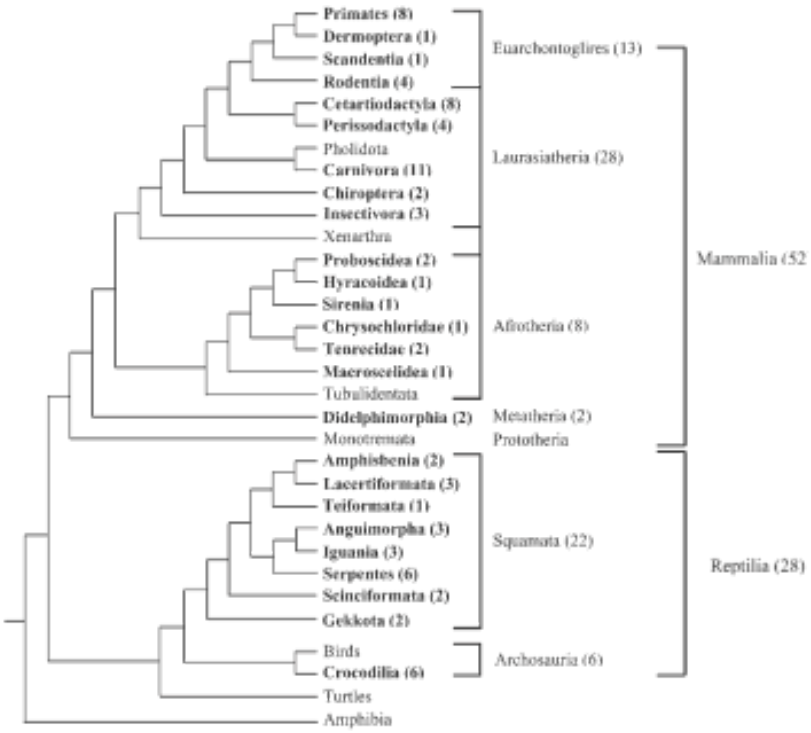


Figure 1. Relationships of the amniote lineages (in bold) for which amelogenin was used in this study (adapted from Madsen *et al.*, 2001; Murphy *et al.*, 2001; Janke *et al.*, 2005; Vidal and Hedges, 2005). The number of species in each clade is indicated between the brackets. See APPENDIX 1 for information on the species and sequences.

Methods

DNA and RNA Extraction

Genomic DNA was extracted (DNeasy tissue kit: Qiagen-GmBH, Ilden, Germany) from soft tissues conserved in ethanol. mRNAs were obtained from 4 lizards (RNeasy kit: Qiagen) and converted into cDNAs (ReverAid kit: MBI Fermentas, Hanover, PA, USA).

Primers

Primers were defined from the alignment of known *AMEL* sequences (see APPENDIX 2).

PCR Amplification

Genomic DNA or cDNA (1 µL) was amplified in a mixture composed of 5 µL Taq buffer (10x) (pH 8.8), 3 µL $MgCl_2$ 2 mM, and 1 µL dNTP 10 mM, in the presence of sense and antisense primers, and 0.3 µL Red Hot polymerase (Advanced Biotechnologies Ltd., Foster City, CA, USA). Amplification was performed in a thermocycler (Genius Techne) for 38 cycles, each cycle consisting of 1 min of denaturation at 94°C, 1 min of annealing at 59°C, and 1 min of extension at 72°C. The final extension was for 20 min at 72°C.

Cloning

One microgram of PCR product was isolated, ligated to pCR 2.1-TOPO plasmid vector (Invitrogen SA, Carlsbad, CA, USA) by the TA-cloning method, then used to transform competent *E. coli* TOP10F bacteria. The transformed bacteria were grown overnight at 37°C in Luria-ampicillin broth, and subjected to lysis in 200 µL of NaOH 0.2 M-SDS 1%, at 0°C for 5 min. Subsequently, a 150-µL quantity of AcK 3 M was added at 0°C for 5 min to precipitate the proteins. The plasmids were purified in a phenol/chloroform mixture. Sequencing was done by Genome Express S.A (Meylan, France).

Molecular Analyses

AMEL sequences were aligned *via* Clustal X 1.81 (Thompson *et al.*, 1997), and checked by hand with Se-Al v2.0 (available at http://evolve.zoo.ox.ac.uk/).

RESULTS

The Mammalian and Reptilian AMEL Datasets

Of the 250 amino acids (aa) in the alignment of the 52

Figure 2. Amino-acid sequence of human amelogenin with indication of important residues inferred from the alignment of 52 mammalian sequences (**A**) and of 80 amniote sequences (52 mammals, 6 crocodiles, and 22 squamates) (**B**). (The alignments are presented in APPENDIX 2.) Exon 4 (14 residues) was not included, because it was absent in most species studied. Signal peptide is on the grey background. The protein sequence (191 amino acids) is numbered from methionine (1). Large characters = residues unchanged; italics = residues that can be substituted for by an amino acid from the same group only. Small characters = residues for which substitution can be made. Boldface characters = the 5 residues known to lead to amelogenesis imperfecta after substitution.

mammalian AMEL sequences (including residue insertions), 77 were unchanged, and 30 were substituted by a residue from the same group (APPENDIX 3A). Most of the conserved amino acids were located in the N- and C-terminal regions [coded by exons 2, 3, 5, and the begining of exon 6, up to the TRAP (tyrosine-rich amelogenin peptide) proteolytic sites (aa 1-64) and the end of exon 6 (aa 218-250), respectively]. In contrast, the central region of exon 6 (aa 65-217) showed numerous variations, with a particular region characterized by large sequence deletions or insertions (aa 130-208). Twelve AMEL sequences possessed triplet (PXQ or PXX) insertions (up to 10 in the water opossum), while 4 other sequences showed deletions (up to 17 in the dolphin). All positions currently considered important were unchanged, including the TRAP proteolytic loci (aa 59 and 61) and the LRAP (leucine-rich amelogenin peptide) intra-exonic splicing site (aa 223).

In crocodiles, the 6 AMEL sequences were highly similar (APPENDIX 3B). Of 199 aa in the alignment, only 11 were substituted, and most of these were by residues from the same group. In squamates, the 22 AMEL sequences showed a high degree of variation (APPENDIX 3C). Of 217 amino acids in the alignment, 53 were unchanged, and 18 were substituted by a residue from the same group. Most unchanged residues were located in the N (aa 1-64) and C (aa 192-217) terminal regions; nearly all positions in the variable region of exon 6 (aa 65-191) were substituted.

When we considered the complete alignment of amniote AMEL, we could not align most parts of exon 6 (from aa 68 onward in our alignment), due to the high number of variations (substitutions, deletions, and insertions) (APPENDIX 3D). Only the N- and C- terminal regions could be aligned. We found 34 unchanged residues in these regions and 15 residues that were substituted by a residue from the same group. The proteolytic loci leading to TRAP were conserved, while the intra-exonic splicing site for LRAP could not be identified in most squamates.

Validation of AIH1 Using Two Sequence Datasets

The results obtained from the analysis of mammalian (52 AMEL) and amniote (80 AMEL) sequences were transposed onto the human AMEL sequence, with indication of residues that were unchanged, substituted by an amino acid from the same group, or variable (Figs. 2A, 2B). Of the 5 residues known to lead to AIH1 when substituted (M1, W4, T37, P56, and H63 in our sequence; p.M1T, p.W4S, p.T51I, p.P70T, p.H77L, respectively, in the AIH1 nomenclature), 4 were validated (*i.e.*, unchanged) in mammalian and amniote sequence datasets, and all when only AMEL sequences were used. Indeed, the p.H77L mutation was not validated by the amniote dataset: Histidine (H: basic group) was substituted by glutamine (Q: polar) in crocodiles and in a snake. In humans, this AIH1 resulted from substitution by a leucine (L: non-polar). Most residues known to be important for a correct function of AMEL were conserved in amniotes. In addition, the datasets revealed a high number of unchanged amino acids.

DISCUSSION

A genetic diagnosis of AIH1 relies, eventually, upon the sequencing of *AMEL* and comparison of the obtained sequence with the reference sequence for humans. When an obvious mutation is found (large deletions, reading frameshift leading to a stop codon, etc.), it is considered to be responsible for the

observed phenotype. When the mutation leads to a single amino acid substitution, the genotype-phenotype relationship is less obvious, and one could envisage this mutation as a polymorphism, *i.e.*, the disorder not being related to this mutation. Of the 14 AMEL mutations identified for X-linked AI (Hart *et al.*, 2002a; Kim *et al.*, 2004), 5 are single-residue substitutions. If the mutation is in a position conserved in other species, this feature supports the genetic diagnosis. Indeed, the sites of crucial importance for AMEL must be kept unchanged during evolution; otherwise, their substitution could lead to a genetic disease. However, given the high sequence similarity of AMEL in closely related mammalian species, it is difficult to decide whether conserved sites are preserved because they are highly constrained or because the evolutionary distance between these lineages is too short to reveal all low-constrained sites. Species that are too closely related are not relevant in a decision of evolutionary conservation. To ensure that residue conservation is related to a functional constraint, one needs to know AMEL sequences in species that are more distantly related. This is the reason we built these sequence datasets based on mammalian and reptilian diversity, to help in AIH1 validation.

We have chosen to present 2 datasets, one based on AMEL sequences of 52 mammals, and the other on a compilation of 80 amniote sequences. Indeed, although enamel structure is roughly similar in mammals and reptiles, some enamel specificities could have been selected for during the long evolutionary period (310 my) that separates these lineages. In contrast to reptiles, in which some ancestral characters, such as polyphyodonty, have been conserved, mammals no longer replace their teeth continuously throughout life. Furthermore, from a structural viewpoint, Tomes' processes, a feature of mammalian ameloblasts related to the prismatic structure of enamel, do not exist in reptiles, in which enamel is non-prismatic (Sander, 2000). These two mammalian novelties could have led to new constraints in the AMEL sequence. We hypothesized that the 34 AMEL residues which are unchanged at the amniote level are essential for the correct formation and mineralization of enamel, *i.e.*, they are important for AMEL interactions with the cell membrane and/or the mineral crystals. This hypothesis was well-supported: All these conserved positions were found at the N- and C-terminal regions, which are known to exert such functions (Paine *et al.*, 2003; Snead, 2003). We hypothesized also that the 43 residues that are conserved only in mammals are related to the peculiar features of enamel that were selected for during mammalian evolution (180 my). Half of the unchanged positions were found in the N- and C-terminal regions, reflecting a possible stronger constraint on the AMEL sequence in these regions in mammals than in reptiles. The other conserved positions were found in the region known to be variable (Delgado *et al.*, 2005; Sire *et al.*, 2005, 2006), either close to the N- and C-terminal regions or in the central region of exon 6. This could also reflect new constraints in this region, but we can also envisage that these positions are not really important for AMEL function. Perhaps 180 my are insufficient for random substitution of amino acids that are not really important.

The 5 amino acid substitutions known to lead to AIH1 were validated by our method with the mammalian dataset, and 4 of them with the amniote dataset. In reptiles, the substitution of H63 in our alignment (p.H77L: Hart *et al.*, 2002b) by a glutamine (Q) could indicate that this locus has probably been constrained during mammalian evolution only. The presence of this basic residue probably plays a role in TRAP proteolysis by enamelysin (MMP20). Does this mean that there is no TRAP in crocodiles, or that a polar residue (Q) could replace a basic one (H)? Amino acids that were replaced by residues from the same group were also indicated in the human sequence. Indeed, if one considers that only the biochemical characteristics of a position are important, there would be no problem if the residue were substituted by an amino acid from the same group.

Our evolutionary analysis of AMEL at the amniote level confirmed our previous findings, inferred from the comparative study of mammalian AMEL, *i.e.*, highly conserved residues in the N- and C-terminal regions, and a variable region in exon 6 (Delgado *et al.*, 2005, 2006; Sire *et al.*, 2006). In exon 6, the intra-exonic splicing site, which releases LRAP (a short peptide involved in cell signaling: Veis *et al.*, 2000), was well-conserved in mammals, but not in reptiles. The 'hot spot' of mutation (*i.e.*, large insertions and/or deletions located in the central region of exon 6) in mammals (Delgado *et al.*, 2005) was found in the present study in a few newly sequenced AMEL of mammalian species, but was absent in reptiles. These features were acquired recently in mammalian evolution.

In addition to proposed sequence dataset, which will help in the diagnosis of AIH1, this analysis has revealed 30 unchanged residues with unknown, but certainly important, function. These amino acids could be good candidates for AIH1 if they were substituted, and their role in AMEL function should be evaluated.

Our study showed how evolutionary analysis, when conducted within a phylogenetic framework, could help both in validating mutations in humans and in revealing amino acids that could play important roles in enamel structure and organization. In dental research, this method could be applied to the study of other genes—for instance, enamelin, which is known to be responsible for autosomal-dominant AI, and dentin sialophosphoprotein, responsible for dentinogenesis imperfecta. The large number of genomes currently being sequenced in mammals could be taken as an opportunity to build datasets that could be used to validate mutations responsible for a genetic disease.

ACKNOWLEDGMENTS

We are grateful to Prof. Ann Huysseune (Ghent University, Belgium) for helpful criticism of the manuscript. We are grateful to the following colleagues for sending either DNA or tissue samples: F. Catzeflis (UMR 5554, Université de Montpellier 2, France); L. Fougeirol and S. Martin (La Ferme des Crocodiles, Pierrelatte, France); A. Lécu and F. Ollivet (Zoo de Vincennes, MNHN, France); G. Véron, V. de Buffrénil and N. Vidal (Muséum national d'Histoire naturelle, France); W. Dabin (Muséum de la Rochelle, France); T. Robinson (Stellenboch University, Afrique du Sud); and D.J. Harris (Centro de Estudos de Ciência Animal, Vila do Conde, Portugal). This work was financially supported by IFRO (Institut Français de Recherche Odontologiques).

Since our article was in press, "A Novel Missense Mutation (p.P52R) in Amelogenin Gene Causing X-linked Amelogenesis Imperfectca" was published in *JDR*, 86:69-72, 2007, by M. Kida *et al.* This substitution is validated by our evolutionary analysis (exon5, position 38 in our alignment).

REFERENCES

Collier PM, Sauk JJ, Rosenbloom SJ, Yuan ZA, Gibson CW (1997). An amelogenin gene defect associated with human X-linked amelogenesis imperfecta. *Arch Oral Biol* 42:235-242.

Delgado S, Girondot M, Sire JY (2005). Molecular evolution of amelogenin in mammals. *J Mol Evol* 60:12-30.

Delgado S, Couble ML, Magloire H, Sire JY (2006). Cloning, sequencing, and expression of the amelogenin gene in two scincid lizards. *J Dent Res* 85:138-143.

Girondot M, Sire JY (1998). Evolution of the amelogenin gene in toothed and toothless vertebrates. *Eur J Oral Sci* 106(Suppl 1):501-508.

Hart PS, Hart TC, Simmer JP, Wright JT (2002a). A nomenclature for X-linked amelogenesis imperfecta. *Arch Oral Biol* 47:255-260.

Hart PS, Aldred MJ, Crawford PJ, Wright NJ, Hart TC, Wright JT (2002b). Amelogenesis imperfecta phenotype-genotype correlations with two amelogenin gene mutations. *Arch Oral Biol* 47:261-265.

Hedges SB (2002). The origin and evolution of model organisms. *Nat Rev Genet* 3:838-849.

Hu CC, Bartlett JD, Zhang CH, Qian Q, Ryu OH, Simmer JP (1996). Cloning, cDNA sequence, and alternative splicing of porcine amelogenin mRNAs. *J Dent Res* 75:1735-1741.

Ishiyama M, Mikami M, Shimokawa H, Oida S (1998). Amelogenin protein in tooth germs of the snake *Elaphe quadrivirgata*, immunohistochemistry, cloning and cDNA sequence. *Arch Histol Cytol* 61:467-474.

Iwase M, Satta Y, Takahata N (2001). Sex-chromosomal differentiation and amelogenin genes in mammals. *Mol Biol Evol* 18:1601-1603.

Iwase M, Satta Y, Hirai Y, Hirai H, Imai H, Takahata N (2003). The amelogenin loci span an ancient pseudoautosomal boundary in diverse mammalian species. *Proc Natl Acad Sci USA* 100:5258-5263.

Janke A, Gullberg A, Hughes S, Aggarwal RK, Arnason U (2005). Mitogenomic analyses place the gharial (*Gavialis gangeticus*) on the crocodile tree and provide pre-K/T divergence times for most crocodilians. *J Mol Evol* 61:620-626.

Kida M, Sakiyama Y, Matsuda A, Takabayashi S, Ochi H, Sekiguchi H, *et al.* (2007). A novel missense mutation (p.P52R) in amelogenin gene causing X-linked amelogenesis imperfecta. *J Dent Res* 86:69-72.

Kim JW, Simmer JP, Hu YY, Lin BP, Boyd C, Wright JT, *et al.* (2004). Amelogenin p.M1T and p.W4S mutations underlying hypoplastic X-linked amelogenesis imperfecta. *J Dent Res* 83:378-383.

Li W, Mathews C, Gao C, DenBesten PK (1998). Identification of two additional exons at the 3′ end of the amelogenin gene. *Arch Oral Biol* 43:497-504.

Madsen O, Scally M, Douady CJ, Kao DJ, DeBry RW, Adkins R, *et al.* (2001). Parallel adaptive radiations in two major clades of placental mammals. *Nature* 409:610-614.

Murphy WJ, Eizirik E, Johnson WE, Zhang YP, Ryder OA, O'Brien SJ (2001). Molecular phylogenetics and the origins of placental mammals. *Nature* 409:614-618.

Paine ML, Luo W, Zhu DH, Bringas P Jr, Snead ML (2003). Functional domains for amelogenin revealed by compound genetic defects. *J Bone Miner Res* 18:466-472.

Ravindranath RM, Moradian-Oldak J, Fincham AG (1999). Tyrosyl motif in amelogenins binds N-acetyl-D-glucosamine. *J Biol Chem* 274:2464-2471.

Salido EC, Yen PH, Koprivnikar K, Yu LC, Shapiro LJ (1992). The human enamel protein gene amelogenin is expressed from both the X and the Y chromosomes. *Am J Hum Genet* 50:303-316.

Sander PM (2000). Prismless enamel in amniotes: terminology, function and evolution. In: Development, function and evolution of teeth. Teaford M, Ferguson MWJ, Smith MM, editors. New York: Cambridge University Press, pp. 92-106.

Sire JY, Delgado S, Fromentin D, Girondot M (2005). Amelogenin: lessons from evolution. *Arch Oral Biol* 50:205-212.

Sire JY, Delgado S, Girondot M (2006). The amelogenin story: origin and evolution. *Eur J Oral Sci* 114(Suppl 1):64-77.

Snead ML (2003). Amelogenin protein exhibits a modular design: implications for form and function. *Connect Tissue Res* 44(Suppl 1):47-51.

Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG (1997). The ClustalX windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res* 24:4876-4882.

Veis A, Tompkins K, Alvares K, Wei K, Wang L, Wang XS, *et al.* (2000). Specific amelogenin gene splicing products have signaling effects on cells in culture and in implants in vivo. *J Biol Chem* 275:41263-41272.

Vidal N, Hedges SB (2005). The phylogeny of squamate reptiles (lizards, snakes, and amphisbaenians) inferred from nine nuclear protein-coding genes. *C R Biol* 328:1000-1008.

Yamamoto K, Tsubota T, Komatsu T, Katayama A, Murase T, Kita I, *et al.* (2002). Sex identification of Japanese black bear, *Ursus thibetanus japonicus*, by PCR based on amelogenin gene. *J Vet Med Sci* 64:505-508.

Yuan ZA, Collier PM, Rosenbloom J, Gibson CW (1996). Analysis of amelogenin mRNA during bovine tooth development. *Arch Oral Biol* 41:205-213.

Yuan ZA, Chen E, Gibson CW (2001). Model system for evaluation of alternative splicing: exon skipping. *DNA Cell Biol* 20:807-813.