Génomique comparée des procaryotes synthétiques marins Prochlorococcus et Synechococcus

Alexis Dufresne

Avertissement

N° ORDRE
de la thèse: 3110

THÈSE

présentée

DEVANT L'UNIVERSITÉ de RENNES 1

pour obtenir

le grade de: ***DOCTEUR DE L'UNIVERSITÉ DE RENNES 1***
Mention : BIOLOGIE

PAR

ALEXIS DUFRESNE

Équipe d'accueil : UMR 7127, CNRS et UPMC, STATION BIOLOGIQUE, ROSCOFF
Ecole Doctorale: Vie, Agronomie, Santé
Composante universitaire: Ecole Interne n°37 de l'Université Pierre & Marie Curie

GÉNOMIQUE COMPARÉE DES PROCARYOTES PHOTOSYNTHÉTIQUES MARINS *PROCHLOROCOCCUS* ET *SYNECHOCOCCUS*

SOUTENUE le 14 décembre 2004 devant la commission d'Examen

COMPOSITION DU JURY

Patrick Forterre, Professeur, Université Paris-Sud	*Rapporteur*
Cheng-Cai Zhang, Professeur, Université d'Aix-Marseille	*Rapporteur*
Frédérique Hubler, Chargée de Recherche, CNRS, Villejean	*Examinateur*
Brice Felden, Professeur, Université de Rennes 1	*Examinateur*
Nicole Tandeau de Marsac, Directeur de Recherche, CNRS, Pasteur	*Examinateur*
Frédéric Partensky, Directeur de Recherche, CNRS, Station Biologique de Roscoff	*Directeur de thèse*

Étant donné que le texte qui suit est relativement sérieux, je propose au lecteur un petit jeu avant de commencer. Le but est, bien évidemment, de retrouver à qui les remerciements sont adressés. Attention, un remerciement peut correspondre à plusieurs personnes. Toute personne ayant plus de 85 % de bonnes réponses se verra remettre un exemplaire dédicacé de cette thèse.

Merci pour / Merci d'avoir...	**Merci à**
l'encadrement	Fred
le chocolat	La Famille
la complicité	Florence
le cinéma	Fabienne
les blagues pourries	Fran
les tartes banane-chocolat	Lôh
les conseils	Isa
...été l'exemple à ne pas suivre	Vincent
le poisson	Gaëtan
...eu peur quand je fais "**BHOU**"	Daniel
...ri à mes blagues	Les membres du Jury
la gentillesse	Pascale
le transport en voiture	Mumu
les concerts	Phinou
les ragots	Nathalie
l'Amour	Domi
le bon vin	Laure
les bandes dessinées	Les green's passés/presents/futurs
les tenues très improbables	Les étudiants roscovites
les pizzas	Marie
les vacances	Stéphane
les larmes	Fabrice
les repas	Julie
les rencontres	Rahel
les soirées	Le comité de thèse
les confidences	Gilles
les critiques	Dorota
...accepté d'évaluer cette thèse	Mathieu
les fleurs	Julia
les cadeaux	Jens
les livres	Tof
les souvenirs	Le trio Manon/Ludo/Nico
les barbecues sur la plage	Olivier
l'amitié	Marie-Noëlle
les recettes	Erwan
la musique	Carole
la chorale	Khadidja
les repas de mariage	Le personnel de la cantine
la motivation	La plupart des membres de la SBR mais pas tout le monde, faut être sérieux aussi...
le sport	
les petites bouffes entre amis	
les conneries	
...été là	
...trouvé les mots qu'il faut	
...su la fermer	
le café	
les petits riens	
les grands moments	
les voyages	
...	

Je tiens à préciser que ces remerciements un peu particuliers ne sont en aucun cas la marque d'une pudeur excessive...

******

SOMMAIRE

Liste des illustrations

Les figures des articles publiés sont numérotés de manière indépendante.

Liste des tableaux

Les tableaux des articles publiés sont numérotés indépendamment.

Liste des abréviations

aa	acide aminé
ABC	ATP binding cassette
ADN	Acide désoxyribonucléique
APC	Allophycocyanine
ARN	Acide ribonucléique
ARNr 16S	ARN de la sous-unité16S du ribosome
ATP	Adenosine triphosphate
BLAST	Basic Local Alignment Search Tool
Chl	Chlorophylle
COG	Clusters of Orthologous Groups
CPD	Cyclobutane pyrimidine dimer
FAD	Flavin adenine dinucleotide
ITS	Internal Transcribed Spacer
kb	Kilobase
kDa	Kilodalton
NADP	Nicotinamide adenine dinucleotide phosphate
NCBI	National Center for Biotechnology Information
nr	banque de séquences protéiques du NCBI non-redondante
pb	paire de bases
PC	Phycocyanine
PE	Phycoérythrine
Pfam	Protein families database of alignments and HMMs
PSI-BLAST	Position-Specific Iterated-BLAST
PSI	Photosystème I
PSII	Photosystème II
ORF	Open Reading Frame
THG	Transfert horizontal de gènes
UV	Ultra Violet

CHAPITRE I

Introduction

Introduction

L'adaptation est une composante fondamentale de l'évolution des êtres vivants. La sélection de caractères avantageux permet aux organismes d'exploiter plus efficacement les ressources de la niche écologique dans laquelle ils vivent.

Au niveau phénotypique, de très nombreux exemples d'adaptation ont été mis en évidence. A l'inverse, l'adaptation est beaucoup plus difficilement identifiable au niveau moléculaire (Golding and Dean 1998). Seul l'emploi combiné de méthodes moléculaires, biochimiques, physiologiques et d'analyses phylogénétiques ou structurales peut permettre d'identifier des cas de substitutions d'acides aminés qui conduisent un changement adaptatif. Par exemple, le polymorphisme très élevé qui est observé dans les séquences de gènes est difficilement explicable par l'adaptation. Cette observation, associée à d'autres comme la constance du taux d'évolution moléculaire, a conduit à la formulation de la théorie de l'évolution moléculaire neutre de Motoo Kimura (Kimura 1983). Celle-ci stipule que la majorité des substitutions observées dans les séquences nucléotides sont neutres (elles n'apportent ni avantage, ni désavantage adaptatif) et sont dues essentiellement à la dérive génétique.

Une autre possibilité pour découvrir les bases moléculaires de l'adaptation consiste à identifier les gènes présents spécifiquement dans une lignée d'organismes adaptés à un environnement particulier. Chez les eucaryotes, l'absence de corrélation entre la complexité des organismes et le nombre de gènes qu'ils possèdent diminue fortement les possibilités offertes par cette approche. Ainsi, le génome humain ne contient que deux fois plus de gènes que celui de *Drosophila melanogaster* (Rubin 2001). Ce paradoxe s'explique par le fait que les gènes peuvent assurer plusieurs fonctions. Enfin, chez les eucaryotes, un caractère est souvent sous le contrôle de plusieurs gènes ce qui accentue la difficulté de relier le phénotype au génotype.

La situation semble, au contraire, beaucoup plus simple chez les procaryotes. La taille des génomes est beaucoup plus variable que chez les eucaryotes et le nombre de gènes est étroitement relié à la complexité physiologique de ces organismes. Ainsi, le jeu de gènes détermine directement le mode de vie et la capacité à utiliser les ressources d'une niche écologique donnée.

L'étude des génomes est donc tout à fait pertinente pour comprendre l'adaptation puisque qu'il s'agit de la cible privilégiée sur laquelle agit la sélection. De plus, le génome d'un organisme contient toute l'information nécessaire à son fonctionnement optimal et à sa reproduction dans une niche écologique particulière. La comparaison d'un nombre toujours plus important de génomes de bactéries et d'archées offre la possibilité de comprendre

comment le répertoire de gènes évolue pour permettre l'adaptation.

La génomique comparée a permis de révéler que les génomes procaryotes sont extrêmement plastiques. Ainsi, *Escherichia coli* a acquis autant de gènes qu'elle en a perdu depuis sa séparation avec le genre *Salmonella* (voir figure 4 dans Lawrence and Roth 1999). Le répertoire de gènes est le résultat d'une combinaison complexe de transmission verticale (d'ancêtre à descendants), de gains, par transferts horizontaux et duplications, et de pertes de gènes non-essentiels. La sélection joue un rôle déterminant puisqu'elle permet le maintien des gènes qui offrent un avantage adaptatif suffisant (Lawrence and Roth 1999).

Les cyanobactéries marines des genres *Prochlorococcus* et *Synechococcus* constituent un très bon modèle pour comprendre comment l'adaptation à des environnements différents se traduit au niveau génomique. Ces deux genres, très proches phylogénétiquement, abondent dans les écosystèmes océaniques. De plus, les études réalisées sur des souches en culture ont montré que celles-ci possèdent des caractéristiques écophysiologiques différentes qui leurs permettent de coloniser des niches écologiques distinctes. Enfin, les génomes de plusieurs souches de *Prochlorococcus* et de *Synechococcus* ont été entièrement séquencés offrant, ainsi, l'accès à une source considérable d'informations. L'annotation et la comparaison des génomes de *Prochlorococcus* et *Synechococcus* peuvent donc permettre d'identifier les gènes spécifiques d'une souche, d'un genre ou d'un écotype. L'analyse de ces génomes offre aussi la possibilité de déterminer les mécanismes évolutifs qui les ont façonnés. La connaissance des gènes responsables des différences adaptatives est essentielle pour comprendre les facteurs qui contrôlent l'abondance et la dynamique des populations naturelles de ces cyanobactéries.

I.1 Présentation des modèles d'étude: *Prochlorococcus* et *Synechococcus*

Dans les communautés phytoplanctoniques des écosystèmes océaniques, une part importante de la biomasse chlorophyllienne et de la production primaire (30 à 50%) est réalisée par deux genres de minuscules cyanobactéries unicellulaires: *Prochlorococcus* et *Synechococcus* (Partensky et al. 1999b; Waterbury et al. 1986).

Ces deux genres sont extrêmement proches et ont divergé relativement récemment à partir d'un ancêtre commun. En effet, le niveau d'identité entre les séquences du gène de l'ARNr 16S des souches de *Prochlorococcus* et *Synechococcus* indique que ces genres se sont différenciés il y a environ 150 à 200 millions d'années (Rappe and Giovannoni 2003) et voir chapitre IV). De plus, la faible résolution des limites entre ces deux groupes (indiquée par de faibles valeurs de "bootstrap" sur les branches basales des arbres phylogénétiques construits à partir du gène de l'ARNr 16S) suggère une diversification quasi-simultanée et relativement rapide (Urbach et al. 1992). Il est important de noter que le genre *Synechococcus sensu lato* est clairement polyphylétique comme le montre l'arbre de la figure I-1. Cependant, les *Synecho*coccus strictement marins (c'est-à-dire qui requièrent de fortes concentrations en Cl^-, Mg^{2+} et Ca^{2+} pour se développer) sont tous regroupés au sein d'un même clade. Celui-ci a d'abord été nommé "Marine Cluster A" (ou MC-A) avant d'être renommé "sous-cluster 5.1" (Fuller et al. 2003; Herdman et al. 2001). Pour des raisons pratiques, le terme *Synechococcus* ne sera plus employé, dans la suite de ce texte, que pour faire référence aux cyanobactéries marines du sous-cluster 5.1.

I.1.2 Distribution géographique et abondance.

Prochlorococcus et *Synechococcus* présentent des distributions géographiques qui se superposent partiellement. *Prochlorococcus* est essentiellement présent entre 40° de latitude Nord et Sud (Partensky et al. 1999b) tandis que *Synechococcus* est présent depuis l'équateur jusqu'aux régions polaires (Liu et al. 2002; Olson et al. 1990; Partensky et al. 1999a). *Prochlorococcus* est plus abondant dans les eaux oligotrophes intertropicales où il atteint typiquement des concentrations de 2 à 300 000 cellules par millilitre. *Synechococcus* domine dans les écosystèmes côtiers plus riches en sels nutritifs ou les abords de zones d'upwelling (zones mésotrophes à faiblement eutrophes). *Prochlorococcus* peut également se développer à une profondeur supérieure à celle de *Synechococcus* (Partensky et al. 1999a). En effet, *Prochlorococcus* est capable de vivre depuis la surface jusqu'à 150 à 200 m de profondeur, où moins de 1 % de la lumière arrivant en surface est disponible.

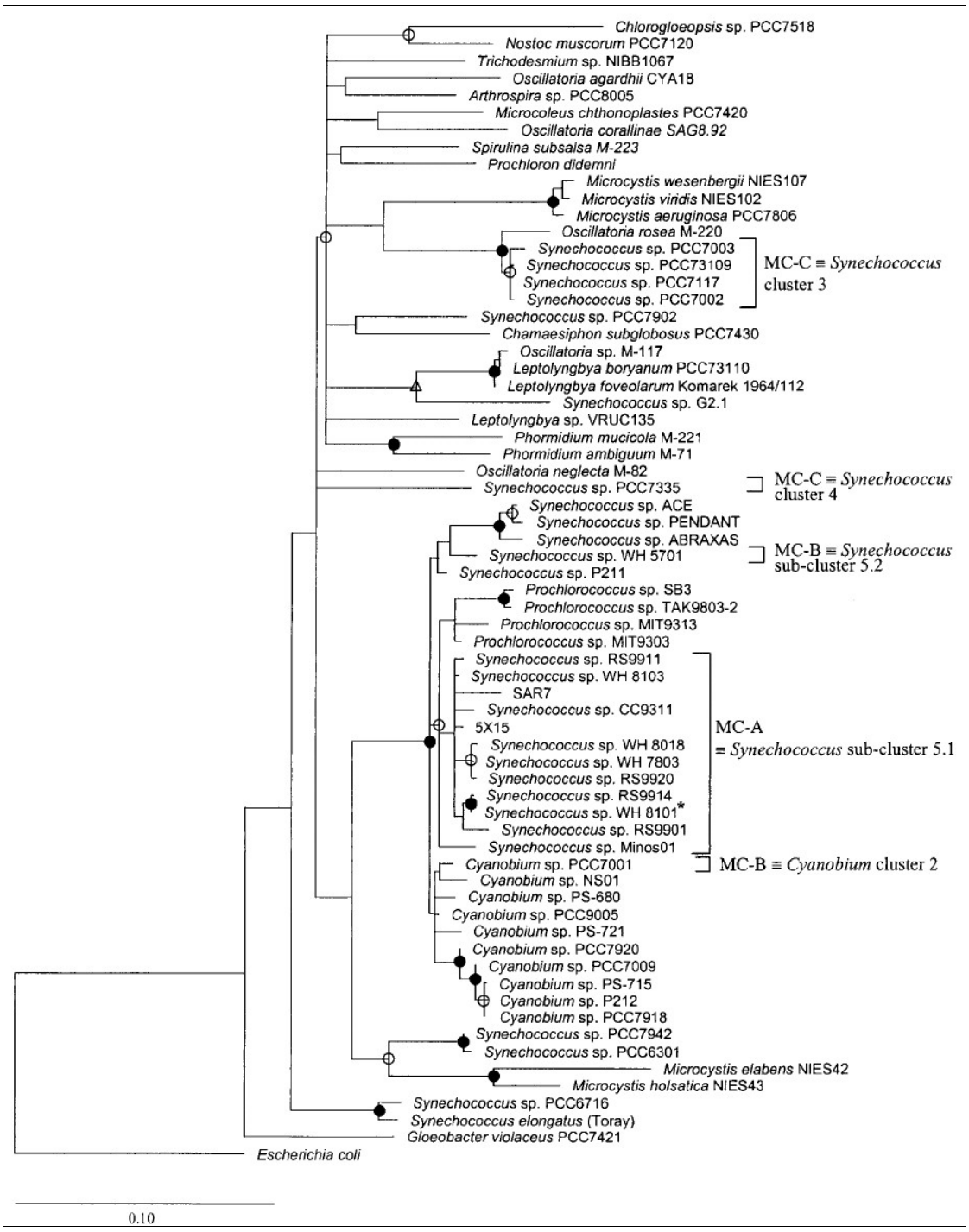

Figure I-1

Arbre phylogénétique des cyanobactéries montrant la position de *Prochlorococcus* et des *Synechococcus* marins du sous-cluster 5.1. Arbre basé sur la séquence du gène de l'ARNr 16S et construit avec la méthode du neighbor-joining. Les distances sont corrigées avec la méthodes de Juke et Cantor. *Escherichia coli* est utilisée comme groupe externe pour enraciné l'arbre. Ronds noirs, valeurs de bootstrap > 95; cercle, valeurs de bootstrap comprises entre 70 et 95. Les triangles indiquent l'utilisation de séquences partielles. D'après Fuller et al. 2003.

Les distributions géographiques et verticales très larges de *Prochlorococcus*, ainsi que les fortes densités de populations qu'il peut atteindre dans le milieu naturel font que cette cyanobactérie est probablement l'organisme photosynthétique le plus abondant sur terre (Partensky et al. 1999b).

I.1.3 Caractéristiques générales

Prochlorococcus et *Synechococcus* appartiennent tous les deux à la fraction picoplanctonique. Celle-ci regroupe les organismes du plancton dont la taille est inférieure à 2 µm. Les estimations de la taille et du contenu en carbone pour ces deux genres sont données dans le tableau I-1. La taille de *Prochlorococcus* en fait le plus petit organisme photosynthétique connu à l'heure actuelle. La souche *P. marinus* MED4 par exemple a une taille de 0,5-0,6 µm en diamètre et 0,7-0,8 µm en longueur (Rippka et al. 2000). Cela correspond à la plus petite taille théorique possible pour un organisme oxyphototrophe (Raven 1994).

Tableau I-1

Taille cellulaire et contenu en carbone estimés pour *Prochlorococcus* et *Synechococcus* (Partensky et al. 1999a).

	Prochlorococcus	***Synechococcus***
Taille cellulaire	0,5-0,8 µm	0,8-1,2 µm
Contenu en carbone	50-130 fg / cell	~250 fg / cel

Dans le milieu naturel, les cellules de *Prochlorococcus* voient leur taille augmenter avec la profondeur. Ainsi, dans la mer des Sargasses, les cellules passent d'une longueur de 0,45 à 0,75 µm depuis la surface jusqu'à 150 m de profondeur (Sieracki et al. 1995). La taille des cellules de *Synechococcus* est supérieure à celles de *Prochlorococcus*, mais reste très inférieure à la taille d'une cyanobactérie d'eau douce telle que *Synechocystis* PCC 6803 (2,3 à 2,5 µm).

Les deux genres de picocyanobactéries ont une organisation cellulaire très proche et similaire à celle des autres cyanobactéries (Fig. I-2). Cependant, les cellules de *Prochlorococcus* semblent avoir une forme un peu plus allongée que celle de *Synechococcus.* Chez ces deux genres, les thylacoïdes (membranes photosynthétiques) sont situés à la périphérie des cellules, parallèles à la membrane plasmique. Le cytoplasme contient des carboxysomes qui sont le lieu d'incorporation du CO_2 par la Rubisco.

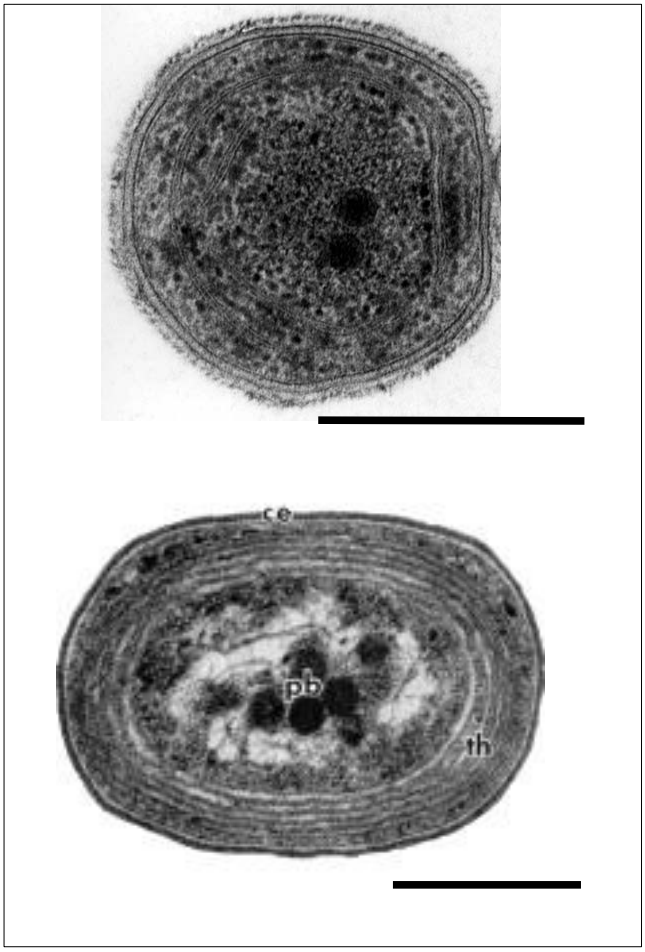

Figure I-2
Photographie en microscopie électronique à transmission de *Synechococcus* (en haut, B. Palenik) et de *Prochlorococcus* (en bas, Johnson et Sieburth, 1979). Echelle = 0.5 µm.

I.1.4 Pigmentation

Prochlorococcus est caractérisé par le présence de dérivés divinylés de la chlorophylle *a* (Chl a_2) et *b* (Chl b_2) (Goericke and Repeta 1992). Ces deux pigments qui sont spécifiques de *Prochlorococcus* permettent à ce dernier d'absorber plus efficacement les longueurs d'ondes situées dans la partie bleue du spectre visible (Morel et al. 1993). Ces longueurs d'ondes sont celles qui pénètrent le plus profondément dans les eaux oligotrophes des régions centrales océaniques. Au contraire, les eaux plus riches des régions côtières sont caractérisées par des longueurs d'ondes situées majoritairement dans la partie verte du spectre à cause des concentrations élevées en matière organique et des fortes densités des populations

phytoplanctoniques. Certaines souches de *Prochlorococcus*, qui ont été isolées en profondeur, sont également capables de synthétiser de la Chl *b* lorsqu'elles sont exposées à de fortes intensités lumineuses probablement par conversion enzymatique d'une partie du pool de divinyle Chl *b* (Moore et al. 1995; Partensky et al. 1993). En plus de ces deux formes majoritaires de Chl (*a* et *b*), *Prochlorococcus* possède également de petite quantité de Chl *c*, ainsi que plusieurs caroténoïdes dont les plus abondants sont la zéaxanthine et l'*α-carotène.*
La plupart des *Synechococcus* marins contiennent de la phycoérythrine comme pigment photosynthétique majeur. Celle-ci confère une coloration orangée aux cellules de *Synechococcus,* ce qui les rend aisément reconnaissables en microscopie à fluorescence.

I.1.5 Appareil photosynthétique

L'appareil photosynthétique de *Prochlorococcus* et de *Synechococcus* se distingue essentiellement par le système utilisé pour collecter l'énergie lumineuse des photons. Chez *Synechococcus,* comme chez la majorité des cyanobactéries, l'antenne photocollectrice est un très gros complexe multiprotéique extrinsèque appelé phycobilisome (Fig. I-3A). Ce complexe est classiquement composé d'un coeur d'allophycocyanine (APC) sur lequel sont fixés des bras formés de phycocyanine (PC). Cependant, la partie la plus distale des bras est formée non pas d'une mais de deux formes distinctes de phycoérythrine (PE I et II) (Ong and Glazer 1991). Cette caractéristique semble être spécifique des *Synechococcus* marins.

Les différentes phycobiliprotéines sont reliées entre elles par des protéines « linker » et fixent des chromophores (pigments solubles) appelés phycobilines. Chez les souches WH8102 et WH8020 de *Synechococcus*, environ 40 gènes, organisés en opérons ou en groupes de gènes (clusters) plus larges, sont requis pour la formation du phycobilisome. Le plus grand de ces groupes (~ 15 kb) est formé d'une vingtaine de gènes nécessaires à la formation de l'APC, de la PC et des deux formes de PE ainsi qu'à la synthèse et à l'attachement de phycobilines (Wilbanks and Glazer 1993).

Chez *Prochlorococcus*, l'utilisation du phycobilisome comme antenne photocollectrice principale a été abandonnée au cours de l'évolution. Cela se traduit par la disparition progressive des gènes nécessaires à la formation de ce complexe et à la synthèse des phycobilines. Chez les souches de profondeur (*P. marinus* SS120 et *Prochlorococcus* sp. MIT9313), le principal cluster observé chez *Synechococcus* est réduit aux seuls gènes codant pour les sous-unités α et β d'une troisième forme de PE (PE III) et pour les enzymes servant à la synthèse et à la fixation du chromophore (phycourobiline) sur la PE (Hess et al. 2001; Hess et al. 1999) (Fig. I-4).

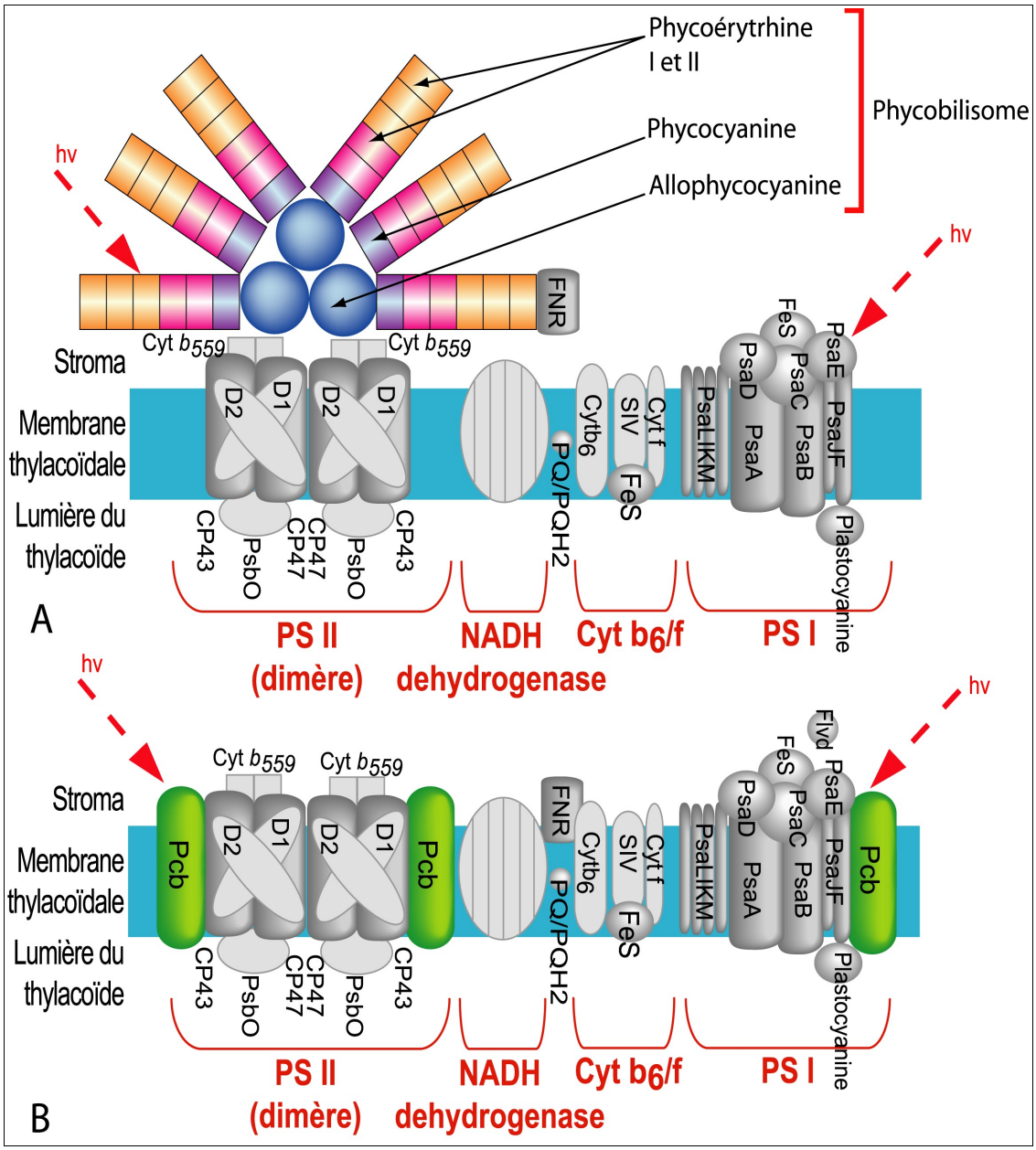

Figure I-3

Schéma simplifié de l'appareil photosynthétique de (A) *Synechococcus* sp. WH8102 et de (B) *Prochlorococcus marinus* SS120. Un seul monomère est représenté pour le PSI au lieu de trois normalement. PSI, photosystème I; PSII, photosystème II; Cyt, cytochrome; PQ, Plastoquinone; FNR, Ferredoxine:$NADP^+$ oxidoreductase; Flvd, Flavodoxine.

La préservation des gènes codant pour la PE chez ces souches est assez surprenante puisque les gènes codant pour la partie la plus interne du phycobilisome (APC et PC) sont tous absents chez *Prochlorococcus* (Hess et al. 2001). Une possibilité est que la PE soit utilisée comme antenne photocollectrice secondaire chez les souches vivant en profondeur. Cette hypothèse est renforcée par le fait que cette phycobiliprotéine fixe comme chromophore

la phycourobiline. En effet, son spectre d'absorption à 495 nm correspond à la longueur d'onde qui pénètre le mieux dans la colonne d'eau au delà de 100 m de profondeur. Cependant, le gène codant pour la seule protéine « linker » de PE chez *P. marinus* SS120 (*ppeC*) est absent du génome de *Prochlorococcus* sp. MIT9313. Or l'absence de ce gène empêche l'association des deux sous-unités de la PE entre elles et donc son fonctionnement en tant que système photocollecteur (Hess et al. 1999; Lokstein et al. 1999).

Chez la souche de surface *P. marinus* MED4, le cluster ne contient plus que deux gènes dont *cpeB* (sous-unité β de la PE) (Hess et al. 2001) (Fig. I-4). La séquence de ce dernier est caractérisée par un taux élevé de mutations et par le remplacement de deux cystéines (sur quatre) servant de sites de fixation pour le chromophore (Ting et al. 2001).

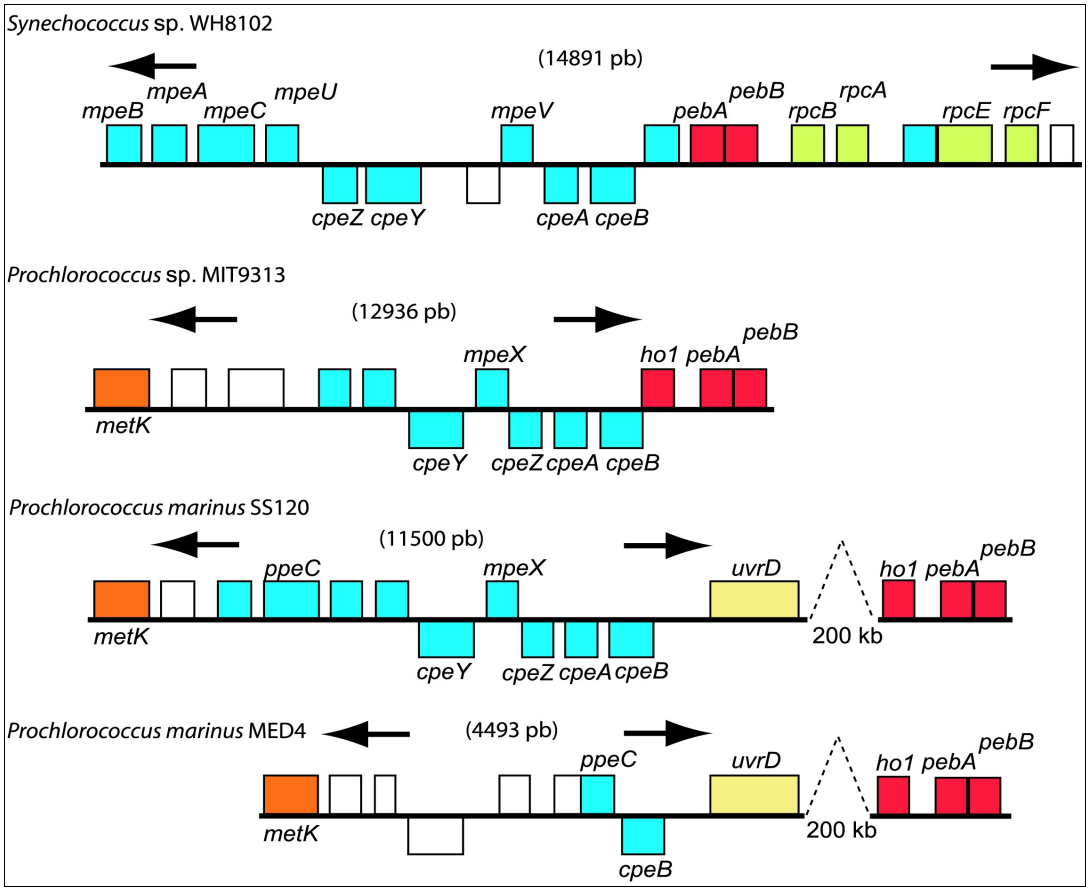

Figure I-4

Réduction progressive du cluster phycoérythrine/phycocyanine chez le genre *Prochlorococcus*. Les gènes codant pour la PE sont en bleu, ceux codant pour la phycocyanine sont en vert. Les gènes en rouge, *ho1* (hème oxygenase), *pebA* (15,16-dihydrobiliverdin:ferredoxin oxidoreductase) et *pebB (*phycoérythrobilin: ferredoxin oxidoreductase), participent à la synthèse des phycobilines. Ces gènes sont situés à environ 200 kb du reste du cluster chez *P. marinus* SS120 et *P. marinus* MED4. La taille du cluster est indiquée entre parenthèse. Adapté de Hess et al. 2001.

Il a été montré que ce gène est exprimé chez *P. marinus* MED4 (Hess et al. 2001). De plus, le rapport du taux de substitutions non-synonymes sur le taux de substitutions synonymes n'est pas significativement plus élevé chez *P. marinus* MED4 que chez les autres souches de *Prochlorococcus* (Ting et al. 2001). Ces observations laissent penser que le gène *cpeB* évolue encore sous l'influence de la sélection purifiante mais que sa fonction est différente chez *P. marinus* MED4. Il a été suggéré que la phycoerythrine β pourrait servir de photorécepteur chez les souches de haute lumière, néanmoins, cette hypothèse reste à valider expérimentalement (Steglich 2003).

La fonction du phycobilisome est assurée par la protéine Pcb (Fig. I-3B) chez *Prochlorococcus*. Cette dernière est une protéine membranaire à 6 hélices α (LaRoche et al. 1996) qui fixe la majorité de la Chl b_2. Elle appartient à la même famille que la protéine IsiA qui fixe la chlorophylle *a* et joue le rôle d'antenne photocollectrice lors de carence en fer chez les cyanobactéries d'eau douce. Cette protéine rappelle aussi, de part sa structure, l'antenne des plantes supérieures (protéines Lhc pour light-harvesting complex) mais ces deux familles de protéines semblent avoir des origines différentes (LaRoche et al. 1996).

Le remplacement du phycobilisome pourrait être le résultat d'une sélection pour un investissement moins lourd en acides aminés par molécule de tétrapyrolle (chlorophylle ou phycobiline) fixé sur l'antenne photocollectrice (Ting et al. 2002). En effet, cet investissement est de 4,3 kDa par tetrapyrolle dans le cas de la protéine Pcb (39 kDa / 9 tetrapyrolles fixés) alors qu'il est de 7,5 kDa dans le cas de la PE I et de 6,5 kDa dans le cas de la PE II (Ting et al. 2002). Ceci indique que la quantité d'azote qui doit être investie par la cellule dans l'antenne photosynthétique est 1,5 à 1,7 fois plus grande dans le cas du phycobilisome que dans le cas de la protéine Pcb (Ting et al. 2002). L'économie réalisée constituerait un avantage important dans le cas d'une cyanobactérie vivant dans un milieu très pauvre en azote.

I.1.6 Diversité génétique et écotypes

Les analyses phylogénétiques réalisées à partir des séquences de souches en culture mais aussi de séquences provenant directement du milieu naturel ont montré que, chez *Prochlorococcus*, la diversité génétique n'est pas corrélée à la distance géographique. Ainsi, la diversité est maximale non pas entre deux populations de surface éloignées géographiquement mais entre les populations de surface et celles de profondeur d'une même zone géographique. Le fort gradient de lumière et de sels nutritifs entre la surface et la base de la couche euphotique apparaît donc comme un facteur déterminant pour la spéciation chez *Prochlorococcus*.

Les souches de profondeur de *Prochlorococcus* présentent des rapports Chl b_2 sur Chl a_2 nettement plus élevés (0,4 à 2,5) que les souches de surface (0,1 à 0,6) (Moore and Chisholm 1999; Moore et al. 1995; Partensky et al. 1993). Les souches de surface et de profondeur se distinguent également par leurs éclairements optimaux de croissance (Moore and Chisholm 1999; Partensky et al. 1993) (Fig. I-5). Ainsi, les souches de profondeur sont capables de croître à des intensités lumineuses beaucoup plus faibles que celles de surface. A l'inverse, elles sont incapables de se développer sous les fortes lumières qui conviennent parfaitement aux souches de surface (Moore and Chisholm 1999).

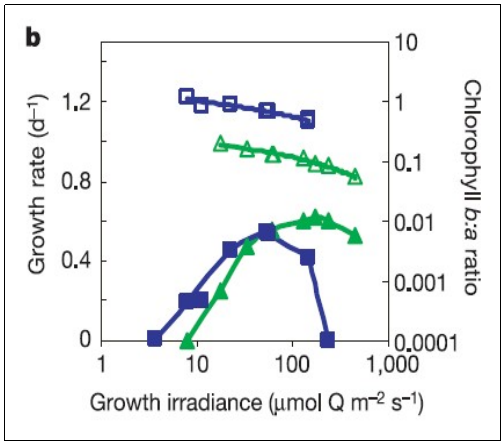

Figure I-5
Taux de croissance (symboles pleins) et rapport Chl b_2 / Chl a_2 (symboles vides) en fonction de l'intensité lumineuse de croissance. Courbes vertes, *P. marinus* MED4 (souche de surface); courbes bleues, *Prochlorococcus* sp. MIT9313 (souche de profondeur). D'après Rocap et al. 2003.

Ces différences physiologiques qui reflètent l'adaptation aux conditions environnementales sont corrélées avec les différences génétiques comme le montrent les arbres phylogénétiques (Moore et al. 1998; Rocap et al. 2003; Urbach et al. 1998) (Fig. I-6). En effet, les souches de surface forment un clade monophylétique (divisé en deux sous-clades). Par contre, la diversité est bien plus élevée entre les souches de profondeur, ces dernières formant plusieurs clades séparés. Cela suggère qu'il existe au moins deux écotypes adaptés à des niches écologiques différentes chez *Prochlorococcus*, un écotype de surface (aussi appelé écotype de forte lumière) et un (ou plusieurs) écotype(s) de profondeur (écotype de faible lumière) (Moore and Chisholm 1999; Moore et al. 1998). De part leur position dans les arbres phylogénétiques, les souches de surfaces sont probablement les plus récentes.

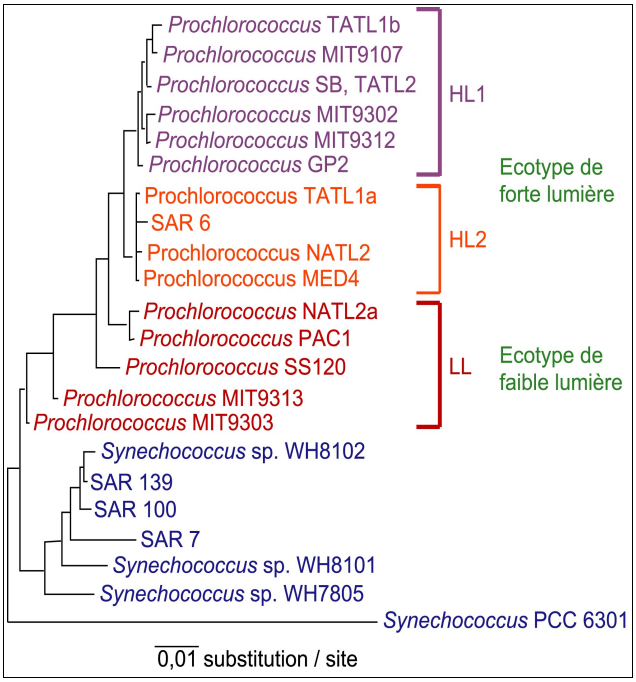

Figure I-6

Arbre basé sur la séquence du gène de l'ARNr 16S de souches de *Prochlorococcus*, de *Synechococcus* et de séquences environnementales (SAR, mer des Sargasses). Arbre construit avec la méthode du maximum de vraissemblance. La cyanobactérie d'eau douce *Synechococcus* PCC 6301 sert de groupe externe pour enraciner l'arbre. HL1, High-light-adapted clade 1; HL2, High-light-adapted clade 2; LL, Low-light-adapted clade. D'après Moore et al. 1998

L'existence de deux écotypes pourrait expliquer la présence de *Prochlorococcus* depuis la surface jusqu'à la base de la couche euphotique (West and Scanlan 1999). Aucune souche cultivée au laboratoire n'est capable de supporter des variations de plus de 2,5 ordres de magnitude de lumière de culture (Moore et al. 1995). Plusieurs études ont permis d'identifier certaines des bases moléculaires expliquant l'adaptation à la niche écologique de surface ou à celle de profondeur. Garczarek et collaborateurs ont montré la présence de plusieurs copies différentes du gène *pcb* (de 2 à 8) chez les souches de profondeur (Garczarek et al. 2000). La plupart des souches de surface ne contiennent qu'une seule copie de ce gène même si deux copies ont été trouvées récemment chez la souche MIT9312 (Claudia Steglich, communication personnelle). La multiplication de ces gènes chez les souches de profondeur permettrait d'augmenter leurs capacités à réaliser la photosynthèse à des profondeurs où la quantité de lumière incidente est très faible.

Une étude plus récente a permis de préciser le rôle des différents gènes *pcb.* Ainsi, la duplication successive d'un gène homologue au gène *isiA,* au cours de l'évolution de genre *Prochlorococcus*, a conduit à la formation de la famille des gènes *pcb* avec la spécialisation des membres de cette famille comme antenne du photosystème I (uniquement chez *P. marinus* SS120) ou du photosystème II (chez *P. marinus* MED4, *P. marinus* SS120 et *Prochlorococcus* sp. MIT9313) (Bibby et al. 2003). De plus, chez *Prochlorococcus* sp. MIT9313 et chez *P. marinus* SS120, un des gènes *pcb* a conservé une fonction homologue à celle d'*isiA*. En effet, chez ces deux souches, le produit de ce gène vient former un anneau de 18 protéines autour des trimères du PSI lorsque les cellules sont exposées à une carence en fer (Bibby et al. 2003).

La situation observée chez *Synechococcus* est beaucoup plus complexe. La diversité génétique est bien plus élevée que chez *Prochlorococcus* et le sous-cluster 5.1 qui regroupe les *Synechococcus* marins contient (au moins) dix clades phylogénétiques différents (Fig. I-7). Contrairement à *Prochlorococcus*, aucune relation nette entre les caractéristiques physiologiques et la diversité génétique n'a pu être mise en évidence. A l'intérieur d'un même clade, certaines souches peuvent utiliser les nitrates comme source d'azote alors que d'autres avec la même séquence d'ARNr 16S en sont incapables (Fuller et al. 2003). De même, aucune différenciation écotypique liée à la lumière n'a été identifiée jusqu'ici (Ferris and Palenik 1998; Fuller et al. 2003). Toutes les souches peuvent être adaptées, en culture, à des intensités lumineuses élevées correspondant à celles que reçoivent les cyanobactéries en surface (Six et al. 2004).

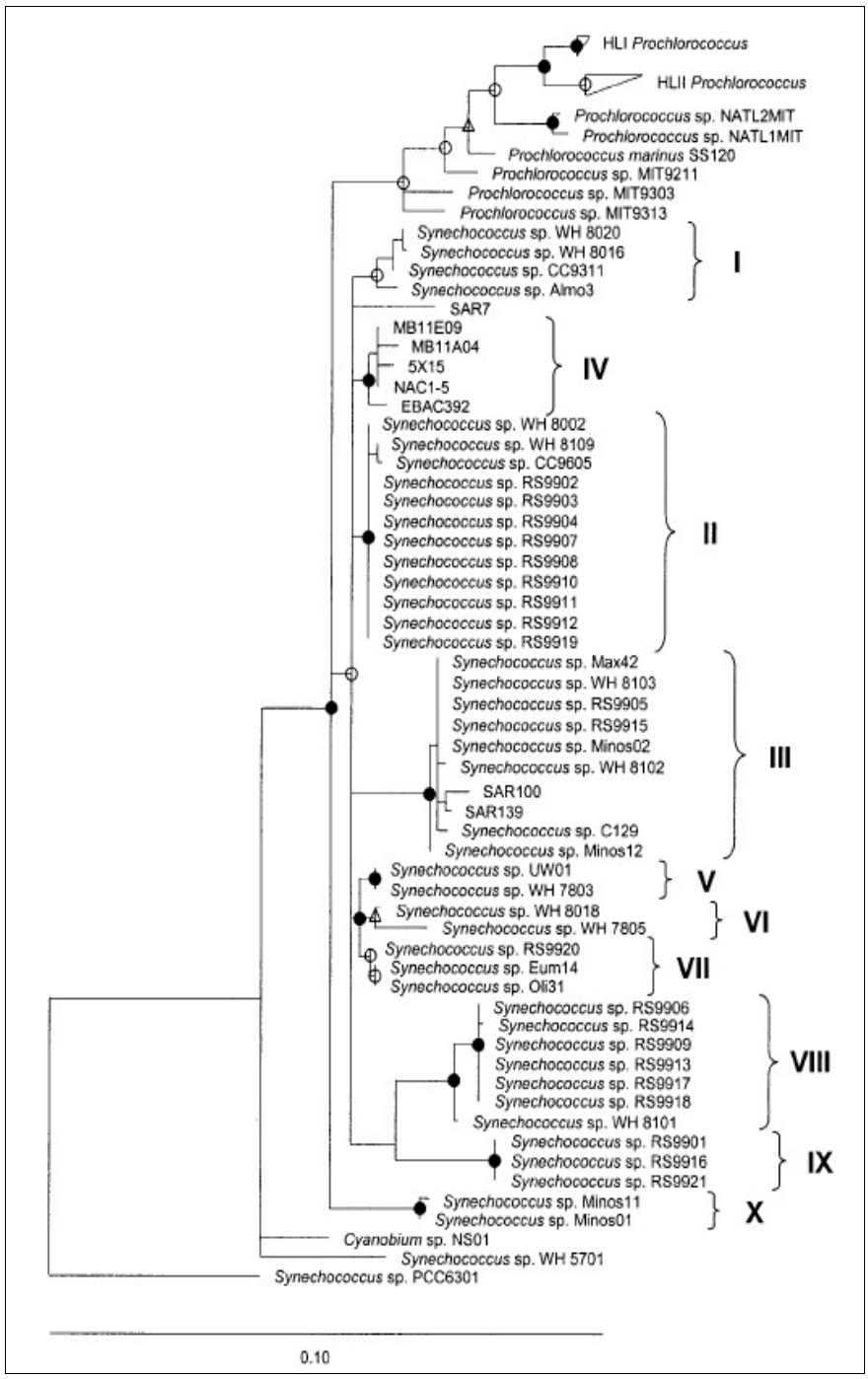

Figure I-7

Arbre phylogénétique montrant la diversité génétique existant au sein du sous-cluster 5.1 de *Synechococcus*. Arbre basé sur la séquence du gène de l'ARNr 16S et obtenu par la méthode du neighbor-joining avec la méthode de correction de Juke et Cantor. Ronds noirs, valeurs de bootstrap > 95; cercle, valeurs de bootstrap comprises entre 70 et 95. Les triangles indiquent l'utilisation de séquences partielles. HLI, High- light-adapted clade 1; HLII, high-light-adapted clade II. D'après Fuller et al. 2003.

I.3 Evolution des génomes de procaryotes

I.3.1 Intérêts de la comparaison de génome complets

La disponibilité de génomes complets d'organismes procaryotes et eucaryotes constitue une source d'information sans précédent pour la compréhension de la biologie de ces organismes. Cependant, ces énormes masses de données n'ont réellement d'intérêt qu'une fois comparées entre elles.

La comparaison du répertoire de gènes de génomes différents est basé sur l'identification des gènes homologues. Par définition, ces gènes dérivent de la séquence d'un même gène. Il est important de différencier ces deux types d'homologie que sont l'orthologie et la paralogie (Fitch 1970). Les gènes orthologues se forment par transmission verticale à la descendance (Fig. I-8).

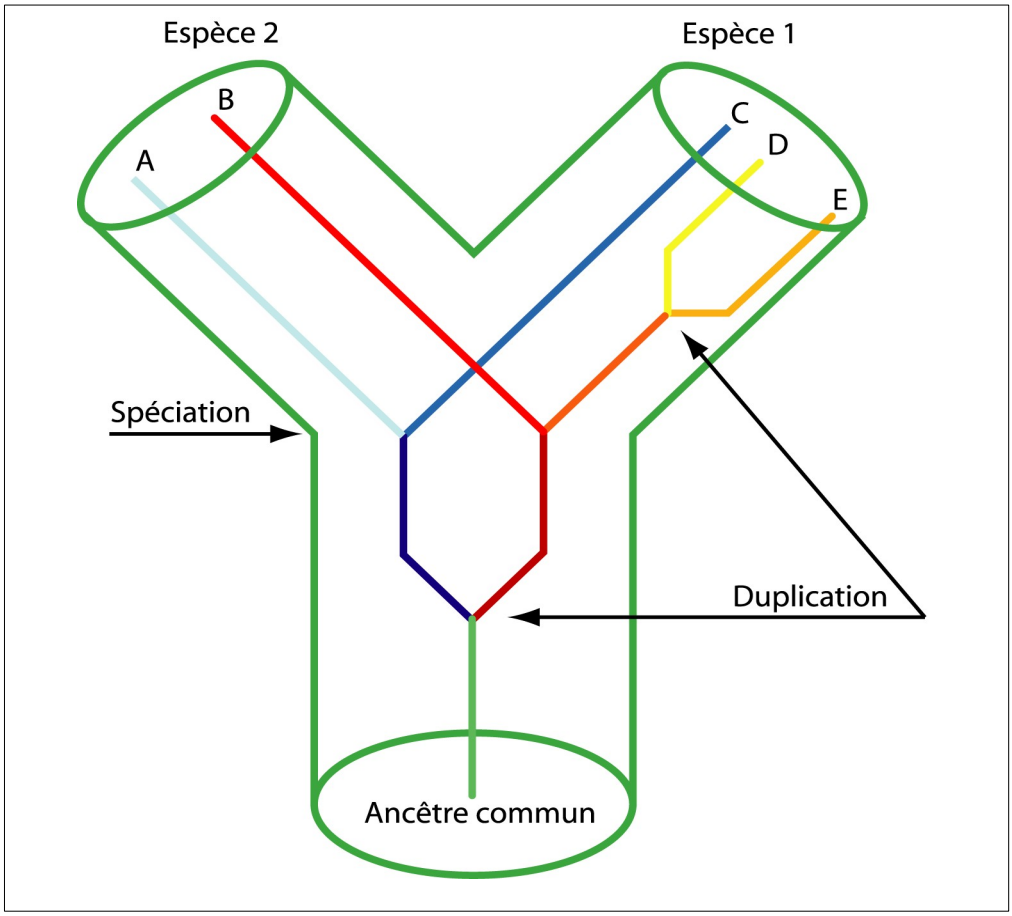

Figure I-8

Exemple de relation d'orthologie et de paralogie. A est orthologue avec C, B est orthologue avec D et E. De même, A est paralogue avec B (out-paralogue), D est paralogue avec E (in-paralogue) et C est paralogue avec D et E.

Ainsi lors d'un événement de spéciation, une copie de chaque gène du génome d'une espèce « parent » se retrouve dans le génome de chaque espèce « fille » issue de cette dernière. En l'absence de duplication et tant qu'ils demeurent essentiels, les gènes orthologues conservent la même fonction. L'évolution des gènes orthologues reflètent la phylogénie des organismes et par conséquent, l'identification de ces gènes est primordiale pour la validité des analyses phylogénétiques.

A l'inverse, les gènes paralogues sont issus de la duplication d'une séquence de gène au sein d'un génome (Fig. I-8). Contrairement aux gènes orthologues, ces gènes peuvent diverger suffisamment pour acquérir de nouvelles fonctions (voir § Duplication génique et formation de familles multigéniques).

On peut distinguer également deux types de gènes paralogues (Fig. I-8). La formation de gènes paralogues peut avoir eu lieu dans le génome de l'ancêtre commun (avant la spéciation) à un groupes d'espèces. Les différentes copies créées (out-paralogues) sont transmises à tous les descendants. Inversement, la formation de gènes paralogues peut aussi intervenir après la différentiation en espèces séparées. Dans ce cas, les gènes paralogues (in-paralogues) ne sont présents que dans le génome d'une espèce. La distinction entre in-paralogues et out-paralogues n'est pas fixe mais dépend entièrement du niveau taxonomique auquel on se place.

I.3.1.1 Génome minimal et remplacement non-orthologue

L'identification du génome minimal est l'une des possibilités les plus intéressantes offertes par la comparaison de génomes. Elle consiste à déterminer la nature et le nombre de gènes essentiels qui seraient suffisants pour permettre le fonctionnement autonome d'un organisme le plus simple possible et présentant des caractéristiques modernes (Mushegian 1999). En d'autres termes, il s'agit de trouver les gènes qui constituent les bases mêmes de la vie.

Le concept de génome minimal, et *a fortiori* de gène essentiel, ne se comprend qu'en fonction des caractéristiques du milieu au sein duquel un organisme théorique possédant un tel génome se développerait. Ainsi, il semble évident que la majorité des gènes ne sont essentiels que dans certaines conditions de milieu bien définies. Dans la plupart des cas, le génome minimal a été déterminé en partant de l'hypothèse d'un organisme se développant dans un milieu stable, riche en éléments nutritifs et sans compétition. Ces conditions sont logiquement les seules qui permettraient le fonctionnement d'un organisme réellement minimal. Elles se rencontrent, aujourd'hui, essentiellement dans le cas des bactéries pathogènes, parasites et symbiotiques qui vivent en association plus ou moins obligatoire avec d'autres organismes généralement eucaryotes. Il est intéressant de noter que les plus petits

génomes connus appartiennent à des bactéries pathogènes ou symbiotiques. Néanmoins, même s'ils s'en approchent, ces génomes ont une taille bien supérieure à celle d'un génome réellement minimal.

En comparant les deux premiers génomes disponibles, ceux de la bactérie Gram-positive *Haemophilus influenza* (Fleischmann et al. 1995) et de la bactérie Gram-négative *Mycoplasma génitalium* (Fraser et al. 1995), Mushegian et Koonin (Mushegian and Koonin 1996) ont estimé que le génome minimal devait contenir 256 gènes. Ces deux organismes pathogènes appartiennent à deux lignées de bactéries ayant divergé il y a environ 1,6 milliard d'années et ont tous les deux subit un processus d'évolution réductive. Les deux auteurs ont donc émis l'hypothèse que les gènes orthologues conservés dans ces deux génomes étaient probablement essentiels et devaient être inclus dans le génome minimal. Le génome défini par Mushegian et Koonin, bien que basé sur la comparaison de deux génomes bactériens, contenait en majorité des gènes ayant des homologues à la fois chez les eucaryotes et les archées. Ainsi, il constituait une approximation correcte du génome minimal pour les trois domaines du vivant. Dans un but similaire à celui des deux auteurs précédents, Gil et collaborateurs ont comparé les génomes de cinq bactéries endocellulaires vivant en symbiose avec différents insectes (Gil et al. 2003). Ces bactéries ont perdu un nombre très important de gènes au cours de leur co-évolution avec leurs hôtes et possèdent toutes des génomes inférieurs à un Mb. Ces comparaisons ont abouti à l'identification de 277 gènes codant pour des protéines et de 36 gènes spécifiant des ARNs, soit 313 gènes orthologues communs aux cinq génomes. Si l'on ne prend en compte que les gènes codant pour des protéines, ce nombre est assez proche de celui estimé par Mushegian et Koonin.

Différentes approches expérimentales ont également été utilisées dans le but d'identifier les gènes essentiels formant le génome minimal. Une des premières tentatives fut réalisée par Itaya (Itaya 1995). Ce dernier a construit aléatoirement 79 mutants "knockout" chez *Bacillus subtilis*. Seulement 6 se révélèrent létaux, et par extrapolation, l'auteur conclut que le génome de *B. subtilis* contenait environ 300 gènes essentiels. Ce résultat, à nouveau très proche de celui de Mushegian et Koonin, est d'autant plus remarquable qu'au moment de cette étude, le génome de *B. subtilis* n'était pas encore disponible. Par la suite, de nombreuses techniques, telles que l'inactivation par insertion de transposons ou l'utilisation d'ARN anti-sens, ont été employées sur différents organismes procaryotes ou eucaryotes et ont donné des estimations assez différentes du nombre de gènes essentiels (Tableau I-2).

Le répertoire de gènes essentiels identifiés avec les deux types d'approches (*in-silico* ou expérimentale) contient surtout des gènes impliqués dans les mécanismes de réplication de transcription et de traduction. Les gènes codant pour les enzymes des voies métaboliques sont logiquement peu représentés étant donné la richesse en composés organiques du milieu considéré. Enfin, très peu de gènes ont une fonction inconnue.

Tableau I-2

Nombre et pourcentage de gènes de gènes essentiels identifiées expérimentalement dans cinq génomes procaryotes et deux génomes eucaryotes. D'après Koonin 2003.

Organisme	***Nb de gènes codant pour des protéines***	***Nb de gènes analysés***	***Nb de gènes essentiels***	***Nb de gènes essentiels extrapolé***	***Méthode d'inactivation***	***Référence***
M. gentialium / M. pneumoniae	480	480	351	265-380 (55-79%)	Insertion de transposons	Hutchison et al. 1999
B. subtilis	4118	3613	192	271 (6,6 %)	Insertion de plasmides	Kobayashi et al. 2003
H influenzae	1714	1272	478	670 (38 %)	ARN anti-sens	Akerley et al. 2002
E. coli	4275	3746	620	708 (17 %)	Insertion de transposons	Gerdes et al. 2003
S. cerevisiae	~ 6000	5916	1105	1124 (19 %)	Deletion par recombinaison mitotique	Giaever et al. 2002
C. elegans	~ 20000	16757	929	1080 (5,4 %)	RNA interference	Kamath et al. 2003

Les deux approches présentent des biais responsables d'une mésestimation du nombre de gènes essentiels:

- l'approche expérimentale entraîne une surestimation du nombre de gènes essentiels puisqu'elle ne prend en compte que les gènes dont l'inactivation se révèle létale mais pas ceux dont l'inactivation ralentit la croissance, donc diminue la valeur sélective du mutant. De plus, certaines mutations ne sont létales que dans le cas de mutations multiples sur d'autres gènes (mutations synthétiques) et ne sont pas considérées dans ce type d'études.

- les estimations obtenues par comparaison de génomes sont aussi probablement sous-estimées puisque, par définition, ce type d'approche ne prend en compte que les gènes orthologues et est fortement dépendante du degré de conservation de ces gènes. Si le niveau de similarité est trop faible, par exemple dans le cas de gènes ayant évolué très rapidement, la relation d'orthologie risque de ne pas être détectée et ces gènes ne seront pas inclus dans le jeu minimal de gènes. Cependant l'une des principales difficultés rencontrées avec cette approche vient du fait qu'une même fonction biologique peut être réalisée par des protéines non-orthologues (non-orthologous gene displacement), voire non-homologues, c'est-à-dire ne présentant aucune similarité de séquence ni de structure (Koonin et al. 1996).

Ainsi parmi les 256 gènes essentiels proposés par Mushegian et Koonin, seuls 240 correspondaient à des orthologues véritables entre les deux génomes. En effet, l'analyse des fonctions de ces orthologues a montré qu'ils étaient insuffisants pour assurer toutes les fonctions biologiques nécessaires à un organisme. Les deux auteurs ont identifié plusieurs cas de remplacement non-orthologue et ont ajouté 16 gènes aux 240 précédemment identifiés afin d'obtenir un jeu de gènes suffisant pour le métabolisme d'une cellule moderne. De même, seuls 179 gènes identifiés par Gil et collaborateurs avaient un orthologue dans le génome de *Mycoplasma genitalium.* Il est assez probable que ces différences proviennent également de la présence dans ces génomes de gènes différents réalisant des fonctions identiques. Le nombre de gènes orthologues tombe à 156 lorsque les génomes de *Rickettsia Prowazekii* et de *Chlamidya trachomotis* sont inclus dans l'analyse (Klasson and Andersson 2004).

Ainsi, plus le nombre de génomes utilisés dans les comparaisons augmente et plus le nombre de gènes orthologues communs diminue. On estime que le nombre de gènes orthologues universellement conservés dans tous les génomes est inférieur à 80 gènes (Koonin 2000). Ce nombre est très inférieur au nombre estimé de fonctions essentielles et il semble maintenant évident que, dans de nombreux cas, plusieurs solutions ont été trouvées indépendamment au cours de l'évolution pour réaliser une même fonction (Galperin et al. 1998). Ainsi, il semble plus intéressant d'envisager le concept de génome et d'organisme minimal en terme de fonctions primordiales et non pas par rapport à la présence ou à l'absence de gènes essentiels. La recherche du jeu de gènes essentiels dans les génomes de différents organismes devrait offrir la possibilité de révéler la diversité des mécanismes inventés pour réaliser des fonctions identiques.

I.3.1.2 Adaptation à la niche écologique

L'étude du génome du point de vue de l'adaptation à la niche écologique apporte de nouveaux éléments permettant de comprendre comment le génome évolue en fonction de l'environnement et des pressions de sélection qu'il exerce. Cela est d'autant plus pertinent que le génome constitue probablement le principal niveau de sélection.

L'adaptation à la niche écologique a été étudiée, notamment, en déterminant la composition en acides aminés de l'ensemble des protéines de procaryotes vivant dans des environnements caractérisés par des paramètres physico-chimiques (pH, température, salinité) différents. Ainsi, il semble que la composition en acides aminés des protéines soit directement liée aux conditions environnementales (Dumontier et al. 2002; Kawashima et al. 2000; Kreil and Ouzounis 2001).

Chez l'archée halophile *Halobacterium* sp. NRC-1, le protéome est très acide (point isoélectrique moyen de 4,9) et ne contient quasiment aucune protéine basique. La comparaison de la structure de plusieurs protéines d'*Halobacterium* avec celles d'organismes non-halophiles comme *Escherichia coli* révèle également qu'un plus grand nombre d'acides aminés acides sont localisés à la surface des protéines de cette archée. Ce biais dans la composition de l'ensemble des protéines d'*Halobacterium* résulte d'une adaptation à un milieu très riche en sel. En effet, *Halobacterium*, accumule le potassium à haute concentration dans son cytoplasme pour équilibrer la pression osmotique avec le milieu extérieur. L'acidification du protéome permet de maintenir la solubilité des protéines dans le milieu interne hypersalé.

Un autre aspect de l'étude de l'adaptation concerne l'identification des gènes qui sont présents dans un génome et absents d'un autre. Cette génomique soustractive (par analogie à l'hybridation soustractive) fournit un outil efficace pour déterminer les bases moléculaires qui peuvent expliquer les différences de phénotypes entre organismes.

Cette approche a surtout été utilisée pour les bactéries pathogènes d'humains et d'animaux (Edwards et al. 2002; Himmelreich et al. 1997; Janssen et al. 2001; Shirai et al. 2000) et de plantes (da Silva et al. 2002) afin d'identifier les gènes de virulences ou encore ceux définissant la spécificité pour une gamme d'hôtes plus ou moins large.

Dans le cas des procaryotes ne dépendant pas d'un hôte, la comparaison des répertoires de gènes a été utilisée pour déterminer les gènes permettant l'adaptation aux fortes valeurs de pH. La comparaison du génome *d'Oceanobacillus iheyensis* (bactérie halotolérante et alcalinophile) avec ceux de *Bacillus halodurans* (alcalinophile et modérément halotolérante) et de *Bacillus subtilis* (neutrophile) a permis l'identification de 243 gènes orthologues spécifiques des deux souches alcalinophiles. Les auteurs ont pu identifier une fonction pour environ la moitié de ces gènes. Plusieurs codent pour des transporteurs ABC qui servent à importer des acides aminés aliphatique (valine, leucine, isoleucine) et des oligopeptides. Ces acides aminés pourraient servir à la synthèse de L-glutamate. Ce composé est chargé négativement à pH élevé et pourrait être utilisé pour maintenir le cytoplasme aux alentours de 8-8,5 (chez *O. iheyensis*) bien que le pH du milieu extracellulaire soit plus élevé (~10,5). De même, un grand nombre de gènes sont impliqués dans l'importation de Na^+ (Na^+ / solute symporteur, canal sodium voltage-dépendant). Le cycle du sodium est critique pour le maintien du pH intracellulaire notamment grâce aux antiporteurs Na^+ / H^+ qui permettent l'accumulation de protons aux dépens des ions sodium. Les deux génomes de souches alcalinophiles contiennent aussi un gène servant à la synthèse de teichunoropeptide (polymère de polyglutamate et d'acide polyglucuronique) qui est un composant essentiel de la paroi cellulaire de *Bacillus halodurans* et qui participe à la régulation du pH chez cette souche.

La comparaison de génomes présentent, néanmoins, plusieurs limitations qui réduisent la possibilité de déterminer les gènes de niches. Les génomes doivent être suffisamment

proches pour pouvoir identifier de manière certaine les gènes orthologues et pour être sûr que les différences observées dans le jeu de gènes correspondent bien à des différences d'adaptation et ne sont pas simplement le résultat de la dérive génétique. Un autre problème vient, comme pour la définition du génome minimal, des remplacements non-orthologues. L'absence d'un gène dans un génome ne veut pas dire forcément que la fonction biologique correspondante est absente.

I.3.2 Mécanismes d'évolution du répertoire de gènes

I.3.2.1 Duplication génique et formation de familles multigéniques

Bien qu'extrêmement compacts, les génomes procaryotes contiennent une information génétique assez redondante. Les gènes paralogues peuvent représenter un pourcentage significatif de l'ensemble du répertoire de gènes d'un génome (Brenner et al. 1995; Koonin et al. 1995). De plus, certaines familles de gènes paralogues ont été amplifiées spécifiquement dans une lignée bactérienne (par exemple uniquement chez les protéobactéries) et représentent à elles seules entre 5 et 33% du nombre total de gènes (Jordan et al. 2001). La formation de gènes paralogues joue ainsi un rôle majeur dans l'augmentation de la taille des génomes procaryotes (Snel et al. 2002).

Plus la taille du génome est grande et plus le nombre de familles multigéniques est important. De même, le nombre de gènes paralogues par famille croit également avec l'augmentation de la taille des génomes. Différentes études ont montré que la distribution de la taille des familles de gènes paralogues chez les procaryotes comme chez les eucaryotes se rapproche de celle décrite par une loi statistique appelée « power-law » (la fréquence d'une famille de k éléments est égale à $P(k) = c.k^{-\gamma}$; où c et γ sont des constantes) (Huynen and van Nimwegen 1998; Qian et al. 2001). Ce type de distribution indique que la probabilité de duplication d'un gène est proportionnelle à la taille de la famille à laquelle il appartient et que la formation des familles multigéniques s'est faite de manière stochastique et non par duplication de génomes entiers (Lespinet et al. 2002).

Dans de nombreux cas, les gènes dupliqués sont impliqués dans l'adaptation aux conditions environnementales (Kondrashov et al. 2002). D'un point de vue physiologique, la présence de copies multiples possédant des fonctions identiques ou du moins similaires offre le moyen de doser l'effet d'un gène et d'obtenir une plus grande plasticité physiologique afin de répondre aux variations des facteurs environnementaux. Ce mécanisme est bien connu chez les cyanobactéries qui doivent s'adapter aux variations constantes de l'intensité lumineuse. Ainsi, la plupart des cyanobactéries contiennent deux copies identiques du gène *psbD* codant pour la protéine D2 du centre réactionnel du photosystème II (voir Fig. I-3**)**. Chez

Synechococcus sp. PCC 7942, ces deux copies sont différentiellement exprimées et seul le gène *psbDII* voit son niveau d'expression augmenter lors d'une élévation de l'intensité lumineuse (Bustos and Golden 1992). L'expression du gène *psbDI* qui est en opéron avec le gène *psbC* (codant pour la protéine CP43 du photosystème II) reste inchangée. Cela permet d'augmenter rapidement la production de la protéine D2 et de maintenir le photosystème II fonctionnel lors d'une exposition à de fortes intensités lumineuses (Bustos and Golden 1992). La nécessité de posséder plusieurs copies d'un gène, pour s'adapter aux variations de l'environnement, peut aussi se déduire inversement de la présence d'un nombre réduit de paralogues par famille de gènes, chez les bactéries vivant dans un milieu relativement stable comme le cytoplasme d'une cellule eucaryote ou encore le milieu marin (Pushker et al. 2004). Cette constatation est valable aussi bien pour les bactéries ayant un tout petit génome que pour celles ayant un génome de très grande taille comme *Pirellula*.

De nombreux exemples d'adaptation à la niche écologique peuvent être expliqués par l'amplification de familles de gènes. Par exemple, l'adaptation à la niche de profondeur chez *Prochlorococcus* est reliée (en partie) à la multiplication des gènes *pcb* chez les souches de profondeur (Garczarek et al. 2000). Chez les bactéries intracellulaires du genre *Buchnera*, qui vivent en symbiose obligatoire avec des insectes, les gènes des voies de biosynthèse des acides aminés essentiels ont été multipliés spécifiquement (Shigenobu et al. 2000). Cela permet à ces bactéries d'alimenter leurs hôtes en acides aminés qu'ils ne peuvent pas synthétiser par eux-mêmes et qui sont absents de leurs sources de nourriture (sève des plantes). Ainsi, la multiplication de ces gènes chez ces bactéries endosymbiotiques pourrait avoir permis non seulement l'adaptation de ces organismes à leur niche écologique (le bactériocyte) mais aussi la conquête d'une nouvelle niche écologique par leurs hôtes (Wernegreen 2002).

La différenciation de familles géniques permet l'émergence de nouvelles fonctions, ce qui constitue un avantage fondamental à plus long terme. Ces familles, qui se sont formées par duplications successives à partir d'une séquence ancestrale, contiennent souvent à la fois des gènes orthologues (ayant conservé la même fonction) et des gènes paralogues (possédant des fonctions différentes). De même, des familles de fonctions différentes peuvent être regroupées en super-familles ayant une origine monophylétique.

La duplication de gènes a très souvent été considérée comme étant sélectivement neutre. Le modèle développé par Susumu Ohno (Ohno 1970) sur l'évolution des gènes paralogues, propose qu'après la duplication, une des deux copies nouvellement créées échappe à la sélection purifiante et accumule des mutations qui normalement auraient été contre-sélectionnées. Ainsi, la redondance apportée par la duplication permettrait le développement d'une nouvelle fonction à partir de la copie évoluant de façon neutre, tout en maintenant la

fonction du gène grâce à l'autre copie. Cependant, plusieurs études ont montré par la suite que les gènes paralogues ne passent pas par une phase de diminution forte de la sélection purifiante. Kondrashov et collaborateurs (Kondrashov et al. 2002) ont identifié les gènes paralogues récemment dupliqués dans les génomes de 26 bactéries, 6 archées et 7 eucaryotes. Ces gènes présentaient un taux de substitutions non-synonymes (qui changent les séquences d'acides aminés) toujours très inférieur au taux de substitutions synonymes (qui ne changent pas les séquences d'acides aminés). Les auteurs ont également mesuré les vitesses relatives d'évolution des gènes paralogues. Dans la très grande majorité des cas, ces gènes ne présentaient pas de différence significative du taux de substitution des acides aminés. Ces résultats montrent que les deux copies issues de la duplication d'un gène évoluent sous l'action de la sélection purifiante et que la force de la sélection est sensiblement la même pour les deux copies. Néanmoins, les auteurs ont observé que les gènes paralogues évoluent plus rapidement que les gènes orthologues correspondants dans les génomes d'organismes proches (appartenant à la même lignée procaryotique). Cette accélération pourrait être due à une diminution de la sélection purifiante ou au contraire à l'action de la sélection positive qui favoriserait la fixation de mutations avantageuses dans les séquences de paralogues.

I.3.2.2 Transferts horizontaux

L'un des résultats les plus remarquables de la génomique comparée est la mise en évidence de l'ampleur des transferts horizontaux de gènes (THG) entre génomes d'organismes appartenant à des taxons différents. La réalité de ce phénomène a été démontrée dès les débuts de la biologie moléculaire mais ce n'est qu'avec la disponibilité des premières séquences de génomes que l'importance de ce mécanisme évolutif a été reconnu (Koonin et al. 2001; Ochman et al. 2000). Ainsi, le génome d'*Escherichia coli* contiendrait jusqu'à 18 % de gènes issus d'un transfert latéral (Lawrence and Ochman 1998). Par ailleurs, de très nombreux gènes d'origine bactérienne ont été détectés dans les génomes d'archées hyperthermophiles (Koonin et al. 1997) alors qu'un grand nombre de gènes issus d'archées ont été identifiés dans le génome de *Thermotoga maritima* ou d'*Aquifex aeolicus* (Aravind et al. 1998; Nelson et al. 1999). La proportion élevée de gènes acquis par transferts horizontaux a conduit certains auteurs à remettre en question, au moins pour les procaryotes, l'idée d'un arbre phylogénétique unique. En conséquence, les relations évolutives entre les organismes ne devraient plus être représentées par un arbre mais plutôt par un réseau (Doolittle 1999).

Le transfert horizontal constitue une source de nouveauté importante pour les génomes receveurs. L'analyse du génome de la cyanobactérie *Synechocystis* sp PCC 6803 a montré la présence de protéines des voies de signalisation qui étaient considérées auparavant comme

étant spécifiques des eucaryotes (Ponting et al. 1999). Contrairement à la perte différentielle de gènes, ce mécanisme jouerait certainement un rôle considérable dans l'adaptation des organismes à leur environnement (Lawrence 1997; Lawrence and Ochman 1998; Ochman et al. 2000).

On peut distinguer trois types distincts de THG en fonction de la nature des gènes transférés (Koonin et al. 2001). Le premier concerne l'acquisition de gènes strictement nouveaux qui n'ont aucun homologue dans le génome receveur. Le deuxième type comprend l'acquisition de gènes paralogues de gènes du génome receveur. Enfin, le troisième type est caractérisé par le transfert de gènes orthologues qui viennent remplacer les gènes présents à l'origine chez le receveur. Dans ce dernier cas, on parlera de remplacement xénologue (xenologous gene displacement) (Gogarten 1994).

Bien que la duplication de gènes et les transferts horizontaux permettent un apport régulier de nouveaux gènes dans les génomes bactériens, la taille de ces génomes ne peut augmenter indéfiniment (Lawrence and Roth 1999). Ainsi, les gènes transférés ne seront fixés dans la population que s'ils apportent un avantage sélectif suffisant à l'organisme receveur et à sa descendance (Lawrence and Roth 1999). Cependant, dans le cas de remplacements xénologues, les gènes transférés étant les orthologues de gènes préexistants dans le génome receveur, la fixation de ces gènes pourrait être sélectivement neutre et se faire aléatoirement.

Le fait qu'un gène transféré confère un phénotype avantageux implique qu'il soit exprimé et traduit. De plus, les gènes codant pour les protéines qui interagissent avec le produit de ce gène doivent également avoir été transférés. D'une manière générale, il semble que les gènes codant pour des protéines appartenant à des systèmes fonctionnels complexes, tels que ceux impliqués dans les mécanismes de transcription et de traduction, ont une probabilité beaucoup plus faible d'être transférés et maintenus sur une longue période dans le génome du receveur (Rivera et al. 1998). Jain et collaborateurs (Jain et al. 1999) ont proposé que le niveau de complexité des interactions soit un facteur limitant pour le transfert des gènes impliqués dans les mécanismes de transcription et de traduction.

Différentes approches peuvent être employées dans le but d'identifier des THG. Ils sont le plus souvent suspectés lors de la détection, par exemple avec BLAST, de séquences présentant un niveau significatif de similarité uniquement avec des séquences d'organismes éloignés phylogénétiquement. Cependant, ce type de distribution peut être aussi interprété par la perte de gènes dans les génomes proches même si cette dernière hypothèse conduit généralement à des scénarios évolutifs moins parcimonieux.

Une autre approche couramment utilisée est basée sur l'analyse de la composition en nucléotides (GC %), de la fréquence des oligonucléotides ou encore de l'usage des codons (Mrazek et al. 2001). Les génomes bactériens ont une composition en bases assez homogène et ces paramètres sont spécifiques de chaque espèce. Par conséquent, la présence de gènes,

ayant des caractéristiques différentes de celles de l'ensemble du génome dont ils font partie, a souvent été interprété comme le résultat d'un THG. En utilisant une combinaison de plusieurs paramètres (GC%, usage des codons et des acides aminés) sur 24 génomes de bactéries et d'archées, Garcia-Vallvé et collaborateurs ont estimé que le pourcentage de THG variait de 1,56% à 14,47% (Garcia-Vallve et al. 2000).

Il faut souligner que ce type d'analyses statistiques présente un grand nombre de limitations et plusieurs études ont mis en évidence le caractère peu résolutif de ces méthodes pour l'identification des THG (Koski et al. 2001). Ainsi, il existe de nombreux autres facteurs qui peuvent faire qu'un gène présente des caractéristiques atypiques. L'usage des codons est par exemple sous l'influence de plusieurs facteurs, tels que le biais dans la pression de mutation et la sélection en faveur de codons optimaux (c'est-à-dire ceux pour lesquels les ARNt sont abondants) dans les séquences de gènes fortement exprimés. D'autre part, les gènes transférés entre organismes proches dont les génomes possèdent des paramètres de composition similaires ne sont pas identifiés par ce type d'approche ; ce qui tendrait à sous-estimer fortement la fréquence des THG. De plus, les gènes transférés possédant au départ les caractéristiques du génome donneur subissent un processus « d'amélioration » (Lawrence and Ochman 1998) qui entraîne le changement de la composition de ces gènes afin de s'adapter à celle du génome receveur.

Les méthodes phylogénétiques qui offrent la possibilité d'inférer réellement le caractère homologue entre séquences similaires constituent par nature le meilleur moyen pour identifier les cas de THG. Ainsi, les exemples de séquences homologues appartenant à des taxons distants et se regroupant dans les arbres phylogénétiques procurent les indices les plus probants de THG. Par exemple, les gènes des deux sous-unités de la Rubisco, de la phosphoribulokinase et de la pentose-5-phosphate epimerase de *Prochlorococcus* et de *Synechococcus* ont des homologues plus proches chez plusieurs genres de protéobacteries marines que chez les cyanobactéries d'eau douce (Fig. I-9) (Badger and Price 2003 ; Hess et al. 2001). Cela pourrait indiquer un transfert de ces gènes entre l'ancêtre commun du groupe *Prochlorococcus*/*Synechococcus* et celui de ces protéobactéries. Cependant, il pourrait également s'agir d'une évolution convergente de ces gènes à cause de contraintes fonctionnelles similaires. De nombreux problèmes peuvent ainsi conduire à l'identification erronée des cas de THG. Les arbres obtenus dépendent fortement de la qualité des alignements utilisés ou encore du modèle évolutif choisi (Moreira and Philippe 2000). De plus, les méthodes phylogénétiques présentent de nombreux artefacts, tels que l'attraction des longues branches (Sanderson et al. 2000) qui peuvent provoquer le regroupement artificiel de séquences ayant évolué rapidement à des vitesses différentes.

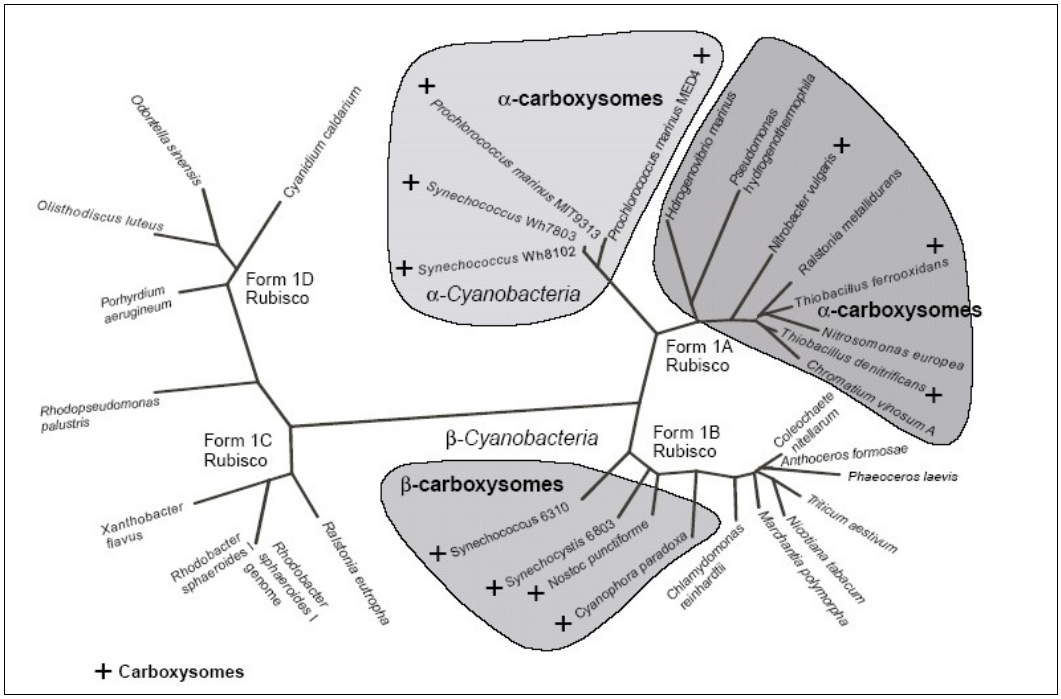

Figure I-9

Arbre phylogénétique fondé sur l'analyse des gènes de la forme 1 de la Rubisco, chez les bactéries photosynthétiques et chemo-autotrophes. Les pico-cyanobactéries marines possèdent la même forme de Rubisco que les protéobactéries. Arbre construit avec la méthode du neighbor-joining. D'après Badger et al. 2002

Dans un article récent, Novichkov et collaborateurs (Novichkov et al. 2004) ont identifié des cas de remplacements xénologues entre les génomes de trois lignées bactériennes en recherchant les gènes orthologues pour lesquels le principe d'horloge moléculaire (corrélation du taux d'évolution pour l'ensemble des gènes des génomes d'une même lignée) n'était pas respecté.

I.3.2.3 Pertes différentielles de gènes

Les génomes procaryotes sont généralement petits en comparaison des génomes eucaryotes. En effet, tous les génomes procaryotes connus ont une taille comprise entre 0,5 et 10 Mb, soit une variation d'un ordre de magnitude. Malgré cela, on observe une variabilité très importante du répertoire de gènes chez ces organismes. Les génomes de souches phylogénétiquement proches présentent souvent des pourcentages significatifs de gènes

spécifiques alors qu'ils ont des tailles très semblables. Les comparaisons de génomes ont montré l'importance de la duplication génique et du transfert de gènes dans l'évolution des génomes procaryotes. Ces deux mécanismes génèrent un gain régulier de gènes dans les génomes qui doit être compensé par une perte équivalente de gènes pour expliquer le maintien de la taille des génomes (Kunin and Ouzounis 2003; Mira et al. 2001). En l'absence de pertes, on observerait certainement une majorité de très grands génomes procaryotes.

Ces trois mécanismes (duplication, transfert et perte) contribuent différemment à la taille des génomes. En analysant les génomes de 41 bactéries et de 10 archées, Kunin et Ouzounins (Kunin and Ouzounis 2003) ont montré que les pertes de gènes étaient trois fois plus fréquentes que les transferts horizontaux, ceux-ci étant deux fois moins fréquents que les duplications de gènes. L'élimination de gènes semble donc être un des principaux facteurs qui déterminent la taille et le contenu en gènes des génomes.

Plusieurs facteurs influent de manière combinée sur le maintien d'un gène dans un génome. Parmi ceux-ci, on peut citer l'efficacité de la sélection sur la fonction du gène, le taux de mutations, la taille de la population ainsi que le taux de transferts horizontaux (Lawrence and Roth 1999). Peu de gènes sont essentiels et contribuent très fortement à la valeur sélective. La majorité des autres gènes peuvent être éliminés sans que cela ne diminue significativement la valeur sélective globale de la cellule. Parmi ces gènes, certains contribuent de manière si faible à la valeur sélective que la sélection ne peut s'opposer à leur perte. De même, l'augmentation du taux de mutations augmente la probabilité de fixation de mutations délétères qui entraînent l'inactivation des gènes. L'augmentation du taux de fixation des mutations délétères survient également lors de la diminution de la taille de la population. Enfin, l'arrivée régulière de nouveaux gènes à forte valeur sélective peut provoquer en retour une perte de gènes de valeurs sélectives plus faibles.

La perte de gènes peut se traduire par de grandes délétions qui éliminent un ou plusieurs gènes en une seule fois. Une autre possibilité est que le gène soit d'abord inactivé à cause de l'accumulation de mutations délétères. Le pseudogène ainsi formé est ensuite éliminé graduellement par de petites délétions. En estimant les taux d'insertions et de délétions dans les séquences de pseudogènes, Mira et collaborateurs (Mira et al. 2001) ont montré que le taux de délétions était supérieur au taux d'insertions. Le biais de délétion qui en résulte tend à éliminer les séquences non-fonctionnelles. Son origine reste mystérieuse mais il pourrait s'agir d'un mécanisme de défense permettant de lutter contre la prolifération des éléments génétiques mobiles (séquences d'insertion, transposons) et des séquences de phages (Lawrence et al. 2001).

Chez certains organismes, la perte est supérieure au gain de gène ; ce qui provoque une élimination massive de gènes et la réduction de la taille du génome. Ce phénomène est particulièrement important chez les procaryotes pathogènes (*Mycoplasma* .spp, *Bordetella* spp., *Rickettsia* spp.) ou symbiotique (*Buchnera aphidicola*, *Wigglesworthia glossinidia*, *Nanoarchaeum equitans*). Ce phénomène de réduction semble être directement le résultat du passage d'un mode de vie libre à un mode de vie où l'association à un hôte (généralement eucaryote) est obligatoire. Les tissus de l'hôte constituent un milieu très riche et tamponné, ce qui rend inutile la fonction de nombreux gènes, notamment ceux des voies de biosynthèse de nombreux composés qui peuvent être obtenus directement de l'hôte. Ainsi, l'efficacité de la sélection purifiante (qui s'oppose au changement) diminue fortement sur ces gènes. De plus, les organismes pathogènes ou symbiotiques ont le plus souvent des populations de très petite taille ; ce qui renforce le processus d'inactivation et de perte de gènes à cause de l'augmentation du taux de fixation des mutations délétères (Moran 2002 ; Moran 2003).

I.4 Contexte scientifique et démarche adoptée au cours de la thèse

Le travail réalisé au cours de cette thèse comporte deux axes principaux qui sont (i) la détermination des bases moléculaires de l'adaptation des picocyanobactéries marines à leurs environnements respectifs et (ii) la compréhension des processus évolutifs qui ont façonné la taille de leur génome et leur contenu en gènes. Ces deux axes complémentaires sont basés sur l'annotation et la comparaison des génomes de cinq picocyanobactéries marines.

Au début de cette thèse, les génomes complets de deux souches de *Prochlorococcus* étaient en cours d'annotation au MIT (Etats-Unis). L'une de ces souches (*P. marinus* MED4) appartient à l'écotype de haute lumière, tandis que l'autre fait partie de l'écotype de basse lumière (*Prochlorococcus* sp. MIT9313). A la même période, la séquence d'un troisième génome de *Prochlorococcus*, celui de la souche de basse lumière *P. marinus* SS120, venait d'être obtenue par le Genoscope à la demande d'un consortium d'équipes européennes, coordonnées par l'équipe Plancton Océanique de Roscoff (Unité des Cyanobactéries de l'Institut Pasteur dirigée par Nicole Tandeau de Marsac, Département de Biologie de l'Université Humboldt de Berlin dirigé par Wolfgang Hess, Département de Sciences Biologique de l'Université de Warwick dirigé par Dave Scanlan).

La première partie de mon travail de thèse a donc consisté à annoter le génome de *Prochlorococcus marinus* SS120 qui est la souche type de ce genre. Ce travail a été réalisé par le consortium d'annotation européen en collaboration avec l'équipe américaine du Prof. Eugene Koonin (NCBI). Parallèlement, j'ai également participé à l'annotation des gènes photosynthétiques de *Synechococcus* sp. WH8102, une souche adaptée au milieu océanique oligotrophe et appartenant à un genre phylogénétiquement très proche de *Prochlorococcus*. Enfin, durant la dernière année de ma thèse, j'ai participé à l'annotation (encore en cours au moment de la rédaction) d'un second génome de *Synechococcus* (souche WH7803) qui est caractéristique de l'environnement marin mésotrophe. Les résultats de l'annotation de ces trois génomes sont présentés dans le **chapitre II**.

Les cinq picocyanobactéries marines qui ont servi de modèles d'étude pour cette thèse sont adaptées à des environnements différents en terme d'intensité lumineuse ou encore de richesse en sels nutritifs. Afin d'identifier les gènes potentiellement impliqués dans l'adaptation à la niche écologique, nous avons entrepris la comparaison systématique des répertoires de gènes de ces procaryotes photosynthétiques. Nous avons abordé cet axe d'étude en nous focalisant sur l'adaptation aux conditions lumineuse des niches de forte et de faible lumière (**chapitre III**).

La disponibilité des génomes de trois souches de *Prochlorococcus,* deux représentant les deux clades les plus distants dans l'arbre phylogénétique de ce genre et le troisième étant intermédiaire, nous a offert la possibilité d'étudier de manière fine les mécanismes évolutifs au sein de ce genre. La comparaison de ces génomes indique clairement que l'évolution chez le genre *Prochlorococcus* a été marquée par une réduction de la taille du génome et par conséquent du nombre de gènes codés dans ces génomes. Aussi, nous nous sommes intéressés aux causes et aux conséquences de cette réduction notamment en nous interrogeant sur le rôle joué par la réduction du génome dans l'extraordinaire succès écologique de *Prochlorococcus* (**chapitre IV**).

*****

CHAPITRE II

Annotation des Génomes de Picocyanobactéries Marines

II.1 Résumé des résultats obtenus

II.1.1 Annotation du génome de *Prochlorococcus marinus* SS120

Les analyses préliminaires (avant la disponibilité du génome) de quelques classes de gènes, séquencées spécifiquement chez *P. marinus* SS120, avaient déjà suggéré la présence d'un nombre plus réduit de gènes chez *Prochlorococcus* que chez d'autres genres de cyanobactéries (Hess et al., 1999). Afin d'identifier systématiquement les gènes absents du génome de SS120, nous l'avons donc comparé à ceux, plus grands et plus complexes, de trois cyanobactéries d'eau douce, *Synechocystis* PCC 6803, *Anabaena* PCC 7120 et *Thermosynechococcus elongatus* BP-1 (Kaneko et al. 2001; Kaneko et al. 1996; Nakamura et al. 2002).

Ces comparaisons ont montré que le génome de SS120, qui est le second plus petit génome de cyanobactérie après *P. marinus* MED4 (Rocap et al., 2003), possède un génome quasiment minimal pour une oxyphototrophe. En effet la majorité des gènes, tels que ceux de l'appareil photosynthétique, est en simple copie. De nombreux gènes impliqués dans la réparation de l'ADN, la photosynthèse, l'importation de solutés, le métabolisme intermédiaire ou encore la mobilité et le phototactisme sont absents. Certaines familles de gènes sont même complètement absentes de ce génome alors que de nombreux exemplaires de ces gènes sont présents dans les génomes de cyanobactéries d'eau douce. Cette tendance à la réduction génomique est particulièrement marquée dans le cas des transposases, des systèmes de transduction des signaux et de réponse aux stress environnementaux. Ainsi, chez les trois cyanobactéries d'eau douce, les gènes codant pour les systèmes à deux composants sont très largement représentés (Mizuno et al, 1996; Ohmori et al, 2001; Meeks et al, 2002) alors que le génome de *P. marinus* SS120 ne contient que 5 senseurs histidines kinases et 6 régulateurs de réponse. L'analyse de ce génome nous apporte donc des renseignements précieux sur le jeu minimal de gènes nécessaires au fonctionnement d'un organisme oxyphototrophe.

Ce travail est présenté sous la forme d'une publication dans la revue *Proceedings of the National Academy of Sciences of the U.S.A*. Les données complémentaires associées à cette publication sont présentées en annexe I.

II.1.2 Annotation du génome de *Synechococcus* sp. WH8102.

L'annotation du génome de *Synechococcus* sp. WH8102 a permis de mettre en évidence, au niveau moléculaire, certaines des adaptations de cette cyanobactérie à l'environnement oligotrophe des régions centrales océaniques.

L'analyse du génome a montré la présence de plusieurs gènes dont les fonctions sont directement liées à l'adaptation à la salinité. Ainsi, *Synechococcus* sp. WH8102 possède, en comparaison des cyanobactéries d'eau douce, un plus grand nombre de transporteurs utilisant le gradient de sodium pour le passage de différents composés au travers de l'enveloppe cellulaire. Le génome de *Synechococcus* sp. WH8102 possède également plusieurs gènes impliqués dans l'import ou dans la synthèse de glycine betaïne. Ce composé intervient dans l'osmorégulation et n'est pas synthétisé par les cyanobactéries d'eau douce.

Un autre aspect intéressant de l'adaptation de *Synechococcus* sp. WH8102 concerne l'utilisation du fer. Cet élément est très peu abondant dans les océans et le génome de cette cyanobactérie contient plusieurs exemples de gènes permettant d'économiser le fer. Des enzymes telles que la plastocyanine, la ribonucléotide réductase ou la superoxyde dismutase utilisent normalement le fer pour fonctionner. L'annotation a permis de montrer que *Synechococcus* sp. WH8102 synthétise des formes différentes de ces enzymes qui utilisent d'autres métaux (cuivre, cobalt ou nickel) comme co-facteurs. Cependant, le génome de *Synechococcus* . WH8102 ne possède aucun des deux gènes de l'opéron *isiAB*, qui chez les cyanobactéries d'eau douce est régulé par la disponibilité du fer, et qui code pour deux protéines importantes:

- CP43', (ou IsiA) forme une couronne de protéines fixant la chlorophylle autour du photosystème I (Bibby et al. 2001; Boekema et al. 2001), un phénomène qui notamment compense la forte diminution du rapport PSI/PSII survenant durant la carence en fer.
- La flavodoxine (IsiB), une protéine qui remplace la ferredoxine comme accepteur d'électrons

L'absence de ces deux gènes est d'autant plus surprenante qu'*isiB* est présent dans les génomes des trois souches séquencées de *Prochlorococcus.* De plus, les protéines de l'antenne majeure du PSI et du PSII de *Prochlorococcus* constituent des formes dérivées de IsiA qui selon les cas sont restées ou non sous le contrôle de la disponibilité en fer (Bibby et al., 2003). L'ancêtre commun de *Prochlorococcus* et *Synechococcus* devait donc bien posséder ces deux gènes associés ou non en opéron.

Comme chez *Prochlorococcus*, le nombre de gènes des voies de signalisation est

particulièrement peu important par rapport aux cyanobactéries d'eau douce. Cela est corrélé avec le fait que *Synechococcus* sp. WH8102 est représentatif de populations de *Synechococcus* adaptées aux environnements océaniques oligotrophes stables, comme *Prochlorococcus*.

Par rapport à *Prochlorococcus*, le génome de *Synechococcus* sp. WH8102 contient un plus grand nombre d'intégrases (DNA recombinases) de phages. La présence de ces enzymes, couplée à d'autres indices compositionnels (GC %, composition en trinucléotides), indique que de nombreux gènes provenant d'autres bactéries ont été transférés horizontalement dans le génome de *Synechococcus* sp. WH8102. Ces transferts horizontaux de gènes sont à l'origine de l'acquisition de nouvelles fonctions par *Synechococcus*. C'est le cas des gènes impliqués dans la mobilité (*swmA* et *swmB*) ou des gènes codant pour des glycosyltransférases. Ces dernières pourraient intervenir dans la modification de l'enveloppe cellulaire pour permettre aux cellules de *Synechococcus* d'échapper aux virus ou aux prédateurs.

Ce travail a fait l'objet d'une publication dans la revue *Nature* (Palenik et al. 2003), présentée en annexe II.

II.1.3 Annotation du génome de *Synechococcus* sp. WH7803.

L'annotation du génome de *Synechococcus* sp. WH7803 est encore en cours. Le paragraphe qui suit présente quelques résultats préliminaires de l'analyse de différentes familles de gènes de ce génome.

Cette deuxième souche de *Synechococcus* a été isolée dans un environnement caractérisé par sa relative richesse en sels nutritifs. La pigmentation de cette souche (riche en phycoérythrobiline qui absorbe à 520 nm) diffère de celle *Synechococcus* sp. WH8102 (riche en phycourobiline qui absorbe à 495 nm), reflétant les différences de propriétés optiques entre les eaux mésotrophes vertes d'où vient la première et les eaux oligotrophes bleues d'où vient la seconde. A cet égard, il est intéressant de noter que *Synechococcus* sp. WH8102 possède un gène de fusion (*SYNW2025*) ayant une forte homologie avec les gènes *pecE* et *pecF* de *Fischerella* sp. Cohn (AAC64647) et *Anabaena* sp. PCC7120 (AAA22019). Ils codent pour une lyase-isomérase qui fixe une phycobiline de type I (la phycocyanobiline) à un site (résidu cystéine) bien particulier de la chaîne α d'une phycobiliprotéine (phycoérythrocyanine) et, concomitamment, l'isomérise en phycobiline de type II (la phycoviolibiline). Les *Synechococcus* marins ne possèdant ni phycoérythrocyanine ni phycoviolibiline, il est

probable, puisque la réaction enzymatique est la même, que le produit de *SYNW2025* fixe plutôt une molécule de phycoérythrobiline à la phycoérythrine (I et/ou II) et l'isomérise en phycourobiline, comme prédit par (Storf et al. 2001). Au contraire, les deux gènes équivalents (non fusionnés) à *SYNW2025* trouvés chez *Synechococcus* sp. WH7803 (ORF0480/0481) sont homologues aux gènes *rpcE* et *rpcF* caractérisés chez la souche côtière *Synechococcus* sp. PCC 7002 comme codant pour une lyase simple (fixant une phycobiline de type I sans isomérisation). Cette observation suggère une différence probable dans la nature des phycobiline-lyases présentes chez *Synechococcus* spp. WH7803 et WH8102 et pourrait peut être expliquer en partie les différences de rapports phycourobiline/phycoérythribiline (0,45 et 1,95, respectivement; Six et al., 2004) observées entre ces souches.

Parmi les picocyanobactéries marines, *Synechococcus* sp. WH7803 possède le plus grand nombre de systèmes à deux composants, impliqués dans la transduction des signaux environnementaux. Pas moins de 10 senseurs histidine kinases et 17 régulateurs de réponse ont été identifiés dans ce génome. De plus, 2 protéines hybrides contenant à la fois le domaine transmetteur des histidine kinase et le domaine receveur des régulateurs de réponse, sont également présentes (Mizuno et al. 1996 ; Parkinson and Kofoid 1992). Ainsi, cette cyanobactérie est probablement capable de répondre aux variations d'un plus grand nombre de paramètres environnementaux, variations qui surviennent plus fréquemment dans un environnement côtier ou mésotrophe que dans un milieu oligotrophe.

Une autre différence avec les génomes de picocyanobactéries marines concerne les gènes de protection contre l'oxydation. WH7803 possède deux superoxides dismutases différentes qui utilisent un co-facteur à base de cuivre et de zinc pour l'une, et à base de fer et de manganèse pour l'autre. Aucune superoxide dismutase fonctionnant avec le nickel, comme chez *Synechococcus* sp. WH8102 et *Prochlorococcus*, n'est présente chez *Synechococcus* sp. WH7803. Cette dernière se développe dans des eaux où le fer est moins limitant, aussi le besoin d'économiser cet élément est peut-être moins déterminant dans le cas de cette souche. Le génome de *Synechococcus* sp. WH7803 contient également un gène codant pour une catalase (peroxidase I) (Yamada et al. 2002). Cette enzyme utilise une molécule d'hème comme groupement prosthétique pour éliminer le peroxyde d'hydrogène. Ce gène est également présent chez *Synechocystis* sp PCC 6803 et *Gloeobacter violaceus* PCC 7421, mais est absent des autres génomes de picocyanobactéries marines. La séquence de la catalase de *Synechococcus* sp. WH7803 est plus proche des séquences de protéobactéries, telle que *Geobacter sulfurreducens,* et d'archées, telle que *Methanosarcina acetivorans,* que de celles des cyanobactéries d'eau douce. Ceci suggère que ce gène est issu d'un transfert latéral. La présence de ces trois gènes chez *Synechococcus* sp. WH7803 contre un seul gène de superoxyde dismutase dans les autres génomes de picocyanobactéries marines fait penser que

WH7803 subit, dans son environnement, un stress oxydatif plus important que les autres picocyanobactéries et/ou est mieux équipé pour faire face à cette éventualité.

Le génome de *Synechococcus* sp. WH7803 est aussi caractérisé par l'absence des gènes *urtBC* et des gènes *ureABCDEFG* qui codent respectivement pour deux sous-unités du transporteur ABC d'urée et pour les différentes sous-unités de l'uréase. L'absence de ces gènes, qui sont présents chez *Synechococcus* sp. WH8102, indique clairement que cette souche a perdu la capacité d'utiliser l'urée comme source d'azote, comme c'est aussi le cas pour *P. marinus* SS120. On ne peut pas exclure que cette particularité résulte de la mise en culture prolongée de ces souches dans un milieu sans urée (depuis 1978 pour la première, depuis 1988 pour l'autre).

A l'instar de *P. marinus* MED4 et de *Synechococcus* sp. WH8102, *Synechococcus* sp. WH7803 possède une cyanate lyase (cyanase) (Palenik et al. 2003). Cette enzyme catalyse la réaction du cyanate avec le bicarbonate pour produire du dioxide de carbone et de l'ammonium (Walsh et al. 2000). Elle permet, ainsi, de dégrader le cyanate, qui est toxique, et d'exploiter une source d'azote supplémentaire. Curieusement, *Synechococcus* sp. WH7803 ne possède pas les gènes codant pour un éventuel transporteur ABC de cyanate, identifiés chez *P. marinus* MED4 et *Synechococcus* sp. WH8102. *Synechococcus* sp. WH7803 possède néanmoins un gène dont la partie N-terminale est similaire au gène codant pour la sous-unité ATPase (fixation et hydrolyse de l'ATP) du transporteur ABC de cyanate. Ce gène est situé juste en aval d'un gène codant pour un transporteur ABC (permease et domaine de fixation du substrat) dont les homologues les plus proches sont présents chez les protéobactéries. De part la similarité entre les sous-unités ATPases, on peut supposer que ce transporteur ABC permet l'importation du cyanate chez *Synechococcus* sp. WH7803. Ce transporteur semble avoir été acquis indépendamment par cette souche et pourrait provenir d'un transfert horizontal.

Les premières analyses du génome de *Synechococcus* sp. WH7803 offrent donc une estimation préliminaire de la variabilité qui existe entre deux souches très proches de *Synechococcus,* notamment pour les gènes du transport de l'azote. Ces résultats montrent clairement que cette souche est adaptée à un environnement plus riche et plus variable que *Synechococcus* sp.WH8102 et les *Prochlorococcus* spp.

Genome sequence of the cyanobacterium *Prochlorococcus marinus* SS120, a nearly minimal oxyphototrophic genome

Alexis Dufresne*, Marcel Salanoubat†, Frédéric Partensky*‡, François Artiguenave†, Ilka M. Axmann§, Valérie Barbe†, Simone Duprat†, Michael Y. Galperin¶, Eugene V. Koonin¶, Florence Le Gall*, Kira S. Makarova¶, Martin Ostrowski‖, Sophie Oztas†, Catherine Robert†, Igor B. Rogozin¶, David J. Scanlan‖, Nicole Tandeau de Marsac, Jean Weissenbach†, Patrick Wincker†, Yuri I. Wolf¶, and Wolfgang R. Hess§††**

*Station Biologique, Unité Mixte de Recherche 7127, Centre National de la Recherche Scientifique et Université Paris 6, BP74, 29682 Roscoff Cedex, France; †Genoscope et Unité Mixte de Recherche 8030, Centre National de la Recherche Scientifique, CP 5706, 91057 Evry Cedex, France; §Department of Biology, Humboldt University of Berlin, Chausseestrasse 117, 10115 Berlin, Germany; ¶National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894; ‖Department of Biological Sciences, University of Warwick, Coventry CV4 7AL, United Kingdom; and **Unité des Cyanobactéries, Unité de Recherche Associée 2172, Centre National de la Recherche Scientifique, Institut Pasteur, 28 Rue Dr. Roux, 75724 Paris Cedex 15, France

Edited by Robert Haselkorn, University of Chicago, Chicago, IL, and approved July 1, 2003 (received for review May 28, 2003)

***Prochlorococcus marinus*, the dominant photosynthetic organism in the ocean, is found in two main ecological forms: high-light-adapted genotypes in the upper part of the water column and low-light-adapted genotypes at the bottom of the illuminated layer. *P. marinus* SS120, the complete genome sequence reported here, is an extremely low-light-adapted form. The genome of *P. marinus* SS120 is composed of a single circular chromosome of 1,751,080 bp with an average G+C content of 36.4%. It contains 1,884 predicted protein-coding genes with an average size of 825 bp, a single rRNA operon, and 40 tRNA genes. Together with the 1.66-Mbp genome of *P. marinus* MED4, the genome of *P. marinus* SS120 is one of the two smallest genomes of a photosynthetic organism known to date. It lacks many genes that are involved in photosynthesis, DNA repair, solute uptake, intermediary metabolism, motility, phototaxis, and other functions that are conserved among other cyanobacteria. Systems of signal transduction and environmental stress response show a particularly drastic reduction in the number of components, even taking into account the small size of the SS120 genome. In contrast, housekeeping genes, which encode enzymes of amino acid, nucleotide, cofactor, and cell wall biosynthesis, are all present. Because of its remarkable compactness, the genome of *P. marinus* SS120 might approximate the minimal gene complement of a photosynthetic organism.**

Marine cyanobacteria of the genus *Prochlorococcus* (1) dominate phytoplankton communities in most tropical and temperate open ocean ecosystems (2). Their tiny cell sizes (0.5–0.7 μm) make *Prochlorococcus* spp. the smallest photosynthetic organisms known to date. Their major pigments are divinyl derivatives of chlorophyll *a* and *b* (Chl a_2 and b_2), which are unique to this genus (3). *Prochlorococcus* lacks phycobilisomes, large extrinsic multisubunit light-harvesting complexes found in typical cyanobacteria. These complexes are replaced by Chl a_2/b_2-binding proteins called Pcbs, which are analogous in function but are structurally and phylogenetically distinct from the light-harvesting complexes of higher plants (4).

In the ocean, *Prochlorococcus* cells face a number of natural constraints, including strong inverse vertical gradients of irradiance and nutrients. A key factor of the adaptation to such variable conditions seems to be the existence of several physiologically and genetically distinct genotypes growing in different ecological niches (5). High-light-adapted genotypes (or "ecotypes") occupy the upper, well illuminated but nutrient-poor 100-m layer of the water column, whereas low-light-adapted genotypes preferentially thrive at the bottom of the euphotic zone (80–200 m) at dimmer light but in a nutrient-rich environment.

In this article we describe the complete genome sequence of the low-light-adapted *Prochlorococcus marinus* type strain SS120 (also known as CCMP1375). Genomes of a high-light-adapted strain, *P. marinus* MED4, and another low-light-adapted strain, *P. marinus* MIT9313, have also recently been sequenced, and their genomes have been compared (6). In terms of photophysiology, *P. marinus* SS120 represents an extreme within the *Prochlorococcus* genus because of its ability to grow at very low light levels (5). Analyses of the genomic information of *P. marinus* SS120 and comparisons with other cyanobacterial genomes available to date (7–9) allowed us to delineate a putative minimal gene set of an oxyphotoautotrophic bacterium.

Materials and Methods

P. marinus SS120 could not be cultured axenically, leading to an ≈5% contamination of the genomic DNA. To construct a plasmid library with a low level of contaminants, we produced five pilot libraries with insert size ranging from 3 to 10 kb. Approximately 100 clones were end-sequenced, and the AT content of each read was calculated. The G+C content of SS120 and the contaminant were ≈40% and ≈60%, respectively, allowing the assignment of low G+C sequences to *Prochlorococcus*. The contamination level was the lowest, with inserts of 7 and 10 kb. Two shotgun genomic libraries were made by mechanical shearing of the DNA, size selection of the fragments, ligation into a low-copy plasmid (pCNS), and electroporation into DH10b cells (Invitrogen, Cergy-Pontoise, France). Plasmid DNAs (11,944 and 26,821 for the 7- and 10-kb insert libraries, respectively) were purified and end-sequenced as described (10) by using dye-primer and dye-terminator chemistries (50/50) on Licor 4200L and ABI3700 sequencers. Data were assembled with PHRAP (www.phrap.org), taking all sequences into account. An additional 789 directed reactions were performed to close the gaps and raise the quality of the sequence to finished standards. The integrity of the assembly was verified by comparing the theoretical lengths of bands obtained with two restriction enzymes with those determined experimentally (11).

ORFs were identified by using Glimmer, GeneMarks, and Critica (12–14). Transfer RNAs were predicted by tRNAscan-SE (15). In addition, transcription start sites were predicted by using the

This paper was submitted directly (Track II) to the PNAS office.

Abbreviations: ABC, ATP-binding cassette; Pro, *Prochlorococcus*; PS, photosystem; Chl, chlorophyll.

Data deposition: The sequence reported in this paper has been deposited in the GenBank database (accession no. AE017126).

‡To whom correspondence should be addressed. E-mail: partensk@sb-roscoff.fr.

††Present address: Ocean Genome Legacy, Beverly, MA 01915.

Table 1. Properties of the genomes of autotrophic microorganisms

Organism	Genome size, kb	Protein-coding genes	rRNA genes	tRNA genes*	Paralog cluster sizes†
A. aeolicus	1,551	1,522	6	44	−5.39
Chlorobium tepidum TLS	2,155	2,252	6	50	ND
Cyanobacteria					
P. marinus SS120	1,751	1,884	3	40	−5.01
Thermosynechococcus elongatus BP-1	2,594	2,475	3	42	−3.98
Synechocystis sp. PCC 6803	3,573	3,169	6	41	−4.44
Anabaena sp. (*Nostoc*) PCC 7120	6,414	6,129	12	48 + 19	−3.22
Archaea					
Methanococcus jannaschii	1,665	1,715	6	37	−4.39
Aeropyrum pernix K1	1,670	≈1,720	5	47	−5.25
Methanopyrus kandleri AV19	1,695	1,691	3	35	ND
Methanobacterium thermoautotrophicum	1,751	1,869	6	39	−4.11

The data are from original articles and from the latest variants of the genome sequences at the NCBI Genomes (www.ncbi.nlm.nih.gov/PMGifs/Genomes/micr.html) and Genomic tRNA Database (http://rna.wustl.edu/GtRDB) web sites. Only those archaeal autotrophs with genome sizes <2 Mb are listed.

*The data for *Anabaena* sp. represent the sum of the chromosomal and plasmid-encoded tRNA genes.

†To assess the propensity of these different prokaryotes for gene duplication, the distribution of sizes of species-specific paralog clusters in each genome was approximated by the power law (20). The steeper the slope of the curve, the fewer large clusters are observed. ND, not determined.

raster-score filter method recently developed for *P. marinus* MED4 (16) with respect to the higher G+C content of 34% within upstream regions of *P. marinus* SS120 for the calculation of the scoring matrix.

Genome annotation was performed by comparing the protein sequences with the Cluster of Orthologous Groups (COG) database (www.ncbi.nlm.nih.gov/COG) (17) and with the National Center for Biotechnology Information (NCBI) protein database (www.ncbi.nlm.nih.gov) by using BLAST and PSI-BLAST (18) with manual verification, as described (19). The number of copies of each particular gene in cyanobacterial genomes was either taken from the COG database (17) or estimated by BLAST searches against cyanobacterial protein databases (7–9). Species-specific expansions of paralogous gene families were determined as described (20). The SS120 genome sequence can be blasted at www.sb-roscoff.fr/Phyto/ProSS120.

Results and Discussion

General Features. The genome of *P. marinus* SS120 is composed of a single circular chromosome of 1,751,080 bp with an average G+C content of 36.4% (Table 1).

The origin of replication (*oriC*) was mapped between the *dnaN* and *thrC* genes on the basis of GC- and AT-skew analyses. The intergenic region between these two genes is AT-rich (73%) and contains six possible DnaA boxes with the consensus sequence 5′-[AT]TTCCACA-3′. The gene arrangement around *oriC* differs from the conserved arrangement found in many bacteria but is similar to the one in *Synechococcus* sp. PCC 7942 (21).

The genome contains 1,884 predicted ORFs with an average size of 825 bp. These ORFs represent 88.5% of the chromosome sequence. There is no significant asymmetry in the distribution of ORFs between the leading strand (50.6% of ORFs) and the lagging strand (49.4%). Biological roles were assigned to 1,254 (66.6%) of these ORFs. Among ORFs with unknown function, 399 (21.2%) have detectable homologs in the National Center for Biotechnology Information nonredundant database, and 231 (12.2%) have no detectable homolog. Three hundred ninety ORFs code for polypeptides of ≤100 aa. Although some had known function, such as the small subunits of photosystem II (PSII; e.g., PsbI, M, T or X) or high-light-inducible proteins (22), the function of many of them is not assigned yet. The SS120 genome contains a single rRNA operon (16S-23S-5S), 40 tRNA genes (which include cognates for all amino acids), and genes for three other RNAs.

Transcription of the genome is regulated by a reduced set of RNA polymerase sigma factors. In addition to SigA, SS120 has four genes encoding putative group 2 (23) sigma factors. A total of 3,130 transcription start sites were predicted by using the raster-score-filter method (16) for 1,289 noncoding upstream regions of the 1,930 predicted protein-coding or RNA genes. Our analyses excluded 641 regions <49 bp. The complete prediction can be downloaded from www.biologie.hu-berlin.de/~genetics/hess/hessproj.html. By analogy to MED4 (16), ≈ 40% of these sites were estimated to be functional. An experimental validation using eight randomly chosen genes (data not shown) showed that promoter elements and transcriptional start sites were correctly predicted for *pcbA*, *psbA*, *psbD*, *petH*, *kaiB*, *cpeB*, *ftsZ*, and *ntcA*.

***P. marinus* SS120 as a Minimal Genome of an Oxyphototrophic Organism.** Although larger than the "minimal" genomes of *Mycoplasma* spp. (24) and other obligate parasites, *P. marinus* SS120 has the smallest genome of all cyanobacteria sequenced to date (7–9), with the sole exception of the closely related 1.66-Mbp *P. marinus* MED4 genome (6). The *P. marinus* SS120 genome is comparable in size to the genomes of *Aquifex aeolicus* and archaeal autotrophs (Table 1). However, as a free-living autotroph, *P. marinus* SS120 has to encode enzymes for complete biosynthetic pathways for amino acids, nucleotides, cell wall carbohydrates, and cofactors, not to mention numerous components of the photosynthetic machinery (Fig. 1). A comparison of the four complete cyanobacterial genomes available to date gives some clues as to which genes are missing or underrepresented in *P. marinus* SS120. These include a few photosynthetic genes, genes involved in DNA repair, solute uptake, intermediary metabolism, and many other systems (see supporting information on the PNAS web site, www.pnas.org). Systems of signal transduction and environmental stress response (e.g., two-component systems) that are widely represented in the genomes of *Synechocystis* sp. PCC 6803 and *Anabaena* (*Nostoc*) sp. PCC 7120 (25–27) show a particularly drastic reduction. The number of encoded components is much larger than expected from the simple difference in size between *P. marinus* and the other cyanobacterial genomes (Table 2). There are no genes encoding light receptors such as bacteriophytochromes, cryptochromes, or rhodopsin. Many classes of signaling proteins (including hybrid histidine kinases, adenylate cyclases, diguanylate cyclases, phosphodiesterases, serine/threonine kinases, and phosphatases) that are found in most microbial genomes (28) are also lacking in SS120.

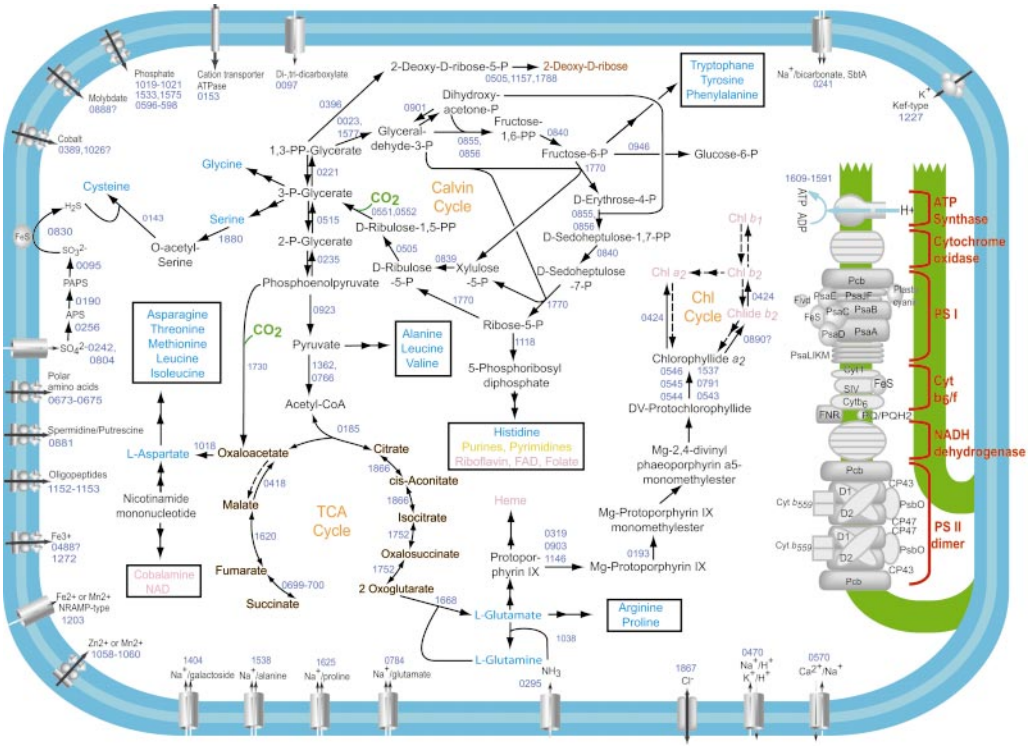

Fig. 1. Schematic organization of a *P. marinus* SS120 cell showing the main metabolic pathways and transporters. Gene identifiers are shown by numbers in blue. Genes with uncertain annotation are shown with a question mark. Reactions for which no candidate enzyme was confidently predicted are indicated by dashed arrows. Pathways that involve multiple reactions are shown by double arrows. Final biosynthetic products are indicated as follows: light blue, amino acids; dark yellow, nucleotides; brown, sugars; pink, cofactors. Chl cycle is based on ref. 37. Cyt, cytochrome; Flvd, flavodoxin; FNR, ferredoxin:NADP$^+$ oxidoreductase; PQ, plastoquinone.

In contrast to almost every other microbial genome, of the five sensor histidine kinases and six response regulators encoded by *P. marinus* SS120 only one pair of genes forms an operon. The absence of diguanylate cyclases and phosphodiesterases that are often involved in regulation of extracellular polysaccharide production and biofilm formation (28) is in agreement with the lack of close association between *P. marinus* cells in their natural habitat.

A low number of lineage-specific duplications in the SS120 genome compared with other cyanobacteria was confirmed by the smaller size of paralog clusters (Table 1). Thus, in agreement with the trend noted previously for other small-genome organisms (29), SS120 has relatively few paralogous genes and encodes few analogous enzymes.

DNA Repair and Genome Rearrangements. Although *P. marinus* SS120 is a low-light-adapted strain, its DNA repair genes are similar in their number and diversity to those from other cyanobacterial genomes, with the exception of a few missing genes (see supporting information). It is so far the only cyanobacterium that encodes a complete exonuclease V/recombinase complex (RecBCD). Although SS120 lacks the gene for deoxyribodipyrimidine photolyase, which is present in other cyanobacteria, it encodes instead a pyrimidine dimer-specific glycosylase, Pro1489, which is absent in other cyanobacteria and might have been horizontally transferred from a viral/phage genome. Another possible example of horizontally transferred genes is the type-I restriction/modification system (Pro0628–0630, Pro1274), which might be responsible for specific DNA methylation. In contrast to freshwater cyanobacteria, which have numerous transposase genes, SS120 does not encode any transposases (see supporting information). Together with the absence of site-specific recombinases (homologs of DNA invertase Pin) and XerC-like integrases, this observation suggests that SS120 is less prone to genome rearrangements than are other cyanobacteria.

Photosynthesis. ***Photosynthetic apparatus.*** The genome of *P. marinus* SS120 contains the whole set of PSII genes, with the exception of *psbU*, encoding the 12-kDa extrinsic subunit of PSII, which acts in the stabilization and protection of the oxygen-evolving complex against inactivation by heat, and *psbV*, encoding a low-potential cytochrome *c* associated with the luminal surface of the PSII reaction center complex.

All PSII genes in SS120 but one (*psbF*) are single-copy, in contrast to other cyanobacterial genomes, which contain multiple copies of *psbA*, *psbC*, and *psbD* genes (see supporting information).

SS120 has 11 genes coding for PSI proteins and 6 genes encoding the subunits of the cytochrome b_6/f complex. All these genes are single-copy. Ferredoxin:NADP$^+$ oxidoreductase (PetH, encoded by Pro1123), which catalyzes electron transfer from photochemically reduced ferredoxin to NADP$^+$, has an N-terminal domain with a predicted transmembrane helix that could anchor PetH to the thylakoid membrane. This contrasts with PetH from freshwater cyanobacteria (30, 31), whose N-terminal part is similar to a phycocyanin-associated linker polypeptide, allowing specific binding of this protein to phycocyanin. This observation is consistent with the different light-harvesting systems found in *P. marinus* and other cyanobacteria (4).

The *pcb* gene family, which encodes the proteins of the light-harvesting antenna complex in SS120, is a rare example of gene amplification in this otherwise compact genome. Compared with only one *pcb* gene in MED4 and all other high-light-adapted strains checked to date and two *pcb* copies in MIT9313 (6, 32), SS120 encodes eight different *pcb* genes, one more copy than previously reported (33).

SS120 has a small set of genes to synthesize the α, β, and γ

Table 2. Signaling and regulatory domains or genes that are absent in *P. marinus* SS120 (Pma) or in lower copy number than in the freshwater cyanobacteria *Thermosynechococcus elongatus* (Tel), *Synechocystis* sp. PCC6803 (Syn), and *Anabaena* sp. PCC7120 (Ana)

Signaling system component or domain	Gene name	COG no.	No. of genes/genome			
			Pma	Tel	Syn	Ana
Signal transduction mechanisms						
Histidine kinases	—	0642	5	16	42	124
Response regulators	—	2197	6	26	41	84
PAS domains	—	2202	1	13	23	57
GAF domains	—	2203	0	15	28	62
HPt domains	—	2198	0	4	7	9
GGDEF domains	—	2199	0	10	23	14
EAL domains	—	2200	0	5	13	7
HD-GYP domains	—	2206	0	1	2	2
Ser/Thr protein kinase	—	0515	0	10	9	35
Ser/Thr protein Phosphatase 1	—	0631	0	3	3	4
Ser/Thr protein Phosphatase 2	—	0639	0	1	1	2
Adenylate cyclase	*cyaA*	2114	0	1	2	7
cAMP-binding domains	*crp*	0664	2	3	12	14
Phototaxis*						
Hybrid histidine kinase	*taxAY*	0643/2198/0784	0	3	3	3
Photoreceptor for positive phototaxis	*taxD*	2203/0840	0	3	3	3
Putative phototaxis protein	*taxP*	0784	0	3	3	3
CheW-like signal transducer	*taxW*	0835	0	3	3	3
CheY-like signal transducer	*taxY*	0784	0	3	3	3
Regulation of RNA polymerase						
Alternative σ28	*fliA*	1191	0	1	1	2
Alternative σ24	*rpoE*	1592	0	3	3	2
Anti-sigma factor	*rsbW*	2172	1	2	2	5
Serine phosphatase regulator of σ	*rsbU*	2208	1	2	5	5
Others						
Phycobilisome degradation proteins	*nblA*	—	0	1	2	2
	nblB	—	0	2	1	1
Universal stress protein	*uspA*	0589	0	4	4	7
Drought-induced stress protein	CDS P34	—	0	1	2	2
Ankyrin repeats	—	0666	0	1	1	2
Forkhead domains	—	1716	0	2	2	2

COG, clusters of orthologous groups.
*Gene designations after ref. 54.

subunits of phycoerythrin type III (PEIII) as the sole phycobiliprotein (32, 34). This also includes three bilin reductases, PcyA, PebA, and PebB (Pro0819, Pro1748–1749), for the biosynthesis of phycobiliprotein chromophores phycoerythrobilin and 3Z-phycocyanobilin (35). Phycoerythrobilin is bound by PEIII, whereas 3Z-phycocyanobilin, which is generated by the activity of PcyA, as shown for *Prochlorococcus* MED4 (35), serves as the chromophore for phycocyanin and allophycocyanin. Thus, questions arise as to which cognate polypeptide binds phycocyanobilin and why PEIII has been conserved in this genus given that the amount of PEIII per cell is very low (34) and light harvesting rests on Pcb proteins.

Chl biosynthesis. The set of Chl biosynthesis genes found in *P. marinus* SS120 is fairly complete (36). It has only one copy of *hemN* (Pro1385) encoding the oxygen-independent coproporphyrinogen III oxidase versus two in freshwater cyanobacteria (see supporting information). Similarly, SS120 has only one copy (compared with several copies in other cyanobacteria) of the *acsF/crd1* gene encoding an aerobic Mg-protoporphyrin IX monomethyl ester oxidative cyclase. Like *P. marinus* spp. MED4 and MIT9313 (32), SS120 has a gene (Pro0890) encoding a non-heme oxygenase with binding domains for a [2Fe-2S] Rieske center and for a mononuclear iron. These properties make it the most plausible candidate for chlorophyllide *a* oxygenase (Cao), an enzyme needed for the biosynthesis of chlorophyllide b_2 and therefore Chl b_2 (37) (see Fig. 1). However, Pro0890 is only distantly related to Caos previously identified in other Chl *b*-containing oxyphotobacteria, green algae and plants (38). Low-light-adapted strains such as SS120 or NATL1 can synthesize monovinyl Chl *b* (Chl b_1) when grown under high-light conditions (39), and Chl b_1 was assumed to derive from Chl b_2 (40). Thus, SS120 is likely to encode a 4-vinyl reductase, but no such enzyme has been found in the genome. Therefore, further studies are needed to elucidate the enzymes for the final phases of the biosynthesis of the specific divinyl-Chls of SS120.

Autotrophic Metabolism. ***Carbon assimilation.*** Despite the key role of carbon dioxide in autotrophic metabolism, the *P. marinus* SS120 genome does not encode any of the three principal CO_2/HCO_3^- uptake systems found in other cyanobacteria (41, 42): the ATP-binding cassette (ABC)-type bicarbonate transporter CmpABCD, the constitutive CO_2 uptake system NdhD4/NdhF4/CupB, or the low-CO_2 induced high-affinity system NdhD3/NdhF3/CupA. The only CO_2 uptake system found in SS120 is the $\Delta\mu Na^+$-dependent transporter SbtA (Pro0241). The *ntpJ* gene, whose product is

MICROBIOLOGY

reportedly required for the SbtA-catalyzed bicarbonate uptake (41), is also present in the SS120 genome (Pro0098). SS120 does not encode any known carbonic anhydrases, so their function is probably fulfilled by an as-yet-unidentified protein. *P. marinus* SS120 assimilates CO_2 via the Calvin cycle and has the complete set of enzymes of this pathway (Fig. 1). However, Rubisco (large and small subunits), phosphoribulokinase, and pentose-5-phosphate epimerase of SS120 all have closer homologs in proteobacteria, such as *Thiobacillus* species, than in freshwater cyanobacteria, a feature shared with other marine picocyanobacteria (32, 42).

Whereas all freshwater cyanobacteria encode two analogous fructose-1,6-bisphosphatases, related, respectively, to *Escherichia coli fbp* and *glpX* genes, *P. marinus* SS120 has only the latter one. This form (F-I) has been shown to hydrolyze both fructose-1,6-bisphosphate and sedoheptulose-1,7-bisphosphate (43). This is yet another case of gene economy in which a gene encoding a monofunctional enzyme (*fbp*, active only with the former substrate) could have been eliminated because a bifunctional enzyme was present.

Nitrogen assimilation. The *P. marinus* SS120 genome does not contain genes for transport systems for nitrate, nitrite, cyanate, and urea, which are present in freshwater cyanobacteria, nor that coding for a nitrate/nitrite permease recently discovered in a marine *Synechococcus* (44). Accordingly, it does not encode nitrate/nitrite reductases or urease. These results indicate that this strain relies for growth on reduced nitrogen compounds, such as NH_4^+ and amino acids. Indeed, SS120 encodes an ammonia transporter (Fig. 1). Field studies recently showed that *Prochlorococcus* can import amino acids (45). There are four unassigned ABC-type transport systems and several Na^+/amino acid symporters in SS120 that could provide that capability.

P. marinus SS120 does not possess enzymes for the synthesis or hydrolysis of cyanophycin (poly-L-arginyl-L-aspartate), which serves as nitrogen reserve in some cyanobacteria (see supporting information); however, it can conserve nitrogen by forming spermidine. Indeed, it possesses the *speE* gene encoding spermidine synthase (Pro1848), which is missing in the freshwater strains *Synechocystis* sp. PCC 6803 and *Anabaena* sp. PCC 7120.

Phosphorus assimilation. *P. marinus* SS120 encodes a typical ATP-dependent *Pst*CAB system (Pro0598–Pro0596) for transporting phosphate, as well as an ABC-type transporter for potential use of phosphonate (Pro1019–1021). The regulatory component PhoU is missing, as are PhoR and PhoB proteins, which form a two-component system responsible for phosphate sensing and regulation in a variety of bacteria. This apparent gene loss goes a step further than even in the *Prochlorococcus* MIT9313 genome, which contains an intact *phoB* gene and a frameshifted *phoR* gene (46).

Sulfur assimilation. The pathway of sulfate uptake and reduction in *P. marinus* SS120 is similar to that in other cyanobacteria, except that the ABC-type sulfate transporter is replaced by two sulfate permeases of the major facilitator superfamily (Pro0242 and Pro0804). SS120 encodes two copies of cysteine synthase, Pro0143 and Pro0403.

Intermediary Metabolism. ***Tricarboxylic acid cycle (TCA) and related reactions.*** Like many other bacteria, *P. marinus* SS120 encodes an incomplete citric acid cycle (47). The *sucA* and *sucB* genes coding for the E1 (dehydrogenase) and E2 (dihydrolipoamide succinyl transferase) components of the 2-oxoglutarate dehydrogenase complex are missing, as are genes for both subunits of succinyl-CoA synthetase. NAD-dependent malate dehydrogenase is also lacking in SS120; its function might be taken over by malate:quinone oxidoreductase (Pro0418), an enzyme not found in other cyanobacteria but one that is highly similar to the TCA enzyme from *Helicobacter pylori* (47). However, the *H. pylori* enzyme could not catalyze the reverse reaction, reduction of oxaloacetate, which is required for the functioning of the reductive branch of the incomplete TCA cycle in SS120. Therefore, it remains to be determined whether Pro0418 actually catalyzes this reaction in SS120.

Amino acid biosynthesis. *P. marinus* SS120 encodes the complete set of enzymes for biosynthesis of all amino acids except for lysine. Lysine biosynthesis in SS120 must occur anyway and possibly proceeds via the diaminopimelate pathway, although the mechanism of conversion of tetrahydrodipicolinate into diaminopimelate remains unclear. The pathway of methionine biosynthesis in SS120 includes three enzymes that are found in *E. coli* and other bacteria but seem to be missing in other cyanobacteria. Conversion of homoserine into homocysteine, catalyzed by these enzymes [homoserine transsuccinylase (MetA; Pro0801), cystathionine γ-synthase (MetB; Pro0405), and β-cystathionase (MetC; Pro0404)], can also be catalyzed by a combination of the homoserine *O*-acetyltransferase and *O*-acetylhomoserine-sulfhydrylase (Pro0800).

In yet another manifestation of gene economy, a dedicated tyrosine aminotransferase is missing, and the last step in the biosynthesis of Phe, Tyr, and Trp is apparently catalyzed by the nonspecific aromatic acid aminotransferase HisC.

Nucleotide biosynthesis. *P. marinus* SS120 encodes a complete set of enzymes for *de novo* purine and pyrimidine biosynthesis. In a rare deviation from its "minimal" gene content, SS120 encodes two variants of phosphoribosylglycinamide formyltransferase, folate-dependent PurN and formate-dependent PurT. Likewise, it encodes two versions of dihydroorotase that have been previously described in *E. coli* and *Bacillus subtilis*, respectively. The glutamine amidotransferase and pyrophosphatase domains of GMP synthase (GuaA), which form a single polypeptide chain in all bacteria, are encoded in SS120 by two separate genes, as is the case in most archaea. Like other cyanobacteria, SS120 encodes the catalytic subunit of aspartate carbamoyltransferase (PyrB) but not the regulatory subunit (PyrI).

In contrast to *de novo* biosynthesis pathways, purine and pyrimidine salvage pathways are poorly represented in SS120. Like other cyanobacteria, SS120 lacks classical thymidylate synthase ThyA and instead encodes the recently described alternative enzyme ThyX (48). Similar to some other cyanobacteria, SS120 does not encode either anaerobic ribonucleoside triphosphate reductase (NrdDG) or ribonucleoside diphosphate reductase (NrdAF). Instead, SS120 encodes the B_{12}-dependent (class-II) ribonucleotide reductase that is also found in *Anabaena* sp. PCC 7120 (49).

Cell wall biosynthesis. As a typical bacterium, SS120 encodes a complete set of enzymes for peptidoglycan synthesis, including MurABCDEFGI, alanine racemase, and D-ala-D-ala ligase.

Cofactor biosynthesis. SS120 encodes the full set of enzymes of the biosynthetic pathways for NAD, FAD, heme, B_{12}, biotin, folate, tetrahydrobiopterin, and phylloquinone. As in other cyanobacteria, the thiamine biosynthetic pathway lacks one enzyme, hydroxymethylpyrimidine/phosphomethylpyrimidine kinase (ThiD). Similarly, the set of *ubi* genes (which in cyanobacteria are likely involved in plastoquinone, not ubiquinone biosynthesis) is also incomplete, lacking *ubiB*. Thus, the reactions catalyzed in bacteria by ThiD and UbiB are likely performed in cyanobacteria by alternative enzymes yet to be characterized. The pathway of CoA biosynthesis lacks two enzymes, aspartate 1-decarboxylase (PanD) and panthothenate kinase (CoaA), whereas pyridoxine biosynthesis is represented only by pyridoxine synthase (PdxA and PdxJ subunits). The well known diversity of enzymes in these pathways (19) also suggests that these "missing" enzymes are encoded in the SS120 genome in alternative versions. In contrast, the genes for the biosynthesis of molybdenum cofactor are all missing, consistent with the absence of nitrate and nitrite reductases and other molybdopterin-containing enzymes.

Adaptation to the marine environment. *P. marinus* SS120 contains several systems that are likely critical in the adaptation to the marine

environment. Based on the sequence of the c subunit (AtpE) of its H^+-ATP synthetase (50), it appears that, in addition to (or instead of) H^+ ions, this enzyme can also transport Na^+. Unlike many marine bacteria, *P. marinus* SS120 does not encode a primary Na^+ pump. Nevertheless, it can extrude Na^+ ions at the expense of the proton gradient, using an NhaP-type Na^+/H^+ antiporter (Pro0470). The resulting sodium gradient is apparently used for export of Ca^{2+} ions via the Ca^{2+}/Na^+ antiporter (Pro0570). Other Na^+-dependent transporters in SS120 include Na^+/bicarbonate symporter SbtA (Pro0241), Na^+/proline symporter PutP (Pro1635), Na^+/alanine symporter AlsT (Pro1538), Na^+/glutamate symporter GltS (Pro0784), Na^+/galactoside symporter MelB (Pro1404), and two SS120-specific predicted Na^+-dependent permeases that belong, respectively, to the divalent anion:sodium symporter family (Pro0097) and the neurotransmitter:sodium symporter family (Pro1452). SS120 also encodes a homolog of a proteobacterial salt-induced outer membrane protein (Pro1529), which is absent in freshwater cyanobacteria.

Conclusions

We show here that *P. marinus* SS120, the type species of the dominant photosynthetic genus in the ocean, has a nearly minimal gene complement of an oxyphototrophic organism. The frugality of the gene repertoire of this organism is manifest at many levels. Signaling systems are either absent or represented by fewer domains in *P. marinus* SS120 than in other cyanobacteria, consistent with the fact that the oligotrophic marine environment where it preferentially thrives is much more stable than fresh waters. This argument is strengthened by the fact that, in the field, *P. marinus* SS120-like cells are restricted to the bottom part of the illuminated layer of oceans (5, 39). It appears likely that the compact genome of *P. marinus* SS120 is maintained by selection and is connected to the small cell volume of this organism ($\approx$0.1 μm^3), which is the theoretical lower limit for an oxyphototroph (51). Small cell size has at least two distinct advantages for a phytoplanktonic organism: (*i*) to increase the *in vivo* absorption coefficient by reducing the package effect (52) and (*ii*) to increase the cell surface to volume ratio and thereby improve nutrient uptake. Thus, *P. marinus* SS120 seems to be a rare case of a free-living organism for which a direct connection between fundamental characteristics of the genome and organism ecology is apparent.

Whether such a reduced genome is a derived state resulting from progressive gene loss or is an ancestral state is as yet unclear. Phylogenies based on 16S rRNA genes (see, for example, refs. 46 and 53) show that, within the *Prochlorococcus* radiation, SS120 is found in a "low-light clade" at an intermediate position between the "high-light clade," represented by MED4 (1) and another "low-light clade" containing MIT9313 (5) that is located near the base of the radiation. The MED4 strain has an even more compact genome than SS120 (1.66 vs. 1.75 Mbp, respectively), whereas that of MIT9313 is larger (2.41 Mbp) (6). The lower diversity within the high-light clade suggests that it has appeared more recently than the more highly divergent low-light clades (46, 53). Thus, evolution in the genus *Prochlorococcus* would have tended toward genome reduction. However, this phenomenon would certainly not be enough to account for the large differences in genome sizes and complexity between marine *P. marinus* SS120 and the freshwater cyanobacteria strains used as references in this paper. Specific genome amplification and diversification also must have taken place during adaptation of the latter to their specific environments. Confirmation of these hypotheses still awaits phylogenetic analysis of large gene regions, but these will be reliable only when more complete cyanobacterial genomes become available.

We thank D. Bhaya for helpful hints about motility and phototaxis genes. This work was supported by the European Union program MARGENES (QLRT-2001-01226) and Genomer (Région Bretagne). A.D. is supported by a doctoral fellowship from Région Bretagne, W.R.H. is supported by Deutsche Forschungsgemeinschaft Grant SFB 429-TPA4, and D.J.S. is a Royal Society University research fellow.

1. Chisholm, S. W., Frankel, S. L., Goericke, R., Olson, R. J., Palenik, B., Waterbury, J. B., West-Johnsrud, L. & Zettler, E. R. (1992) *Arch. Microbiol.* **157,** 297–300.
2. Partensky, F., Hess, W. R. & Vaulot, D. (1999) *Microbiol. Mol. Biol. Rev.* **63,** 106–127.
3. Goericke, R. & Repeta, D. J. (1992) *Limnol. Oceanogr.* **37,** 425–433.
4. LaRoche, J., van der Staay, G. W., Partensky, F., Ducret, A., Aebersold, R., Li, R., Golden, S. S., Hiller, R. G., Wrench, P. M., Larkum, A. W. *et al.* (1996) *Proc. Natl. Acad. Sci. USA* **93,** 15244–15248.
5. Moore, L. R. & Chisholm, S. W. (1999) *Limnol. Oceanogr.* **44,** 628–638.
6. Rocap, G., Larimer, F. W., Lamerdin, J., Malfatti, S., Chain, P., Ahlgren, N. A., Arellano, A., Coleman, M., Hauser, L., Hess, W. R., *et al.* (2003) *Nature*, 10.1038/nature01947.
7. Kaneko, T., Nakamura, Y., Wolk, C. P., Kuritz, T., Sasamoto, S., Watanabe, A., Iriguchi, M., Ishikawa, A., Kawashima, K., Kimura, T. *et al.* (2001) *DNA Res.* **8,** 205–213.
8. Kaneko, T., Sato, S., Kotani, H., Tanaka, A., Asamizu, E., Nakamura, Y., Miyajima, N., Hirosawa, M., Sugiura, M., Sasamoto, S. *et al.* (1996) *DNA Res.* **3,** 109–136.
9. Nakamura, Y., Kaneko, T., Sato, S., Ikeuchi, M., Katoh, H., Sasamoto, S., Watanabe, A., Iriguchi, M., Kawashima, K., Kimura, T. *et al.* (2002) *DNA Res.* **9,** 123–130.
10. Artiguenave, F., Wincker, P., Brottier, P., Duprat, S., Jovelin, F., Scarpelli, C., Verdier, J., Vico, V., Weissenbach, J. & Saurin, W. (2000) *FEBS Lett.* **487,** 13–16.
11. Strehl, B., Holtzendorff, J., Partensky, F. & Hess, W. R. (1999) *FEMS Microbiol. Lett.* **181,** 261–266.
12. Delcher, A. L., Harmon, D., Kasif, S., White, O. & Salzberg, S. L. (1999) *Nucleic Acids Res.* **27,** 4636–4641.
13. Besemer, J., Lomsadze, A. & Borodovsky, M. (2001) *Nucleic Acids Res.* **29,** 2607–2618.
14. Badger, J. H. & Olsen, G. J. (1999) *Mol. Biol. Evol.* **16,** 512–524.
15. Lowe, T. M. & Eddy, S. R. (1997) *Nucleic Acids Res.* **25,** 955–964.
16. Vogel, J., Axmann, I. M., Herzel, H. & Hess, W. R. (2003) *Nucleic Acids Res.* **31,** 2890–2899.
17. Tatusov, R. L., Galperin, M. Y., Natale, D. A. & Koonin, E. V. (2000) *Nucleic Acids Res.* **28,** 33–36.
18. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zheng, Z., Miller, W. & Lipman, D. J. (1997) *Nucleic Acids Res.* **25,** 3389–3402.
19. Koonin, E. V. & Galperin, M. Y. (2002) *Sequence–Evolution–Function: Computational Approaches in Comparative Genomics* (Kluwer, Boston).
20. Lespinet, O., Wolf, Y. I., Koonin, E. V. & Aravind, L. (2002) *Genome Res.* **12,** 1048–1059.
21. Liu, Y. & Tsinoremas, N. F. (1996) *Gene* **172,** 105–109.
22. Bhaya, D., Dufresne, A., Vaulot, D. & Grossman, A. (2002) *FEMS Microbiol. Lett.* **215,** 209–219.
23. Imamura, S., Yoshihara, S., Nakano, S., Shiozaki, N., Yamada, A., Tanaka, K., Takahashi, H., Asayama, M. & Shirai, M. (2003) *J. Mol. Biol.* **325,** 857–872.
24. Fraser, C. M., Gocayne, J. D., White, O., Adams, M. D., Clayton, R. A., Fleischmann, R. D., Bult, C. J., Kerlavage, A. R., Sutton, G., Kelley, J. M. *et al.* (1995) *Science* **270,** 397–403.
25. Mizuno, T., Kaneko, T. & Tabata, S. (1996) *DNA Res.* **3,** 407–414.
26. Ohmori, M., Ikeuchi, M., Sato, N., Wolk, P., Kaneko, T., Ogawa, T., Kanehisa, M., Goto, S., Kawashima, S., Okamoto, S. *et al.* (2001) *DNA Res.* **8,** 271–284.
27. Meeks, J. C., Campbell, E. L., Summers, M. L. & Wong, F. C. (2002) *Arch. Microbiol.* **178,** 395–403.
28. Galperin, M. Y., Nikolskaya, A. N. & Koonin, E. V. (2001) *FEMS Microbiol. Lett.* **203,** 11–21.
29. Galperin, M. Y., Walker, D. R. & Koonin, E. V. (1998) *Genome Res.* **8,** 779–790.
30. Schluchter, W. M. & Bryant, D. A. (1992) *Biochemistry* **31,** 3092–3102.
31. Fillat, M. F., Flores, E. & Gomez-Moreno, C. (1993) *Plant Mol. Biol.* **22,** 725–729.
32. Hess, W. R., Rocap, G., Ting, C. S., Larimer, F., Stilwagen, S., Lamerdin, J. & Chisholm, S. W. (2001) *Photosynth. Res.* **70,** 53–71.
33. Garczarek, L., Hess, W. R., Holtzendorff, J., van der Staay, G. W. & Partensky, F. (2000) *Proc. Natl. Acad. Sci. USA* **97,** 4098–4101.
34. Hess, W. R., Steglich, C., Lichtle, C. & Partensky, F. (1999) *Plant Mol. Biol.* **40,** 507–521.
35. Frankenberg, N., Mukougawa, K., Kohchi, T. & Lagarias, J. C. (2001) *Plant Cell* **13,** 965–978.
36. Suzuki, J. Y., Bollivar, D. W. & Bauer, C. E. (1997) *Annu. Rev. Genet.* **31,** 61–89.
37. Oster, U., Tanaka, R., Tanaka, A. & Rudiger, W. (2000) *Plant J.* **21,** 305–310.
38. Tanaka, A., Ito, H., Tanaka, R., Tanaka, N. K., Yoshida, K. & Okada, K. (1998) *Proc. Natl. Acad. Sci. USA* **95,** 12719–12723.
39. Moore, L. R., Goericke, R. & Chisholm, S. W. (1995) *Mar. Ecol. Prog. Ser.* **116,** 259–275.
40. Partensky, F. & Garczarek, L. (2003) in *Photosynthesis in Algae*, ed. Larkum, A. W. D. (Kluwer, Dordrecht, The Netherlands), Vol. 14, in press.
41. Shibata, M., Ohkawa, H., Katoh, H., Shimoyama, M. & Ogawa, T. (2002) *Funct. Plant Biol.* **29,** 123–129.
42. Badger, M. R. & Price, G. D. (2003) *J. Exp. Bot.* **54,** 609–622.
43. Tamoi, M., Murakami, A., Takeda, T. & Shigeoka, S. (1998) *Biochim. Biophys. Acta* **1383,** 232–244.
44. Sakamoto, T., Inoue-Sakamoto, K. & Bryant, D. A. (1999) *J. Bacteriol.* **181,** 7363–7372.
45. Zubkov, M. V., Fuchs, B. M., Tarran, G. A., Burkill, P. H. & Amann, R. (2003) *Appl. Environ. Microbiol.* **69,** 1299–1304.
46. Scanlan, D. J. & West, N. J. (2002) *FEMS Microbiol. Ecol.* **40,** 1–12.
47. Kather, B., Stingl, K., van der Rest, M. E., Altendorf, K. & Molenaar, D. (2000) *J. Bacteriol.* **182,** 3204–3209.
48. Myllykallio, H., Lipowski, G., Leduc, D., Filee, J., Forterre, P. & Liebl, U. (2002) *Science* **297,** 105–107.
49. Gleason, F. K. & Olszewski, N. E. (2002) *J. Bacteriol.* **184,** 6544–6550.
50. Dzioba, J., Hase, C. C., Gosink, K., Galperin, M. Y. & Dibrov, P. (2003) *J. Bacteriol.* **185,** 674–678.
51. Raven, J. A. (1994) *J. Plankton Res.* **16,** 565–580.
52. Morel, A., Ahn, Y.-W., Partensky, F., Vaulot, D. & Claustre, H. (1993) *J. Mar. Res.* **51,** 617–649.
53. Urbach, E., Scanlan, D. J., Distel, D. L., Waterbury, J. B. & Chisholm, S. W. (1998) *J. Mol. Evol.* **46,** 188–201.
54. Bhaya, D., Takahashi, A. & Grossman, A. R. (2001) *Proc. Natl. Acad. Sci. USA* **98,** 7540–7545.

CHAPITRE III

Comparaison des Répertoires de Gènes et Adaptation à la Niche Écologique

III.1 Introduction

La quantité et de la qualité de la lumière sont des facteurs écologiques primordiaux pour les organismes phytoplanctoniques. L'absorption des longueurs d'ondes par les molécules d'eau, la matière organique et les organismes phytoplanctoniques eux-mêmes entraînent une modification de la composition spectrale et surtout de l'intensité du rayonnement lumineux en fonction de la profondeur (Kirk 1994). Ainsi on observe des gradients verticaux de lumière dans la couche euphotique des océans, couche dont la profondeur peut varier de 30-50 m en zone eutrophe ou mésotrophe (zones côtières ou d'upwelling) à plus de 150 m en zone oligotrophe (zones pélagiques intertropicales). Ces gradients de lumière sont souvent associés à des gradients opposés de sels nutritifs ainsi qu'à des gradients physico-chimiques (de température, de salinité, etc.), ces derniers pouvant constituer de véritables barrières physiques pour les populations phytoplanctoniques. Ainsi, malgré l'apparente continuité de l'écosystème océanique, on constate que celui-ci peut abriter des niches écologiques distinctes, propices à l'apparition "d'écotypes" (c'est-à-dire de souches d'une même espèce occupant la même niche écologique ou des niches très similaires), voire à des processus de spéciation.

Ainsi, les organismes phytoplanctoniques présents en surface doivent faire face à des intensités lumineuses pouvant être très élevées (jusqu'à 2 000 µmol photon m^{-2} s^{-1}). De plus, ils sont soumis à des doses significatives de rayonnement UV (Helbling et al. 2001) ayant d'éventuels effets mutagènes et/ou cytotoxiques (Ravanat et al. 2001). Au contraire, les organismes phytoplanctoniques vivant en profondeur doivent s'adapter à la faible quantité de lumière disponible. Le gradient de lumière est donc extrêmement contraignant et impose à ces organismes l'existence de mécanismes moléculaires, physiologiques et cellulaires leur permettant de faire face à ces contraintes différentes.

Les cyanobactéries marines *Prochlorococcus* et *Synechococcus* représentent la grande majorité de la fraction procaryotique du phytoplancton océanique. Les analyses moléculaires, phylogénétiques et physiologiques montrent que la structure des populations de *Prochlorococcus* reflète bien le gradient de lumière vertical (Moore et al. 1998; Rocap et al. 2002; West and Scanlan 1999; West et al. 2001). Ainsi, il semble que la large distribution verticale de *Prochlorococcus* soit due à l'existence d'au moins deux écotypes au sein de ce genre. Le premier est présent en surface dans une niche dite « de forte lumière » (0-100 m de profondeur). Le second se développe à plus grande profondeur dans une niche « de faible lumière »

La disponibilité de plusieurs génomes de *Synechococcus* et de *Prochlorococcus* offre la possibilité d'appréhender le degré de corrélation entre la diversité des génomes de ces

cyanobactéries et les caractéristiques des niches écologiques (de basse ou de forte lumière). Dans ce chapitre, nous nous sommes focalisés sur l'identification des gènes responsables de l'adaptation à ces niches. La détermination des gènes potentiellement impliqués dans l'adaptation à la niche de forte lumière est basée sur deux hypothèses. *P. marinus* MED4 et *Synechococcus* spp. WH8102 et WH7803 sont toutes les trois capables de pousser aux fortes intensités lumineuses (Six et al. 2004) et on peut donc s'attendre à ce que les génomes de ces trois cyanobactéries possèdent des gènes leur permettant de faire face au stress lumineux (visible et UV). De plus, le génome de *P. marinus* MED4 est réduit par rapport à celui des deux *Synechococcus*. Ainsi, le fait qu'un gène présent chez les deux souches de *Synechococcus* ait été conservé uniquement par *P. marinus* MED4 constitue un bon indice de l'importance de ces gènes dans l'adaptation à la niche forte lumière.

De même, les gènes spécifiques de *P. marinus*. SS120 et *Prochlorococcus* sp. MIT9313 peuvent être considérés comme potentiellement impliqués dans l'adaptation à la niche basse lumière. Là encore, le fait que *P. marinus* SS120 ait un génome réduit est un avantage car la préservation dans son génome d'un gène qui aurait été éliminé au cours de l'évolution des génomes de picocyanobactéries vivant en surface peut logiquement laisser supposer que ce gène apporte un avantage adaptatif significatif dans la niche de faible lumière.

III.2 Méthodes d'analyse.

Les séquences de protéines des cinq picocyanobactéries marines ont été extraites des fichiers d'annotation des génomes. Dans le cas de *Prochlorococcus marinus* MED4 (NC_005072), *Prochlorococcus* sp. MIT9313 (NC_005071) et *Synechococcus* sp. WH8102 (NC_005070), ces fichiers ont été téléchargés depuis la section “Genome” du système “ENTREZ” du NCBI. Les caractéristiques des cinq génomes sont résumées dans l'annexe IV. Les cinq jeux de séquences ont ensuite été concaténés dans un même fichier contenant au final 10 970 séquences de protéines.

Les régions de faible complexité ont été masquées dans les séquences de protéines en utilisant l'algorithme CAST. Toutes les protéines ont été comparées contre elles-mêmes avec le programme BlastP. Le programme TribeMCL a été utilisé pour regrouper les protéines en cluster à partir des résultats des recherches de similarité. La classification en clusters dépend fortement de la valeur seuil utilisée pour considérer que les séquences sont réellement similaires et que la ressemblance trouvée n'est pas due au hasard. Une valeur seuil trop stricte donnera un nombre de clusters trop réduit alors qu'une valeur seuil plus élevée risque de regrouper ensemble des séquences non-homologues. Aussi, l'étape de comparaison a été

répétée trois fois en utilisant, à chaque fois, une valeur seuil différente (e-value égale à 10^{-10}, 10^{-5} et 10^{-3}).

Tableau III-1

Exemple de cluster obtenu avec TribeMCL. PMM, *P. marinus* MED4; Pro, *P. marinus* SS120; PMT, *Prochlococococcus* sp. MIT9313, SYNW, *Synechococcus* sp. WH8102; ORF, *Synechococcus* sp. WH7803, Chl; Chlorophylle.

						Prochlorococcus			***Synechococcus***	
N° cluster	**ORF**	**Produit**	**Taillle (aa)**	**Nb de séquences / cluster**	**Nb de génomes / cluster**	**PMM**	**Pro**	**PMT**	**SYNW**	**ORF**
19	PMM0627	Chl *a*/*b* binding light-harvesting antenna protein	353	11	3	1	8	2	0	0
19	PMT0496	Chl *a*/*b* binding light-harvesting antenna protein	369	11	3	1	8	2	0	0
19	PMT1046	Chl *a*/*b* binding light-harvesting antenna protein	351	11	3	1	8	2	0	0
19	Pro0783	Chl *a*/*b* binding light harvesting antenna protein PcbA	352	11	3	1	8	2	0	0
19	Pro0885	Chl *a*/*b* binding light harvesting antenna protein PcbC	352	11	3	1	8	2	0	0
19	Pro0892	Chl *a*/*b* binding light harvesting antenna protein PcbG	354	11	3	1	8	2	0	0
19	Pro1167	Chll *a*/*b* binding light harvesting antenna protein PcbD	362	11	3	1	8	2	0	0
19	Pro1169	Chl *a*/*b* binding light harvesting antenna protein PcbB	350	11	3	1	8	2	0	0
19	Pro1174	Chl *a*/*b* binding light harvesting antenna protein	353	11	3	1	8	2	0	0
19	Pro1288	Chl *a*/*b* binding light harvesting antenna protein PcbF	355	11	3	1	8	2	0	0
19	Pro1450	Chl *a*/*b* binding light harvesting antenna protein PcbE	362	11	3	1	8	2	0	0

Les résultats du clustering ont été formatés en utilisant un script Perl afin de compter le nombre de gènes par cluster, le nombre de génomes représentés dans chaque cluster, et le nombre de gènes par génome et par cluster (voir l'exemple du tableau III-1). Les résultats du clustering ont ensuite été insérés dans une base de données MySQL pour identifier rapidement les gènes spécifiques de chaque écotype.

Parmi les séquences identifiées, celles sans fonction connue ont été comparées, avec PSI-BLAST, contre la base de séquences protéiques "nr" du NCBI afin de rechercher des séquences homologues dans d'autres génomes de bactéries et ainsi obtenir des informations supplémentaires (par exemple sur la fonction biochimique) sur ces séquences.

III.3 Résultats et Discussion.

III.3.1 Classification en clusters de protéines

L'inspection des résultats obtenus avec trois valeurs seuils différentes a montré qu'une valeur de "e-value" égale à 10^{-5} constitue un compromis relativement bon entre sensibilité et spécificité. Avec une valeur de 10^{-10}, plusieurs familles de gènes, homologues mais avec un niveau de similarité peu élevé, se retrouvent séparées en plusieurs clusters. Au contraire, avec une valeur de 0,001, un certain nombre de clusters apparaissent formés par le regroupement artificiel de plusieurs familles.

La valeur choisie (10^{-5}) ne permet, cependant pas, d'identifier correctement toutes les familles de gènes. Plusieurs gènes codant pour des protéines du photosystème I et II (psbL, psbX-Z, psaM) ont des tailles très petites (inférieures à 200 nucléotides) et sont classés dans des clusters séparés. De même, d'autres gènes de petite taille ont été regroupés avec des gènes beaucoup plus grands. Ces petits gènes présentent un niveau de conservation extrêmement fort avec une toute petite partie des séquences des gènes plus grands. La très grande majorité de ces petits gènes se trouvent chez *Prochlorococcus* sp. MIT9313 et chez les deux *Synechococcus*. Il pourrait s'agir pour certains de fragments de pseudogènes dont le reste de la séquence a été éliminée ou a tellement évoluée qu'elle n'est plus distinguable des régions intergéniques environnantes. Cette hypothèse est renforcée par le fait que, dans certains cas, ces gènes sont groupés côte à côte sur la séquence du génome et présentent le même environnement génique que les gènes entiers avec lesquels ils ont été associés. Ces pseudogènes potentiels ont été placés dans des clusters différents.

Au total, et après un ajustement manuel (46 clusters ont été séparés en plusieurs clusters différents), 4231 clusters ont été obtenus. Ces clusters correspondent plus ou moins

à des familles de gènes. Dans certains cas, ils peuvent être assimilés à des superfamilles. Ainsi, les deux plus grands clusters (84 et 41 gènes) regroupent la majorité des gènes appartenant à la superfamille des transporteurs ABC.

Pour chaque génome, une très grande majorité des clusters obtenus (~90%) ne contiennent qu'un seul gène. Cela indique un faible pourcentage de gènes paralogues dans ces génomes. Le pourcentage de gènes paralogues est ainsi bien plus faible que celui estimé chez *Escherichia coli* (30%) (Blattner et al. 1997) ou *Streptomyces coelicolor* (50%) (Pushker et al. 2004). Le cluster contenant le plus grand nombre de gènes paralogues d'un seul génome (cluster 1, voir figure III-1), contient 31 gènes de *Prochlorococcus* sp. MIT9313 et deux gènes de *Synechococcus* sp. WH7803.

Figure III-1

Alignement des séquences protéiques du cluster 1 et des séquences homologues à celles de ce cluster. Symboles: *, cluster 1; +, cluster 1360; **“**, cluster 1434; ¤, cluster 1564; #, cluster 1828.

Chez *Prochlorococcus* sp. MIT9313, ces gènes ont des séquences extrêmement conservées entre elles et sont répartis en plusieurs groupes le long de la séquence du génome. Cela indique que cette famille de gènes a été amplifiée spécifiquement, et relativement récemment, chez cette souche. La recherche de séquences similaires avec PSI-BLAST montre que ces gènes, dont la fonction est totalement inconnue, ne sont pas spécifiques de *Prochlorococcus* sp. MIT9313 et de *Synechococcus* sp. WH7803. Les trois autres génomes de picocyanobactéries marines contiennent au moins un homologue (plus ou moins distant) de ces gènes. De même, deux gènes homologues sont présents chez *Anabaena* sp. PCC 7120.

III.3.2 Distribution des gènes dans les cinq génomes.

Dans les trois génomes de *Prochlorococcus*, la majorité des gènes appartiennent à des clusters qui sont représentés chez les cinq picocyanobactéries marines (Fig. III-2). La proportion de gènes communs aux cinq génomes est maximale chez *P. marinus* MED4 (68%) et minimale chez *Synechococcus* spp. WH8102 et WH7803 (49%). Il faut noter que pourcentage est identique chez les deux *Synechococcus*, alors que le génome de *Synechococcus* sp. WH7803 est plus petit que celui de *Synechococcus* sp. WH8102. Ceci est vraisemblablement du au fait que *Synechococcus* sp. WH7803 a une densité génique plus forte que *Synechococcus* sp. WH8102 et un nombre d'ORFs prédits très similaire (2572 et 2526, respectivement).

La comparaison du nombre de gènes entre les trois génomes de *Prochlorococcus* suggère que la taille du génome s'est fortement réduit au cours de l'évolution de ce genre (Dufresne et al., 2003; Rocap et al., 2003; Hess, 2004). La classification en familles de gènes montre l'existence d'un noyau de gènes très peu affecté par cette diminution de la taille du génome.

Les gènes communs aux cinq génomes se répartissent dans 1148 familles distinctes. La majorité d'entre elles (94%) ne contiennent qu'un seul gène de chaque génome. Le nombre de familles contenant au moins deux gènes de chaque génome est extrêmement réduit (seulement 22). Le noyau de ces 1148 groupes est donc formé de gènes présents en exemplaire unique dans les cinq génomes. De plus, la plupart de ces gènes ont une fonction bien définie. Ce noyau de gènes contient, par exemple, ceux codant pour la majorité des sous-unités des photosystèmes I et II ainsi que ceux codant pour les enzymes de la voie de biosynthèse des chlorophylles.

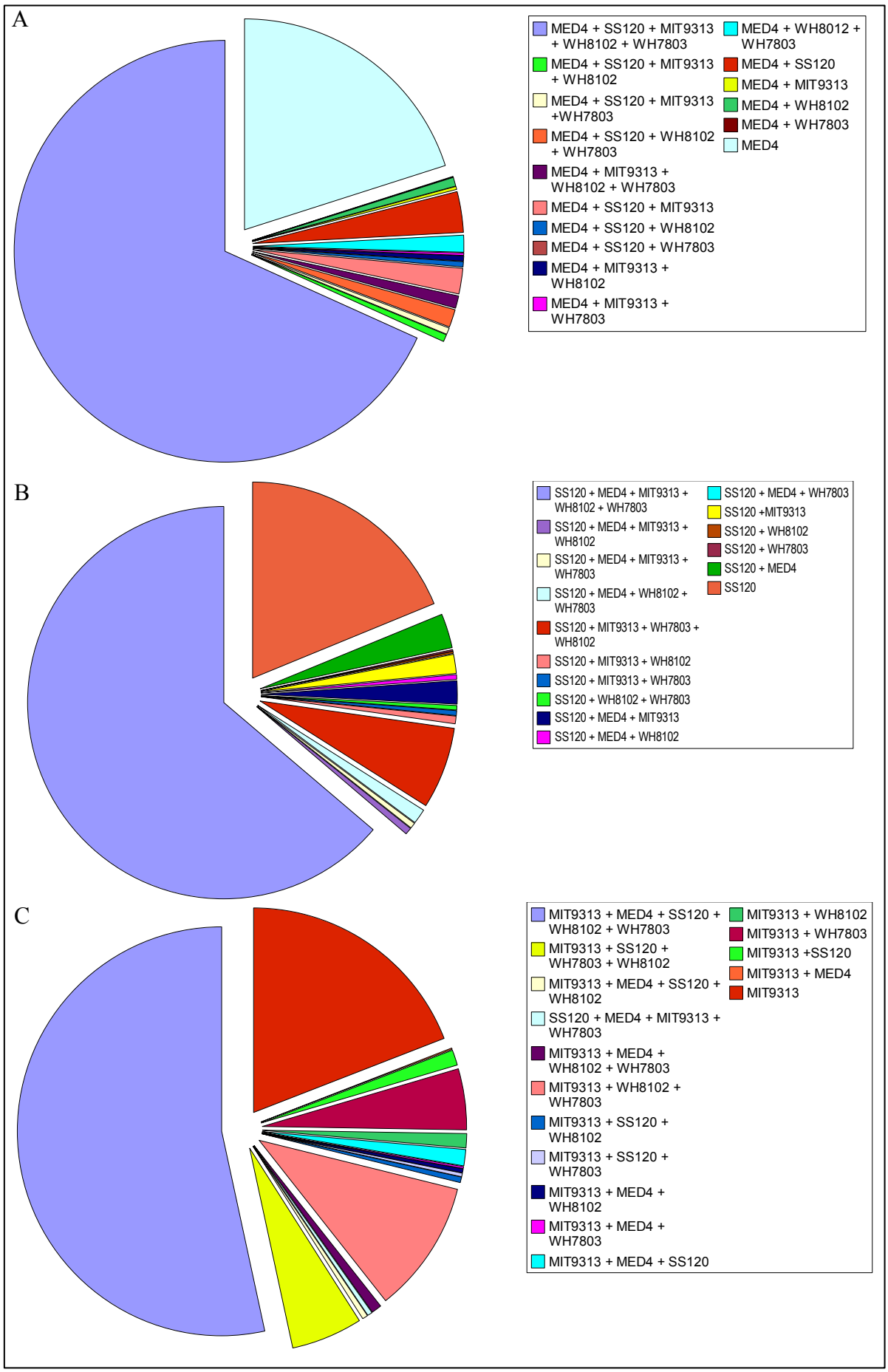
A
MED4 + SS120 + MIT9313 + WH8102 + WH7803
MED4 + SS120 + MIT9313 + WH8102
MED4 + SS120 + MIT9313 +WH7803
MED4 + SS120 + WH8102 + WH7803
MED4 + MIT9313 + WH8102 + WH7803
MED4 + SS120 + MIT9313
MED4 + SS120 + WH8102
MED4 + SS120 + WH7803
MED4 + MIT9313 + WH8102
MED4 + MIT9313 + WH7803
MED4 + WH8012 + WH7803
MED4 + SS120
MED4 + MIT9313
MED4 + WH8102
MED4 + WH7803
MED4
B
SS120 + MED4 + MIT9313 + WH8102 + WH7803
SS120 + MED4 + MIT9313 + WH8102
SS120 + MED4 + MIT9313 + WH7803
SS120 + MED4 + WH8102 + WH7803
SS120 + MIT9313 + WH7803 + WH8102
SS120 + MIT9313 + WH8102
SS120 + MIT9313 + WH7803
SS120 + WH8102 + WH7803
SS120 + MED4 + MIT9313
SS120 + MED4 + WH8102
SS120 + MED4 + WH7803
SS120 +MIT9313
SS120 + WH8102
SS120 + WH7803
SS120 + MED4
SS120
C
MIT9313 + MED4 + SS120 + WH8102 + WH7803
MIT9313 + SS120 + WH7803 + WH8102
MIT9313 + MED4 + SS120 + WH8102
SS120 + MED4 + MIT9313 + WH7803
MIT9313 + MED4 + WH8102 + WH7803
MIT9313 + WH8102 + WH7803
MIT9313 + SS120 + WH8102
MIT9313 + SS120 + WH7803
MIT9313 + MED4 + WH8102
MIT9313 + MED4 + WH7803
MIT9313 + MED4 + SS120
MIT9313 + WH8102
MIT9313 + WH7803
MIT9313 +SS120
MIT9313 + MED4
MIT9313

Figure III-2

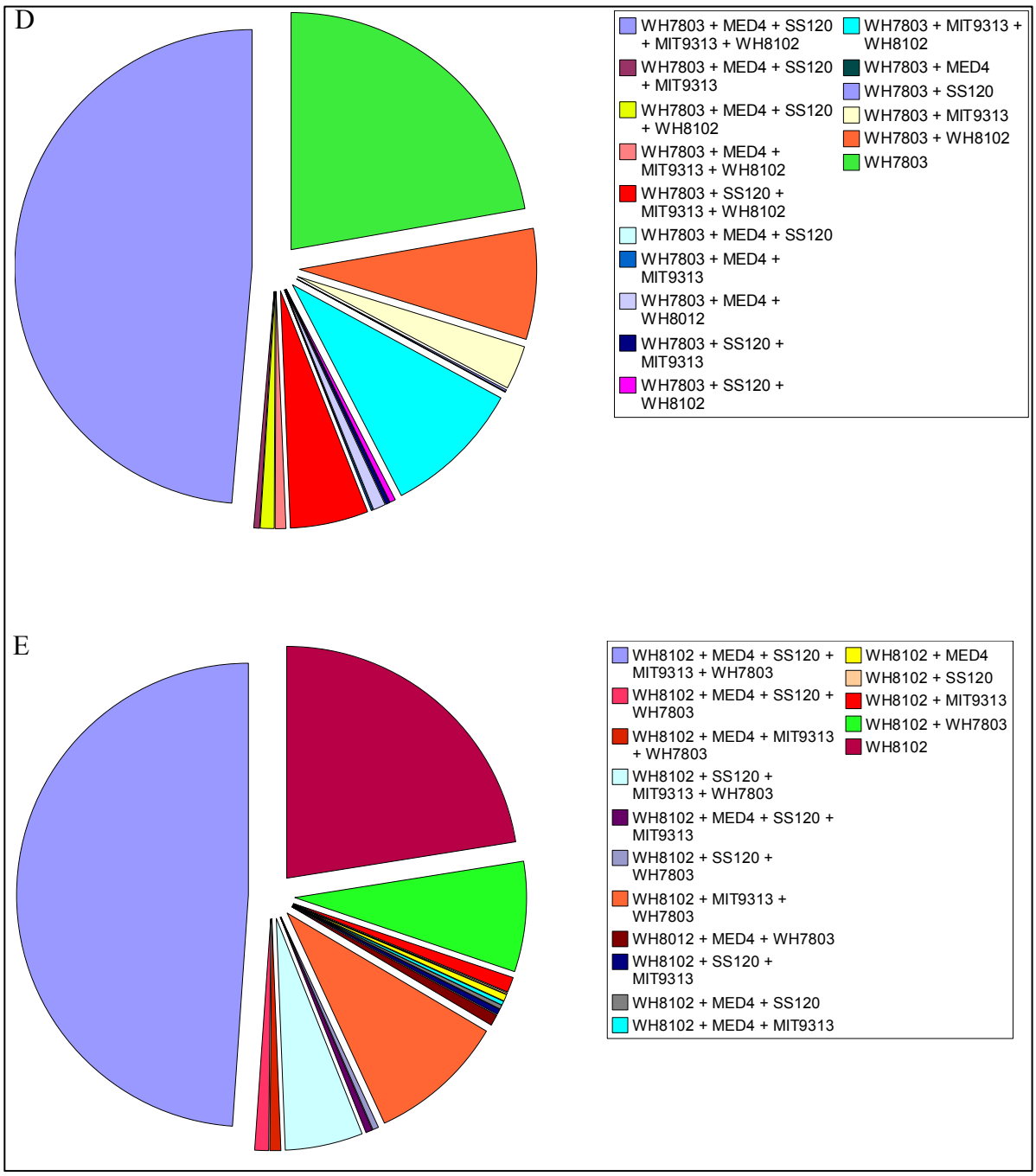

Figure III-2 (suite)

Pourcentage de gènes spécifiques de chaque picocyanobactérie marine ou communs à cinq, quatre, trois, ou seulement deux génomes. A, *P. marinus* MED4; B, *P. marinus* SS120; C, *Prochlorococcus* sp. MIT9313; D, *Synechococcus* sp. WH7803; E, *Synechococcus* sp. WH8102.

Ce noyau n'est pas spécifique des cyanobactéries puisqu'il contient aussi les gènes des voies de biosynthèse des acides aminés, des nucléotides, des lipides, des cofacteurs enzymatiques ainsi que ceux des systèmes de réplication, de transcription et de traduction.

La nature des gènes de ce noyau et le fait qu'ils soient présents en simple copie explique pourquoi ils ont été conservés au cours de l'évolution du genre *Prochlorococcus* vers un génome réduit. En effet, le mode de vie libre et autotrophe de *Prochlorococcus* l'oblige à posséder un répertoire de gènes le plus complet possible pour la synthèse des composés du métabolisme intermédiaire.

La fraction de gènes spécifiques d'un seul génome est la plus importante après celle des gènes communs au cinq génomes (Fig. III-2). Cette fraction représente, à peu près, le même pourcentage de gènes dans les cinq génomes (~ 21%) et par conséquent, le nombre de gènes spécifiques varie fortement entre le plus petit génome (*P. marinus* MED4, 359 gènes) et le plus grand (*Synechococcus* sp. WH8102, 572 gènes). Comme cela a été observé pour la plupart des autres génomes de bactéries, la plupart de ces gènes n'ont pas de fonction connue. Enfin, il est assez probable que la fraction de gènes spécifiques de chaque génome soit surestimée. En effet, un grand nombre de ces gènes ont une petite taille (< 450 nucléotides). Par ailleurs, il est possible que, comme dans le cas des petits gènes photosynthétiques, il existe d'autres gènes homologues qui ont été placés dans des clusters différents à cause de leurs petites tailles et de leurs séquences peu conservées.

Les fractions qui diminuent le plus avec la taille des génomes sont celles contenant les gènes communs à deux, trois et quatre génomes (Fig. III-2). Chez *Prochlorococcus* sp. MIT9313, *Synechococcus* spp. WH8102 et WH7803, la somme de ces gènes communs représente un pourcentage plus important que celui des gènes spécifiques. Par contre, ces gènes occupent une part plus réduite dans les génomes de *P. marinus* MED4 et SS120. Cette observation est cohérente avec l'hypothèse d'évolution réductive du génome chez *Prochlorococcus* et suggère que le phénomène de perte de gènes touche essentiellement les gènes situés dans ces catégories.

III.3.3 Gènes de la niche de forte lumière.

Etonnamment, les clusters ne contenant que des gènes de *P. marinus* MED4, *Synechococcus* sp. WH8102 et *Synechococcus* sp. WH7803 sont très peu nombreux puisque 20 clusters seulement correspondent à ce critère (Tableau III-2). Ces clusters représentent 21 gènes chez *P. marinus* MED4 et chez *Synechococcus* sp. WH8102 et 20 gènes chez *Synechococcus* sp. WH7803.

Environ la moitié de ces gènes n'ont pas de fonction connue. Cependant il est intéressant de noter que parmi les gènes sans fonction connue, près de la moitié code pour des protéines de petite taille (80 à 180 aa), caractérisées par la présence d'une ou deux hélices transmembranaires. Ces protéines pourraient donc être localisées soit dans la

membrane plasmique, soit dans la membrane thylacoïdale. Cela est particulièrement intéressant puisque les membranes constituent une des premières barrières de défense contre l'excès de lumière.

Tableau III-2

Gènes spécifiques de la niche de forte lumière. MED4, *P. marinus* MED4; WH7803, *Synechococcus* sp. WH7803; WH8102, *Synechococcus* sp. WH8102.

Cluster	Produit	MED4	WH7803	WH8102
491	Conserved hypothetical protein	PMM0872, PMM1028	ORF0838	SYNW1060, SYNW1942
1409	Deoxyribodipyrimidine photolyase	PMM0285	ORF0264	SYNW0219
1462	Carotenoid isomerase / phytoene dehydrogenase	PMM0339	ORF0809	SYNW0761
1463	Conserved hypothetical protein	PMM1642	ORF0829	SYNW1462
1464	Predicted membrane protein	PMM1359	ORF0830	SYNW1452
1486	Predicted membrane protein	PMM1420	ORF1022	SYNW0980
1490	Putative arsenite transporter, ACR3 family	PMM0716	ORF1088	SYNW1039
1492	Conserved hypothetical protein with a glutathione synthetase ATP-binding domain-like (SSF56059)	PMM1038	ORF1099	SYNW2456
1495	Conserved hypothetical protein	PMM1174	ORF1109	SYNW3653n
1497*	Mg-protoporphyrin IX monomethyl ester oxidative cyclase	PMM0844	ORF1115	SYNW1198
1518	Putative ATP-dependent helicase	PMM0728	ORF1203	SYNW1322
1519	ATP-dependent DNA ligase	PMM0729	ORF1209	SYNW1321
1520	Predicted exonuclease of the beta-lactamase fold involved in RNA processing (COG1236) / Predicted metal-dependent RNase, (COG1782)	PMM0730	ORF1211	SYNW1317
1533	Deoxyribodipyrimidine photolyase-related protein	PMM0425	ORF1286	SYNW1244
1534	Dehydrogenase with different specificities (COG1028)	PMM0414	ORF1296	SYNW1234
1540	Conserved hypothetical protein	PMM1106	ORF1386	SYNW0911
1541	Predicted membrane protein	PMM1037	ORF1406	SYNW0803
1585	Predicted membrane protein	PMM1015	ORF1848	SYNW1817
1586	Predicted membrane protein	PMM0476	ORF1856	SYNW1823
1643	Cyanate lyase	PMM0373	ORF2536	SYNW2490

* également présent chez *P. marinus* SS120 et *Prochlorococcus* sp. MIT9313

Parmi les gènes codant pour une protéine dont la fonction est bien caractérisée, on peut remarquer la présence d'une hélicase ATP-dépendante et d'une ligase ATP-dépendante. Ces deux protéines pourraient servir à la réparation des lésions induites par le fort rayonnement lumineux.

Les génomes des trois souches de forte lumière possèdent tous un gène codant pour une CPD photolyase de classe I (cluster 1409) (Kanai et al. 1997). Cette enzyme de réparation de l'ADN a des représentants dans les trois domaines du vivant. Elle est capable d'éliminer les dimères de thymines en utilisant l'énergie de la lumière visible qu'elle capte grâce à deux chromophores (le FAD et le méthenyltetrahydrofolate ou le 8-hydroxy-5-deazariboflavin) (Sancar 1994). Chaque génome contient également un gène codant pour une protéine apparentée à la famille des photolyases mais dont la fonction exacte n'est pas connue (cluster 1533). Il faut noter également que les deux génomes de *Synechococcus* contiennent un gène supplémentaire appartenant à la famille des photolyases. Cependant, ces deux gènes ont des séquences assez divergentes des autres photolyases de *Prochlorococcus* et *Synechococcus* et forment un cluster à part (cluster 1777). Enfin, chacun des trois génomes possède aussi un gène codant pour une protéine qui correspond au domaine de fixation du FAD des photolyases. Le domaine de fixation du chromophore secondaire est absent de cette protéine et il est probable qu'elle ait une fonction différente de celles des photolyases. Ces trois gènes sont répartis dans deux clusters différents, le gène de *P. marinus* MED4 (cluster 2894) possédant une séquence très différente de celles des deux gènes de *Synechococcus* (cluster 1734).

La distribution de ces gènes, qui sont tous absents des souches de profondeur, de *Prochlorococcus* laisse entrevoir une histoire évolutive complexe. Ces gènes pourraient avoir été perdus indépendamment et conservés uniquement chez *P. marinus* MED4. Une autre possibilité, tout aussi parcimonieuse, est que ces gènes aient été perdus par l'ancêtre commun à tous les *Prochlorococcus* et transférés à nouveau dans le génome de *P. marinus* MED4. Cependant, dans le cas des gènes des clusters 1409, 1777 (CPD photolyase) et 1533 (protéine apparentée aux photolyases), l'environnement génique est conservé dans les trois génomes, ce qui plaide en faveur de la première possibilité. A l'inverse, dans le cas des gènes des clusters 1734 et 2894 (domaine de fixation du FAD), l'environnement des gènes codant pour ces protéines est conservé chez *Synechococcus* sp. WH8102 et *Synechococcus* sp. WH7803, mais pas chez *P. marinus* MED4. Il est donc fort probable que ce gène ait été transféré horizontalement lors de l'évolution récente de ce génome.

Les analyses phylogénétiques (Fig. III-3) montrent que les gènes du cluster 1777 pourraient avoir une fonction différente de la réparation des mutations due aux UV. En effet ces gènes ne se regroupent pas avec les CPD photolyases de classe 1 mais avec les séquences de photorécepteurs, sensibles à la lumière bleue, appelés cryptochromes. Ces protéines appartiennent à la famille des photolyases et ont évolué à partir de ces dernières (Todo et al. 1996). Elles ont été identifiées chez une grande variété d'organismes eucaryotes mais aussi chez une cyanobactérie d'eau douce (Hitomi et al. 2000).

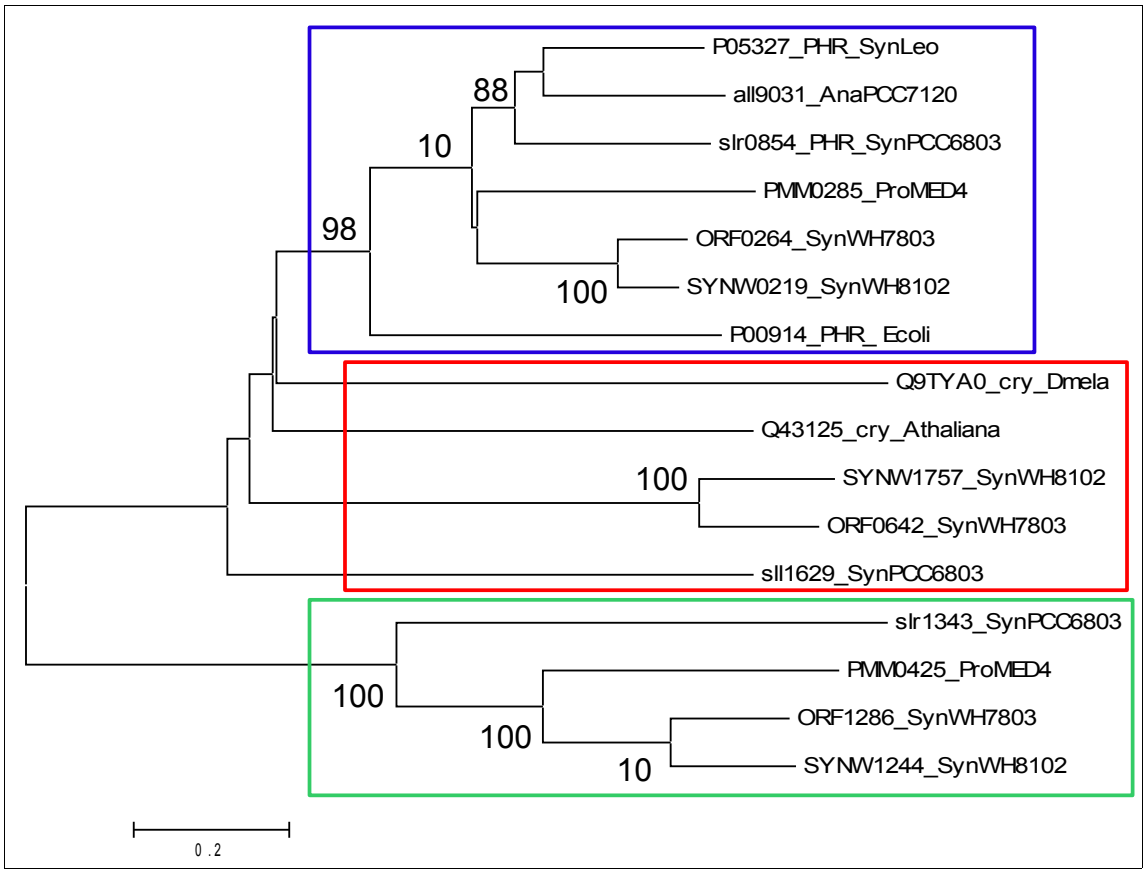

Figure III-3

Arbre phylogénétique fondé sur l'analyse des séquences protéiques des CPD photolyases de classe I (cluster 1409, cadre bleu), des cryptochromes (cluster 1777, cadre rouge) et des protéines apparentées aux photolyases (cluster 1533, cadre vert). Arbre construit avec la méthode du Neighbor-Joining. Les valeurs de bootstrap (1000 réplications) supérieures à 70 sont indiquées dans l'arbre. AnaPCC7120, *Anabaena* sp. PCC 7120; Athaliana, *Arabidopsis thaliana*; Ecoli, *Escherichia coli*; ProMED4, *P. marinus* MED4; ProSS120, *P. marinus* SS120; ProMIT9313, *Prochlorococcus* sp. MIT9313; SynLeo, *Synechococcus leopoliensis*; SynWH8102, *Synechococcus* sp. WH8102; SynWH7803, *Synechococcus* sp. WH7803.

L'adaptation à la niche de forte lumière pourrait également être due à la présence d'un gène supplémentaire codant pour une phytoène déhydrogénase (ou phytoène désaturase). Cette enzyme intervient dans la conversion du phytoène en zeta-carotène qui est la seconde étape de la voie de biosynthèse des caroténoïdes. Il s'agit de molécules hydrocarbonées qui jouent le rôle de pigments photosynthétiques et dont les dérivés oxygénés (xanthophylles) interviennent également dans les mécanismes de photoprotection.

Par exemple, la zéaxanthine intervient dans la protection contre le rayonnement UVB chez la cyanobactérie d'eau douce *Synechococcus* sp. PCC 7942 (Gotz et al. 1999). Les cinq génomes de picocyanobactéries marines contiennent trois autres gènes qui appartiennent, également, à la famille des phytoène déhydrogénases et qui peuvent prendre en charge la synthèse de zeta-carotène à partir de phytoène. Cependant la spécificité de chacun de ces gènes n'est pas connue. On peut imaginer que la présence d'un gène supplémentaire chez les trois cyanobactéries de surface permet d'accroître la production de carotenoïdes en cas de stress lumineux.

Afin d'élargir le panel de gènes potentiellement impliqués dans l'adaptation aux fortes lumières, nous avons aussi recherché les gènes présents uniquement chez *P. Marinus* MED4 et chez *Synechococcus* sp. WH8102 ou chez *P. marinus* MED4 et *Synechococcus* sp. WH7803. Cependant, ce nombre est, là encore, extrêmement réduit. Ainsi, il n'existe ainsi qu'un seul gène (sans fonction connue) présent uniquement chez *P. marinus* MED4 et *Synechococcus* sp. WH7803. *P. marinus* MED4 et *Synechococcus* sp. WH8102 partagent spécifiquement 12 gènes. Le produit d'un de ces gènes est une protéine de petite taille homologue au domaine GAF. Celui-ci est présent dans les phytochromes de plantes et de cyanobactéries, où il sert à la fixation des chromophores, et dans les cGMP phosphodiesterases d'animaux, où il constitue le domaine de fixation du cGMP (Aravind and Ponting 1997). Aucune autre protéine de cyanobactéries marines ne contient ce domaine alors que celui ci est présent dans un grand nombre de protéines chez les cyanobactéries d'eau douce. Une recherche dans la base de donnée Pfam (http://www.sanger.ac.uk/Software/Pfam/) en utilisant ce domaine comme requête montre la présence de protéines présentant la même architecture mono-domaine que celles des trois picocyanobactéries marines. Malheureusement, la fonction de ces protéines n'est pas connue.

L'adaptation à une niche forte lumière peut aussi passer par l'amplification de familles de gènes. C'est le cas par exemple des gènes codant pour les protéines HLIP qui ont été multipliés spécifiquement chez *P. marinus* MED4 par rapport aux deux autres génomes de *Prochlorococcus*. Cependant, il faut noter que ces gènes n'ont pas été particulièrement amplifiés chez les *Synechococcus* spp. Ainsi, s'il est très probable que la multiplication des gènes *hli* chez *P. marinus* MED4 est bien le résultat de l'adaptation de cette souche à la niche forte lumière, il est possible que celle-ci ait été rendue nécessaire à cause du type d'antenne particulier de *Prochlorococcus*. En effet, l'antenne majeure de *Prochlorococcus* est incluse dans les membranes thylacoïdales (Bibby et al. 2003; LaRoche et al. 1996; Partensky and Garczarek 2003), au contraire des phycobilisomes qui sont externes, et la multiplication des gènes d'HLIPs pourrait donc permettre une meilleure protection de cette antenne spécifique contre le stress lumineux. L'analyse de cette famille de gènes a fait l'objet

d'un article spécifique , présenté en annexe III, en collaboration avec l'équipe d'Arthur Grosmann à Stanford (Bhaya et al. 2002).

III.3.4 Gènes de la niche de faible lumière.

Le nombre de clusters identifiés avec cette approche (22) est très similaire à celui trouvé pour les gènes de la niche de forte lumière (Tableau III-3). Cela correspond à seulement 24 et 25 gènes respectivement chez *Prochlorococcus* sp. MIT9313 et *P. marinus* SS120. La majorité de ces gènes n'ont pas de fonction connue et codent pour des protéines de petite taille (< 150 aa). La proportion de gènes codant pour de possibles protéines membranaires est aussi plus faible que dans le cas des gènes de la niche de forte lumière.

Tableau III-3

Gènes spécifiques de la niche de faible lumière. SS120, *P. marinus* SS120; MIT9313, *Prochlorococcus* sp. MIT9313.

Cluster	Produit	SS120	MIT9313
1689	Putative potassium chanel	Pro1527	PMT0217, PMT1024
1697	Predicted membrane protein	Pro0646, Pro0668	PMT1027
1703	Conserved hypothetical protein	Pro1454	PMT3895n, PMT3896n
1704	Conserved hypothetical protein	Pro1458, Pro1889n	PMT3897n
1705	Predicted membrane protein	Pro0718, Pro1559	PMT3909n
2025	Conserved hypothetical protein	Pro0661, Pro1198	PMT0278
2028	Conserved hypothetical protein	Pro1270	PMT0285
2029	2OG-Fe(II) oxygenase superfamily enzyme	Pro1271	PMT0286
2039	Pyrimidine dimer DNA glycosylase/Endonuclease V	Pro1489	PMT0842
2041	Conserved hypothetical protein	Pro1473	PMT0883
2047	Conserved hypothetical protein	Pro1569	PMT0965
2048	Predicted phosphoesterase	Pro1568	PMT0966
2049	Predicted ATPase	Pro1567	PMT0967
2050	Conserved hypothetical protein	Pro1525	PMT1002
2058	Conserved hypothetical protein	Pro1448	PMT1448
2065	Conserved hypothetical protein	Pro0216	PMT2113

Tableau III-3 (suite)

2067	Conserved hypothetical protein	Pro0994	PMT2194
2068*	Mg-protoporphyrin IX monomethyl ester oxidative cyclase	Pro0992	PMT2196
2071	Conserved hypothetical protein	Pro0599	PMT3871n
2072	Predicted membrane protein	Pro1481	PMT3877n
2074	Conserved hypothetical protein	Pro1461	PMT3898n
4220	Rubredoxin	Pro0379	PMT3869n

* également présent chez *P.marinus* MED4 et *Synechococcus* spp. WH7803 et WH8102

L'un de ces clusters contient deux gènes codant pour la « Mg-protoporphyrin IX monomethyl ester oxidative cyclase ». Cette protéine, nommée AcsF, permet la conversion de la Mg-protoporphyrin IX monomethyl ester en divinyl protochlorophyllide. Ce composé sera, à son tour, transformé en chlorophyllide *a*2 puis en chlorophylle *a*2. Cette protéine est essentielle à la synthèse des pigments photosynthétiques et le gène qui lui correspond est aussi présent chez les trois autres cyanobactéries marines. Plusieurs copies de ce gène sont aussi présentes chez les cyanobactéries d'eau douce ainsi que chez les eucaryotes photosynthétiques. Il est assez surprenant que les gènes des cinq génomes de cyanobactéries marines ne se regroupent pas ensemble dans un même cluster. L'alignement de l'ensemble des séquences protéique de ces gènes révèle un fort niveau de conservation aussi bien entre les cyanobactéries marines qu'entre les cyanobactéries marines et celles d'eau douce. Néanmoins, les séquences des deux souches de profondeur ont des séquences nettement différentes de toutes les autres, comme le montre la très longue branche portant ces deux séquences dans l'arbre phylogénétique de la figure III-4. Etant donné sa longueur, le placement de cette branche dans l'arbre est assez douteux.

Cette situation pourrait être interprétée comme étant le résultat d'une paralogie cachée. L'ancêtre commun de *Prochlorococcus* et *Synechococcus* pourrait avoir possédé plusieurs copies de ce gène (comme chez les cyanobactéries d'eau douce). Des copies différentes auraient, par la suite, été perdues au cours de l'évolution des cyanobactéries de surface et de profondeur. Il faut cependant remarquer que l'environnement génique de ces gènes est conservé dans les cinq génomes. Ceci laisse supposer que ces gènes ne sont pas des paralogues mais de véritables orthologues qui ont évolué dans des directions différentes chez les cyanobactéries de surface et de profondeur. La signification de ces différences reste mystérieuse pour le moment.

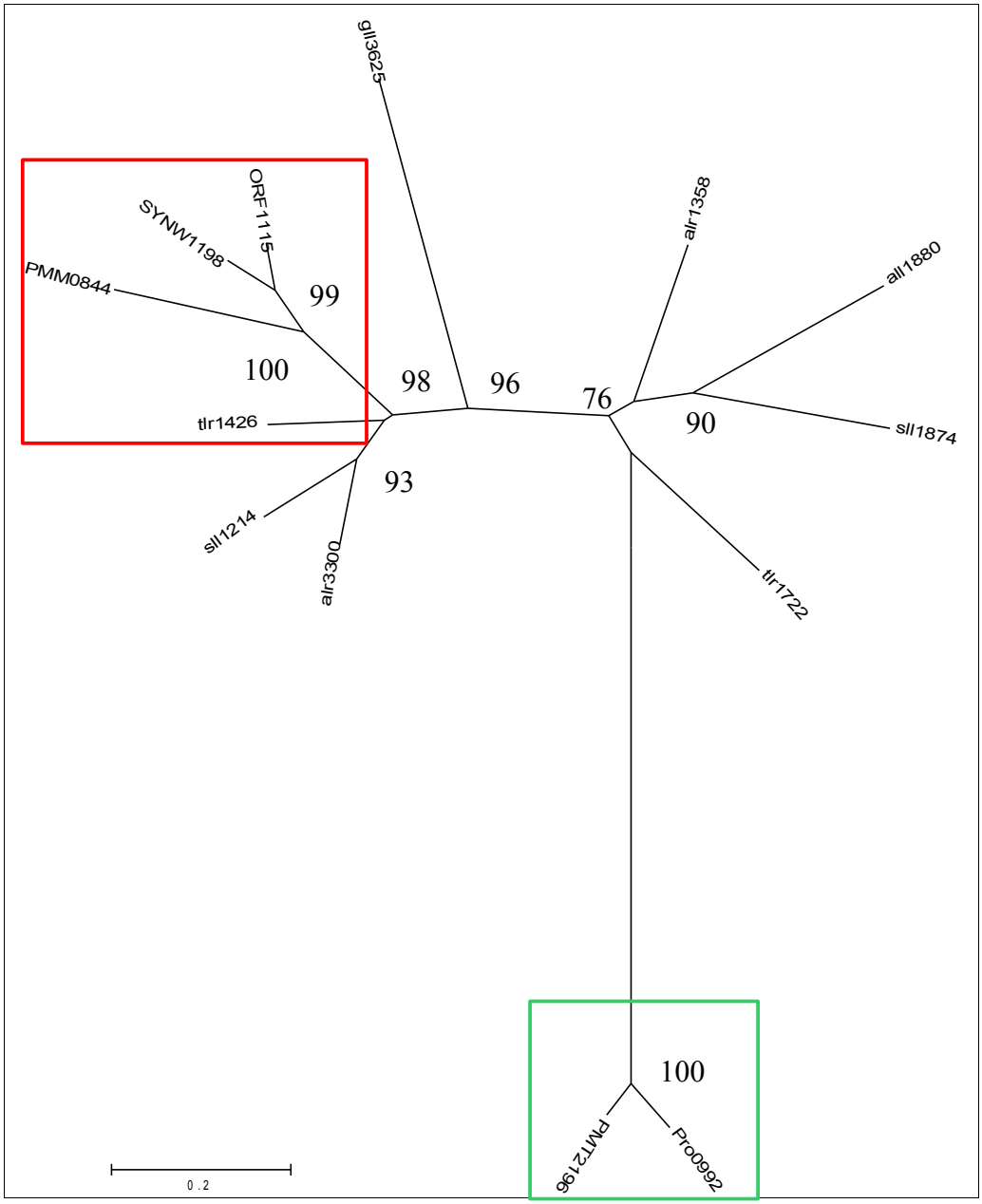

Figure III-4

Arbre phylogénétique basé sur les séquences protéiques du gène acsF. Les souches de profondeur de Prochlorococcus sont encadrées en vert. La souche de surface de Prochlorococcus et celles de Synechococcus sont encadrées en rouge. Arbre construit avec la méthode du Neighbor-Joining. Les valeurs de bootstrap (1000 réplications) supérieures à 70 sont indiquées dans l'arbre. PMM, Prochlorococcus marinus MED4; Pro, Prochlorococcus marinus SS120; PMT, Prochlorococcus sp. MIT9313; SYNWH, Synechococcus sp. WH8102; ORF Synechococcus sp. WH7803.

Un autre cluster contient deux gènes codant pour « Pyrimidine dimer DNA glycosylase » (McCullough et al. 1998). Cette enzyme vient initier la réparation des lésions provoquées par les UV en excisant les dimères de thymines. Son mécanisme d'action est néanmoins différent de celui des photolyases (Garvish and Lloyd 1999). Ce gène a probablement une origine virale étant donnée sa présence dans un grand nombre de génomes viraux et son absence de tous les génomes de cyanobactéries et de la majorité des génomes de bactéries. L'existence de ce gène, dans les génomes de ces deux souches de *Prochlorococcus,* est assez intrigante puisque celles-ci ne sont normalement pas exposées aux rayonnements UV. La perte de tous les gènes codant pour des photolyases ou pour des protéines apparentées aux photolyases est attribuée à l'inutilité de ces gènes dans un milieu sans UV. Le gène est particulièrement proche de celui présent dans le génome du bactériophage infectant l'algue d'eau douce *Chlorella* (celle-ci vivant en symbiose avec le protozoaire *Paramecium bursaria*). Chez ce bactériophage, l'enzyme intervient dans la réparation des lésions de l'ADN dues aux UV mais aussi dans celles provenant de l'oxydation des bases par des groupements hydroxyles. On peut penser que c'est cette seconde fonction qui est privilégiée dans les génomes de *Prochlorococcus* de profondeur.

Comme pour les gènes de la niche de forte lumière, l'adaptation à la niche de faible lumière peut être due à la multiplication spécifique de certains gènes. Le présence de huit gènes *pcb* chez la souche *P. marinus* SS120 en est le meilleur exemple. Malheureusement, la plupart des gènes présents en plusieurs copies dans ces génomes n'ont pas de fonction connue ou des fonctions qui sont très difficilement reliables aux caractéristiques de la niche de faible lumière. De plus, aucun de ces clusters ne contient des gènes qui ont été multipliés à la fois chez *P. marinus* SS120 et chez *Prochlorococcus* sp. MIT9313. Cependant, on peut remarquer la présence d'un second gène (*hemL*) codant pour la glutamate-1-semialdehyde 2,1-aminomutase. Cette enzyme permet la synthèse du 5-aminolevulinate qui est le premier composé de la voie de biosynthèse des chlorophylles.

III. 4 Conclusions

La comparaison des cinq génomes de picocyanobactéries marines dévoile l'existence d'un gros noyau de gènes communs. Il ressort de ces analyses que les répertoires de gènes de *P. marinus* MED4 et SS120 ont évolué principalement sous l'influence d'un processus de réduction génomique. La conséquence de cette réduction est une augmentation de la proportion de gènes communs aux cinq génomes dans les génomes de *Prochlorococcus marinus* MED4 et SS120. La majorité de ces gènes sont en simple copie, révélant, par là même, une faible redondance de l'information génétique dans ces génomes. Ceci pourrait

être le résultat de la relative simplicité de la morphologie (ce sont toutes des bactéries unicellulaires) et surtout de l'habitat marin de ces cyanobactéries, un milieu très tamponné par définition.

La comparaison de ces génomes offre la possibilité d'identifier les gènes impliqués dans l'adaptation aux forts rayonnements visibles et UV de la niche forte lumière ou ceux permettant d'optimiser la capture de photons et la photosynthèse dans la niche de basse lumière. Malheureusement, ces analyses n'ont permis d'identifier qu'un petit nombre de gènes. Très peu d'entre eux ont une fonction interprétable en fonction des conditions lumineuses particulières que ces organismes rencontrent en surface ou en profondeur. Il faut noter aussi que, parmi ces gènes, ceux qui pourrait jouer un rôle dans l'adaptation à la niche forte lumière (caroténoïdes isomérase, photolyase, hélicase et ADN ligase, protéines membranaires) sont plus nombreux que ceux potentiellement impliqués dans l'adaptation à la niche faible lumière. De plus, ces dernières correspondent plutôt à la multiplication (*pcb*) ou la différenciation (*acsF*) de gènes présents aussi dans les génomes d'une ou de plusieurs souches de surface. Ce résultat n'est pas inattendu puisque les contraintes exercées par les forts rayonnements visibles et UV sont plus probablement plus variées (dommages de l'appareil photo synthétique et à l'ADN, production de composés oxydants...) que celles exercées par le manque de lumière.

La recherche de gènes de la niche forte lumière a été réalisée en analysant les gènes communs aux génomes de *P. marinus* MED4 et aux deux *Synechococcus*. Or il est probable que des mécanismes différents de protection contre les fortes intensités lumineuses aient été développés chez ces cyanobactéries. Ainsi l'utilisation d'une antenne interne de type Pcb comme antenne photosynthétique chez *Prochlorococcus* impose l'existence de systèmes de protection particuliers, tels qu'un type spécifique de HLIP, par rapport à ceux utilisés par *Synechococcus* pour protéger ses phycobilisomes. Le nombre de gènes identifiés comme potentiellement impliqués dans l'adaptation à la niche forte lumière est donc une estimation basse du nombre réel de gènes.

CHAPITRE IV

Évolution Réductive chez *Prochlorococcus*

IV. 1 Résumé des résultats obtenus

Trois génomes complets de *Prochlorococcus* spp., le plus petit et le plus abondant des organismes photosynthétiques dans les océans, ont été publiés récemment. Les analyses de génomique comparée révèlent qu'un processus de réduction génomique a eu lieu au sein de ce genre, associé à une forte diminution du pourcentage en G+C. Tous les exemples de réduction génomique connus ont, jusqu'ici, été reliés à un mode de vie caractérisé par l'association obligatoire à un hôte, tel que le parasitisme ou la symbiose. *Prochlorococcus* constitue donc le premier exemple d'évolution réductive chez un organisme ayant un mode de vie complètement libre.

Nos résultats indiquent clairement que la réduction du génome a été accompagnée par une accélération de l'évolution des protéines chez *P. marinus* SS120 et plus particulièrement chez *P. marinus* MED4. Cette accélération touche toutes les catégories fonctionnelles de gènes codant pour des protéines. Au contraire, le gène de l'ARNr 16S semble avoir évolué en suivant l'horloge moléculaire. Nous avons observé également que *P. marinus* MED4 et *P. marinus* SS120 ont perdu plusieurs gènes intervenant dans les mécanismes de réparation de l'ADN. L'absence de ceux-ci pourrait être reliée au biais de mutations et à l'augmentation du taux de substitutions des acides aminés.

Dans cet article, nous avons étudié les mécanismes évolutifs impliqués dans ce processus, qui sont différents de ceux connus chez les organismes dépendants d'un hôte. En effet, la plupart des substitutions que l'on observe chez *Prochlorococcus* doivent être sélectivement neutres, puisque la grande taille des populations de *Prochlorococcus* impose une faible dérive génétique et une forte sélection purifiante. Notre hypothèse est que la réduction du génome chez *Prochlorococcus* pourrait correspondre à une adaptation de cet organisme l'environnement, stable et très pauvre, au sein duquel il vit. En nous basant sur cette hypothèse, nous proposons également un scénario pour l'évolution du génome de *Prochlorococcus*. Cet article a été soumis à *Genome Biology*.

IV. 2 Article

Research

Accelerated evolution associated with genome reduction in a free-living prokaryote

Alexis Dufresne, Laurence Garczarek and Frédéric Partensky

Address: Station Biologique, UMR 7127 CNRS et Université Paris 6, BP74, 29682 Roscoff Cedex, France.

Correspondence: Frédéric Partensky. E-mail: partensky@sb-roscoff.fr

Published: 14 January 2005

Genome ***Biology*** 2005, **6:**R14

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2005/6/2/R14

Received: 5 October 2004
Revised: 2 December 2004
Accepted: 7 December 2004

Abstract

Background: Three complete genomes of *Prochlorococcus* species, the smallest and most abundant photosynthetic organism in the ocean, have recently been published. Comparative genome analyses reveal that genome shrinkage has occurred within this genus, associated with a sharp reduction in G+C content. As all examples of genome reduction characterized so far have been restricted to endosymbionts or pathogens, with a host-dependent lifestyle, the observed genome reduction in *Prochlorococcus* is the first documented example of such a process in a free-living organism.

Results: Our results clearly indicate that genome reduction has been accompanied by an increased rate of protein evolution in *P. marinus* SS120 that is even more pronounced in *P. marinus* MED4. This acceleration has affected every functional category of protein-coding genes. In contrast, the 16S rRNA gene seems to have evolved clock-like in this genus. We observed that MED4 and SS120 have lost several DNA-repair genes, the absence of which could be related to the mutational bias and the acceleration of amino-acid substitution.

Conclusions: We have examined the evolutionary mechanisms involved in this process, which are different from those known from host-dependent organisms. Indeed, most substitutions that have occurred in *Prochlorococcus* have to be selectively neutral, as the large size of populations imposes low genetic drift and strong purifying selection. We assume that the major driving force behind genome reduction within the *Prochlorococcus* radiation has been a selective process favoring the adaptation of this organism to its environment. A scenario is proposed for genome evolution in this genus.

Background

The size of bacterial genomes is primarily the result of two counteracting processes: the acquisition of new genes by gene duplication or by horizontal gene transfer; and the deletion of non-essential genes. Genomic flux created by these gains and losses of genetic information can substantially alter gene content. This process drives divergence of bacterial species and eventually adaptation to new ecological niches [1]. In some cases, gene deletion may prevail over gene acquisition, leading to genome reduction. This process has occurred several times during evolution and has been well documented for cellular organelles [2,3], obligate pathogens such as *Mycoplasma genitalium* [4] or phytoplasmas [5] and symbionts such as the insect endosymbiont *Buchnera* [6-8] or the

hyperthermophile *Nanoarchaeum equitans* [9]. In the case of organelles, the degree of genome reduction can be extensive as a result of massive gene transfer into the host nucleus, allowing maintenance of the corresponding functions in the resulting composite organism. Mitochondrial or chloroplast genomes, for instance, can be as small as 6 kilobases (kb) [10] and 35 kb [11], respectively. In the case of obligate host-dependent bacteria, the reduction is more limited because the relationships with their hosts are less intimate than for organelles in eukaryotic cells. Thus, obligatory pathogens need to retain a minimum of functions that allow them to infect new hosts and to avoid host defenses, and obligate endosymbionts carry genes which are absolutely necessary for host survival. For instance, a substantial part (approximately 10 %) of the *Buchnera* genome is devoted to biosynthesis of amino acids which are essential to its host [6].

So far, all characterized examples of genome reduction have been associated with a change from a free-living to a host-dependent lifestyle [12]. It is therefore intriguing that a similar phenomenon of genome reduction has occurred within the free-living marine cyanobacterial genus *Prochlorococcus* [13-15]. The latter is present at high abundance (often over 10^5 cells/ml) in all nutrient-poor areas of the world's oceans between 40°N and 40°S and is probably the most abundant photosynthetic organism on Earth [16,17]. It has been shown that two major ecotypes exist within this genus [18]. The first is adapted to grow at the base of the illuminated layer and displays a high divinyl-chlorophyll *b* to *a* ratio; the second inhabits the upper layer of the ocean and has a low divinyl-chlorophyll *b* to *a* ratio [19]. The genome of one high-light-adapted (HL) strain, *Prochlorococcus marinus* MED4 [14], and of two low-light-adapted (LL) strains, *P. marinus* SS120 [13] and *Prochlorococcus* species MIT9313 [14], have recently been sequenced and annotated.

Phylogenetic trees based on 16S rRNA sequences [18] or 16S-23S ribosomal internal transcribed spacer sequences [20] show that *Prochlorococcus* sp. MIT9313 branches at the base of the *Prochlorococcus* radiation, close to the *Synechococcus* group [21]. In contrast, the *Prochlorococcus* HL clade, encompassing the MED4 strain, appears to be the most recently evolved *Prochlorococcus* group, consistent with the fact that this clade is much less diversified than are the LL clades.

Despite the close relatedness of these strains, their genomes vary widely in terms of size, G+C content and the number of protein-coding genes (Table 1). While the general characteristics of the MIT9313 genome are very similar to those of the *Synechococcus* sp. WH8102 genome [22], MED4 has the smallest genome for a photosynthetic organism known to date and the SS120 genome is only 90 kb larger. Furthermore, this genome reduction is clearly accompanied by a drift in G+C content, a phenomenon that commonly occurs during the evolution of host-dependent genomes [23]. However, the evolutionary mechanisms involved in the genome reductive process are most probably different from those that have occurred in host-dependent organisms. Using comparative sequence analyses of the four genomes of marine picocyanobacteria published to date, we have attempted to better understand the causes and consequences of this phenomenon and to address the relationships between genome reduction and niche adaptation in marine picocyanobacteria.

Table 1

General features of the genomes of the four marine picocyanobacteria used in this study

Genome	Size (Mbp)	GC%	Number of protein-coding genes
P. marinus MED4	1.66	30.8	1,716
P. marinus SS120	1.75	36.4	1,882
Prochlorococcus sp. MIT9313	2.41	50.7	2,273
Synechococcus sp. WH8102	2.43	59.4	2,525

Results

Synteny and genome stability

Alignments of whole genomes show a strong conservation of the gene order between MED4 and SS120 (Figure 1a). There are only five inversions larger than 20 kb between these two genomes. In contrast, the large number of inversions and translocations and the shorter size of the colinear segments between SS120 and MIT9313 on the one hand and MIT9313 and WH8102 on the other hand (Figure 1b,c) indicate that extensive genome rearrangements have occurred not only between *Synechococcus* and *Prochlorococcus* but also between MIT9313 and the two other *Prochlorococcus* strains (see also Figure 2 in [14]). The degree of synteny observed between the four marine picocyanobacteria genomes strengthens the hypothesis of a more recent divergence of the clades containing MED4 and SS120 than of the clade containing MIT9313.

Overall genome composition

The downsizing of MED4 and SS120 genomes during evolution is associated with a genome-wide adenine (A) and thymine (T) enrichment (Table 1). The bias is most pronounced at neutral sites such as intergenic regions (MED4, 76.6% A+T; SS120, 69.3% A+T) and third-codon positions of protein-coding genes (MED4, 79.7% A+T; SS120, 73.85% A+T). This bias has little effect on ribosomal RNA genes (5S, 16S and 23S) which have a G+C content greater than 50% in all four picocyanobacterial genomes. In both MED4 and SS120, the single rRNA gene cluster can easily be spotted as a G+C-rich anomaly compared to the rest of the genome (see for example, Figure 1 in [15]). In direct contrast, protein-coding genes are

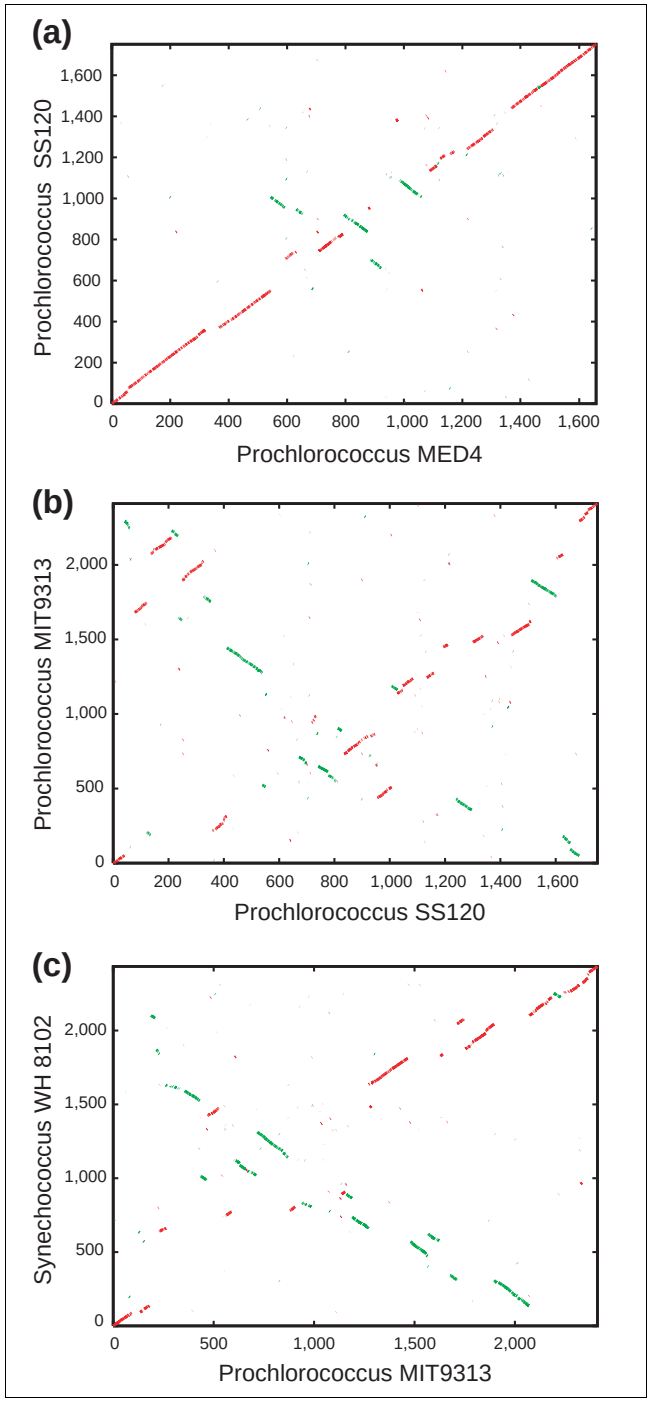

Figure 1
Alignments of complete genome sequences of marine picocyanobacteria. Genome sequences are translated in their six reading frames. **(a)** Comparison of the MED4 and SS120 genomes; **(b)** comparison of the SS120 and MIT9313 genomes; **(c)** comparison of the MIT9313 and WH8102 genomes. Colinear segments are shown in red and inversions in green. Translocated segments are above or below the diagonal.

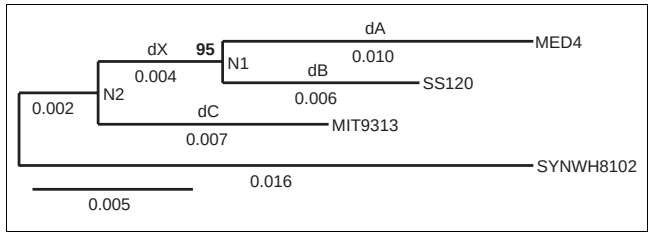

Figure 2
Phylogenetic tree of 16S rRNA genes from the four marine picocyanobacteria. Neighbor-joining tree with Kimura 2-parameter correction. The bootstrap value (1,000 replications) is shown in boldface. Lengths of the branches dA, dB, dC and dX (see text) are given below the branches. N1, node 1, branchpoint between MED4 and SS120; N2, node 2, branchpoint between MIT9313 and Node 1.

strongly affected by the extreme base composition of these genomes. First, the bias influences codon usage since, for a given amino acid, AT-rich codons are preferentially used (Figure 3a). Second, the amino-acid composition of the proteins themselves is affected (Figure 3b). Indeed, when compared to *Prochlorococcus* sp. MIT9313 and *Synechococcus* sp. WH8102, the genes of *P. marinus* MED4 and SS120 contain fewer amino acids encoded by G+C-rich codons (for example, alanine or arginine) and more amino acids encoded by A+T-rich codons (for example, isoleucine or lysine).

Orthologous gene pool size

A total of 1,306 orthologs belonging to all major functional categories are common to the four genomes (see Additional data file 1) and probably constitute an estimate of the core of genes conserved in all marine picocyanobacteria. This is sensibly more than the pool of around 1,000 orthologs identified by W.R. Hess [15]. The difference certainly results from the use by the latter author of a low E-value threshold ($10e^{-12}$) for BLAST comparisons. In contrast, our analysis is based on identification of reciprocal best hits without the use of any particular threshold (apart from the default BLAST threshold) and consequently allows the detection of orthologous relationships whatever the gene lengths or the level of similarity. Still, our ortholog identification process is rather strict and the set of orthologs identified in this study probably corresponds to a lower estimate of the actual number of orthologs shared by the four genomes. This set of genes represents a substantial percentage of the total pool of all protein-coding genes in *P. marinus* MED4 (73.2%) and SS120 (69.2%) and about half of the gene set in *Prochlorococcus* sp. MIT9313 (56.2%) and *Synechococcus* sp. WH8102 (51.1%). These percentages are consistent with the differences in the respective number of genes within these genomes (Table 1) and are compatible with the assumption that a massive gene loss has occurred in MED4 and SS120 during their evolution from a *Prochlorococcus* ancestor with a larger genome [13-15].

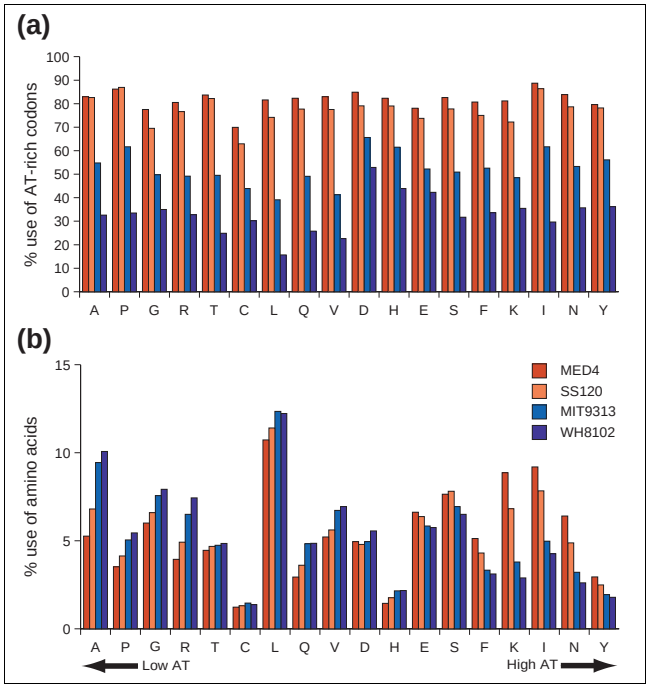

Figure 3
Influence of mutational bias in codon usage and amino-acid usage. **(a)** Percentage use of AT-rich codons in the four marine picocyanobacteria. Amino acids are ranked according to AT content of their respective codons. Methionine and tryptophan, which are both encoded by only one codon, have been discarded from the analysis. **(b)** Percentage use of amino acids in marine picocyanobacteria.

Accelerated rate of evolution of protein-coding genes in *Prochlorococcus*

Because biased base composition seems to constrain amino-acid usage in the *Prochlorococcus* genomes, we have investigated whether it also affects the rate of protein sequence evolution in these genomes. We used the 1,306 orthologs common to the four genomes to estimate the amino-acid substitution rate in each genome. Branch lengths calculated for a given tree topology (the same topology as for the 16S rRNA gene tree; see Figure 2) are 0.46, 0.22, 0.16 and 0.14 amino acid substitutions per site for branches dA, dB, dC and dX, respectively. Using *Synechococcus* sp. WH8102 as the outgroup, we tested the rate-constancy hypothesis and computed the ratios of branch lengths. Relative rate tests (two-cluster and branch length tests) indicate that protein sequences evolved at significantly different rates ($P < 0.001$) between MED4, SS120 and MIT9313. Therefore the hypothesis of a constant evolutionary rate between these strains can be rejected for protein-coding genes. The calculation of branch-length ratios reveals that the amino-acid substitution rate is 2.04-fold higher inMED4 than in SS120 (dA/dB) and 3.81-fold higher in MED4 than in MIT9313 ((dA+dX)/dC). This rate is also 2.31-fold higher for SS120 than for MIT9313 ((dB+dX)/dC). Computation of branch lengths for each functional category shows that the increased rate of amino-acid replacement in protein sequences concerns every category (Figure 4 and Table 2). These results imply that the rate of amino-acid substitution increased during evolution of the *Prochlorococcus* genus concomitantly with genome reduction and increase in A+T content.

Synonymous and nonsynonymous substitutions

The ratio of the rate of nonsynonymous substitutions (d_N) to the rate of synonymous substitutions (d_S) is commonly used to measure the relative rate of purifying selection acting at the protein level. We determined d_S and d_N for each gene pair of every group of orthologs and their values were averaged for each genome. Surprisingly, we observed saturation at synonymous sites for all genome pairs ($d_S > 2$) and the calculation of the d_N/d_S ratio was thus impossible. Still, the average d_N was higher between MED4 and SS120 (0.36) than between SS120 and MIT9313 (0.32). The lowest d_N was observed between MIT9313 and WH8102 (0.24), a finding which is consistent with the relative acceleration of amino-acid substitutions in MED4 and in SS120.

DNA-repair systems

A shift in base composition may reflect the loss of DNA-repair genes and we therefore determined the presence or absence of genes involved in these mechanisms. As the mutational pressure is toward a high A+T content in both MED4 and SS120, we looked more closely at those genes whose absence could increase the frequency of G:C to A:T mutations. Among the genes putatively encoding DNA-repair enzymes identified in MIT9313 and WH8102, a few are missing in SS120 and/or MED4 (Table 3). Both MED4 and SS120 lack the *ada* gene, which encodes 6-O-methylguanine-DNA methyltransferase, which repairs alkylated forms of guanine and thymine in DNA. Such alkylations generate lesions that can lead to G:C to A:T transversions [24]. Interestingly, the MED4 genome is the only one among the four picocyanobacteria not to encode the A/G-specific DNA glycosylase MutY, as previously noted by Rocap and co-workers [14]. This enzyme acts with MutT (NTP pyrophosphohydrolase) and MutM (formamido-pyrimidine-DNA glycosylase) in the GO system to avoid misincorporation of oxidized guanine (8-oxoG) in DNA and to repair the base mismatches A:8-oxoG [25]. In *Escherichia coli*, knocking out both *mutM* and *mutY* translates into a 1,000-fold increase of G:C to A:T transversions in comparison to the wild-type strain [26]. In addition to MutT and MutY, MIT9313 and WH8102 encode a third enzyme of the NUDIX hydrolase family that is missing in MED4 and SS120. This hydrolase could act to prevent mutations. However because of the broad substrate specificity of this family, one cannot know with certainty the function of this protein. Likewise, two genes coding for enzymes of the RecF pathway have been lost either by both MED4 and SS120 (DNA helicase RecQ) or only by MED4 (exonuclease RecJ).

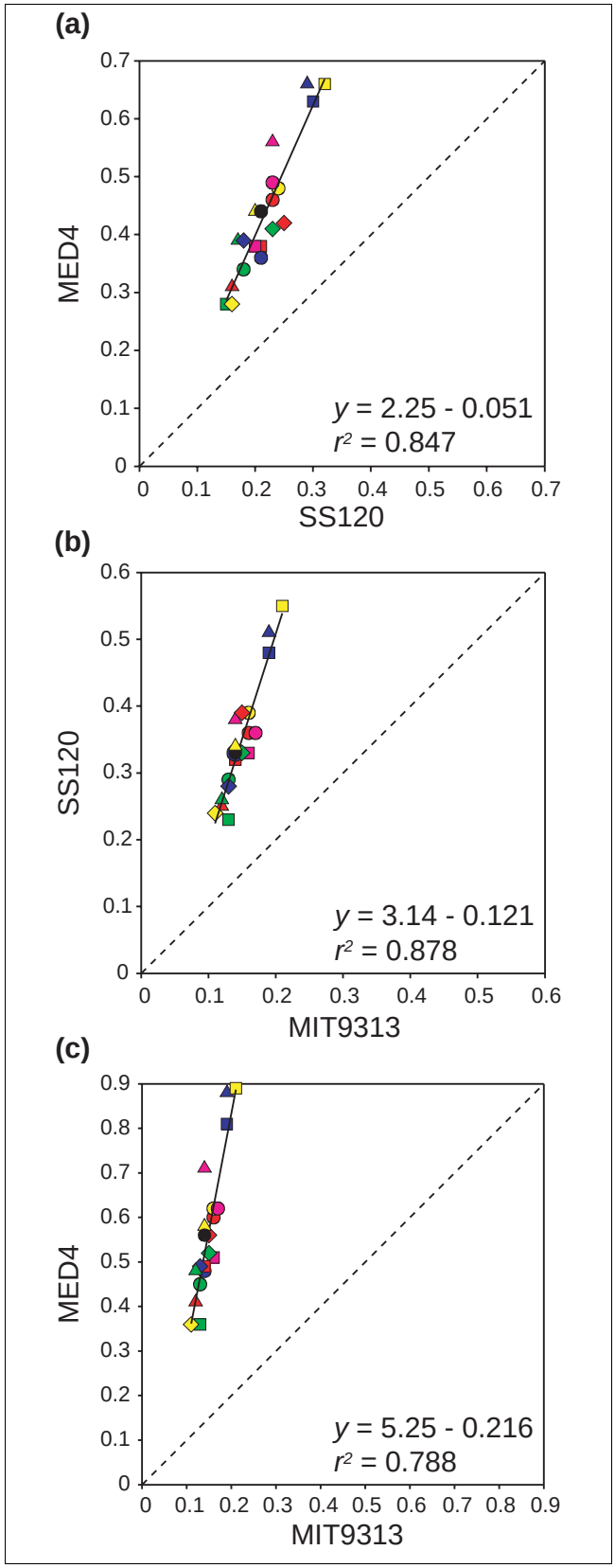

Figure 4

Figure 4
Amino-acid substitution rate per functional category. Branch lengths computed for each functional category between **(a)** MED4 and SS120, **(b)** SS120 and MIT9313 and **(c)** MED4 and MIT9313. In the three comparisons, branch-length values are aligned along a line with a slope much greater than 1, indicating that acceleration of the substitution rates occurs in every functional category. Axes represent the number of amino-acid substitutions per site. Red circle, amino-acid transport and metabolism; green circle, carbohydrate transport and metabolism; yellow circle, cell-cycle control; blue triangle, cell wall/membrane biogenesis; pink circle, coenzyme metabolism; red square, defense mechanisms; green square, energy production and conversion; yellow square, function unknown; blue square, general function prediction only; pink square, inorganic ion transport and metabolism; red triangle, intracellular trafficking; green triangle, lipid transport and metabolism; yellow triangle, nucleotide transport and metabolism; blue circle, posttranslational modification, protein turnover; pink triangle, replication, recombination and repair; red diamond, secondary metabolite biosynthesis, transport and catabolism; green diamond, signal transduction mechanisms; yellow diamond, transcription; blue diamond, translation; black circle, miscellaneous.

Discussion

The process of genome reduction which has occurred within the *Prochlorococcus* radiation has to our knowledge never been observed so far in any other free-living prokaryote. Since *Prochlorococcus* sp. MIT9313 has a genome size very similar to that of *Synechococcus* sp. WH8102 (2.4 megabase-pair (Mbp)), as well as several other marine *Synechococcus* spp. (M. Ostrowski and D. Scanlan, personal communication), it is reasonable to assume that the common ancestor of all *Prochlorococcus* species also had a genome size around 2.4 Mbp. Under this hypothesis, the genome reduction which has occurred in MED4 would correspond to around 31%. By comparison, the extent of genome reduction in the insect endosymbiont *Buchnera*, as compared to a reconstructed ancestral genome, is around 77% [27]. The genome of *P. marinus* SS120 - and *a fortiori* the MED4 genome - is considered to be near minimal for a free-living oxyphototrophic organism [13]. It would seem that genome reduction in these organisms probably cannot proceed below a certain limit, corresponding to a gene pool containing all the essential genes of biosynthetic pathways and housekeeping functions (probably including most of the 1,306 four-way orthologous genes identified in this study) plus a number of other genes, including genus-specific as well as niche-specific genes. For instance, MED4 encodes a number of photolyase-related proteins, a few specific ABC transporters (for cyanate, for example; [14] and data not shown). These specific compounds might be critical for survival in the upper water layer, which receives high photon fluxes, UV light and is nutrient-depleted, but less so for life deeper in the water column.

If both *Prochlorococcus* lineages and host-dependent organisms have undergone genome reduction associated with accelerated substitution rates, these phenomena must have arisen from very different causes as the resulting gene

Table 2

Number and percentage of orthologous genes per functional category

Category	Number of genes	% of orthologs
Amino-acid transport and metabolism	94	7.2
Carbohydrate transport and metabolism	50	3.8
Cell-cycle control	17	1.3
Cell wall/membrane biogenesis	55	4.2
Coenzyme metabolism	99	7.6
Defense mechanisms	14	1.1
Energy production and conversion	106	8.1
Function unknown	269	20.6
General function prediction only	116	8.9
Inorganic ion transport and metabolism	47	3.6
Intracellular trafficking	13	1.0
Lipid transport and metabolism	25	1.9
Nucleotide transport and metabolism	39	3.0
Posttranslational modification, protein turnover	59	4.5
Replication, recombination and repair	51	3.9
Secondary metabolite biosynthesis, transport and catabolism	6	0.5
Signal transduction mechanisms	11	0.8
Transcription	26	2.0
Translation	127	9.7
Miscellaneous	82	6.3

repertoires of the two types of organisms differ tremendously. Indeed, the genome evolution of endosymbionts and obligatory pathogens is driven by two main processes which have mutually reinforcing effects on genome size and evolutionary rates. Being confined inside their host, these bacteria have tiny population sizes and are regularly bottlenecked at each host generation or at each new host infection. Consequently, they experience a strong genetic drift [28] involving an increase in substitution rate. This acceleration results in the accumulation at random of slightly deleterious mutations in protein-coding genes [8,29] as well as in rRNA genes [29,30]. This genetic drift enhances the downsizing of the genome through inactivation and then elimination of potentially beneficial but dispensable genes. Among these, there have been a number of DNA-repair genes, the disappearance of which could have further increased the mutation rate [6,31-33]. Furthermore, a number of genes may be subject to a relaxation of purifying selection which is therefore rendered less effective in maintaining gene function. This relaxation particularly affects genes which have become useless because they are redundant in their host genome, such as genes involved in the biosynthesis of amino acids, nucleotides, fatty acids and even ATP [4-6,8,9,32]. Selection pressure is also reduced for genes involved in environmental sensing and regulatory systems, such as two-component systems, because of the much buffered environment offered by the host [6].

In the free-living genus *Prochlorococcus*, the very large size of field populations [34] means that these populations are subject to much lower genetic drift and their genomes are subject to much stronger purifying selection than are those of endosymbionts and pathogens [35]. Consequently, the observed accelerated rate of evolution probably results merely from the increase in the mutation rate, which in turn is probably due to the loss of DNA-repair genes, even if one should note that, in *P. marinus* SS120 only two such genes are missing (Table 3). We observed a similar acceleration of amino-acid substitutions for all functional categories (Figure 4). This finding is more consistent with a global increase in the mutation rate than with relaxed selection, the latter being unlikely to occur to the same extent at all loci. We also assume that most amino-acid substitutions that have occurred in *Prochlorococcus* proteins are neutral; that is, they have not altered protein function. Indeed, populations of the HL clade which, like MED4, have the most derived protein sequences of all *Prochlorococcus* species, appear to be the most abundant photosynthetic organisms in the upper layer of the temperate and inter-tropical oceans [16]. Such an ecological success would hardly be possible for organisms handicapped by a large number of slightly deleterious mutations, especially given the fact that most genes are single copy, and so compensation of gene function is generally not possible. The effect of the maintenance of a high level of purifying selection on coun-

Table 3

DNA-repair genes missing only in *P. marinus* MED4 or in both MED4 and SS120

Gene	COG	Product	MED4	SS120	MIT9313	WH8102
ada/ogt	0350	6-O-methylguanine-DNA methyltransferase	-	-	PMT0269	SYNW1680
mutY	1194	A/G-specific DNA glycosylase	-	Pro1789	PMT0135	SYNW0115
recQ	0514	Superfamily II DNA helicase	-	-	PMT0189	SYNW1958
recJ	0608	Single-stranded DNA-specific exonuclease	-	Pro0984	PMT0761	SYNW1206
exoI/xseA	1570	Exonuclease VII large subunit	-	Pro0111	PMT1641	SYNW2181
xseB	1722	Exonuclease VII small subunit	-	Pro0112	PMT1642	SYNW2182
-	0494	NUDIX hydrolase family	-	-	PMT1026	SYNW1334

Genes in bold are involved in repair of G:C to A:T mutations.

teracting deleterious substitutions is particularly obvious in the rRNA genes. Contrary to the protein-coding genes, relative rate tests did not show any significant differences in the rates of evolution of the 16S rRNA genes in the four marine picocyanobacterial genomes, and thus there is no evidence that either SS120 or MED4 could have accumulated mutations destabilizing the secondary structure of their 16S rRNA molecule. One noteworthy consequence of the acceleration in the rates of evolution of protein-coding genes in *Prochlorococcus* is that phylogenetic reconstructions based on protein sequences are biased. Indeed, this leads to much longer branches for these two strains than for MIT9313. The resulting tree topology most often does not support that obtained with the 16S rRNA gene, for which the molecular clock hypothesis holds true according to our analyses. Thus, rRNA genes are likely to be among the few genes that will give reliable estimates of the phylogenetic distances between *Prochlorococcus* strains.

If it is neither the relaxation of purifying selection nor an increase in genetic drift that has been the main factor causing *Prochlorococcus* genome reduction, an alternative possibility is that the latter could be the result of a selective process favoring the adaptation of *Prochlorococcus* to its environment. The apparently better ecological success in oligotrophic areas of *Prochlorococcus* species compared to their close relative *Synechococcus* [16,34], strongly suggests that the reduction of *Prochlorococcus* genome size could provide a competitive advantage to the former. Indeed, extensive comparisons of the gene complements of these two organisms show very few examples - at least among genes for which function is known - of the occurrence of specific genes in MED4 which could explain its better adaptation (data not shown). One noteworthy exception is the presence in *Prochlorococcus*, but not *Synechococcus*, of flavodoxin and ferritin, two proteins that possibly give *Prochlorococcus* a better resistance to iron stress. Apart from that, *Synechococcus* appears more like a generalist, in particular with regard to nitrogen or phosphorus uptake and assimilation [22], and should *a priori* be more suited to sustain competition. Hence, we assume that the key to the success of *Prochlorococcus* resides less in the development of a specific complex or pathway to cope better with unfavorable conditions than in the simplification of its genome and cell organization, which can allow this organism to make substantial economies in energy and material for cell maintenance.

The mere reduction in genome size *per se* is a potential source of substantial economies for the cell, as it reduces the amount of nitrogen and phosphorus, two particularly limiting elements in the upper part of the ocean, which are necessary, for instance, in DNA synthesis. Another advantage is that it allows a concomitant reduction in cell volume. It has been previously suggested (see, for example [36]) that, for a phytoplanktonic organism, a small cell volume confers two selective advantages by reducing self-shading (the package effect) and by increasing the cell surface-to-volume ratio, which can improve nutrient uptake. The first advantage would improve the fitness of the LL strains, whereas the second would offer an advantage to the HL strains living in nutrient-depleted surface waters. Finally, cell division is less costly for a small than for a large cell. On the basis of these observations, we assume that the major driving force for genome reduction within the *Prochlorococcus* radiation has been the selection for a more economical lifestyle. The bias toward an A+T-rich genome in MED4 and SS120 is also consistent with this hypothesis, as it can be seen as a way to economize on nitrogen. Indeed, an AT base-pair contains seven atoms of nitrogen, one less than a GC base-pair.

With this hypothesis in mind, we propose a possible scenario for the evolution of *Prochlorococcus* genomes. Using a rate of 16S rRNA divergence of 1% per 50 million years [37], one can estimate that the differentiation of these two genera is as recent as 150 million years, as the molecular clock hypothesis holds for this gene in *Prochlorococcus* and *Synechococcus*. The ancestral *Prochlorococcus* cells must have developed in the LL niche, a niche probably left free by other picocyanobac-

teria. Given the considerable difference in genome size between the LL strains MIT9313 and SS120, it appears that genome reduction itself must have started in one (or possibly several) lineage(s) within the LL niche some time after *Prochlorococcus* differentiation from its common ancestor with marine *Synechococcus* species. Why the selection has affected only one (or some?) and not all *Prochlorococcus* lineages remains unclear. Examination of the gene repertoire of *P. marinus* SS120 [13] suggests that this genome reduction must have concerned the random loss of dispensable genes from many different pathways. At some point during evolution, some genes involved in DNA repair have been affected; these would include the *ada* gene, which may be responsible for the shift in base composition, but also possibly several others, not necessarily involved in GC to AT mutation repair (see Table 3). Loss of these genes may have led to an increase in the mutation rate and therefore in the rate of evolution of protein-coding genes, accompanied by a more rapid genome shrinkage and a shift of base composition toward AT. It is worth noting that one likely consequence of this genome-wide compositional shift is the absence of the adaptive codon bias in the genomes of *Prochlorococcus* species MED4 and SS120. AT-rich codons are preferentially used whatever the amino acid (Figure 3a). Thus, codon usage in these genomes appears to reflect more the local base-composition bias than the selection for a more efficient translation through the use of optimal codons. The same conclusion has been drawn for other small genomes with high A+T content [28,38].

Later during evolution (around 80 million years ago, according to the degree of 16S rRNA sequence divergence between MED4 and SS120) one LL population which probably already had a significantly reduced cell and genome size must have progressively adapted to the HL niche and eventually recolonized the upper layer. How this change in ecological niche was possible is still hard to define. Comparison of the gene set that differs between the LL-adapted SS120 and the HL-adapted MED4 shows that very few genes might be sufficient to shift from one to the other niche, including a multiplication of *hli* genes [39] and the differential retention of genes which were present in the common ancestor of *Prochlorococcus* and *Synechococcus*, (such as the photolyases and cyanate transporters mentioned above) and were secondarily lost in the LL-adapted lineages.

Conclusions

Genome evolution in the free-living genus *Prochlorococcus* has similar features to that in host-dependent prokaryotes: genome reduction, bias toward a low G+C content, acceleration in the evolution rate of protein-coding genes, and loss of DNA-repair genes. In contrast to the latter organisms, however, in *Prochlorococcus* this evolution does not appear to be the result of genetic drift or relaxed selection being exerted on some gene categories. Indeed, purifying selection is very efficient in *Prochlorococcus*, as rRNA genes have evolved at a similar rate in all genomes. Despite the decrease in G+C content and an accelerated rate of evolution of protein-coding genes, purifying selection must also act on these genes and avoid potentially deleterious mutations. We hypothesize that a reduction in genome size (which allows a concomitant reduction in cell size and substantial economies in energy and nutrients) can constitute a selective advantage for life in the open ocean, both at depths where photon energy is low and in surface waters where nutrients are scarce.

Genome shrinkage in *Prochlorococcus* has led to populations highly specialized to narrow ecological niches, at the expense of versatility and competitiveness in changing conditions. Indeed, not only is the distribution of the *Prochlorococcus* genus limited to low latitudes (40°N and 40°S, see [34]) but the different ecotypes are themselves more or less confined to a restricted part of the euphotic layer [40]; for example, they experience only limited changes in temperature and salinity. Paradoxically, because warm oligotrophic areas constitute a very large part of the world's oceans, the ecological niches (both LL and HL) occupied by *Prochlorococcus* species are huge, and thus this organism appears globally, despite its specialization, as one of the most successful oxyphototrophs on Earth.

Materials and methods

Genome sequence data

The complete genome sequences and annotations of *Prochlorococcus marinus* MED4, *P. marinus* SS120, *Prochlorococcus* sp. MIT9313 and *Synechococcus* sp. WH8102 (accession numbers: NC_005071, NC_005072, NC_005042 and NC_005070 respectively) were downloaded from the Genome division of the NCBI Entrez system. A few additional genes which were modeled in at least one genome and were present in the other genomes but not modeled (because of their small size, for example) were included in our dataset (see Additional data file 2).

Alignment of whole genomes

Genome sequences translated in their six reading frames were aligned with the Promer program of the MUMmer 3.0 system [41].

Codon and amino-acid usage

Codon usage was computed for every open reading frame (ORF) of each genome with the EMBOSS program cusp. Amino-acid usage was derived from the results produced by cusp.

Identification of orthologous proteins

We used a sequence-similarity based approach which is similar to the procedure used for the cluster of orthologous groups (COGs [42]). For each genome pair, all-against-all BLAST [43] comparisons were performed using protein sequences and reciprocal genome-specific best hits were identified. We

considered genes as being probable orthologs when they were included in groups of size four in which each gene was the best hit of the three others. From similarity searches against the COG database, orthologs were assigned to functional categories according to those defined for the COG system. Because of the lack of a particular category for photosynthesis genes, the latter were assigned to the 'energy production and conversion' COG category. Other genes which fell into more than one of the 19 COG categories have been assigned to a supplementary category called 'miscellaneous'.

Phylogenetic branch length estimations

Protein sequences from each of the groups of four orthologous genes were aligned using ClustalW [44] with default parameters. After exclusion of all gap sites, individual alignments were concatenated in one super-alignment of 388,120 sites. Gamma distances [45] with an alpha parameter of 1 were estimated between each pair of sequences of the super-alignment. Phylogenetic branch lengths were calculated from distances with the ordinary least-squares method [45]. Relative rate tests (two-cluster test and Branch length test) were applied in order to test the constancy of amino-acid substitution rates between the three *Prochlorococcus* genomes (hypothesis of the molecular clock). The same analysis was applied to orthologs of each functional category.

Estimate of synonymous and nonsynonymous substitution rates

Nucleotide sequences of each group of orthologs were aligned with Protal2dna according to alignments of their corresponding amino-acid sequences [46]. Pairwise estimates of the synonymous (d_S) and non-synonymous (d_N) substitution rates were obtained from the Yn00 program of the PAML 3.13 package [47].

Additional data files

The following additional data are available with the online version of this article. Additional data file 1 lists the orthologous genes classified by functional category. Orthologous genes were assigned to the functional categories of COG system. Photosynthesis genes were assigned to the 'energy production and conversion' COG category. Genes falling in more than one of the 19 COG categories have been assigned to a supplementary category called 'miscellaneous'. Additional data file 2 is a fasta file of orthologous genes which were modeled in at least one genome and present but not modeled in the other genomes.

Acknowledgements

We are very grateful to Martin Ostrowski and Dave Scanlan for their critical reading of the manuscript. This work was supported by the European Union Program MARGENES (QLRT-2001-01226), the EU FP6 Network of Excellence 'Marine Genomics Europe' and by the French programs Genomer (Région Bretagne) and Ouest-Genopole. AD is supported by a doctoral fellowship from Région Bretagne.

References

1. Lawrence JG, Roth JR: **Genomic flux: genome evolution by gene loss and acquisition.** In *Organization of the Prokaryotic Genome* Edited by: Charlebois RL. Washington, DC: American Society for Microbiology; 1999:263-289.
2. Andersson SG, Kurland CG: **Reductive evolution of resident genomes.** *Trends Microbiol* 1998, **6:**263-268.
3. Martin W: **Gene transfer from organelles to the nucleus: frequent and in big chunks.** *Proc Natl Acad Sci USA* 2003, **100:**8612-8614.
4. Fraser CM, Gocayne JD, White O, Adams MD, Clayton RA, Fleischmann RD, Bult CJ, Kerlavage AR, Sutton G, Kelley JM, *et al.*: **The minimal gene complement of *Mycoplasma genitalium*.** *Science* 1995, **270:**397-403.
5. Oshima K, Kakizawa S, Nishigawa H, Jung HY, Wei W, Suzuki S, Arashida R, Nakata D, Miyata S, Ugaki M, Namba S: **Reductive evolution suggested from the complete genome sequence of a plant-pathogenic phytoplasma.** *Nat Genet* 2004, **36:**27-29.
6. Shigenobu S, Watanabe H, Hattori M, Sakaki Y, Ishikawa H: **Genome sequence of the endocellular bacterial symbiont of aphids *Buchnera* sp. APS.** *Nature* 2000, **407:**81-86.
7. Tamas I, Klasson L, Canback B, Naslund AK, Eriksson AS, Wernegreen JJ, Sandstrom JP, Moran NA, Andersson SG: **50 million years of genomic stasis in endosymbiotic bacteria.** *Science* 2002, **296:**2376-2379.
8. van Ham RC, Kamerbeek J, Palacios C, Rausell C, Abascal F, Bastolla U, Fernandez JM, Jimenez L, Postigo M, Silva FJ, *et al.*: **Reductive genome evolution in *Buchnera aphidicola*.** *Proc Natl Acad Sci USA* 2003, **100:**581-586.
9. Waters E, Hohn MJ, Ahel I, Graham DE, Adams MD, Barnstead M, Beeson KY, Bibbs L, Bolanos R, Keller M, *et al.*: **The genome of *Nanoarchaeum equitans*: insights into early archaeal evolution and derived parasitism.** *Proc Natl Acad Sci USA* 2003, **100:**12984-12988.
10. Conway DJ, Fanello C, Lloyd JM, Al-Joubori BM, Baloch AH, Somanath SD, Roper C, Oduola AM, Mulder B, Povoa MM, *et al.*: **Origin of *Plasmodium falciparum* malaria is traced by mitochondrial DNA.** *Mol Biochem Parasitol* 2000, **111:**163-171.
11. Kohler S, Delwiche CF, Denny PW, Tilney LG, Webster P, Wilson RJ, Palmer JD, Roos DS: **A plastid of probable green algal origin in Apicomplexan parasites.** *Science* 1997, **275:**1485-1489.
12. Moran NA: **Microbial minimalism: genome reduction in bacterial pathogens.** *Cell* 2002, **108:**583-586.
13. Dufresne A, Salanoubat M, Partensky F, Artiguenave F, Axmann IM, Barbe V, Duprat S, Galperin MY, Koonin EV, Le Gall F, *et al.*: **Genome sequence of the cyanobacterium *Prochlorococcus marinus* SS120, a nearly minimal oxyphototrophic genome.** *Proc Natl Acad Sci USA* 2003, **100:**10020-10025.
14. Rocap G, Larimer FW, Lamerdin J, Malfatti S, Chain P, Ahlgren NA, Arellano A, Coleman M, Hauser L, Hess WR, *et al.*: **Genome divergence in two *Prochlorococcus* ecotypes reflects oceanic niche differentiation.** *Nature* 2003, **424:**1042-1047.
15. Hess WR: **Genome analysis of marine photosynthetic microbes and their global role.** *Curr Opin Biotechnol* 2004, **15:**191-198.
16. Partensky F, Blanchot J, Vaulot D: **Differential distribution and ecology of *Prochlorococcus* and *Synechococcus* in oceanic waters: a review.** In *Marine Cyanobacteria* Edited by: Charpy L, Larkum AWD. Monaco: Musée Océanographique; 1999:457-475.
17. Garcia-Pichel F, Belnap J, Neuer S, Schanz F: **Estimates of cyanobacterial biomass and its distribution.** *Archiv Hydrobiol* 2003, **109(Suppl 148):**213-228.
18. Moore LR, Rocap G, Chisholm SW: **Physiology and molecular phylogeny of coexisting *Prochlorococcus* ecotypes.** *Nature* 1998, **393:**464-467.
19. Moore LR, Chisholm SW: **Photophysiology of the marine cyanobacterium *Prochlorococcus*: ecotypic differences among cultured isolates.** *Limnol Oceanogr* 1999, **44:**628-638.
20. Rocap G, Distel DL, Waterbury JB, Chisholm SW: **Resolution of *Prochlorococcus* and *Synechococcus* ecotypes by using 16S-23S ribosomal DNA internal transcribed spacer sequences.** *Appl Environ Microbiol* 2002, **68:**1180-1191.
21. Fuller NJ, Marie D, Partensky F, Vaulot D, Post AF, Scanlan DJ: **Clade-specific 16S ribosomal DNA oligonucleotides reveal the predominance of a single marine *Synechococcus* clade throughout a stratified water column in the Red Sea.** *Appl Environ Microbiol* 2003, **69:**2430-2443.

22. Palenik B, Brahamsha B, Larimer FW, Land M, Hauser L, Chain P, Lamerdin J, Regala W, Allen EE, McCarren J, *et al.*: **The genome of a motile marine *Synechococcus*.** *Nature* 2003, **424:**1037-1042.
23. Moran NA: **Tracing the evolution of gene loss in obligate bacterial symbionts.** *Curr Opin Microbiol* 2003, **6:**512-518.
24. Mackay WJ, Han S, Samson LD: **DNA alkylation repair limits spontaneous base substitution mutations in *Escherichia coli*.** *J Bacteriol* 1994, **176:**3224-3230.
25. Michaels ML, Cruz C, Grollman AP, Miller JH: **Evidence that MutY and MutM combine to prevent mutations by an oxidatively damaged form of guanine in DNA.** *Proc Natl Acad Sci USA* 1992, **89:**7022-7025.
26. Horst JP, Wu TH, Marinus MG: ***Escherichia coli* mutator genes.** *Trends Microbiol* 1999, **7:**29-36.
27. Moran NA, Mira A: **The process of genome shrinkage in the obligate symbiont *Buchnera aphidicola*.** *Genome Biol* 2001, **2:**research0054.1-0054.12.
28. Wernegreen JJ, Moran NA: **Evidence for genetic drift in endosymbionts (*Buchnera*): analyses of protein-coding genes.** *Mol Biol Evol* 1999, **16:**83-97.
29. Moran NA: **Accelerated evolution and Muller's rachet in endosymbiotic bacteria.** *Proc Natl Acad Sci USA* 1996, **93:**2873-2878.
30. Lambert JD, Moran NA: **Deleterious mutations destabilize ribosomal RNA in endosymbiotic bacteria.** *Proc Natl Acad Sci USA* 1998, **95:**4458-4462.
31. Koonin EV, Mushegian AR, Rudd KE: **Sequencing and analysis of bacterial genomes.** *Curr Biol* 1996, **6:**404-416.
32. Glass JI, Lefkowitz EJ, Glass JS, Heiner CR, Chen EY, Cassell GH: **The complete sequence of the mucosal pathogen *Ureaplasma urealyticum*.** *Nature* 2000, **407:**757-762.
33. Akman L, Yamashita A, Watanabe H, Oshima K, Shiba T, Hattori M, Aksoy S: **Genome sequence of the endocellular obligate symbiont of tsetse flies, *Wigglesworthia glossinidia*.** *Nat Genet* 2002, **32:**402-407.
34. Partensky F, Hess WR, Vaulot D: ***Prochlorococcus*, a marine photosynthetic prokaryote of global significance.** *Microbiol Mol Biol Rev* 1999, **63:**106-127.
35. Ohta T: **The nearly neutral theory of molecular evolution.** *Annu Rev Ecol Syst* 1992, **23:**263-286.
36. Chisholm SW: **Phytoplankton size.** In *Primary Productivity and Biogeochemical Cycles in the Sea* Edited by: Falkowski PG, Woodhead AD. New York: Plenum Press; 1992:213-237.
37. Ochman H, Wilson AC: **Evolution in bacteria: evidence for a universal substitution rate in cellular genomes.** *J Mol Evol* 1987, **26:**74-86.
38. Andersson SG, Sharp PM: **Codon usage and base composition in *Rickettsia prowazekii*.** *J Mol Evol* 1996, **42:**525-536.
39. Bhaya D, Dufresne A, Vaulot D, Grossman A: **Analysis of the *hli* gene family in marine and freshwater cyanobacteria.** *FEMS Microbiol Lett* 2002, **215:**209-219.
40. West NJ, Scanlan DJ: **Niche-partitioning of *Prochlorococcus* populations in a stratified water column in the eastern North Atlantic Ocean.** *Appl Environ Microbiol* 1999, **65:**2585-2591.
41. Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, Salzberg SL: **Versatile and open software for comparing large genomes.** *Genome Biol* 2004, **5:**R12.
42. Tatusov RL, Koonin EV, Lipman DJ: **A genomic perspective on protein families.** *Science* 1997, **278:**631-637.
43. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Res* 1997, **25:**3389-3402.
44. Thompson JD, Higgins DG, Gibson TJ: **CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice.** *Nucleic Acids Res* 1994, **22:**4673-4680.
45. Nei M, Kumar S: *Molecular Evolution and Phylogenetic* Oxford: Oxford University Press; 2000.
46. **protal2dna** [http://bioweb.pasteur.fr/seqanal/interfaces/protal2dna.html]
47. Yang Z: **PAML: A program package for phylogenetic analysis by maximum likelihood.** *Comput Appl Biosci* 1997, **13:**555-556.

CHAPITRE V

Conclusions et Perspectives

Les génomes de *Prochlorococcus marinus* SS120 et MED4 contiennent un répertoire de gènes très restreint par rapport aux autres cyanobactéries marines et surtout d'eau douce. Ces génomes n'en restent pas moins nettement plus grands que les plus petits génomes connus (par exemple celui de *Mycoplasma genitali*um, 580 Kb) et ils contiennent un nombre de gène très supérieur à celui estimé pour un génome minimal théorique (~250). Pourtant, ils semblent posséder à peine plus du minimum nécessaire au développement, au fonctionnement et à la reproduction d'un organisme oxyphototrophe libre. Ainsi, l'étude de ces génomes permet d'aborder sous un angle différent l'idée de génome minimal, c'est à dire dans des conditions autres que celles d'un milieu riche, stable et sans compétiteurs.

La petite taille des génomes de *Prochlorococcus* est directement le résultat d'un processus d'évolution réductive qui n'avait jusqu'à présent jamais été observé chez une bactérie à mode de vie libre. Ainsi, il semble que ce processus évolutif ne soit pas simplement limité aux organismes qui vivent en association avec un hôte. On peut s'interroger sur l'importance de ce phénomène dans l'évolution des procaryotes à cycle de vie libre. Par exemple, il est intéressant de se demander si d'autres bactéries, et plus particulièrement celles présentes dans le même environnement oligotrophe que *Prochlorococcus,* ont évolué de manière similaire. De façon intéressante, cela ne semble pas être le cas pour *Synechococcus* du sous-cluster 5.1 (souches marines strictes) puisque tous les génomes étudiés à ce jour ont une taille aux alentours de 2,4 Mbp, et cela quel que soit le clade (M. Ostrowski, pers. comm). A un niveau évolutif plus large, on peut aussi s'interroger sur l'existence de ce phénomène chez les organismes eucaryotes du picophytoplancton tel qu'*Ostreococcus*. Celui-ci possède le plus petit génome et la plus petite taille de cellule connue chez un organisme eucaryote à l'heure actuelle. La position phylogénétique de cette picoalgue indique que cette taille minuscule est un caractère dérivé et non ancestral (Courties et al. 1998). Toutefois, les différences importantes d'habitat entre *Prochlorococcus* et *Ostreococcus*, ce dernier étant quasiment absent du milieu oligotrophe et ne proliférant apparemment qu'en milieu riche (ex: dans l'étang de Thau), laisse penser que les causes de la réduction pourraient être différentes de celles identifiées chez *Prochlorococcus*.

V.1 Différenciation écotypique chez *Prochlorococcus*

Les deux écotypes de *Prochlorococcus* ont tout d'abord été définis en fonction des différences d'intensité lumineuse à laquelle leur croissance était optimale et leur différence de rapport Chl *b* / Chl a_2, un paramètre lié à la capacité d'absorption des photons bleus (Moore and Chisholm 1999; Moore et al. 1995; Moore et al. 1998). Par la suite, il a été avancé que la

différenciation écotypique devait aussi inclure le mode de nutrition, puisque dans les eaux océaniques oligotrophes, les concentrations de sels nutritifs inorganiques sont toujours très faibles dans la couche supérieure (80-120 m), alors qu'ils sont abondants dans le bas de la couche euphotique, domaine exclusif des souches dites "de basse lumière". Toutefois, les résultats obtenus en comparant les cinq génomes suggèrent que la situation est bien plus complexe que l'on ne l'avait envisagé au départ. En effet, les analyses réalisées montrent une grande variabilité dans la capacité d'utilisation des sources d'azote et de phosphate. Par exemple, Rocap et collaborateurs (Rocap et al. 2003) ont émis l'hypothèse, en limitant leur analyse aux génomes de *P. marinus* MED4 et *Prochlorococcus* sp. MIT9313, que les deux écotypes posséderaient les gènes permettant l'utilisation de la forme d'azote qui prédomine dans leur niche écologique respective, les nitrites pour la niche de profondeur et l'urée pour la souche de surface. Cependant l'absence des gènes permettant l'utilisation des nitrates, des nitrites et de l'urée chez la souche SS120 (Dufresne et al. 2003) démontre bien la grande variabilité qui existe entre souches de profondeur et la difficulté de relier ces variations aux paramètres environnementaux. Ainsi, autant la lumière apparaît comme un paramètre primordial pour expliquer l'évolution de *Prochlorococcus* (multiplication du nombre de gènes *pcb* dans certaines souches de profondeur comme SS120 ou NATL2, réduction dans certaines souches de surface comme MED4 ; multiplication des gènes *hli* et retention différentielle de gènes de réparation de l'ADN chez les souches de surface, etc.), autant la disponibilité en sels nutritifs n'a pas eu un impact aussi clair. Cependant, ce dernier paramètre pourrait quand même avoir joué un rôle indirect sur l'évolution du clade "forte lumière" vers une petite taille cellulaire, un caractère qui semble important pour l'adaptation à la niche de surface, car il permet une optimisation "mécanique" (sans dépenser d'énergie spécifique) de l'absorption des sels nutritifs (chapitre IV).

Il apparaît clairement au regard des caractéristiques de leurs génomes (taille, % GC, nombre de gènes) que la souche de profondeur *P. marinus* SS120 est beaucoup plus proche de la souche de surface *P. marinus* MED4 que de la seconde souche de profondeur *Prochlorococcus* sp. MIT9313. Il ressort donc de cette étude que, bien qu'appartenant tous les deux à l'écotype de basse lumière, *P. marinus* SS120 et *Prochlorococcus* sp. MIT9313 ont suivi des trajectoires évolutives séparées en réponse à des pressions évolutives différentes.

Comme cela a été évoqué dans l'introduction, les souches de basse lumière ne forment pas un clade monophylétique comme c'est le cas pour l'écotype de haute lumière. Au contraire, plusieurs clades différents peuvent être identifiés sur la base des analyses phylogénétiques (Rocap et al. 2002). Cette polyphylie des souches de profondeur semble indiquer l'existence de plusieurs écotypes se développant dans des niches écologiques séparées. En effet, si ces souches ont rencontrées des conditions environnementales identiques

ou du moins très similaires, la sélection devrait avoir pour effet de purger périodiquement la diversité entre les souches de basse lumière en sélectionnant les souches porteuses d'une mutation avantageuse (Cohan 2004). La répétition de ce processus aurait assurément conduit à la formation d'un clade monophylétique, ce qui n'est pas la situation observée aujourd'hui.

Il est intéressant de noter que la majorité des souches cultivées de l'écotype de basse lumière ont probablement un génome réduit, similaire à celui de *P. marinus* SS120. En effet, la taille et le pourcentage en GC de la région ITS (situé entre les gènes ARNr 16S et ARNr 23S) diminue avec la réduction de la taille du génome entre *Prochlorococcus* sp. MIT9313 et *P. marinus* MED4. L'analyse de la séquence des ITS d'un grand nombre de souches de *Prochlorococcus* et de *Synechococcus* montre que les souches *Prochlorococcus* spp. MIT9313 et MIT93103, ont un ITS dont la longueur et le pourcentage en GC sont très similaires à ceux des souches de *Synechococcus* (Rocap et al. 2002). Ces deux souches sont situées à la base de la radiation des *Prochlorococcus* dans l'arbre phylogénétique basé sur la séquence des ITS. Les autres souches de l'écotype de basse lumière présentent toutes des ITS plus petits et moins riches en GC et dont les caractéristiques sont très similaires à celle de SS120 et de *P. marinus* NATL2A, qui vient d'être séquencée aux Etats-Unis, et dont la taille de génome est très proche de celle de SS120.

Les caractéristiques « ancestrales » de la souche *Prochlorococcus* sp. MIT9313 font penser que les premiers *Prochlorococcus* sont d'abord apparus dans la niche de profondeur. Au cours de l'évolution de ce genre, une réduction du génome a eu lieu, entraînant la formation de souches à « petit génome » Grâce par exemple à la réduction concomitante du volume cellulaire, cette réduction pourrait constituer une adaptation à la faible luminosité de la niche de profondeur. Ces souches se sont ensuite diversifiées pour occuper des niches écologiques différentes. *A contrario*, certaines souches ont conservé un génome de grande taille. Il est possible que ces souches occupent une niche écologique où l'avantage adaptatif apporté par la réduction de génome n'existe pas. *Prochlorococcus* sp. MIT9313 a été isolée d'une station située dans le courant du Gulf Steam qui est sensiblement plus riche en sels nutritifs que la Mer des Sargasses où a été isolée la souche *P. marinus* SS120. Même si les caractéristiques exactes de ces niches ne sont pas connues à l'heure actuelle, celles-ci ont en commun la rareté des photons disponibles pour la photosynthèse. Ainsi, contre toute attente, la niche de basse lumière semble correspondre à une super-niche englobant différents types de niches plus petites occupées par des écotypes distincts.

V.2 Conséquences de la réduction du génome chez *Prochlorococcus*

La perte de gènes chez *Prochlorococcus* se traduit par un réarrangement important du génome des souches *P. marinus* MED4 et *P. marinus* SS120 par rapport à celui de *Prochlorococcus* sp. MIT9313. En conséquence, un certain nombre d'unités transcriptionnelles se retrouvent séparées en plusieurs parties chez les deux génomes réduits. De même, la réduction du génome est aussi caractérisée par la diminution de la taille moyenne des régions intergéniques. Dans de nombreux cas, ces régions ont disparu presque complètement chez *P. marinus* MED4 et SS120. Les régions intergéniques contiennent les séquences régulatrices contrôlant l'expression des gènes. L'effet de la réduction du génome sur la régulation de l'expression des gènes chez *Prochlorococcus* mériterait d'être étudiée plus particulièrement. Un phénomène similaire a été observé chez la bactéries endosymbiotique *Buchnera* (Moran and Mira 2001) et il semble que, chez cette dernière, l'expression des gènes soit relativement indépendante des conditions environnementales.

V.3 Evolution de *Prochlorococcus* et de *Synechococcus*: deux stratégies différentes ?

La caractéristique principale de l'évolution de *Prochlorococcus* est la réduction du génome associée à une réduction de la taille cellulaire et à l'utilisation d'un système plus économique en azote pour la capture de l'énergie des photons. Cette évolution peut être vue comme une spécialisation permettant à *Prochlorococcus* d'exploiter au mieux les ressources limitées des écosystèmes océaniques oligotrophes. Les densités de populations très élevées que *Prochlorococcus* atteint dans les milieux extrêmement oligotrophes viennent corroborer cette hypothèse. Cependant, cette stratégie évolutive a un coût qui est la perte de la capacité d'adaptation à d'autres environnements. En conséquence, *Prochlorococcus* possède une distribution géographique plus réduite que celle de *Synechococcus*.

La stratégie suivie par ce dernier apparaît radicalement différente de celle de *Prochlorococcus*. Comme mentionné plus haut, la taille du génome semble relativement constante (~2,4 Mb) au sein de ce groupe. Ainsi, ces cyanobactéries ne semblent pas avoir subi de pertes massives de gènes au cours de leur évolution. *Synechococcus* a également conservé un système antennaire plus coûteux en azote (phycobilisome).

Ainsi, *Synechococcus* est plus généraliste et a développé une stratégie plus opportuniste (capable de répondre rapidement à des changements des conditions environnementales) que celle de *Prochlorococcus*. La stratégie de *Synechococcus* se révèle moins efficace en milieu très oligotrophe, comme semble l'indiquer sa plus faible abondance par rapport à *Prochlorococcus* dans ces écosystèmes. Par contre, ce genre a pu coloniser un plus grand

nombre d'environnements et ses possibilités d'adaptation semblent plus importantes que celles de *Prochlorococcus*. En effet, *Synechococcus* apparaît mieux équipé pour faire face aux variations de conditions environnementales à plus ou moins longs termes, et son répertoire de gènes plus large offre un potentiel plus développé pour l'acquisition de nouvelles fonctions.

Annexe I

Table 3. Genes involved in structure or major metabolic processes which are absent or in lower copy numbers in *P. marinus* SS120 (Pma) than in the freshwater cyanobacteria *Thermosynechococcus elongatus* (Tel), *Synechocystis* sp. PCC6803 (Syn), and *Anabaena* sp. PCC7120 (Ana).

Functional system, protein	Gene name	COG no.	Number of genes per genome			
			Pma	Tel	Syn	Ana
PHOTOSYNTHESIS						
Allophycocyanin	*apcA*	-	0	2	1	2
	apcBCDE		0	1	1	1
	apcF		0	2	1	1
Phycocyanin	*cpcABDEF*	-	0	1	1	1
	cpcC		0	1	2	1
	cpcG		0	3	2	4
PSII reaction center D1	*psbA*	-	1	3	3	5
PSII reaction center D2	*psbD*	-	1	2	2	2
PSII 12 kDa extrinsic protein	*psbU*	-	0	1	1	1
Cytochrome c_{550}	*psbV*	-	0	1	1	1
Oxygen-independent coproporphyrinogen III oxidase	*hemN*	0635	1	2	2	2
Mg-protoporphyrin IX monomethylester aerobic cyclization system	*acsF, crd1*	-	1	2	2	3
Heme oxygenase	*ho*	5398	1	2	2	2
INFORMATIONAL SYSTEMS						
Translation factor SUA5	*SUA5*	0009	1	2	3	2
Asparaginyl-tRNA synthetase	*asnS*	0017	0	1	1	1
Ribosomal protein L25	*rplY*	1825	0	1	1	1
DNA polymerase III alpha subunit	*dnaE*	0587	1	2*	2*	2*
DNA repair ATPase	*sbcC*	0419	1	2	3	4
Superfamily I DNA helicase	*uvrD*	0210	1	2	2	3
DNA mismatch repair protein	*mutL*	0323	0	1	1	1
Deoxyribodipyrimidine photolyase/cryptochrome	*phrA*	0415	0	2	2	2
Ribonuclease HI/HII	*rnhA*	0328	1	2	2	2
DNA-binding protein HupA	*hupA*	0783	0	2	1	4
16.6 kDa heat shock protein	*ipbA*		0	1	1	2
ATPases involved in chromosome partitioning	*soj*	1192	0	3	4	9
DnaJ class chaperone	*dnaJ*	2214	2	5	5	11
Zn-dependent protease with chaperone function	*htpX*	501	0	2	3	6
Site-specific recombinases	*pin*	1961	0	2	1	2
Circadian clock protein KaiA	*kaiA*	-	0	1	1	1
Circadian clock protein KaiB	*kaiB*	0526	1	3	2	2
TRANSPORT SYSTEMS						
CO_2 uptake protein CupA/CupB	*cupA*	-	0	2	2	2
NAD(P)H-quinone oxidoreductase, NdhF subunit	*ndhF*	1009	1	3	3	3
NAD(P)H-quinone oxidoreductase, NdhD subunit	*ndhD*	1008	2	4	4	5
Ammonia permease	*amt*	0004	1	2	3	3
ABC-type urea transporter	*urtABCDE*	-	0	1	1	1
Oxyanion-transporting ATPase	*arsA*	0003	0	2	2	3
Cu/Zn-transport ATPase	*zntA*	2217	1	3	5	11
Permeases of the major facilitator superfamily	*proP*	0477	4	12	9	20
Ferrous iron transport protein B	*feoB*	0370	0	2	1	1
METABOLISM						
Nitrate reductase	*narB*	0243	0	1	1	1
Nitrite reductase	*nirA*	0155	0	1	1	1
6-Phosphofructokinase	*pfkA*	0205	0	1	2	1
Glycerol-3-phosphate dehydrogenase	*gpsA*	0240	0	1	1	1
Acetate kinase	*ackA*	0282	0	1	1	1

Fructose-2,6-bisphosphatase	*gpmB*	0406	1	4	3	6
Malic enzyme	*sfcA*	0281	0	1	1	1
Phosphoenolpyruvate synthase/ pyruvate phosphate dikinase	*ppsA*	0574	0	2	1	5
Xylulose 5-phosphate phosphoketolase	*xpkA*	3957	0	2	2	3
ADP-ribose pyrophosphatase	*nudF*	1051	0	2	3	5
Urease subunits and accessory proteins	*ureABC*	-	0	1	1	1
	ureDEFG	-	0	1	1	1
Cytochrome bd oxidase subunit I	*cydA*	1271	0	1	1	1
Cytochrome bd oxidase subunit II	*cydB*	1294	0	1	1	1
Cyanophycin synthetase	*cphA*	-	0	1	1	3
Cyanophycinase	*cphB*	-	0	1	1	3
Cell wall amidohydrolases	*nlpD*	0739	0	4	4	8
PILUS BIOGENESIS**						
Pilin	*pilA*	2165	0	2	1	1
Membrane protein with Walker box motif	*pilB*	2804	0	1	1	1
Inner membrane protein	*pilC*	1459	0	2	1	1
ABC transporter ATP-binding protein	*pilH*	1131	0	1	1	1
Pilus assembly protein PilM	*pilM*	4972	0	1	1	1
Pilus assembly protein PilN	*pilN*	3166	0	1	1	1
Pilus assembly protein PilO	*pilO*	3167	0	1	1	1
Secretin	*pilQ*	4786	0	1	1	1
Membrane protein with Walker box motif	*pilT*	2805	0	2	2	2
OTHER MOTILITY-RELATED PROTEINS**						
Motility-related protein kinase	*spkA*	0515	0	1	2	3
Pentapeptide repeats-containing protein	*ppr1*	1357	0	2	1	1
	ppr2	1357	0	6	9	17

*Two separate *dnaE* genes give rise to N- and C-terminal parts of DNA polymerase III alpha subunit that are joined posttranslationally by trans-splicing.
Taken from Bhaya, D., Takahashi, A., Shahi, P. & Grossman, A. R. (2001) *J. Bacteriol.* **183, 6140–6143 and D. Bhaya (personal communication).

Annexe II

Annexe II (article 3):

Analyse de la famille des gènes *hli* chez les cyanobactéries marines et d'eau douce

Résumé

Certaines cyanobactéries sont capables de se développer dans des habitats où l'intensité lumineuse peut atteindre 2000 µmol photon m^{-2} s^{-1} et où les concentrations en sels nutritifs sont extrêmement faibles. Récemment, il a été démontré qu'une famille de gènes, dénommés *hli*, est importante pour la survie des cyanobactéries lorsqu'elles sont exposées à de fortes intensités lumineuses.

Dans cette étude, nous avons identifié les membres de cette famille dans sept génomes de cyanobactéries :

- celui d'une cyanobactérie marine adaptée aux fortes lumières de la niche de surface (*Prochlorococcus marinus* MED4),
- ceux de trois cyanobactéries marines adaptées à des intensités lumineuses moyennes ou basses (*Prochlorococcus marinus* SS120, *Prochlorococcus* sp. MIT9313 et *Synechococcus* sp. WH8102)
- ceux de trois cyanobactéries d'eau douce (la cyanobactérie unicellulaire *Synechocystis* sp. PCC 6803 et les cyanobactéries filamenteuses *Nostoc punctiforme* ATCC29133 et *Anabaena* sp. PCC 7120).

La souche de haute lumière *P. marinus* MED4 a le plus petit génome (1,66 Mb). Cependant, elle possède deux fois plus de gènes *hli* que n'importe laquelle des six autres espèces de cyanobactéries. Certains de ces gènes semblent être le résultat de duplications récentes. La classification de ces gènes en clusters indique que certains sont spécifiques des cyanobactéries marines ou de celles d'eau douce. Ces résultats sont interprétés en fonction du rôle des gènes *hli* dans l'acclimatation des cyanobactéries aux intensités lumineuses élevées ainsi qu'en fonction des relations évolutives possibles existant entre les membres de cette famille très variable.

ELSEVIER

FEMS Microbiology Letters 215 (2002) 209–219

FEMS MICROBIOLOGY Letters

www.fems-microbiology.org

Analysis of the *hli* gene family in marine and freshwater cyanobacteria

Devaki Bhaya [a,*], Alexis Dufresne [b], Daniel Vaulot [b], Arthur Grossman [a]

[a] *Department of Plant Biology, Carnegie Institution of Washington, 260 Panama Street, Stanford, CA 94305, USA*
[b] *Station Biologique, UMR 7127, CNRS, INSU et Université Pierre et Marie Curie, 29682 Roscoff Cedex, France*

Received 23 July 2002; received in revised form 15 August 2002; accepted 16 August 2002

First published online 11 September 2002

Abstract

Certain cyanobacteria thrive in natural habitats in which light intensities can reach 2000 μmol photon m^{-2} s^{-1} and nutrient levels are extremely low. Recently, a family of genes designated *hli* was demonstrated to be important for survival of cyanobacteria during exposure to high light. In this study we have identified members of the *hli* gene family in seven cyanobacterial genomes, including those of a marine cyanobacterium adapted to high-light growth in surface waters of the open ocean (*Prochlorococcus* sp. strain Med4), three marine cyanobacteria adapted to growth in moderate- or low-light (*Prochlorococcus* sp. strain MIT9313, *Prochlorococcus marinus* SS120, and *Synechococcus* WH8102), and three freshwater strains (the unicellular *Synechocystis* sp. strain PCC6803 and the filamentous species *Nostoc punctiforme* strain ATCC29133 and *Anabaena* sp. {*Nostoc*} strain PCC7120). The high-light-adapted *Prochlorococcus* Med4 has the smallest genome (1.7 Mb), yet it has more than twice as many *hli* genes as any of the other six cyanobacterial species, some of which appear to have arisen from recent duplication events. Based on cluster analysis, some groups of *hli* genes appear to be specific to either marine or freshwater cyanobacteria. This information is discussed with respect to the role of *hli* genes in the acclimation of cyanobacteria to high light, and the possible relationships among members of this diverse gene family.

1. Introduction

The major peripheral light-harvesting complex (LHC) of cyanobacteria is the water-soluble phycobilisome, which is comprised of tetrapyrrole-bound phycobiliproteins and non-pigmented linker polypeptides [1]. In vascular plants, the major LHC is composed primarily of the integral membrane Lhc polypeptides that contain three transmembrane helices and bind chlorophylls *a* and *b*. The Lhc polypeptides are encoded by a family of genes that generally contains more than 10 individuals [2–4]. However, there are more distantly related *Lhc* genes that comprise the *Lhc* extended gene family. Polypeptides encoded by this extended family include the early light-inducible proteins (ELIPs), the four transmembrane helix-containing polypeptide PsbS or PSII-S [5], and polypeptides that have one or two putative transmembrane helices [6,7]. Several genes have been identified on cyanobacterial genomes that encode single-helix members of the *Lhc* extended gene family [8,9]. These genes have been designated *hli* (high light inducible; protein designation HLIPs) [8] or *scp* (small cab-like proteins; protein designation Scps) [9]. While *hli* genes were first noted in *Synechococcus* sp. strain PCC7942 [10], they were subsequently identified in other cyanobacteria [11], red algae [12] and vascular plants [6].

The pattern of expression of *hli* genes in cyanobacteria and vascular plants is similar to that of the genes encoding ELIPs; *hli* mRNAs and encoded polypeptides accumulate under conditions that result in the absorption of excess excitation energy, including exposure to high irradiance, nitrogen starvation and low temperature [1,6,8]. *Synechocystis* sp. strain PCC6803 deleted for all four of its *hli* genes was shown to be photosensitive, and under strong illumination the cells lost all variable fluorescence and died [13]. By analogy to vascular plant ELIPs, a number of functions have been suggested for the cyanobacterial HLIPs. They may associate with pigments, perhaps transiently, serving as chlorophyll carriers [11], function in the dissipation of excess absorbed light energy within antennae complexes [6,14], or modulate the biosynthesis of chlorophyll [15]. Essentially all evidence suggests that photo-

* Corresponding author. Tel.: +1 (650) 325 1521, ext. 282; Fax: +1 (650) 325 6857.
E-mail address: devaki@andrew2.stanford.edu (D. Bhaya).

0378-1097/02/$22.00
PII: S0378-1097(02)00934-5

synthetic organisms need HLIPs under stressful, often growth-limiting conditions.

Specific *Prochlorococcus* species thrive in high-light, nutrient-poor environments that characterize the surface waters of the open oceans, while others grow at greater depths in lower-light and higher-nutrient environments [16,17]. Although phylogenetic analyses, based on 16S rDNA and *rpoC* sequences, have established that *Prochlorococcus* is a cyanobacterial genus, it does not contain the light harvesting phycobilisomes typical of most cyanobacteria [18]. Instead, the major antennae pigment complex contains chlorophylls *a* and *b* associated with polypeptides that are similar to CP43, a polypeptide integral to the core of photosystem II that binds chlorophyll *a* [19].

An understanding of the light and nutrient habitats in which specific cyanobacterial strains thrive, and recent acquisition of complete or near complete sequence information for several cyanobacterial genomes, have made it attractive to explore both intra- and inter-species relationships among cyanobacterial *hli* genes. The genomes of seven cyanobacterial strains have been sequenced at the Kazusa DNA Research Institute, Japan; the Joint Genome Institute (JGI), USA and the Genoscope, France, and the sequences of other cyanobacterial genomes are nearly complete. Three of these cyanobacteria, the unicellular *Synechocystis* strain PCC6803 (**SC**), the filamentous *Anabaena (Nostoc)* sp. strain PCC7120 (**AN**) and *Nostoc punctiforme* strain ATCC29133 (**NT**) grow in freshwater habitats. The marine species for which complete genome information is available are represented by the high-light ecotype *Prochlorococcus marinus* strain MED4 (**PM**) and three species that grow in low/moderate light, *P. marinus* strain MIT9313 (**PL**), *P. marinus* strain SS120 (**PS**) and *Synechococcus* sp. strain WH8102 (**SN**) [17,20]. In this report we analyze the cyanobacterial *hli* gene family and discuss the differences in this gene family among the seven different cyanobacterial strains for which complete genome sequences are available.

2. Materials and methods

2.1. Genome data

Sequences of the genomes of *Synechocystis* and *Anabaena* were downloaded from the EMBL web site (http://www.ebi.ac.uk/genomes/); those of the *Prochlorococcus* strains MED4 (version of 12/19/2001), MIT9313 (version of 01/25/2002), and *Synechococcus* strain WH8102 (version of 01/26/2002) from ftp sites of the JGI (ftp://ftp.jgi-psf.org/pub/JGI_data/Microbial/prochlorococcus/final.011129; ftp://ftp.jgi-psf.org/pub/JGI_data/Microbial/prochlorococcusII/final.010823; ftp://ftp.jgi-psf.org/pub/JGI_data/Microbial/synechococcus/final.010910). Contig sequences of the genome of *N. punctiforme* (version of 01/25/2002) were downloaded from the JGI ftp site: (ftp://ftp.jgi-psf.org/pub/JGI_data/Microbial/nostoc/010409/). The genomic sequence of *Prochlorococcus* SS120 was downloaded from http://www.sb-roscoff.fr/Phyto/ProSS120. The data has been provided freely by the US DOE Joint Genome Institute for use in this publication/correspondence only.

2.2. Gene detection

The *hli* genes were identified by similarity searches using the four *hli* genes of SC as query sequences (gene identifiers: ssl1633{*hliC*}, ssr1789{*hliD*}, ssl2542{*hliA*}, ssr2595{*hliB*} in Cyanobase) against complete cyanobacterial genome sequences translated in their six reading frames using the tBLASTn program [21]. Because of the very small size of these genes, the tBLASTn program was operated with a default *E*-value threshold of 10 and no filter for low-complexity regions. In a second step, a multiple alignment of the *hli* genes was performed using ClustalX software [22] in order to build a profile Hidden Markov Model (profile HMM) with the hmmbuild program of the HMMER 2 package (http://hmmer.wustl.edu/) [23]. This profile HMM was calibrated with the hmmcalibrate program and used with the hmmsearch program (HMMER 2 package) to identify *hli* genes that were not detected by tBLASTn. All three of these programs were operated with the default options. Using the HMMER package five new *hli* genes were detected in PL (*hli05, 06, 07, 08, 09*) and two in SN (*hli07* and *hli08*).

2.3. Sequence clustering

Clustering of *hli* genes was achieved using the GeneRAGE algorithm (http://www.ebi.ac.uk/research/cgg/services/rage/) [24], which groups sequences based on their similarity. We chose to use BLASTp to detect similarity between Hli polypeptide sequences. The choice of a threshold value is critical to obtain the optimal ratio between specificity and sensitivity. To determine the optimal threshold, an initial all-against-all comparison of protein sequences was made using BLASTp. After analysis of these results, an *E*-value cut-off of 10^{-13} was chosen as the threshold.

2.4. Conservation of regions neighboring hli genes

To examine whether regions surrounding the *hli* genes were conserved among the different genomes, orthologous relationships between open reading frames (ORFs) flanking the *hli* genes were investigated. Initially, ORFs flanking the *hli* genes were identified using the Artemis software package (http://www.sanger.ac.uk/Software/Artemis/). Each of these flanking ORFs was then compared against the complete cyanobacterial genome sequences using tBLASTn. Whenever the *hli* genes of two different genomes were found to have similar neighboring ORFs,

the neighboring ORFs were used for reciprocal comparisons using tBLASTn; ORFs that satisfied the criterion of reciprocal best hit for each other were considered to be orthologs.

3. Results and discussion

Multiple *hli* genes are present on the genomes of all seven cyanobacterial strains examined in this study (Table 1). Based on this analysis, the four marine cyanobacteria PM, PL, PS and SN have 22, 9, 13 and 8 *hli* genes, respectively, while the three freshwater cyanobacteria SC, AN and NT have 4, 8 and 9 *hli* genes, respectively. These 73 *hli* genes (Table 1) were analyzed to help understand the significance of the large number of *hli* genes on the genome (especially in PM), to determine possible relationships among the different *hli* genes and to evaluate whether the related clusters of *hli* genes represent group-specific (e.g. present only in the marine cyanobacteria or *Prochlorococcus* species etc.) and/or possible functional classes.

The genome of PM, the high-light ecotype of marine *Prochlorococcus*, encodes at least 22 *hli* genes. The number of *hli* genes on the PM genome is significantly greater than the number present on the genomes of the low-light ecotypes PL and PS (which have nine and 13 *hli* genes, respectively), the marine *Synechococcus* SN (which has eight *hli* genes) and the freshwater species SC, AN and NT (which have between four and nine *hli* genes). In SC, the *hli* gene family is required for survival in high light. A mutant of SC lacking one or two copies of the *hli* gene survives, but it is at a disadvantage relative to wild-type cells as evaluated by growth competition experiments performed in high light. If all four of the SC *hli* genes are disrupted, the cells die following exposure to high light [13]. These results suggest that HLIPs act in a cumulative manner to sustain the cells in high light, although there may also be requirements for particular gene products under specific environmental conditions. The significant increase in the number of *hli* genes on the genome of the *Prochlorococcus* high-light ecotype, in spite of the fact that this strain has the smallest genome amongst the seven genomes analyzed, may reflect a requirement for the additional gene products in coping with the persistent high-light conditions associated with the ocean surface [17]. Conversely, the smaller number of *hli* genes in PL, PS, SN and the freshwater strains, which are adapted to low/moderate-light growth conditions, is consistent with an important role for HLIPs in habitats in which the organisms are under persistent excitation pressure [25,26].

The arrangement of 22 *hli* genes in PM is shown in Fig. 1; the *hli* genes are scattered throughout the genome, with some gene clustering in particular regions. There are two instances in which two *hli* genes are contiguous (*hli*11 and *hli*12, and *hli*21 and *hli*22). In addition, two other regions of the genome contain four tandemly arranged *hli* genes (*hli*06-09 shown as A and *hli*16-19 shown as B in Fig. 1). The four tandemly arranged genes in region A are flanked by *hli*05 and *hli*10 (these genes are 4.5 kb and 3.7 kb distant from region A, respectively). Strikingly, the two clusters of tandemly arranged genes each cover 1.3 kbp and represent exact duplications; i.e. the sequence of *hli*06-09 is identical to that of *hli*16-19 at the nucleotide level. The duplicated 1.3-kbp region is comprised exclusively of four *hli* genes plus a small ORF of 59 codons (Fig. 1). This ORF has no similarity to other ORFs in the public databases, and thus may not represent a protein product. If this is the case, the duplication has resulted in the exclusive doubling of just the *hli* genes. A cursory examination of the PM genome indicates that this 1.3-kb region is the only exact duplication in excess of 1 kb in the genome. Interestingly, while two other *hli* genes (*hli*04 and *hli12*) are identical in their predicted amino acid sequences, they are not identical at the nucleotide level.

Since identification of the duplication of the *hli* gene clusters was based on genome sequence information, it was possible that the exact nucleotide match observed

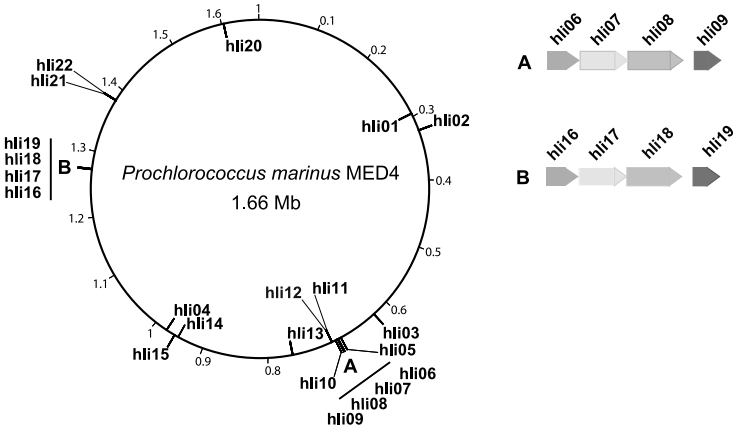

Fig. 1. Position of the 22 *hli* genes detected in the genome of *P. marinus* strain MED4. Gene groups A and B correspond to exactly duplicated regions containing the four *hli* genes (*hli*06-09 and *hli*16-19). The arrangement of the duplicated genes in groups A and B are shown on the right. Note that *hli*06, *hli*07 and *hli*08 (group A) and *hli*16, *hli*17 and *hli*18 (group B) are overlapping (see text for details).

Table 1
hli genes detected in seven fully sequenced cyanobacterial genomes

Genome	Gene	Start	Stop	Size (aa)	GeneRAGE cluster
PM	hli01	292 721	292 575	49	6
PM	hli02	320 885	321 115	77	8
PM	hli03	634 018	634 170	51	7
PM	hli04	713 298	713 194	35	14
PM	hli05	697 543	697 695	51	12
PM	hli06	702 237	702 341	35	14
PM	hli07	702 344	702 550	69	10
PM	hli08	702 553	702 813	87	12
PM	hli09	702 844	702 969	42	15
PM	hli10	706 726	706 851	42	16
PM	hli11	713 054	713 191	46	12
PM	hli12	983 410	983 306	35	14
PM	hli13	777 735	777 469	89	17
PM	hli14	962 255	962 106	50	10
PM	hli15	968 466	968 720	85	24
PM	hli16	1 274 111	1 274 215	35	14
PM	hli17	1 274 218	1 274 427	70	10
PM	hli18	1 274 427	1 274 687	87	12
PM	hli19	1 274 718	1 274 843	42	15
PM	hli20	1 600 841	1 600 662	60	5
PM	hli21	1 388 883	1 389 014	44	12
PM	hli22	1 389 022	1 389 276	85	15
PL	hli01	413 647	413 847	67	5
PL	hli02	72 191	72 409	73	6
PL	hli03	228 052	228 201	50	7
PL	hli04	120 778	121 041	88	8
PL	hli05	746 073	745 927	49	9
PL	hli06	572 474	572 674	67	10
PL	hli07	744 236	744 114	41	11
PL	hli08	572 319	572 459	47	12
PL	hli09	572 833	572 937	35	13
PS	hli01	1 388 013	1 388 129	39	15
PS	hli02	1 489 980	1 489 831	50	7
PS	hli10	1 387 372	1 387 512	47	19
PS	hli11	1 387 549	1 387 653	35	14
PS	hli12	1 397 334	1 397 438	35	14
PS	hli13	1 387 253	1 387 369	39	20
SN	hli01	489 613	489 413	67	21
SN	hli02	1 840 235	1 840 432	66	8
SN	hli03	2 114 476	2 114 625	50	7
SN	hli04	2 299 673	2 299 924	84	2
SN	hli05	1 389 330	1 389 683	118	22
SN	hli06	1 152 224	1 152 361	46	6
SN	hli07	403 771	403 553	73	23
SN	hli08	830 619	830 443	59	5
SC	hli01	701 350	701 138	71	2
SC	hli02	982 968	983 180	71	2
SC	hli03	398 188	398 361	58	3
SC	hli04	1 141 803	1 142 015	71	4
AN	hli01	1 006 982	1 006 782	67	1
AN	hli02	607 714	607 499	72	2
AN	hli03	2 836 843	2 836 676	56	3
AN	hli04	6 277 367	6 277 543	59	4
AN	hli05	3 686 003	3 686 203	67	1
AN	hli06	3 686 251	3 686 451	67	1
AN	hli07	4 499 702	4 499 526	59	4
AN	hli08	531 645	531 526	40	4
NT[a]	hli01	81 354	81 175	60	4
NT[a]	hli02	20 142	20 354	71	2
NT[a]	hli03	181 687	181 854	56	3
NT[a]	hli04	15 331	15 122	70	2
NT[a]	hli05	77 669	77 878	70	1
NT[a]	hli06	77 923	78 123	67	1

Table 1 (*Continued*).

Genome	Gene	Start	Stop	Size (aa)	GeneRAGE cluster
NT[a]	hli07	238 527	238 324	68	1
NT[a]	hli08	95 265	95 089	59	4
NT[a]	hli09	107 200	107 024	59	4
PS	hli03	1 426 441	1 426 352	30	18
PS	hli04	118 186	118 332	49	6
PS	hli05	88 528	88 283	82	8
PS	hli06	1 297 723	1 297 986	88	17
PS	hli07	1 097 580	1 097 786	69	10
PS	hli08	1 097 458	1 097 562	35	14

Start and stop positions for each gene are specified. Abbreviations used are: PM, *P. marinus* strain MED4; PL, *P. marinus* strain MIT9313; PS, *P. marinus* strain SS120; SN, *Synechococcus* sp. strain WH8102; SC, *Synechocystis* sp. strain PCC6803; AN, *Anabaena* sp. strain PCC7120; NT, *N. punctiforme* strain ATCC29133. Marine cyanobacteria are in bold. For each gene the size of the corresponding protein and the GeneRAGE cluster number is shown.

[a]The Nostoc sequence is incomplete, so start and stop data represent information from individual contigs (hli01: contig 483; hli02: contig 397; hli03: contig 507; hli04: contig 485; hli05: contig 480; hli06: contig 480; hli07: contig 509; hli08: contig 476; hli09: contig 486).

was generated by a computational artifact that yielded improper assembly results. To establish whether or not the putative duplication was an artifact, primers within the duplicate regions paired with specific primers flanking the duplications were constructed and used for PCR. The results of these experiments demonstrated the presence of the duplicate sequences at two distinct genomic locations (data not shown, Stephanie Stillwagen, DOE JGI, Walnut Creek, CA, USA, personal communication). Identical sequences within these *hli* gene clusters may reflect a very

Table 2
Cluster analysis of the 73 *hli* genes using the GeneRAGE program (see Section 2)

Cluster	Genome	Gene
1	AN	hli01
1	AN	hli05
1	AN	hli06
1	NT	hli05
1	NT	hli07
2	SN	hli04
2	SC	hli01
2	SC	hli02
2	AN	hli02
2	NT	hli02
2	NT	hli04
3	SC	hli03
3	AN	hli03
3	NT	hli03
4	SC	hli04
4	AN	hli04
4	AN	hli07
4	AN	hli08
4	NT	hli01
4	NT	hli08
4	NT	hli09
5	PM	hli20
5	PS	hli09
5	PL	hli01
5	SN	hli08
6	PM	hli01
6	PS	hli04
6	PL	hli02
6	SN	hli06
7	PM	hli03
7	PS	hli02
7	PL	hli03
7	SN	hli03
8	PM	hli02
8	PS	hli05
8	PL	hli04
8	SN	hli02
9	PL	hli05
10	PM	hli07/17
10	PM	hli14
10	PS	hli07
10	PL	hli06
11	PL	hli07
12	PM	hli05
12	PM	hli08/18
12	PM	hli11
12	PM	hli21
12	PL	hli08
13	PL	hli09
14	PM	hli04/12
14	PM	hli06/16
14	PS	hli08/11
14	PS	hli12
15	PM	hli09/19
15	PM	hli22
15	PS	hli01
16	PM	hli10
17	PM	hli13
17	PS	hli06
18	PS	hli03
19	PS	hli10
20	PS	hli13
21	SN	hli01
22	SN	hli05
23	SN	hli07
24	PM	hli15

Table 2 (*Continued*).

Genes were aligned using the BIOEDIT program. Marine cyanobacteria are in bold.

recent duplication of this locus, or the occurrence of a copy correction mechanism in the cell that maintains identity between the two sequences (although there is no experimental evidence to support such a mechanism). As more bacterial genomes are being sequenced and analyzed, it will be interesting to examine them for the presence of exact sequence duplications and gene duplications. The significance and mechanisms for creation and maintenance of these duplications is not yet understood although genome-wide studies suggest that duplicate genes are subject to specific selection and may not evolve at the same rate [27].

The arrangement of the clustered genes in the 1.3-kbp duplicate region is striking; the last nucleotide of the stop codon (TAA) for *hli06* and *hli07* is the first nucleotide of the start codon (ATG) of *hli07* and *hli08*, respectively (Fig. 1). This type of overlapping gene arrangement was demonstrated to be important for coordinating the expression of the overlapping *trpA* and *trpB* genes in the *Escherichia coli trp* operon [28]. Other examples of translational coupling have also recently been noted in *Prochlorococcus* MED4 (*phoB–PhoR*) and in the *pta–ack* bicistronic operon of *Corynebacterium glutamicum* [16,29]. Analysis of expression from the *hli* gene clusters would identify populations of polycistronic mRNAs that are transcribed from these genes, and may reveal how environmental conditions influence both the levels and distribution of distinct polycistronic transcripts.

The small size of the *hli* genes, the finding that some members of the gene family represent exact duplications, and the somewhat low degree of conservation among the different Hli proteins make classical sequence-based phylogenetic approaches difficult. In particular, bootstrap values of phylogenetic trees obtained by various methods (e.g. maximum likelihood) are very low (data not shown), raising serious concerns about the robustness of the observed relationships. To gain insights into the relationships among the 73 putative *hli* genes, we used the GeneRAGE program (Table 2). This program is a robust algorithm for quickly and accurately clustering large protein datasets into families and subfamilies [24]. Although no single program can give definitive answers regarding the relationship between genes and organisms, it does set the stage for a more complete analysis and provides information for hypothesis generation and evaluation. A number of observations resulting from these analyses are discussed below.

The 73 *hli* genes were separated into 24 clusters (these clusters contain up to seven genes, although 11 of the clusters contain a single gene representative). The clustering analysis clearly shows a strong divergence between marine and freshwater species (Fig. 2). This may indicate an early separation of the marine and freshwater cyanobacteria and the generation of divergent *hli* gene clusters within these environmentally distinct groups. However the strong divergence may also reflect very distinct evolutionary pressures that are associated with the markedly different environments in which these organisms are able to thrive.

A

FRESHWATER SPECIES

Cluster1

```
AN_hli01  ---MELY-PTDKTETA--YNGKDRNAFEFGFTPQSELWNGRLAMLGFLAYLLWDLNGYSVVRDVLHLVAYNAG 67
AN_hli05  .....TRST..LPKV.TE...V.....L..WN....I.......I..........A....L......IG.    67
AN_hli06  ....QTRPS..LPPV.PA...V.....L..W.....I.......I..........A....L......IR.    67
NT_hli06  .....TRSS..LPPV.KA...V.....L..W...A.I......AI...G......A....L.....IIS.    67
NT_hli07  MAT-.TNKVVKLSTE.KA...V....WI..WN..Q.........I..VS......A...LL......FR     68
NT_hli05  MATQ.TRSS..LPPV.PE...V.....L..W...A.I......AI..........A....L.......G.    70
```

Cluster2

```
SN_hli04  MAQTPSTDAPVIRGATVTTE-DGGRLNAFASEPRMQVVEAEQGWGFHERAEKLNGRMAMLGFIALLATEIALG-GEAFTHGLLG-LG 84
SC_hli01  -----------MTTRGFRLDQ.N-...N..I..EVY.DSSV.-A.WTKY...M...F..I..AS..IM.VVT.H.VI---.W.NS.  70
SC_hli02  -----------MTSRGFRLDQ.N-...N..I..PVY.DSSV.-A.WT.Y...M...F..I..VS...M.VIT.H.IV---.W.LS.  70
NT_hli04  -----------MTNKGF.IN.ER.Q..R..I..KIY.D.TP-RI..T.Y.......L..I...S.I.L.VFT.N.LI---.W.TSF  70
AN_hli02  -----------M.TNNAIVD..Q.LM.N..I..KVY.D.QGDRT..TPY..I....L..I...S.I.L.VFT.K.IF---...TN.Q 72
NT_hli02  -----------M.TS.AIID..Q.K..N..I..KVYID.QGDRT..TPY..M....L..I...S.I.L.VFT.H.IV---.V.AN.  71
```

Cluster3

```
SC_hli03  MSEELQPNQTPVQEDPKFGFNNYAEKLNGRAAMVGFLLILVIEYFTNQGVLAWLGLR 57
AN_hli03  ..QT-..TV..KL.E.....E...R.......I....MV....A......S....K 56
NT_hli03  .TQT-..TI..KL.E.....E...R.......I..A.M.....V......S....K 56
```

Cluster4

```
AN_hli08  --------------------------------MGFNHQSESWNGRLAMIGFLAAIAIEFFSGQGFLHFWNILIL 42
NT_hli01  ..........MTNASTTKVTTPVIEDRNAWRW..TP.A.I..........S..VLV.L..........G..    61
AN_hli04  ..........MTD..TTKISASVVEDRNSWRW..TP.A.I............TL..L..........G..D   60
NT_hli09  ..........MAD..VKKTTGSVPEDPNALRW..TP...N....F......S.V.L.V.....I....G..    59
NT_hli08  ..........MTGFK..NPAPIVSEDPNAVRF..TP...N...........S..L..A.....L....G..    59
AN_hli07  ..........MSGFK...........PNAVRF..TSE.......F......SIVL..A..........G..    50
SC_hli04  MGAILCYIYLHRQPSQLVITFLTMNNENSK.F..TAFA.N..........SS.LIL.LV....V...FG..    70
```

Fig. 2. Alignment of *hli* genes in specific clusters of freshwater species (A) and marine species (B) using the GeneRAGE program. Dots mark residues that are identical to the top sequence in each cluster and dashes represent gaps. Clusters that have only a single representative are not shown.

3.1. Fresh water species

Clusters 1–4 contain all 22 *hli* genes of the freshwater species, AN, NT and SC (Fig. 2A and Table 2). Of these, clusters 2, 3 and 4 contain at least one representative from each species; while cluster 1 contains three genes each from NT and AN (AN_*hli 01*, AN_*hli05* AN_*hli06* and NT_*hli05*, NT_*hli06*, NT_*hli07*). Cluster 2 contains two representatives from NT and SC each (NT_ *hli02* and NT_*hli04*, SC_*hli01* and SC_*hli02*) and one from AN (*hli02*) and SN (*hli04*), cluster 3 contains a single representative from each of the freshwater species (SC_*hli03*, AN_*hli03*, NT_*hli03*) and cluster 4 contains seven genes, with three each from NT and AN and one from SC (*hli*04). Based on nearest neighbor analysis, AN_*hli02* and NT_*hli*02 (both in cluster 2), AN_*hli04* and NT_*hli01* (both in cluster 4), and AN_*hli03* and NT_*hli03* (both in cluster 3) all share neighboring genes, as do NT_*hli04* and SC_*hli02* (both in cluster 2) (Fig. 3). These results suggest the following features of the *hli* gene families:

1. There may be three basic *hli* gene 'forms' in the freshwater species represented by the sequences in clusters 2–4. Furthermore, strong sequence similarity among pairs of polypeptides representative of the different freshwater species, encoded by genes within these groups, combined with nearest neighbor analyses with respect to these genes, maybe suggestive of orthologous relationships among specific members of these gene clusters. Whether the genes in clusters 2, 3, and 4 have distinct functions is still unclear. Recent evidence suggests that a severe phenotype is associated only with a mutant in which all four *hli* genes are inactivated;

B

MARINE SPECIES

Cluster 5

```
PL_hli01  ---------MASESPLDSNTSAEPVS--SEELNAWRRGFTPQAEIWNGRMAMAGLIIGISVLLLLRLVMPADCRAWLN 67
SN_hli08  ...........-.Q-.EK-.GGVAEPVG.D.....K.............L..I..SA.LA.V..V.VF-AGN       59
PS_hli09  .MNSQSTNKEKK..---------TQ.VEKS.....K..............SI...L.LI..I.INKFYG         60
PM_hli20  ..........KK..KINLK--ETKKVVDKQ...L.K..............TI.IG.ILI.IA.ISKF-SSI       60
```

Cluster 6

```
PM_hli01  -------------------------MNEDN-QPRFGFVNFAETWNGRMAMMGILIGLGTELITGQSILRQIGIG 48
PL_hli02  MRIYCHQDGQAISMLCYIDEILESP.A-.KP.............L....VI...S...L......S.M.L. 73
PS_hli04  ........................SPEDIE..Y....Y..I....L..L..V...S...L...G..G...F. 49
SN_hli06  ........................S-...-A.............L....FV.......L...G..S...L. 46
```

Cluster 7

```
PM_hli03  MIKPDIVPKRKLPRYGFHFYNEKLNGRMAMIGFIALILTELFLKHGLLLW 50
PS_hli02  ..D.K.I.E....S....NHT.N....W........VIV.FK.G..I.IR 50
PL_hli03  .LE.T.I.Q.RK......SH..........L.....MVV.AT.G....I. 50
SN_hli03  .LE.TDI.Q.R...F...GHT......A..L.....LAV.IK.G....I. 50
```

Cluster 8

```
PL_hli04  MTSPKQNLPGDQDLPSEQAVFEGSEQGSESSEVQPPINSATTGDPPTFGWSAYAERVNGRFAMIGLAAVLLIEVVSRDTFVHWAGLVS 88
SN_hli02  ....SEPPATASVPET----------------------....S.V.A....G...........V.FT.I.V..AI.G...L.....LP 66
PS_hli05  ...-NNPELSKVESSKSESQENND.TNDVQMTP-----....P.I.S....G.............FI.I....TI.KSG.L......P 82
PM_hli02  ...--NQEQNN.EAMELEKTNSEEIKIE.Q--------.IEIE.RYE....N.S.IT......L.FL.II...LI.QKS.LN...IF  77
```

Cluster 10

```
PM_hli07/17  MSNSSYT--TTESGGRQNMFPSETRPYIDESVSYDGYPQNAEKVNGRWAMIGFVALLGAYVTTGQIIPGIF 69
PM_hli14     .A..---QV.........................S..K.................................. 68
PL_hli06     .TS.--.NVI..D......YA..P.MQ..PE--.TAFSKE..LA...G.....LSAVV..LF....L.... 67
PS_hli07     .TS.AQAQI.....N......V.AQ.QLV.N--.S..IED...A.........I......L.S........ 69
```

Cluster 12

```
PM_hli05     ----------------------------------MNSKKVKVLETKTVEKEKVVAEKLNGRFAMIGFIAAIGAYLTTGQIIPGFV 51
PL_hli08     .................................----MK.TPK.NR..NQ.LT..RV..MA..M..W..V.............V. 47
PM_hli11     .................................-----M.NN.P.L.....I...........M..V.LV.............I 46
```

6

Cluster 14

```
PM_hli04/12  MTPEAERFNGWAAMLGFVAAVGAYVTTGQIIPGWF 35
PM_hli06/16  ...D......L.....................F. 35
PS_hli08/11  ......K.............F...A........I. 35
PS_hli12     ...Q..K.......I...C...S.A........I. 35
```

Cluster 15

```
PM_hli09/19  ----------------------------------------MENSKPNYWQNAERTNGRMAMMGFFALVVNYGLFGWIIPGIF 42
PS_hli01     ........................................N.N---..TI.......L..I.L...II...F........Y 39
PM_hli22     MSPLTGFIIVVIAITLQFTLYTIKRLQEPLDPNLFDSQKSPK.N.R.KSF.K...I....L..V.LL.......F.......FI 84
```

Cluster 17

```
PM_hli13  MKEEKPPL-KNSDNSPTENLKEETNNTSSDNEYSKWVDNQGDEVKDVFGFNSSAELVNGRAAMIGFLMLLLTELVFKGRPVTSSIFGIN 88
PS_hli06  .RM.DNLNQ..EEDRFD...IGSRKEITGTSD-A.....NDN..TQ.....EN.....S.......I..I....I.N.K...L...... 88
```

Fig. 2 (*Continued*).

mutants lacking one or two *hli* genes do not exhibit this phenotype which may be indicative of a redundancy or overlapping gene functions [13].

2. AN and NT have multiple copies of closely related *hli* genes within clusters 1 and 4 (and cluster 2 for NT). However, in SC there appears to have been only one recent duplication (*hli01* and *hli02* for which the encoded amino acid sequences are 87% identical; 94% similar). Clusters 4 and 1 have three representatives each from AN and NT, but one or none from SC, respectively. The significance of this is unclear since all three species grow in relatively low-light environments. However, both NT and AN are multi-cellular and developmentally complex (both species can differentiate nitrogen-fixing heterocysts which contain a highly modified photosynthetic apparatus that does not evolve oxygen). It would be particularly interesting to follow expression patterns of the *hli* genes under different growth conditions (e.g. low-nitrogen, high-light) as well as to examine regions upstream of the *hli* coding regions to identify conserved sequence elements.
3. Only one marine cyanobacterial HLIP sequence, SN_*hli04*, clusters with the sequences from the freshwater organisms (within cluster 2). The amino acid similarity between SN_*hli04* and the most closely related gene in cluster 2, NT_*hli02*, is 43%. The significance of this clustering is not apparent at this time.
4. A comparison of sequences in NT, AN and SC demonstrates that there is little conservation of genes that flank the *hli* genes between the filamentous and unicellular cyanobacteria (Fig. 3). There is evidence that there have been rearrangements of genomes in some freshwater cyanobacteria, including SC and there is also evidence for recent transposition events in SC [30,31]. This makes it less surprising that neighborhood conservancy is not maintained between NT/AN and SC.

3.2. Marine species

Clusters 5–8 all contain a single representative from each of the four marine species analyzed (Fig. 2B). Furthermore, these genes also all share flanking gene neighbors (Fig. 3). This may indicate that these four gene clusters are essential for all marine species (somewhat similar to the case in freshwater species where there are three clusters that have at least one representative from each of the species). Although the phylogenetic relationships among the gene groupings are difficult to evaluate, based on sequence similarity, the marine gene clusters 6–8 are possibly most closely related to the freshwater gene clusters 2 and 3. It is unclear if the apparent similarities between these groups reflect the specific functions of the proteins, but it would be interesting to explore this possibility by examining the influence of diverse environmental conditions on the expression of genes within these clusters.

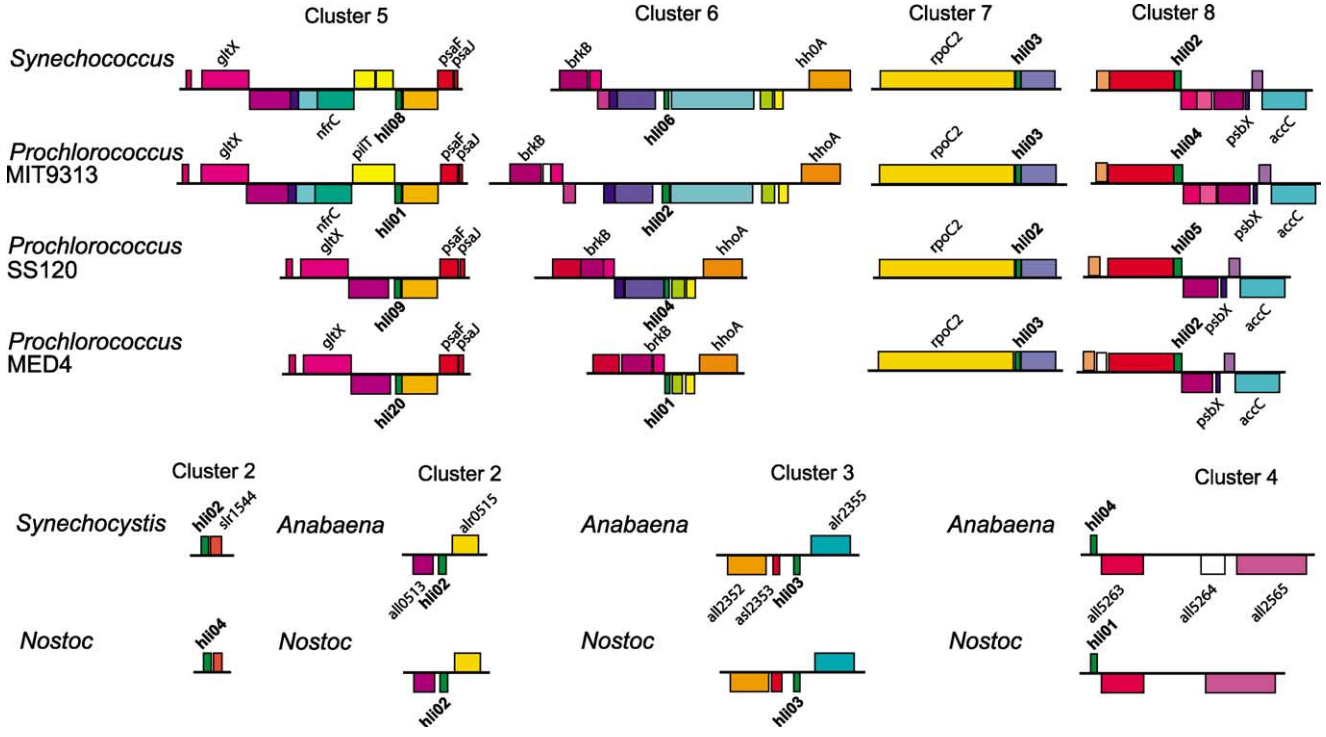

Fig. 3. Conserved genes in the neighborhood of *hli* genes arranged according to the GeneRAGE clusters. ORFs that are painted with the same color are considered orthologs. For clusters 5–8 only genes for which the assignment of a gene name is unambiguous have been labeled (e.g. *rpoC2*). For clusters 2–4 the Cyanobase Gene Identifier for flanking genes of *Synechocystis* or *Anabaena* have been labeled for identification of flanking genes.

Five clusters (10, 12, 14, 15 and 17) contain multiple genes from the marine species but not all of the species are represented in these clusters (Fig. 2B). Furthermore, 11 clusters (9, 11, 13, 16 and 18–24) each contain only one representative gene. This indicates that there are several marine *hli* genes (three from PL {*hli05, hli07, hli09*}; two from PM {*hli10* and *hli15*}; three from PS {*hli3, hli10* and *hli13*}; three from SN {*hli01, hli05* and *hli07*}) that have diverged to the point of not being grouped together using the GeneRAGE program. Clusters with multiple family members from one species may represent an evolutionary trend toward duplication of specific genes; however until more experimental evidence is available, it is not possible to draw any direct conclusions based simply on sequence similarities.

In Fig. 4, 23 out of the 44 *hli* genes within the *Prochlorococcus* species (PL, PM and PS) are aligned using ClustalX. It is quite striking that within this group the C-terminus of the HLIPs maintain a strongly conserved motif TGQIIPGF/IF. This motif is not conserved in clusters 5, 6, 7 and 8 (which contain one representative from each of the marine cyanobacterial species) or in clusters 17 and 18. It is also notably missing in all of the freshwater strains. The fact that this motif is only present in a subset of the *Prochlorococcus* HLIPs may be indicative of specialized function.

3.3. Concluding remarks

We have attempted an analysis of the large *hli* gene family that is ubiquitous in all cyanobacterial species examined so far (as well as in other groups). This study was motivated by the initial observation that there was an apparent over-representation of *hli* genes in PM, the high-light adapted marine strain. The presence of a very large *hli* gene family in PM is consistent with recent results showing that a mutant of SC lacking all four copies of the *hli* gene was unable to survive in high light [13], raising the obvious question of whether the number of *hli* genes in an organism could be correlated with adaptation/acclimation to the light environment.

To attempt to answer this question, we took a bioinformatics approach to analyze the 73 *hli* genes identified in seven recently sequenced cyanobacterial strains. There are a number of problems associated with the use of small genes to construct a phylogeny. The construction of cyanobacterial phylogenies, is particularly problematic since there are issues related to the ancient history of this group; over the course of evolution cyanobacteria may have experienced both lateral gene transfer and the formation of gene mosaics [32,33]. Furthermore, it is even difficult to determine which bacteria are most closely related to cyanobacteria; some sister groups are considered to be the *Deinococcales* and spirochetes and more recent analyses suggests a relationship with low GC Gram-positive bacteria such as *Halobacterium* and *Aquifex aeolicus* [34–36]. To avoid these potential pitfalls we used a clustering analysis method (GeneRAGE) that does not make any assumptions about phylogenetic relatedness between genes.

One of the most obvious results from the analyses presented above suggests that there is a significant distinction between *hli* genes in the marine and freshwater strains. Since there has not yet been an extensive analysis of genes across various cyanobacterial species, it is useful to compare this data in the context of some recent molecular phylogenetic studies of various cyanobacterial species using 16S rRNA sequence data [37,38]. Honda et al. used 16S rRNA sequences from a variety of 44 different freshwater and marine strains to determine evolutionary lineages within the cyanobacteria. Based on maximum likelihood and neighborhood joining trees generated with

```
PM_hli04/12  ----------------------------------------------------MTPEAERFNGWAAMLGFVAAVGAYVTTGQIIPGWF
PM_hli05     ................................MNSKKVKVLETKTVEKEKVV..KL..RF..I..I..I...L........FV
PM_hli06/16  ..........................................................D......L.....................F.
PM_hli07/17  ................MSNSSYTTTESGGRQNMFPSETRPYIDESVSYDGYPQN..KV..RW..I....LL............I.
PM_hli08/18  MSPLAVFLILIVSLTALLVASLTKQFQEENLIYSNKNQMTNSNTKTKTIEKEKVV..TL..RF..I.LI......L........FV
PM_hli09/19  ......................................MENSKPNYWQN...T..RM..M..F.L.VN.GLF.W....I.
PM_hli10     ......................................MEFVKKF..EK..K...K.....MF.LI...YF....V..I.
PM_hli11     .....................................MKNNEPKLVEKEKIV..KL..RF..M....L....L........FI
PM_hli14     ................MANSQVTTESGGRQNMFPSETRPYIDESVSYDSYPKN..KV..RW..I....LL............I.
PM_hli15     ..MIEKKGDNIRSENFYPDSNYYLDQDNTPEETTLPEDQIFNTKKFEWPNSYWFI...T..RL..I..M.VIIN.TLF.W.AYPIL
PM_hli21     .....................................MAKIKSVEKEKIV..KL..RF..I..I......L........FV
PM_hli22     ..MSPLTGFIIVVIAITLQFTLYTIKRLQEPLDPNLFDSQKSPKMNNRKKSFWKN..IT..RL..V.LL.L.VN.GFF.W....FI
PS_hli01     ..........................................MNNNYWTI...T..RL..I.LF.LIIN.GFF.W....IY
PS_hli07     ................MTSSAQAQITTESGNRQNMFPVEAQPQLVENYSGYIED..KA..RW..I..I.LL...L.S......I.
PS_hli08/11  ..........................................................K.............F...A........I.
PS_hli10     ....................................MKTSTSSTKVETSKVL..KI..R..LI.VI.LL...SA.......YL
PS_hli12     ..........................................................Q..K.......I...C...S.A........I.
PS_hli13     ..................................................MDQAN.SVLDIAF.RP..I..ILLL.T.LV.......T.
PL_hli05     ..................................MPISDFLKETINDC..NK..SL..RI..V.ML.LMVT.LA..D....V.
PL_hli06     ................MTSSTNVITEDGGRQNMYASEPRMQIDPEYTAFSK...LA..RG..I..LS..V..LF....L..I.
PL_hli07     ......................................MTIADF.SNK..TW..RV.....LV.I.T..V..E....I.
PL_hli08     ....................................MKKTPKTNRVENQKLT...V..M...M..W......L........VV
PL_hli09     ..................................................NEN..LQ..RW..I..IG.LAS.AA.......L.
```

Fig. 4. Comparison of all the *hli* sequences of *Prochlorococcus* strains that share the conserved C-terminus motif (TGQIIPGI/FF). Dots mark residues that are identical to the top sequence and dashes represent gaps.

these data, they suggested that there were at least seven different evolutionary lineages within the cyanobacteria. These trees also indicate that unicellular and filamentous species may be closely arranged on the same branch of the tree, but that freshwater and marine strains are often well separated. This is consistent with our results where, for example, unicellular SC *hli* genes are much more closely related (based on clustering) to the *hli* genes from freshwater, filamentous species than to the *hli* genes from unicellular marine strains. The generation of phylogenies based on 16S rRNA sequences was often taken as the benchmark for phylogenetic analyses. However with the explosion of information associated with the generation of complete genome sequences from a variety of prokaryotes, a number of individual genes within these genomes can be used to help establish phylogenetic relationships among the cyanobacteria. Differences in phylogenetic relationships that are obtained when different genes are used to evaluate such relationships suggest that the generation of a single, consistent evolutionary tree may not be easy, especially since multiple pressures imposed by specific environments may differentially influence the apparent rates at which specific genes evolve. In the case of the *hli* genes, the evolutionary pressure for this gene family to evolve and adapt to high-light conditions may create a phylogeny that is not necessarily consistent with a 16S rRNA tree. This raises the interesting possibility of analyzing and classifying a variety of molecular markers potentially indicative of specific environmental pressures (high-light or nutrient-stress). Analyses which focus on the proliferation or drastic reduction of genes in a particular adapted ecotype or species (for instance, the proliferation of *hli* genes in a high-light-adapted strain versus in low-light-adapted ecotypes or the steep reduction in two-component regulatory systems in marine species relative to freshwater species may allow us to gain an insight into environmental selection pressures, and the extent to which such pressures has shaped the individual cyanobacterial species. Since we now have a large data base of information from a range of different cyanobacteria that have adapted to very different ecological niches, this orientation represents an attractive approach for future work.

Acknowledgements

We thank F. Partensky the coordinator of the *Prochlorococcus* SS120 genome sequencing project and the Genoscope, France (M. Salanoubat) for providing us with access to the genome sequence prior to publication. This study is partly supported by the MARGENES program (EU contract xxx). A.D. is supported by a doctoral fellowship from Région Bretagne. The cooperation between Stanford and Roscoff is supported by an NSF–CNRS bilateral grant.

References

[1] Grossman, A.R., Bhaya, D. and He, Q. (2001) Tracking the light environment by cyanobacteria and the dynamic nature of light harvesting. J. Biol. Chem. 276, 11449–11452.
[2] Green, B.R., Durnford, D.G. and Jones, R.L. (1996) The chlorophyll-carotenoid proteins of oxygenic photosynthesis. Annu. Rev. Plant Physiol. Plant Mol. Biol. 47, 685–714.
[3] Sandona, D., Croce, R., Pagano, A., Crimi, M. and Bassi, R. (1998) Higher plants light harvesting proteins. Structure and function as revealed by mutation analysis of either protein or chromophore moieties. Biochim. Biophys. Acta 1365, 207–214.
[4] Jansson, S. (1994) The light-harvesting chlorophyll *a*/*b*-binding proteins. Biochim. Biophys. Acta 1184, 1–19.
[5] Kim, S., Sandusky, P., Bowlby, N.R., Aebersold, R., Green, B.R., Vlahakis, S., Yocum, C.F. and Pichersky, E. (1992) Characterization of a spinach psbS cDNA encoding the 22 kDa protein of photosystem II. Fed. Eur. Biochem. Soc. Lett. 314, 67–71.
[6] Jansson, S., Andersson, J., Jung-Kim, S. and Jackowski, G. (2000) An *Arabidopsis thaliana* protein homologous to cyanobacterial high-light-inducible proteins. Plant Mol. Biol. 42, 345–351.
[7] Heddad, M. and Adamska, I. (2000) Light stress regulated two helix proteins in Arabidopsis thaliana related to the chlorphyll *a*/*b*-binding gene family. Proc. Natl. Acad. Sci. USA 97, 3741–3746.
[8] Dolganov, N.A.M., Bhaya, D. and Grossman, A.R. (1995) Cyanobacterial protein with similarity to the chlorophyll *a*/*b*-binding proteins of higher plants: evolution and regulation. Proc. Natl. Acad. Sci. USA 92, 636–640.
[9] Funk, C. and Vermaas, W. (1999) A cyanobacterial gene family coding for single-helix proteins resembling part of the light-harvesting proteins from higher plants. Biochemistry 38, 9397–9404.
[10] Dolganov, N. and Grossman, A.R. (1999) A polypeptide with similarity to phycocyanin a subunit phycocyanobilin lyase involved in degradation of phycobilisomes. J. Bacteriol. 181, 610–617.
[11] Kaneko, T., Sato, S., Kotani, H., Tanaka, A., Asamizu, E., Nakamura, T., Miyajima, N., Hirosawa, M., Sugiura, M., Sasamoto, S., Kimura, T., Hosouchi, T., Matsuno, A., Muraki, A., Nakazaki, N., Naruo, K., Okumura, S., Shimpo, S., Takeuchi, C., Wada, T., Watanabe, A., Yamada, M., Yasuda, M. and Tabata, S. (1996) Sequence analysis of the genome of the unicellular cyanobacterium *Synechocystis* sp. strain PCC 6803. II. Sequence determination of the entire genome and assignment of potential protein-coding regions. DNA Res. 3 (Suppl.), 185–209.
[12] Reith, M.E. and Munholland, J. (1995) Complete nucleotide sequence of the *Porphyra purpurea* chloroplast genome. Plant Mol. Biol. 13, 333–335.
[13] He, Q., Dolganov, N., Bjorkman, O. and Grossman, A.R. (2001) The high light-inducible polypeptides in *Synechocystis* PCC6803. Expression and function in high light. J. Biol. Chem. 276, 306–314.
[14] Montane, M.H. and Kloppstech, K. (2000) The family of light-harvesting-related proteins (LHCs, ELIPs, HLIPs): was the harvesting of light their primary function? Gene 258, 1–8.
[15] Xu, H., Vavilin, D., Funk, C. and Vermaas, W. (2002) Small cab-like proteins regulating tetrapyrrole biosynthesis in the cyanobacterium *Synechocystis* sp. PCC6803. Plant Mol. Biol. 49, 149–160.
[16] Scanlan, D.J. and West, N.J. (2002) Molecular ecology of the marine cyanobacterial genera *Prochlorococcus* and *Synechococcus*. FEMS Microbiol. Ecol. 40, 1–12.
[17] Partensky, F., Hess, W.R. and Vaulot, D. (1999) Prochlorococcus, a marine photosynthetic prokaryote of global significance. Microbiol. Mol. Biol. Rev. 63, 106–127.
[18] Urbach, E., Robertson, D.L. and Chisholm, S.W. (1992) Multiple evolutionary origins of prochlorophytes within the cyanobacterial radiation. Nature 355, 267–270.
[19] Ting, C.S., Rocap, G., King, J. and Chisholm, S.W. (2002) Cyano-

bacterial photosynthesis in the oceans: the origins and significance of divergent light-harvesting strategies. Trends Microbiol. 10, 134–142.
[20] Moore, L.R., Goericke, R. and Chisholm, S.W. (1995) Comparative physiology of *Synechococcus* and *Prochlorococcus*: Influence of light and temperature on growth, pigments, fluorescence and absorptive properties. Mar. Ecol. Prog. Ser 116, 259–275.
[21] Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. (1997) 'Gapped BLAST and PSI-BLAST:' a new generation of protein database search programs. Nucleic Acids Res. 25, 3389–3402.
[22] Thompson, J.D., Gibson, T.J., Plewniak, F., Jeanmougin, F. and Higgins, D.G. (1997) The ClustalX windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res. 24, 4876–4882.
[23] Eddy, S.R. (1998) Profile hidden Markov models. Bioinformatics 14, 755–763.
[24] Enright, A.J. and Ouzounis, C.A. (2000) GeneRAGE: a robust algorithm for sequence clustering and domain detection. Bioinformatics 16, 451–457.
[25] Wyman, M., Fay, P. (1987) Acclimation to the natural light climate. In: The Cyanobacteria (Fay, P., Baalen, C.V., Eds.), pp. 347–376. Elsevier, Amsterdam.
[26] Fogg, G.E. (1987) Marine planktonic cynaobacteria. In: The Cyanobacteria (Fay, P., Baalen, C.V., Eds.), pp. 393–414. Elsevier, Amsterdam.
[27] Wagner, A. (2002) Selection and gene duplication: a view from the genome. Genome Biol. 3, 1012–1013.
[28] Das, A. and Yanofsky, C. (1989) Restoration of a translation stop-start overlap reinstates translational coupling in a mutant trpB′-trpA gene pair of the *Escherichia coli* tryptophan operon. Nucleic Acids Res. 17, 9333–9340.
[29] Reinscheid, D.J., Schnicke, S., Rittmann, D., Zahnow, U., Sahm, H. and Eikmanns, B.J. (1999) Cloning, sequence analysis, expression and inactivation of the *Corynebacterium glutamicum* pta-ack operon encoding phosphotransacetylase and acetate kinase. Microbiology 145, 503–513.
[30] Kaneko, T. and Tabata, S. (1997) Complete genome structure of the unicellular cyanobacterium *Synechocystis* sp. PCC6803. Plant Cell Physiol. 38, 1171–1176.
[31] Okamoto, S., Ikeuchi, M. and Ohmori, M. (1999) Experimental analysis of recently transposed insertion sequences in the cyanobacterium *Synechocystis* sp. PCC 6803. DNA Res. 6, 265–273.
[32] Lawrence, J.G. and Ochman, H. (2002) Reconciling the many faces of lateral gene transfer. Trends Microbiol. 10, 1–4.
[33] Juhala, R.J., Ford, M.E., Duda, R.L., Youlton, A., Hatfull, G.F. and Hendrix, R.W. (2000) Genomic sequences of bacteriophages HK97 and HK022: pervasive genetic mosaicism in the lamboid bacteriophages. J. Mol. Evol. 299, 27–51.
[34] Zhaxybayeva, O. and Gogarten, J.P. (2002) Bootstrap, Bayesian probability and maximum likelihood mapping: exploring new tools for comparative genome analyses. BMC Genomics 3, 4.
[35] Gupta, R.S. (1998) Protein phylogenies and signature sequences: A reappraisal of evolutionary relationships among archaebacteria, eubacteria, and eukaryotes. Microbiol. Mol. Biol. Rev. 62, 1435–1491.
[36] Gupta, R.S. and Golding, G.B. (1996) The origin of the eukaryotic cell. Trends Biochem. Sci. 21, 166–171.
[37] Honda, D., Yokota, A. and Sugiyama, J. (1999) Detection of seven major evolutionary lineages in cyanobacteria based on 16S rRNA gene analysis with new sequences of five marine *Synechococcus* strains. J. Mol. Evol. 48, 723–739.
[38] Ishida, T., Watanabe, M.M., Sugiyama, J. and Yokota, A. (2001) Evidence for polyphyletic origin of the members of the orders of *Oscillatoriales* and *Pleurocapsales* as determined by 16S rDNA analysis. FEMS Micobiol. Lett. 201, 79–82.

Annexe III

Annexe III (article 4):

Le génome d'une souche marine et mobile de *Synechococcus*

Résumé

Les cyanobactéries unicellulaires marines représentent entre 20 et 40 % de la biomasse chlorophyllienne et de la fixation du carbone dans les océans. Dans cet article, nous présentons les résultats du séquençage et de l'analyse du génome de *Synechococcus* sp. WH8102 (2,4 Mb).

Ces résultats ont permis de révéler certaines des voies empruntés par ces organismes pour s'adapter à leur environnement très oligotrophe. *Synechococcus* sp. WH8102 est capable d'utiliser des sources d'azote et de phosphore organiques et possède plus de transporteurs dépendants du gradient de sodium qu'une cyanobactérie d'eau douce. De plus, il semble avoir adopté une stratégie lui permettant d'économiser le fer, particulièrement limitant, en utilisant des enzymes fonctionnant plutôt avec le nickel ou le cobalt. Il a également réduit son système de régulation (en accord avec le fait que le milieu océanique constitue un environnement beaucoup plus stable et tamponné que les milieux d'eaux douces) et a développé un type de mobilité unique.

L'évolution du génome de *Synechococcus* sp. WH8102 semble avoir été fortement influencée par les transferts horizontaux de gènes, notamment par l'intermédiaire des phages. La matériel génétique apporté par les transferts horizontaux inclus plusieurs gènes permettant de modifier la paroi cellulaire ou rendant possible la mobilité de cette cyanobactérie. L'analyse du génome de *Synechococcus* sp. WH8102 indique que ce dernier est plus généraliste que les cyanobactéries du genre *Prochlorococcus.*

processes because of their inherent noisiness. However, simulations of the neutral theory are no longer necessary, and all problems with simulations are moot, because an analytical solution is now available.

The lognormal distribution is biologically less informative and mathematically less acceptable as a dynamical null hypothesis for the distribution of RSA than the neutral theory. The parameters of the neutral theory or RSA are directly interpretable in terms of birth and death rates, immigration rates, size of the metacommunity, and speciation rates. A dynamical model of a community cannot yield a lognormal distribution with finite variance because in its time evolution, the variance increases through time without bound. However, as shown in ref. 18, the lognormal distribution can arise in static models, such as those based on niche hierarchy.

The steady-state deficit in the number of rare species compared to that expected under the log series can also occur because rare species grow differentially faster than common species and therefore move up and out of the rarest abundance categories owing to their rare-species advantage[19]. Indeed, it is likely that several different models (such as an empirical lognormal distribution, niche hierarchy models[18] or the theory presented here) might provide comparable fits to the RSA data (we have found that the lognormal does slightly better than the neutral theory for the Pasoh data set[20], obtained in a tropical tree community in Malaysia). Such fitting exercises in and of themselves, however, do not constitute an adequate test of the underlying theory. Neutral theory predicts that the degree of skewing of the RSA distribution ought to increase as the rate of immigration into the local community decreases. Dynamic data on rates of birth, death, dispersal and immigration are needed to evaluate the assumptions of neutral theory and determine the role played by niche differentiation in the assembly of ecological communities.

Our analysis should also apply to the field of population genetics in which the mutation-extinction equilibrium of neutral allele frequencies at a given locus has been studied for several decades[21–26]. □

Received 7 May; accepted 9 June 2003; doi:10.1038/nature01883.

1. MacArthur, R. H. & Wilson, E. O. *The Theory of Island Biogeography* (Princeton Univ. Press, Princeton, NJ, 1967).
2. Hubbell, S. P. *The Unified Neutral Theory of Biodiversity and Biogeography* (Princeton Univ. Press, Princeton, NJ, 2001).
3. McGill, B. J. A test of the unified neutral theory of biodiversity. *Nature* **422,** 881–885 (2003).
4. Condit, R. *et al.* Beta-diversity in tropical forest trees. *Science* **295,** 666–669 (2002).
5. Preston, F. W. The commonness and rarity of species. *Ecology* **29,** 254–283 (1948).
6. May, R. M. *Ecology and Evolution of Communities* 81–120 (Harvard Univ. Press, Cambridge, MA, 1975).
7. Bell, G. Neutral macroecology. *Science* **293,** 2413–2418 (2001).
8. Diamond, J. & Case, T. J. (eds) *Community Ecology* (Harper and Row, New York, NY, 1986).
9. Tilman, D. *Plant Strategies and the Dynamics and Structure of Plant Communities* (Princeton Univ. Press, Princeton, NJ, 1988).
10. Weiher, E. & Keddy, P. *Ecological Assembly Rules: Perspectives, Advances, and Retreats* (Cambridge Univ. Press, Cambridge, UK, 1999).
11. Boswell, M. T., Ord, J. K. & Patil, G. P. *Statistical Distributions in Ecological Work* 3–157 (International Co-operative Publishing, Fairland, MD, 1979).
12. Caraco, T. *Statistical Distributions in Ecological Work* 371–387 (International Co-operative Publishing, Fairland, MD, 1979).
13. Feller, W. *An Introduction to Probability Theory and Its Applications* Vol. 1 (John Wiley & Sons, Hoboken, NJ, 1968).
14. Van Kampen, N. G. *Stochastic Processes in Physics and Chemistry* (Amsterdam, North-Holland, 2001).
15. Fisher, R. A., Corbet, A. S. & Williams, C. B. The relation between the number of species and the number of individuals in a random sample of an animal population. *J. Anim. Ecol.* **12,** 42–58 (1943).
16. Morse, P. M. & Feshbach, H. *Methods of Theoretical Physics* Part 1 (McGraw-Hill, New York, NY, 1953).
17. Press, W. H., Flannery, B. P., Teukolsky, S. A. & Vetterling, W. T. *Numerical Recipes in C: The Art of Scientific Computing* (Cambridge Univ. Press, Cambridge, UK, 1993).
18. Sugihara, G., Bersier, L., Southwood, T. R. E., Pimm, S. L. & May, R. M. Predicted correspondence between species abundances and dendrograms of niche similarities. *Proc. Natl Acad. Sci. USA* **100,** 5246–5251 (2003).
19. Chave, J., Muller-Landau, H. C. & Levin, S. A. Comparing classical community models: Theoretical consequences for patterns of diversity. *Am. Nat.* **159,** 1–23 (2002).
20. Manokaran, N. *et al. Stand Tables and Species Distributions in the Fifty Hectare Plot at Pasoh Forest Reserve* (Forest Research Institute Malaysia, Kuala Lumpur, 1992).
21. Kimura, M. Evolutionary rate at the molecular level. *Nature* **217,** 624–626 (1968).
22. Ewens, W. J. The sampling theory of selectively neutral alleles. *Theor. Pop. Biol.* **3,** 87–112 (1972).
23. Karlin, J. & McGregor, J. Addendum to a paper of W. Ewens. *Theor. Pop. Biol.* **3,** 113–116 (1972).
24. Watterson, G. A. Models for the logarithmic species abundance distributions. *Theor. Pop. Biol.* **6,** 217–250 (1975).
25. Kimura, M. & Ohta, T. *Theoretical Aspects of Population Genetics* (Princeton Univ. Press, Princeton, NJ, 1971).
26. Kimura, M. *The Neutral Theory of Molecular Evolution* (Cambridge Univ. Press, Cambridge, UK, 1983).
27. McKane, A., Alonso, D. & Solé, R. V. Mean-field stochastic theory for species-rich assembled communities. *Phys. Rev. E* **62,** 8466–8484 (2000).
28. Rao, C. R. *Statistical Ecology* Vol. 1 *Spatial Patterns and Statistical Distributions* 131–142 (The Penn. State Univ. Press, University Park, PA, 1971).

Acknowledgements We are grateful to O. Kargaltsev for a careful reading of the manuscript. This work was supported by NASA, by grants from the NSF, and by the Department of Plant Biology, University of Georgia.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to J.R.B. (banavar@psu.edu) or S.P.H. (shubbell@dogwood.botany.uga.edu).

The genome of a motile marine *Synechococcus*

B. Palenik[1], B. Brahamsha[1], F. W. Larimer[2,3], M. Land[2,3], L. Hauser[2,3], P. Chain[3,4], J. Lamerdin[3,4], W. Regala[3,4], E. E. Allen[1]*, J. McCarren[1], I. Paulsen[5], A. Dufresne[6], F. Partensky[6], E. A. Webb[7] & J. Waterbury[7]

[1]*Marine Biology Research Division, Scripps Institution of Oceanography, University of California, San Diego, La Jolla, California 92093-0202, USA*
[2]*Computational Biology, Oak Ridge National Laboratory, Oak Ridge, Tennessee 37831-6480, USA*
[3]*Joint Genome Institute, Walnut Creek, California 94598, USA*
[4]*Lawrence Livermore National Laboratory, Livermore, California 94550-9234, USA*
[5]*TIGR, 9712 Medical Center Drive, Rockville, Maryland 20850, USA*
[6]*UMR 7127 CNRS Station Biologique de Roscoff, 29682 Roscoff, France*
[7]*Biology Department, Woods Hole Oceanographic Institution, Woods Hole, Massachusetts 02543, USA*

* Present address: Department of Earth and Planetary Science, University of California, Berkeley, California 94720, USA

Marine unicellular cyanobacteria are responsible for an estimated 20–40% of chlorophyll biomass and carbon fixation in the oceans[1]. Here we have sequenced and analysed the 2.4-megabase genome of *Synechococcus* sp. strain WH8102, revealing some of the ways that these organisms have adapted to their largely oligotrophic environment. WH8102 uses organic nitrogen and phosphorus sources and more sodium-dependent transporters than a model freshwater cyanobacterium. Furthermore, it seems to have adopted strategies for conserving limited iron stores by using nickel and cobalt in some enzymes, has reduced its regulatory machinery (consistent with the fact that the open ocean constitutes a far more constant and buffered environment than fresh water), and has evolved a unique type of swimming motility. The genome of WH8102 seems to have been greatly influenced by horizontal gene transfer, partially through phages. The genetic material contributed by horizontal gene transfer includes genes involved in the modification of the cell surface and in swimming motility. On the basis of its genome, WH8102 is more of a generalist than two related marine cyanobacteria[2].

Most species of picoplanktonic marine cyanobacteria currently known belong to two genera: *Synechococcus* and *Prochlorococcus*. Members must have the ability to acquire major nutrients and trace metals at the submicromolar concentrations found in the oligotrophic open seas. Their light-harvesting apparatus is uniquely

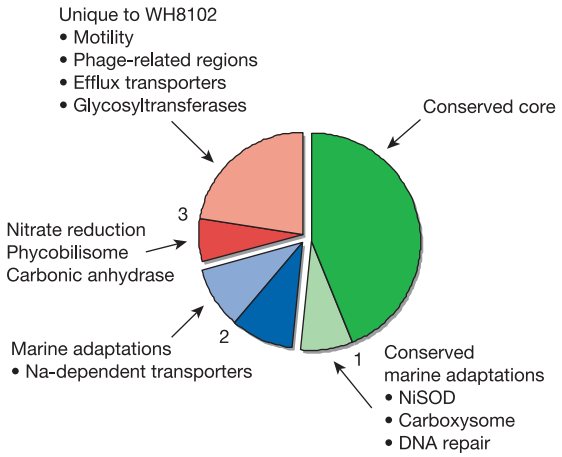

Figure 1 The genome of WH8102 can be divided into three categories: ORFs found in WH8102 and both related *Prochlorococcus* genomes (region 1); ORFs found in WH8102 and only one *Prochlorococcus* genome (region 2); and ORFs not found in the two *Prochlorococcus* genomes (region 3). These regions can be subdivided based on whether or not an ORF has a BLAST hit to freshwater *Synechocystis* sp. strain PCC6803. The lighter shading of each region represents the fraction not in PCC6803. As shown, examination of ORFs in each subcategory has provided insights into the evolutionary adaptation of marine *Synechococcus*.

adapted to the spectral quality of light in the ocean[3,4]. Of the two major marine unicellular genera, *Synechococcus* is usually less abundant in very oligotrophic environments, but has a broader global distribution[1,4]. A great deal of genetic diversity exists within the genus *Synechococcus*, with strains probably adapted to specific ecological niches[3,5]. Furthermore, one group of strains seems to have adapted to the oligotrophic marine environment by developing a new form of swimming motility not seen so far in any other prokaryotic group[6].

On the basis of the genome of *Synechococcus* sp. strain WH8102 (hereafter referred to as WH8102) and comparing it to the related genomes of two *Prochlorococcus* strains[2] we were able to define 1,314 open reading frames (ORFs) common to all three genomes (about half of the genome) and 736 ORFs found only in WH8102 (about a third of the genome) that indicate the specific ecological strategies of WH8102 relative to coexisting *Prochlorococcus* (see Methods, Fig. 1 and Supplementary Table 1). Here we show how the genome reveals some of the interactions of WH8102 with its environment (nutrients, light, toxins) and with other organisms, especially phages.

The WH8102 genome contains 16 probable or possible phage integrases—enzymes that function as site-specific DNA recombinases[7] (Supplementary Table 2). In WH8102, many of these occur adjacent to or near transfer RNAs and in regions with an anomalously low percentage of G + C content (Table 1a and Fig. 2). These regions of low G + C percentage also show atypical trinucleotide composition (data not shown). In addition, possible phage integrase regulators (SYNW2105, SYNW1660, SYNW1665) are also found. Thus, *Synechococcus* has regions that greatly resemble pathogenicity islands—regions that are often mobilized between strains of pathogenic bacteria[7]. Hence, although the genome of WH8102 does not contain prophages or plasmids, it does seem to have been, in its evolutionary past, extensively altered through horizontal gene transfer, possibly due to phages or plasmids. In contrast, far fewer potential phage integrases are found in *Prochlorococcus* (four in MIT9313 and one in MED4).

Other regions of low G + C percentage not associated with phage integrases also show atypical trinucleotide composition, suggestive of recent horizontal gene transfer (Table 1b). These regions encode genes involved in broadening the range of nitrogen substrates that can be used by WH8102, as well as some encoding transport capabilities. Furthermore, several of the genes found in these regions are homologues of ORFs involved in the carbohydrate modification of the cell envelope (including glycosyltransferases, and homologues of genes involved in the synthesis of sialic acid). One hypothesis for the function of these glycosyltransferases is that they may be required in constructing the motility apparatus of this organism, as at least one of its components is glycosylated[8]. Another possibility is that WH8102 may use these envelope-modifying genes to change its cell surface characteristics to help it evade grazers and other predators such as phages. Cell surface properties are known to affect grazing rates in the marine environment[9].

An examination of the genome of WH8102 provides an indication of the uniqueness of *Synechococcus* swimming motility. None of the proteins (motor, flagellar) associated with other forms of prokaryotic motility was found, with the exception of six ORFs associated with type IV-pilus-dependent motility (homologues of *pilB*, *-C*, *-D*, *-Q* and *-T*). Orthologues of these are also present in MIT9313, but not MED4. Nevertheless, these ORFs in WH8102 do not encode the full complement of genes required for pilus assembly and function, and pilin subunit homologues are absent. Pili or surface-associated twitching have not been observed in WH8102.

Recent studies using transposon mutagenesis (J.M. and B.B., manuscript in preparation) coupled with that of motility mutant *swmA*[8] indicate that genes required for motility are found in at least

Table 1 **Atypical regions of G + C per cent in the WH8102 genome**

Region	Size (kb)	No. ORFs	G + C (%)	Acquisition
(a) Low G + C per cent regions associated with predicted phage integrases				
1124989–1158432	33.4	27	53.6	–
1982294–1996886	14.6	14	50.7	Efflux
1606831–1591392	15.4	12	47.8	–
1335428–1344827	9.4	7	49.9	–
2312441–2322542	10.1	12	50.7	–
1488335–1524808	36.0	37	49.8	–
1183175–1171388	11.8	13	51.7	–
857915–842611	15.3	14	49.3	Na/Glut. symporter
2138960–2144868	5.9	4	40.9	–
353145–383689	30.0	29	52.0	–
(b) Low G + C per cent regions not associated with predicted phage integrases				
427233–465883	38.7	37	38.8	Modification of cell envelope, multidrug efflux
622199–633146	10.9	7	43.96	Modification of cell envelope
2379778–2394189	14.4	15	39.0	Cyanate usage, metal uptake
912098–954990	42.9	12	42.2	Motility

two widely separated regions (Fig. 3). The second region contains SwmB (SYNW0953), a very large ORF (10,791 amino acids) that constitutes more than 1% of the genome size and is currently one of the longest bacterial ORFs ever reported. Notably, it is found in one of the unusually low G + C percentage regions (Table 1b).

Transport in WH8102 accounts for about 5–6% of the predicted ORFs, similar to most other bacterial genomes[10]. Compared with other genomes[10], transporter capability is heavily biased towards the use of ABC transporters with about 60% of the ORFs encoding ABC transporter components. A distinct bias against P-type ATPase transporters is found, with only one such transporter, for copper, compared with nine in PCC6803, a model unicellular freshwater cyanobacterium. This one P-type ATPase may be conserved due to the use of copper in plastocyanin, an electron transfer protein that can substitute for an iron-containing cytochrome in photosynthesis.

Notably, WH8102 has multiple channels for transporting major seawater ions and multiple transporters or ABC-type solute-binding proteins for several major nutrients; for example, there are multiple solute-binding proteins for phosphate and two for urea. WH8102 seems to have an independent transporter for urea (SYNW2455), reinforcing its importance as a nitrogen source for cyanobacterial growth in oligotrophic environments. The multiple transporters in WH8102 may have different affinities and be regulated differently depending on nutrient concentrations.

One of the surprises from our analyses of the genome of WH8102 is the prediction that *Synechococcus* can use some new organic compounds as nitrogen and phosphorus sources. As inorganic nitrogen and phosphorus are often thought to be limiting in the marine environment, these potential sources are of particular interest. Amino acid and oligopeptide transporters are found, suggesting that *Synechococcus* may have the ability to use these ubiquitous compounds in sea water; transport of a few amino acids has also been demonstrated[11]. In addition, genes for the transport of cyanate and its breakdown by cyanase appear to be present in WH8102 and in *Prochlorococcus* MED4—in fact, WH8102 grows on

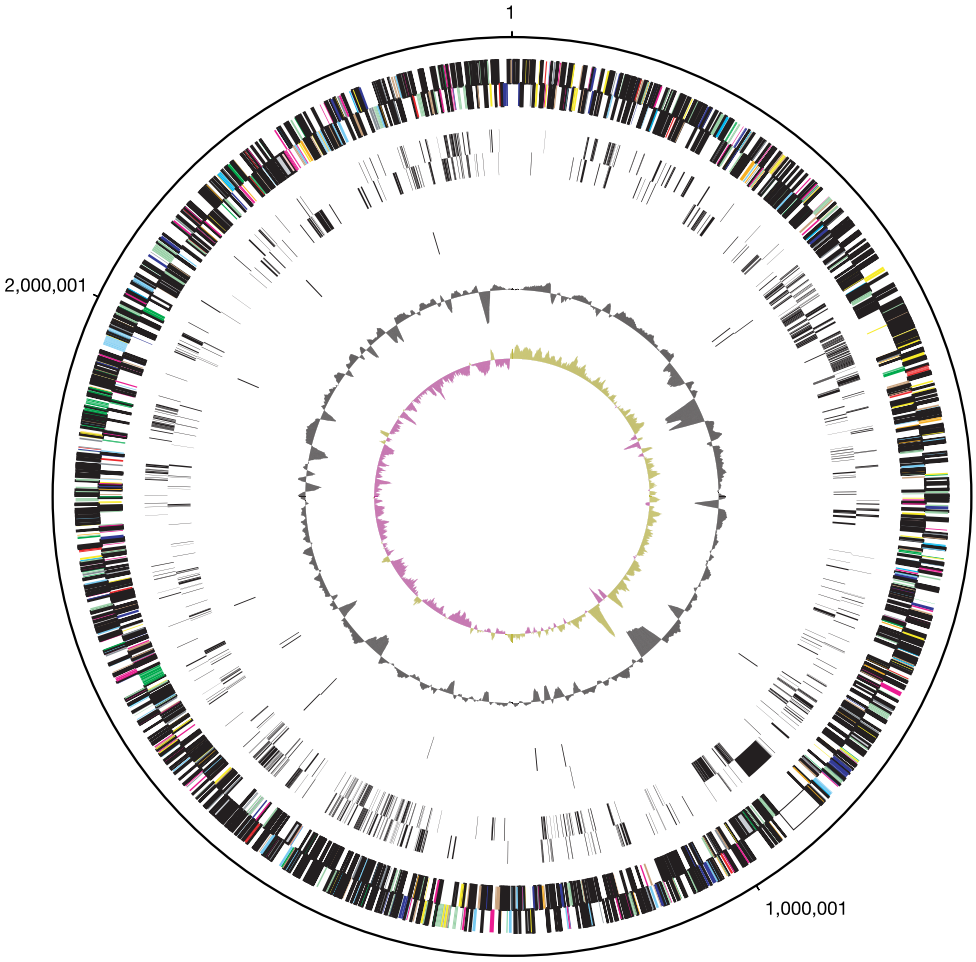

Figure 2 The chromosome of *Synechococcus* sp. strain WH8102. The following description is of the eight circles, with the first circle being the outer circle, and the others progressing inwards. First circle, predicted coding regions on the plus strand coloured by functional category: white, hypothetical; black, unassigned and other; dark red, energy metabolism; green, photosynthesis; blue, DNA replication and repair; cyan, fatty acid metabolism; magenta, biosynthesis of cofactors; yellow, cellular processes; pale green, transport and binding; sky blue, translation; orange, regulatory functions; brown, amino acid biosynthesis; pink, cell envelope; grey, conserved hypothetical; medium red, transcription; light red, purines and pyrimidines; pale pink, central metabolism. Second circle, predicted coding regions on minus strand (same colour scheme as in the first circle). Third and fourth circles, 736 'characteristic' genes not found in two *Prochlorococcus* strains (region 3 in Fig. 1), plus and minus strands, respectively. Fifth and sixth circles, predicted phage integrases on plus and minus strands, respectively. Seventh circle, G + C content (deviation from average); eighth circle, G + C skew curve in purple and olive. Phage integrases are often associated with low G + C regions. Low G + C regions often contain WH8102 characteristic ORFs. Scale (in bp) is indicated along the outside of the circle.

cyanate as a sole nitrogen source (B.P., unpublished data). The genes for cyanate use have been characterized for the freshwater cyanobacterium *Synechococcus elongatus* PCC7942 but cyanate remains uncharacterized as a nitrogen source in aquatic environments[12].

Similarly, genes for the transport of phosphonates (compounds with C–P bonds) are present in WH8102 and are found in the two sequenced *Prochlorococcus* genomes as well. WH8102 grows on phosphonate as a sole phosphorous source (B.P., unpublished data). Phosphonates are known to be produced by some major eukaryotic phytoplankton groups such as the coccolithophorids, and were recently reported to be an important fraction of total phosphate in sea water[13]. WH8102 also has multiple alkaline phosphatases (SYNW0120, SYNW0196, SYNW2391 and SYNW2390) that could be used to obtain phosphate from other organic phosphorus sources in its environment. Thus genome analyses are further dispelling the classical concept of cyanobacteria as being plant-like and dependent solely on inorganic forms of nutrients[14].

In addition, a number of conserved systems for exporting compounds (for example, multidrug efflux systems) are found both in the ABC transporter family and the MFS transporter family. WH8102 has a larger number of efflux transporters in the ABC family compared with *Prochlorococcus*. These results suggest that marine cyanobacteria, despite living in extremely oligotrophic conditions, may still find themselves in the position of needing to export 'toxins' produced by other microorganisms. Antagonistic interactions between pelagic bacteria have been reported recently[15]. Exposure to toxins may be greater for motile *Synechococcus* than for other marine cyanobacteria, as they may be chemotactic towards marine particles where higher localized concentrations of heterotrophic bacteria release nitrogenous coumpounds[11].

WH8102 also has efflux pumps for metals (SYNW1472 and SYNW0900) that are lacking in both *Prochlorococcus* strains. Characterizing these further may put a mechanistic basis behind previous observations[16] that *Synechococcus* seems to be more resistant to copper compared with *Prochlorococcus*, and that this resistance may help explain the seasonal cycles of these organisms in the Sargasso Sea. WH8102 also has predicted genes for the reduction of arsenate to arsenite (SYNW1767) and its efflux (SYNW1039). It has been hypothesized that arsenate is a competitor for phosphate and that systems would be needed to deal with this compound, especially in low-phosphate waters[17].

WH8102 has more capacity for sodium-driven transport than freshwater cyanobacteria such as PCC6803 (see Methods and Fig. 1), with transporters of the alanine/glycine:cation (sodium) symporter family (SYNW0828) and of the neurotransmitter:sodium symporter family (SYNW0699). It also has two transporters from the solute:sodium symporter family (SYNW2455, SYNW0619) compared with one in PCC6803.

In contrast to freshwater cyanobacteria, WH8102 has two potential transporters (SYNW1915, SYNW1916 and SYNW1917, and SYNW0229) for glycine betaine and related compounds found in marine waters. Adjacent to the ABC transporter but on the opposite strand are enzymes predicted to synthesize glycine betaine from glycine (SYNW1914, SYNW1913) using a pathway only reported before from an extremely halophilic proteobacterium[18]. When a freshwater *Synechococcus*, strain PCC7942, was genetically engineered to make glycine betaine, it became more halotolerant[19].

Despite the importance of iron as a limiting nutrient in the oceans, WH8102 does not have a detectable system for siderophore synthesis and uptake. However, it does have strategies for iron conservation such as using plastocyanin (copper) for electron transport and a cobalt-dependent ribonucleotide reductase (SYNW1692) rather than the Fe-containing one found in many other cyanobacteria. Another example of iron conservation in WH8102 is its predicted nickel superoxide dismutase (SOD). Multiple SOD types exist for removing photosynthetically produced superoxide radicals including ones using iron, manganese or copper-zinc as metal cofactors. Unlike the freshwater PCC6803, the marine cyanobacteria WH8102, both *Prochlorococcus* species and *Trichodesmium*, a marine N_2-fixing cyanobacterium (http://www.jgi.doe.gov/JGI_microbial/html/index.html), are predicted to use a new nickel SOD—seen recently in *Streptomyces*—as their only SOD, thus saving iron and manganese for other uses (see Supplementary Fig. 1).

In comparison with the sequenced *Prochlorococcus* strains, WH8102 is a transport generalist. It has predicted transporters for the efflux of chromate (SYNW1323) and arsenite (SYNW1039) that are found in MED4 but not MIT9313. It shares with MIT9313, but not MED4, the ability to use the sodium symporters mentioned above. In addition, WH8102 has transporters that are not found in *Prochlorococcus*, and that are predicted to be involved in the uptake of nitrate, a quaternary ammonium group ($R-N+(CH_3)3$) compound such as sarcosine, another nitrate-like compound, metals (magnesium/cobalt/nickel), and in cation efflux. This may be a characteristic of marine *Synechococcus* in general or it may be a characteristic of motile *Synechococcus*.

In WH8102 the major components of photosynthesis and respiration are well conserved and are usually most closely related to those of other cyanobacteria. Notable exceptions are genes in WH8102 and *Prochlorococcus* implicated in carboxysome structure and assembly, including those encoding the subunits of ribulose-1,5-bisphosphate carboxylase. This is thought to be due to a horizontal gene transfer event[20].

As in most cyanobacteria (other than marine *Prochlorococcus* and two other known prochlorophytes), *Synechococcus* harvests light using phycobilisomes, which are multisubunit complexes binding different types of phycobilins. Two interesting observations differentiate the use of phycobilisomes by WH8102 from those of other cyanobacteria analysed so far. Nowhere in the WH8102 genome are there homologues of the *cpcC* and *cpcD* genes, which in freshwater cyanobacteria are known to encode two types of phycocyanin-

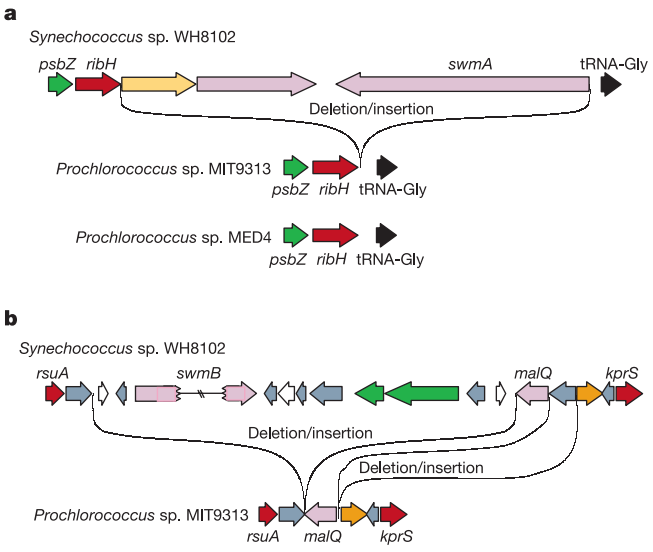

Figure 3 Organization of two chromosomal regions in WH8102 that contain motility genes. **a**, The region containing *swmA* and two other ORFs (a predicted glycosyltransferase and a sulphotransferase) appears to have been inserted in the *ribH*/tRNA-Gly region, or conversely, deleted in the two *Prochlorococcus* genomes. **b**, *swmB* and several other ORFs have been inserted between the *rsuA* and *malQ* regions. The double lines in *swmB* indicate that it is not drawn to scale, as it is approximately 20 times larger than *malQ*. The ORFs are colour-coded by predicted function as described in Fig. 2.

associated L_R linker polypeptides. These linkers are necessary for the correct assembly of phycocyanin discs in the phycobilisome rods[21]. Their absence in WH8102 suggests that there is a single disc of phycocyanin, as is the case in mutants of *Synechococcus* PCC7002 in which the *cpcC* gene has been inactivated[21]. The genome thus provides a basis for the interpretation of absorbance spectra, where reduced phycocyanin (orange-light absorbing) relative to phycoerythrin (blue-light absorbing) probably represents an adaptation to the oligotrophic marine environment where blue light is particularly important[3]. In addition, the genome of WH8102 also lacks homologues of *nblA* and *nblB*, two genes implicated in the degradation of phycobilisomes during nutrient stress in cyanobacteria[22,23]. Thus, phycobilisome degradation may not occur or may be under the control of other genes in WH8102.

Whereas *Synechococcus* possesses homologues of the low-affinity bicarbonate transport mechanism in PCC6803, it lacks homologues of *ndhD3*, *ndhF3* and *chpY*, genes implicated in high-affinity transport in the same organism[20]. Their absence might be an adaptation to the marine habitat, where bicarbonate (2 mM) is probably rarely limiting. Notably, WH8102 has two predicted carbonic anhydrases (SYNW0897 and SYNW2467) whereas *Prochlorococcus* has none, although these genes can be highly divergent and difficult to predict. SYNW2467 is adjacent to the genes encoding nitrate reductase. *Synechococcus* WH8102 can use nitrate for growth in contrast to the two *Prochlorococcus* strains[2], and this carbonic anhydrase may have been lost with the loss of nitrate usage. Although intriguing, a specific connection between nitrate usage and carbonic anhydrase has not been shown.

One way that bacteria sense and respond to their environment is by using two-component regulatory systems consisting of a sensor kinase and a response regulator. In PCC6803 there are nearly 40 sensor kinase and response regulator pairs (http://www.kazusa.or.jp/cyano/index.html). In contrast, WH8102 has only five sensor histidine kinases and nine response regulators, of which one, SYNW1598, may be a pseudogene as it is missing conserved functional residues[24]. Even accounting for a smaller genome size, WH8102 as well as the two *Prochlorococcus* species have fewer systems for responding to environmental changes using these gene families. Furthermore, as there are fewer sensors than response regulators, there seems to be an economy of regulation in which some sensors may transmit information to more than one response regulator.

In addition to the principal RNA polymerase sigma factor *sigA* (SYNW1783), WH8102 encodes five type II sigma factors, typical of cyanobacteria in general[25]. WH8102 however has only one homologue of the type III sigma factor (SYNW1232). This is a low number compared with the three to five seen in other sequenced cyanobacteria (PCC6803, PCC7120 and *Thermosynechococcus*; http://www.kazusa.or.jp/cyano/index.html). One hypothesis for the minimal regulatory machinery (two-component systems and sigma factors) in *Synechococcus* and *Prochlorococcus* is that they have evolved in an open ocean environment that is relatively constant, thus they do not need a regulatory system that could modulate their gene expression to a more variable environment. Alternatively, a minimal regulatory system could be the result of an ecological strategy of only some marine cyanobacteria.

On the basis of its genome, *Synechococcus* WH8102 is clearly more nutritionally versatile and a 'generalist' compared with its *Prochlorococcus* relatives. As the genus *Prochlorococcus* seems to have evolved only once, it may have gone through an evolutionary 'bottleneck' in which its capabilities were originally limited to those of a particular strain followed by subsequent acquisition of new abilities. Alternatively, *Synechococcus* may be more subject than *Prochlorococcus* to horizontal gene transfer from phages, as seen by the presence of more phage integrases. It is possible that not all *Synechococcus* are more versatile in their transport abilities, just the strains that are motile. Partial or complete genomes of additional marine cyanobacteria from this group will help answer these questions. □

Methods

Genome sequencing

Genomic DNA was isolated from WH8102 as reported previously[26]. Whole-genome shotgun libraries were obtained by fragmenting genomic DNA using mechanical shearing and cloning 2–3-kilobase fragments into pUC18. Double-ended plasmid sequencing reactions were carried out using PE BigDye Terminator chemistry (Perkin Elmer) and sequencing ladders were resolved on PE 377 Automated DNA Sequencers (Perkin Elmer). As the first genome drafted during the start-up of the microbial sequencing effort at the J.G.I. Production Sequencing Facility in Walnut Creek, California, this genome was sequenced to unusually high coverage. The whole-genome sequence of WH8102 was obtained from 66,550 reads with an average read length for this project of >575 base pairs (bp) per read for 16-fold redundancy. Sequence assembly was accomplished using PHRAP (P. Green). All gaps were closed by primer walking on gap-spanning library clones or PCR products. The overall genome structure was verified by long-range genomic PCR reactions. The two tandem repeats were resolved by combining information from individual clones, single-nucleotide polymorphism analysis and PCR. Only after this region was finished was it discovered that a single, long ORF was preserved.

Genome analysis

For genome analyses, the combination of three gene modelling programs—Critica[27], Glimmer[28] and Generation (http://compbio.ornl.gov/generation/index.shtml)—was used in the determination of potential coding sequences. These assignments were further checked manually. A revised gene/protein set was searched against the KEGG GENES, Pfam, PROSITE, PRINTS, ProDom and COGs databases, in addition to BLASTP against the NCBI non-redundant database. From these results, categorizations were developed using the KEGG and COGs hierarchies. Transfer RNAs were identified using tRNAscan-SE[29]. To identify regions of atypical nucleotide composition, the trinucleotide composition was determined.

Manual annotation of ORFs was carried out using Artemis, the Artemis Comparison Tool (http://www.sanger.ac.uk/Software/ACT/) and Clustal W[30]. The results of the KEGG and other comparisons described above were examined manually to check automated product assignments and make additional assignments. The proteome sequences of WH8102 and *Prochlorococcus* (MED4 and MIT9313)[2] were compared using the Artemis Comparison Tool. This program, in conjunction with Clustal W, was used for refining predicted start sites, adding ORFs not predicted by the gene modelling programs, and obtaining consistent annotation across three genomes. Manual annotation was done in conjunction with the *Prochlorococcus* annotation team. Transporters were analysed and annotated using methods described in ref. 10.

Pairwise BLAST analyses of three marine cyanobacterial genomes (WH8102 and *Prochlorococcus marinus* strains MED4 and MIT9313 (ref. 2)) against each other and a cut-off *e*-value of e^{-6}, followed by additional manual curation including examination of gene context, were used to partition the genome of WH8102 into three categories: 1,314 ORFs found in all three genomes and predicted to be orthologues; 476 predicted orthologous ORFs found in WH8102 and one other *Prochlorococcus* genome; and 736 ORFs characteristic of WH8102 (not found in either of the other *Prochlorococcus* genomes). The latter category partially represents the ecological capabilities of this organism compared with *Prochlorococcus* (Supplementary Table 3).

Using pairwise BLAST analyses, the three categories of WH8102 ORFs were further subdivided based on whether or not an ORF was found in a model freshwater cyanobacterium *Synechocystis* PCC6803 (hereafter termed as PCC6803; see http://www.kazusa.or.jp/cyano/index.html). After examination of different cut-offs, BLAST analyses with a cut-off *e*-value of e^{-10} were used for this assignment. For the 'core' marine cyanobacterial genome of 1,314 ORFs, 1,112 (85%) are also found in PCC6803 (Fig. 1). This provides an estimate of the portion of the WH8102 genome that has been conserved in all cyanobacterial genomes so far from a primal cyanobacterial ancestor (and includes ORFs conserved in all bacterial taxa). This portion is different from a minimal bacterial genome or a minimal cyanobacterial genome, as horizontally acquired genes could carry out functions required for cell viability. The 15% of the marine cyanobacterial core not in PCC6803 was found to include some of the adaptations and evolutionary events that distinguish the marine *Synechococcus*/*Prochlorococcus* cyanobacterial lineage from other cyanobacterial lineages.

Of the 736 WH8102 characteristic ORFs not found in the *Prochlorococcus* genomes, 23% have related ORFs in PCC6803, using a BLAST cut-off *e*-value of e^{-10}. This is not surprising as these partly represent the ability of both PCC6803 and WH8102, but not the *Prochlorococcus* strains, to create a functional phycobilisome for harvesting light, and a functional nitrate reductase with molybdenum cofactors for using nitrate as a nitrogen source. Forty-five per cent of these characteristic ORFs are hypothetical.

Received 9 May; accepted 28 July 2003; doi:10.1038/nature01943.
Published online 13 August 2003.

1. Partensky, F., Hess, W. R. & Vaulot, D. *Prochlorococcus*, a marine photosynthetic prokaryote of global significance. *Microbiol. Mol. Biol. Rev.* **63,** 106–127 (1999).
2. Rocap, G. *et al.* Genome divergence in two *Prochlorococcus* ecotypes reflects oceanic niche differentiation. *Nature* **424,** 1042–1047 (2003).
3. Waterbury, J. B., Watson, F. W., Valois, F. W. & Franks, D. G. in *Photosynthetic Picoplankton* (eds Platt, T. & Li, W. K. W.) 71–120 (Canadian Department of Fisheries and Oceans, Ottawa, 1986).

4. Scanlan, D. J. & West, N. J. Molecular ecology of the marine cyanobacterial genera *Prochlorococcus* and *Synechococcus*. *FEMS Microbiol. Ecol.* **40,** 1–12 (2002).
5. Ferris, M. J. & Palenik, B. Niche adaptation in ocean cyanobacteria. *Nature* **396,** 226–228 (1998).
6. Toledo, G., Palenik, B. & Brahamsha, B. Swimming strains of marine *Synechococcus* with widely different photosynthetic pigment ratios form a monophyletic group. *Appl. Environ. Microbiol.* **65,** 5247–5251 (1999).
7. Bushman, F. *Lateral DNA Transfer Mechanisms and Consequences* (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York, 2002).
8. Brahamsha, B. An abundant cell-surface polypeptide is required for swimming by the nonflagellated marine cyanobacterium *Synechococcus*. *Proc. Natl Acad. Sci. USA* **93,** 6504–6509 (1996).
9. Monger, B. C., Landry, M. R. & Brown, S. L. Feeding selection of heterotrophic marine nanoflagellates based on the surface hydrophobicity of their picoplankton prey. *Limnol. Oceanogr.* **44,** 1917–1927 (1999).
10. Paulsen, I. T., Nguyen, L., Sliwinski, M. K., Rabus, R. & Saier, M. H. Jr Microbial genome analyses: comparative transport capabilities in eighteen prokaryotes. *J. Mol. Biol.* **301,** 75–101 (2000).
11. Willey, J. M. & Waterbury, J. B. Chemotaxis toward nitrogenous compounds by swimming strains of marine *Synechococcus* spp. *Appl. Environ. Microbiol.* **55,** 1888–1894 (1989).
12. Harano, Y. *et al.* Identification and nitrogen regulation of the cyanase gene from the cyanobacteria *Synechocystis* sp. strain PCC6803 and *Synechococcus* sp. strain PCC7942. *J. Bacteriol.* **179,** 5744–5750 (1997).
13. Clark, L. L., Ingall, E. D. & Benner, R. Marine phosphorus is selectively remineralized. *Nature* **393,** 426 (1998).
14. Zehr, J. P. & Ward, B. B. Nitrogen cycling in the ocean: new perspectives on processes and paradigms. *Appl. Environ. Microbiol.* **68,** 1015–1024 (2002).
15. Long, R. A. & Azam, F. Antagonistic interactions among marine pelagic bacteria. *Appl. Environ. Microbiol.* **67,** 4975–4983 (2001).
16. Mann, E. L., Ahlgren, N., Moffett, J. W. & Chisholm, S. W. Copper toxicity and cyanobacteria ecology in the Sargasso Sea. *Limnol. Oceanogr.* **47,** 976–988 (2002).
17. Palenik, B. & Dyhrman, S. T. in *Phosphorus in Plant Biology: Regulatory Roles in Molecular, Cellular, Organismic, and Ecosystem Processes* (eds Lynch, J. P. & Deikman, J.) 26–38 (American Society of Plant Physiologists, Rockville, MD, 1998).
18. Nyyssola, A., Kerovuo, J., Kaukinen, P., von Weymarn, N. & Reinikainen, T. Extreme halophiles synthesize betaine from glycine by methylation. *J. Biol. Chem.* **275,** 22196–22201 (2000).
19. Nomura, M., Ishitani, M., Takabe, T., Rai, A. K. & Takabe, T. *Synechococcus* sp. PCC7942 transformed with *Escherichia coli bet* genes produces glycine betaine from choline and acquires resistance to salt stress. *Plant Physiol.* **107,** 703–708 (1995).
20. Badger, M. R., Hanson, D. & Price, G. D. Evolution and diversity of CO2 concentrating mechanisms in cyanobacteria. *Funct. Plant Biol.* **29,** 161–173 (2002).
21. de Lorimier, R., Guglielmi, G., Bryant, D. A. & Stevens, S. E. J. Structure and mutation of a gene encoding a M_r 33 000 phycocyanin-associated linker polypeptide. *Arch. Microbiol.* **153,** 541–549 (1990).
22. Collier, J. L. & Grossman, A. R. A small polypeptide triggers complete degradation of light-harvesting phycobiliproteins in nutrient-deprived cyanobacteria. *EMBO J.* **13,** 1039–1047 (1994).
23. Dolganov, N. & Grossman, A. R. A polypeptide with similarity to phycocyanin alpha-subunit phycocyanobilin lyase involved in degradation of phycobilisomes. *J. Bacteriol.* **181,** 610–617 (1999).
24. Volz, K. in *Two-Component Signal Transduction* (eds Hoch, J. A. & Silhavy, T. J.) 53–64 (American Society for Microbiology, Washington DC, 1995).
25. Goto-Seki, A., Shirokane, M., Masuda, S., Tanaka, K. & Takahashi, H. Specificity crosstalk among group 1 and group 2 sigma factors in the cyanobacterium *Synechococcus* sp PCC7942: *in vitro* specificity and a phylogenetic analysis. *Mol. Microbiol.* **34,** 473–484 (1999).
26. Brahamsha, B. A genetic manipulation system for oceanic cyanobacteria of the genus *Synechococcus*. *Appl. Environ. Microbiol.* **62,** 1747–1751 (1996).
27. Badger, J. H. & Olsen, G. J. CRITICA: Coding region identification tool invoking comparative analysis. *Mol. Biol. Evol.* **16,** 512–524 (1999).
28. Delcher, A. L., Harmon, D., Kasif, S., White, O. & Salzberg, S. L. Improved microbial gene identification with GLIMMER. *Nucleic Acids Res.* **27,** 4636–4641 (1999).
29. Lowe, T. M. & Eddy, S. R. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequences. *Nucleic Acids Res.* **25,** 955–964 (1997).
30. Thompson, J. D., Higgins, D. G. & Gibson, T. J. CLUSTAL W:. improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22,** 4673–4680 (1994).

Supplementary Information accompanies the paper on **www.nature.com/nature**.

Acknowledgements We thank the members of the *Prochlorococcus* annotation team: G. Rocap, S. W. Chisholm, D. Lindell, N. Algren, M. Coleman, W. Hess, A. Post, S. Shaw, C. Steglich, C. Ting, M. Sullivan, A. Tolonen, Z. Johnson and E. Zinser. We also thank T. Lane for discussions about carbonic anhydrases. This research was funded by the Biological and Environmental Research Program and the US Department of Energy's Office of Science. Sequencing was carried out and managed at the Joint Genome Institute. Computational analysis was performed at Oak Ridge National Laboratory, managed by UT-BATTELLE for the US Department of Energy. Additional support was provided by a DOE grant to B.P, B.B. and I.P., and an NSF grant to B.B. F.P. and A.D. were supported by the EC program Margenes, and by the Region Bretagne.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to B.P. (bpalenik@ucsd.edu). The sequence for the chromosome of *Synechococcus* sp. strain WH8102 is deposited in GenBank under accession number BX548020.

Genome divergence in two *Prochlorococcus* ecotypes reflects oceanic niche differentiation

Gabrielle Rocap[1], Frank W. Larimer[2,3], Jane Lamerdin[3], Stephanie Malfatti[3], Patrick Chain[3,4], Nathan A. Ahlgren[1], Andrae Arellano[3], Maureen Coleman[5], Loren Hauser[2,3], Wolfgang R. Hess[9]*, Zackary I. Johnson[5], Miriam Land[2,3], Debbie Lindell[5], Anton F. Post[10], Warren Regala[3], Manesh Shah[2,3], Stephanie L. Shaw[6]*, Claudia Steglich[9], Matthew B. Sullivan[7], Claire S. Ting[8], Andrew Tolonen[7], Eric A. Webb[11], Erik R. Zinser[5] & Sallie W. Chisholm[5,8]

[1]*School of Oceanography, University Of Washington, Seattle, Washington 98195, USA*
[2]*Computational Biology, Oak Ridge National Laboratory, Oak Ridge, Tennessee 37831, USA*
[3]*Joint Genome Institute, Walnut Creek, California 94598, USA*
[4]*Lawrence Livermore National Laboratory, Livermore, California 94550, USA*
[5]*Department of Civil and Environmental Engineering,* [6]*Department of Earth, Atmospheric and Planetary Sciences,* [7]*Joint Program in Biological Oceanography, Woods Hole Oceanographic Institution, and* [8]*Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA*
[9]*Institute of Biology, Humboldt-University, D-10115 Berlin, Germany*
[10]*Interuniversity Institute of Marine Science, 88103 Eilat, Israel*
[11]*Biology Department, Woods Hole Oceanographic Institution, Woods Hole, Massachusetts 02543, USA*

* Present addresses: Ocean Genome Legacy, Beverly, Massachusetts 01915, USA (W.R.H.); Department of Environmental Science Policy and Management, University of California, Berkeley, California 94720, USA (S.L.S.)

The marine unicellular cyanobacterium *Prochlorococcus* is the smallest-known oxygen-evolving autotroph[1]. It numerically dominates the phytoplankton in the tropical and subtropical oceans[2,3], and is responsible for a significant fraction of global photosynthesis. Here we compare the genomes of two *Prochlorococcus* strains that span the largest evolutionary distance within the *Prochlorococcus* lineage[4] and that have different minimum, maximum and optimal light intensities for growth[5]. The high-light-adapted ecotype has the smallest genome (1,657,990 base pairs, 1,716 genes) of any known oxygenic phototroph, whereas the genome of its low-light-adapted counterpart is significantly larger, at 2,410,873 base pairs (2,275 genes). The comparative architectures of these two strains reveal dynamic genomes that are constantly changing in response to myriad selection pressures. Although the two strains have 1,350 genes in common, a significant number are not shared, and these have been differentially retained from the common ancestor, or acquired through duplication or lateral transfer. Some of these genes have obvious roles in determining the relative fitness of the ecotypes in response to key environmental variables, and hence in regulating their distribution and abundance in the oceans.

As an oxyphototroph, *Prochlorococcus* requires only light, CO_2 and inorganic nutrients, thus the opportunities for extensive niche differentiation are not immediately obvious—particularly in view of the high mixing potential in the marine environment (Fig. 1a). Yet co-occurring *Prochlorococcus* cells that differ in their ribosomal DNA sequence by less than 3% have different optimal light intensities for growth[6], pigment contents[7], light-harvesting efficiencies[5], sensitivities to trace metals[8], nitrogen usage abilities[9] and cyanophage specificities[10] (Fig. 1b, c). These 'ecotypes'—distinct genetic lineages with ecologically relevant physiological differences—would be lumped together as a single species on the basis of their rDNA similarity[11], yet they have markedly different distributions within a stratified oceanic water column, with high-

Annexe IV

Annexe IV

Propriétés générales des cinq génomes de picocyanobactéries marines utilisés pour cette thèse. MED4, *P. marinus* MED4 ; SS120, *P. Marinus* SS120 ; MIT9313, *Prochlorococcus* sp. MIT9313 ; WH8102, *Synechococcus* sp. WH8102 ; WH7803, *Synechococcus* sp. WH7803.

	MED4	**SS120**	**MIT9313**	**WH8102**	**WH7803**
Longueur du génome (pb)	1657990	1751081	2410873	2434428	2366980
Contenu en G + C	30,8%	36,4%	50,7%	59,4%	60,2%
Densité de gènes	1,04	1,07	0,94	1,03	1,09
ORFs prédits	1716	1884	2272	2526	2532
ORFs sur le brin direct	48%	50,6%	53%	53,2%	49%
ORFs sur le brin complémentaire	52%	49,4%	47%	46,8%	51%
ORFs avec fonction	66%	66,6%	60%	53,3%	61,8%
Gènes hypothétiques conservés	29%	21,1%	31%	33,2%	29,6%
Gènes hypothétiques	5%	12,3%	9%	13,5%	8,6%
ARNt	37	40	43	42	44
Opéron ARNr	1	1	2	2	2

Bibliographie

Akerley, B.J., E.J. Rubin, V.L. Novick, K. Amaya, N. Judson, and J.J. Mekalanos. 2002. A genome-scale analysis for identification of genes required for growth or survival of *Haemophilus influenzae*. *Proc Natl Acad Sci U S A* **99:** 966-971.

Aravind, L. and C.P. Ponting. 1997. The GAF domain: an evolutionary link between diverse phototransducing proteins. *Trends Biochem Sci* **22:** 458-459.

Aravind, L., R.L. Tatusov, Y.I. Wolf, D.R. Walker, and E.V. Koonin. 1998. Evidence for massive gene exchange between archaeal and bacterial hyperthermophiles. *Trends Genet* **14:** 442-444.

Badger, M.R., D. Hanson, and G. Dean Price. 2002. Evolution and diversity of CO2 concentrating mechanisms in cyanobacteria. *Funct Plant Biol* **29:** 161-173.

Badger, M.R. and G.D. Price. 2003. CO2 concentrating mechanisms in cyanobacteria: molecular components, their diversity and evolution. *J Exp Bot* **54:** 609-622.

Bhaya, D., A. Dufresne, D. Vaulot, and A. Grossman. 2002. Analysis of the *hli* gene family in marine and freshwater cyanobacteria. *FEMS Microbiol Lett* **215:** 209-219.

Bibby, T.S., I. Mary, J. Nield, F. Partensky, and J. Barber. 2003. Low-light-adapted *Prochlorococcus* species possess specific antennae for each photosystem. *Nature* **424:** 1051-1054.

Bibby, T.S., J. Nield, F. Partensky, and J. Barber. 2001. Oxyphotobacteria - Antenna ring around photosystem I. *Nature* **413:** 590.

Blattner, F.R., G. Plunkett, 3rd, C.A. Bloch, N.T. Perna, V. Burland, M. Riley, J. Collado-Vides, J.D. Glasner, C.K. Rode, G.F. Mayhew. *et al.* 1997. The complete genome sequence of *Escherichia coli* K-12. *Science* **277:** 1453-1474.

Boekema, E.J., A. Hifney, A.E. Yakushevska, M. Piotrowski, W. Keegstra, S. Berry, K.P. Michel, E.K. Pistorius, and J. Kruip. 2001. A giant chlorophyll-protein complex induced by iron defciency in cyanobacteria. *Nature* **412:** 745-748.

Brenner, S.E., T. Hubbard, A. Murzin, and C. Chothia. 1995. Gene duplications in *H. influenzae*. *Nature* **378:** 140.

Bustos, S.A. and S.S. Golden. 1992. Light-regulated expression of the *psbD* gene family in *Synechococcus* sp. strain PCC 7942: evidence for the role of duplicated *psbD* genes in cyanobacteria. *Mol Gen Genet* **232:** 221-230.

Cohan, F.M. 2004. Periodic selection and ecological diversity in bacteria. In *Selective sweep* (ed. D. Nurminsky). Landes Biosciences.

Courties, C., R. Perasso, M. Chrétiennot-Dinet, M. Gouy, L. Guillou, and M. Troussellier. 1998. Phylogenetic analysis and genome size of *Ostreococcus tauri* (chlorophyta, prasinophyceae). *J Phycol* **34**.

da Silva, A.C., J.A. Ferro, F.C. Reinach, C.S. Farah, L.R. Furlan, R.B. Quaggio, C.B. Monteiro-Vitorello, M.A. Van Sluys, N.F. Almeida, L.M. Alves. *et al.* 2002. Comparison of the genomes of two *Xanthomonas* pathogens with differing host specificities. *Nature* **417:** 459-463.

Doolittle, W.F. 1999. Lateral genomics. *Trends Cell Biol* **9:** M5-8.

Dufresne, A., M. Salanoubat, F. Partensky, F. Artiguenave, I.M. Axmann, V. Barbe, S. Duprat, M.Y.

Galperin, E.V. Koonin, F. Le Gall. *et al.* 2003. Genome sequence of the cyanobacterium *Prochlorococcus marinus* SS120, a nearly minimal oxyphototrophic genome. *Proc Natl Acad Sci U S A* **100:** 10020-10025.

Dumontier, M., K. Michalickova, and C.W. Hogue. 2002. Species-specific protein sequence and fold optimizations. *BMC Bioinformatics* **3:** 39.

Edwards, R.A., G.J. Olsen, and S.R. Maloy. 2002. Comparative genomics of closely related salmonellae. *Trends Microbiol* **10:** 94-99.

Ferris, M.J. and B. Palenik. 1998. Niche adaptation in ocean cyanobacteria. *Nature* **396:** 226-228.

Fitch, W.M. 1970. Distinguishing homologous from analogous proteins. *Syst Zool* **19:** 99-113.

Fleischmann, R.D., M.D. Adams, O. White, R.A. Clayton, E.F. Kirkness, A.R. Kerlavage, C.J. Bult, J.F. Tomb, B.A. Dougherty, J.M. Merrick. *et al.* 1995. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* **269:** 496-512.

Fraser, C.M., J.D. Gocayne, O. White, M.D. Adams, R.A. Clayton, R.D. Fleischmann, C.J. Bult, A.R. Kerlavage, G. Sutton, J.M. Kelley. *et al.* 1995. The minimal gene complement of *Mycoplasma genitalium*. *Science* **270:** 397-403.

Fuller, N.J., D. Marie, F. Partensky, D. Vaulot, A.F. Post, and D.J. Scanlan. 2003. Clade-specific 16S ribosomal DNA oligonucleotides reveal the predominance of a single marine *Synechococcus* clade throughout a stratified water column in the Red Sea. *Appl Environ Microbiol* **69:** 2430-2443.

Galperin, M.Y., D.R. Walker, and E.V. Koonin. 1998. Analogous enzymes: independent inventions in enzyme evolution. *Genome Res* **8:** 779-790.

Garcia-Vallve, S., A. Romeu, and J. Palau. 2000. Horizontal gene transfer in bacterial and archaeal complete genomes. *Genome Res* **10:** 1719-1725.

Garczarek, L., W.R. Hess, J. Holtzendorff, G.W. van der Staay, and F. Partensky. 2000. Multiplication of antenna genes as a major adaptation to low light in a marine prokaryote. *Proc. Natl. Acad. Sci. USA* **97:** 4098-4101.

Garvish, J.F. and R.S. Lloyd. 1999. The catalytic mechanism of a pyrimidine dimer-specific glycosylase (pdg)/abasic lyase, *Chlorella* virus-pdg. *J Biol Chem* **274:** 9786-9794.

Gerdes, S.Y., M.D. Scholle, J.W. Campbell, G. Balazsi, E. Ravasz, M.D. Daugherty, A.L. Somera, N.C. Kyrpides, I. Anderson, M.S. Gelfand. *et al.* 2003. Experimental determination and system level analysis of essential genes in Escherichia coli MG1655. *J Bacteriol* **185:** 5673-5684.

Giaever, G., A.M. Chu, L. Ni, C. Connelly, L. Riles, S. Veronneau, S. Dow, A. Lucau-Danila, K. Anderson, B. Andre. *et al.* 2002. Functional profiling of the *Saccharomyces cerevisiae* genome. *Nature* **418:** 387-391.

Gil, R., F.J. Silva, E. Zientz, F. Delmotte, F. Gonzalez-Candelas, A. Latorre, C. Rausell, J. Kamerbeek, J. Gadau, B. Holldobler. *et al.* 2003. The genome sequence of *Blochmannia floridanus*: comparative analysis of reduced genomes. *Proc Natl Acad Sci U S A* **100:** 9388-9393.

Goericke, R. and D.J. Repeta. 1992. The pigments of *Prochlorococcus marinus* : the presence of divinyl chlorophyll *a* and *b* in a marine prochlorophyte. *Limnol Oceanogr* **37:** 425-433.

Gogarten, J.P. 1994. Which is the most conserved group of proteins? Homology-orthology, paralogy, xenology, and the fusion of independent lineages. *J Mol Evol* **39:** 541-543.

Golding, G.B. and A.M. Dean. 1998. The structural basis of molecular adaptation. *Mol Biol Evol* **15:** 355-369.

Gotz, T., U. Windhovel, P. Boger, and G. Sandmann. 1999. Protection of photosynthesis against ultraviolet-B radiation by carotenoids in transformants of the cyanobacterium *synechococcus* PCC 7942. *Plant Physiol* **120:** 599-604.

Helbling, E.W., A.G.J. Buma, M. Karin de Boer, and V.E. Villafañe. 2001. *In situ* impact of solar ultraviolet radiation on photosynthesis and DNA in temperate marine phytoplankton. *Mar Ecol Prog Ser* **211:** 43-49.

Herdman, M., R.W. Castenholz, J.B. Waterbury, and R. Rippka. 2001. Form-genus XIII. *Synechococcus*. In *Bergey's Manual of Systematic Bacteriology* (eds. D.R. Boone and R.W. Castenholz), pp. 508-512. Springer-Verlag, New York.

Hess, W.R., G. Rocap, C.S. Ting, F. Larimer, S. Stilwagen, J. Lamerdin, and S.W. Chisholm. 2001. The photosynthetic apparatus of *Prochlorococcus*: Insights through comparative genomics. *Photosynt Res* **70:** 53-71.

Hess, W.R., C. Steglich, C. Lichtlé, and F. Partensky. 1999. Phycoerythrins of the oxyphotobacterium *Prochlorococcus marinus* are associated to the thylakoid membrane and are encoded by a single large gene cluster. *Plant Mol Biol* **40:** 507-521.

Himmelreich, R., H. Plagens, H. Hilbert, B. Reiner, and R. Herrmann. 1997. Comparative analysis of the genomes of the bacteria *Mycoplasma pneumoniae* and *Mycoplasma genitalium*. *Nucleic Acids Res* **25:** 701-712.

Hitomi, K., K. Okamoto, H. Daiyasu, H. Miyashita, S. Iwai, H. Toh, M. Ishiura, and T. Todo. 2000. Bacterial cryptochrome and photolyase: characterization of two photolyase-like genes of *Synechocystis* sp. PCC 6803. *Nucleic Acids Res* **28:** 2353-2362.

Hutchison, C.A., S.N. Peterson, S.R. Gill, R.T. Cline, O. White, C.M. Fraser, H.O. Smith, and J.C. Venter. 1999. Global transposon mutagenesis and a minimal *Mycoplasma* genome. *Science* **286:** 2165-2169.

Huynen, M.A. and E. van Nimwegen. 1998. The frequency distribution of gene family sizes in complete genomes. *Mol Biol Evol* **15:** 583-589.

Itaya, M. 1995. An estimation of minimal genome size required for life. *FEBS Lett* **362:** 257-260.

Jain, R., M.C. Rivera, and J.A. Lake. 1999. Horizontal gene transfer among genomes: the complexity hypothesis. *Proc Natl Acad Sci U S A* **96:** 3801-3806.

Janssen, P.J., B. Audit, and C.A. Ouzounis. 2001. Strain-specific genes of *Helicobacter pylori*: distribution, function and dynamics. *Nucleic Acids Res* **29:** 4395-4404.

Johnson, P.W. and J.M. Sieburth. 1979. Chroococcoid cyanobacteria in the sea: a ubiquitous and diverse phototrophic biomass. *Limnol Oceanogr* **24:** 928-935.

Jordan, I.K., K.S. Makarova, J.L. Spouge, Y.I. Wolf, and E.V. Koonin. 2001. Lineage-specific gene expansions in bacterial and archaeal genomes. *Genome Res* **11:** 555-565.

Kamath, R.S., A.G. Fraser, Y. Dong, G. Poulin, R. Durbin, M. Gotta, A. Kanapin, N. Le Bot, S. Moreno, M. Sohrmann. *et al.* 2003. Systematic functional analysis of the *Caenorhabditis elegans* genome using RNAi. *Nature* **421:** 231-237.

Kanai, S., R. Kikuno, H. Toh, H. Ryo, and T. Todo. 1997. Molecular evolution of the photolyase-blue-light photoreceptor family. *J Mol Evol* **45:** 535-548.

Kaneko, T., Y. Nakamura, C.P. Wolk, T. Kuritz, S. Sasamoto, A. Watanabe, M. Iriguchi, A. Ishikawa, K. Kawashima, T. Kimura. *et al.* 2001. Complete genomic sequence of the filamentous nitrogen-fixing cyanobacterium *Anabaena* sp. strain PCC 7120. *DNA Res* **8:** 205-213; 227-253.

Kaneko, T., S. Sato, H. Kotani, A. Tanaka, E. Asamizu, Y. Nakamura, N. Miyajima, M. Hirosawa, M. Sugiura, S. Sasamoto. *et al.* 1996. Sequence analysis of the genome of the unicellular cyanobacterium *Synechocystis* sp. strain PCC6803. II. Sequence determination of the entire genome and assignment of potential protein-coding regions. *DNA Res* **3:** 109-136.

Kawashima, T., N. Amano, H. Koike, S. Makino, S. Higuchi, Y. Kawashima-Ohya, K. Watanabe, M. Yamazaki, K. Kanehori, T. Kawamoto. *et al.* 2000. Archaeal adaptation to higher temperatures revealed by genomic sequence of *Thermoplasma volcanium*. *Proc Natl Acad Sci U S A* **97:** 14257-14262.

Kimura, M. 1983. *The neutral theory of molecular evolution*. Cambridge University Press, Cambridge.

Kirk, J.T.O. 1994. The nature of the underwater light field. In *Light and photosynthesis in aquatic ecosystems* (ed. J.T.O. Kirk), pp. 129. Cambridge University Press.

Klasson, L. and S.G. Andersson. 2004. Evolution of minimal-gene-sets in host-dependent bacteria. *Trends Microbiol* **12:** 37-43.

Kobayashi, K., S.D. Ehrlich, A. Albertini, G. Amati, K.K. Andersen, M. Arnaud, K. Asai, S. Ashikaga, S. Aymerich, P. Bessieres. *et al.* 2003. Essential *Bacillus subtilis* genes. *Proc Natl Acad Sci U S A* **100:** 4678-4683.

Kondrashov, F.A., I.B. Rogozin, Y.I. Wolf, and E.V. Koonin. 2002. Selection in the evolution of gene duplications. *Genome Biol* **3:** RESEARCH0008.

Koonin, E.V. 2000. How many genes can make a cell: the minimal-gene-set concept. *Annu Rev Genomics Hum Genet* **1:** 99-116.

Koonin, E.V. 2003. Comparative genomics, minimal gene-sets and the last universal common ancestor. *Nat Rev Microbiol* **1:** 127-136.

Koonin, E.V., K.S. Makarova, and L. Aravind. 2001. Horizontal gene transfer in prokaryotes: quantification and classification. *Annu Rev Microbiol* **55:** 709-742.

Koonin, E.V., A.R. Mushegian, and P. Bork. 1996. Non-orthologous gene displacement. *Trends Genet* **12:** 334-336.

Koonin, E.V., A.R. Mushegian, M.Y. Galperin, and D.R. Walker. 1997. Comparison of archaeal and bacterial genomes: computer analysis of protein sequences predicts novel functions and suggests a chimeric origin for the archaea. *Mol Microbiol* **25:** 619-637.

Koonin, E.V., R.L. Tatusov, and K.E. Rudd. 1995. Sequence similarity analysis of *Escherichia coli* proteins: functional and evolutionary implications. *Proc Natl Acad Sci U S A* **92:** 11921-11925.

Koski, L.B., R.A. Morton, and G.B. Golding. 2001. Codon bias and base composition are poor indicators of horizontally transferred genes. *Mol Biol Evol* **18:** 404-412.

Kreil, D.P. and C.A. Ouzounis. 2001. Identification of thermophilic species by the amino acid compositions deduced from their genomes. *Nucleic Acids Res* **29:** 1608-1615.

Kunin, V. and C.A. Ouzounis. 2003. The balance of driving forces during genome evolution in prokaryotes. *Genome Res* **13:** 1589-1594.

LaRoche, J., G.W. van der Staay, F. Partensky, A. Ducret, R. Aebersold, R. Li, S.S. Golden, R.G. Hiller, P.M. Wrench, A.W. Larkum. *et al.* 1996. Independent evolution of the prochlorophyte and green plant chlorophyll *a/b* light-harvesting proteins. *Proc Natl Acad Sci U S A* **93:** 15244-15248.

Lawrence, J.G. 1997. Selfish operons and speciation by gene transfer. *Trends Microbiol* **5:** 355-359.

Lawrence, J.G., R.W. Hendrix, and S. Casjens. 2001. Where are the pseudogenes in bacterial genomes? *Trends Microbiol* **9:** 535-540.

Lawrence, J.G. and H. Ochman. 1998. Molecular archaeology of the *Escherichia coli* genome. *Proc Natl Acad Sci U S A* **95:** 9413-9417.

Lawrence, J.G. and J.R. Roth. 1999. Genomic flux: genome evolution by gene loss and acquisition. In *Organization of the Prokaryotic genome* (ed. R.L. Charlebois), pp. 263-289. American society for Microbiology, Washington, D.C.

Lespinet, O., Y.I. Wolf, E.V. Koonin, and L. Aravind. 2002. The role of lineage-specific gene family expansion in the evolution of eukaryotes. *Genome Res* **12:** 1048-1059.

Liu, H.B., K. Suzukil, C. Minami, T. Saino, and M. Watanabe. 2002. Picoplankton community structure in the subarctic Pacific Ocean and the Bering Sea during summer 1999. *Mar Ecol Prog Ser* **237:** 1-14.

Lokstein, H., C. Steglich, and W.R. Hess. 1999. Light-harvesting antenna function of phycoerythrin in *Prochlorococcus marinus*. *Biochim Biophys Acta* **1410:** 97-98.

McCullough, A.K., M.T. Romberg, S. Nyaga, Y. Wei, T.G. Wood, J.S. Taylor, J.L. Van Etten, M.L. Dodson, and R.S. Lloyd. 1998. Characterization of a novel cis-syn and trans-syn-II pyrimidine dimer glycosylase/AP lyase from a eukaryotic algal virus, *Paramecium bursaria* chlorella virus-1. *J Biol Chem* **273:** 13136-13142.

Mira, A., H. Ochman, and N.A. Moran. 2001. Deletional bias and the evolution of bacterial genomes. *Trends Genet* **17:** 589-596.

Mizuno, T., T. Kaneko, and S. Tabata. 1996. Compilation of all genes encoding bacterial two-component signal transducers in the genome of the cyanobacterium, *Synechocystis* sp. strain PCC 6803. *DNA Res* **3:** 407-414.

Moore, L.R. and S.W. Chisholm. 1999. Photophysiology of the marine cyanobacterium *Prochlorococcus*: Ecotypic differences among cultured isolates. *Limnol Oceanogr* **44:** 628-

638.

Moore, L.R., R. Goericke, and S.W. Chisholm. 1995. Comparative physiology of *Synechococcus* and *Prochlorococcus*: influence of light and temperature on growth, pigments, fluorescence and absorptive properties. *Mar Ecol Prog Ser* **116:** 259-275.

Moore, L.R., G. Rocap, and S.W. Chisholm. 1998. Physiology and molecular phylogeny of coexisting *Prochlorococcus* ecotypes. *Nature* **393:** 464-467.

Moran, N.A. 2002. Microbial minimalism: genome reduction in bacterial pathogens. *Cell* **108:** 583-586.

Moran, N.A. 2003. Tracing the evolution of gene loss in obligate bacterial symbionts. *Curr Opin Microbiol* **6:** 512-518.

Moran, N.A. and A. Mira. 2001. The process of genome shrinkage in the obligate symbiont *Buchnera aphidicola*. *Genome Biol* **2:** RESEARCH0054.

Moreira, D. and H. Philippe. 2000. Molecular phylogeny: pitfalls and progress. *Int Microbiol* **3:** 9-16.

Morel, A., Y.-W. Ahn, F. Partensky, D. Vaulot, and H. Claustre. 1993. *Prochlorococcus* and *Synechococcus* : a comparative study of their size, pigmentation and related optical properties. *J Mar Res* **51:** 617-649.

Mrazek, J., D. Bhaya, A.R. Grossman, and S. Karlin. 2001. Highly expressed and alien genes of the *Synechocystis* genome. *Nucleic Acids Res* **29:** 1590-1601.

Mushegian, A. 1999. The minimal genome concept. *Curr Opin Genet Dev* **9:** 709-714.

Mushegian, A.R. and E.V. Koonin. 1996. A minimal gene set for cellular life derived by comparison of complete bacterial genomes. *Proc Natl Acad Sci U S A* **93:** 10268-10273.

Nakamura, Y., T. Kaneko, S. Sato, M. Ikeuchi, H. Katoh, S. Sasamoto, A. Watanabe, M. Iriguchi, K. Kawashima, T. Kimura. *et al.* 2002. Complete genome structure of the thermophilic cyanobacterium *Thermosynechococcus elongatus* BP-1. *DNA Res* **9:** 123-130.

Nelson, K.E., R.A. Clayton, S.R. Gill, M.L. Gwinn, R.J. Dodson, D.H. Haft, E.K. Hickey, J.D. Peterson, W.C. Nelson, K.A. Ketchum. *et al*. 1999. Evidence for lateral gene transfer between Archaea and bacteria from genome sequence of *Thermotoga maritima*. *Nature* **399:** 323-329.

Novichkov, P.S., M.V. Omelchenko, M.S. Gelfand, A.A. Mironov, Y.I. Wolf, and E.V. Koonin. 2004. Genome-wide molecular clock and horizontal gene transfer in bacterial evolution. *J Bacteriol* **186:** 6575-6585.

Ochman, H., J.G. Lawrence, and E.A. Groisman. 2000. Lateral gene transfer and the nature of bacterial innovation. *Nature* **405:** 299-304.

Ohno, S. 1970. *Evolution by gene duplication*. Springer-Verlag, Berlin-Heidelberg-New york.

Olson, R.J., E.R. Zettler, E.V. Armbrust, and S.W. Chisholm. 1990. Pigment, size and distribution of *Synechococcus* in the North Atlantic and Pacific oceans. *Limnol Oceanogr* **35:** 45-58.

Ong, L.J. and A.N. Glazer. 1991. Phycoerythrins of marine unicellular cyanobacteria. I. Bilin types and locations and energy transfer pathways in *Synechococcus* spp. phycoerythrins. *J Biol Chem* **266:** 9515-9527.

Palenik, B., B. Brahamsha, F.W. Larimer, M. Land, L. Hauser, P. Chain, J. Lamerdin, W. Regala, E.E. Allen, J. McCarren. *et al.* 2003. The genome of a motile marine *Synechococcus*. *Nature* **424:**

1037-1042.

Parkinson, J.S. and E.C. Kofoid. 1992. Communication modules in bacterial signaling proteins. *Annu Rev Genet* **26:** 71-112.

Partensky, F., J. Blanchot, and D. Vaulot. 1999a. Differential distribution and ecology of *Prochlorococcus* and *Synechococcus* in oceanic waters: a review. In *Marine Cyanobacteria* (eds. L. Charpy and A.W.D. Larkum), pp. 457-475. Musée Océanographique, Monaco.

Partensky, F. and L. Garczarek. 2003. The photosynthetic apparatus of chlorophyll *b*- and *d*-containing Oxychlorobacteria. In *Photosynthesis in Algae* (ed. A.W.D. Larkum). Kluwer Academic Publishers, Dordrecht.

Partensky, F., W.R. Hess, and D. Vaulot. 1999b. *Prochlorococcus*, a marine photosynthetic prokaryote of global significance. *Microbiol Mol Biol Rev* **63:** 106-127.

Partensky, F., N. Hoepffner, W.K.W. Li, O. Ulloa, and D. Vaulot. 1993. Photoacclimation of *Prochlorococcus* sp. (Prochlorophyta) strains isolated from the North Atlantic and the Mediterranean Sea. *Plant Physiol* **101:** 295-296.

Ponting, C.P., L. Aravind, J. Schultz, P. Bork, and E.V. Koonin. 1999. Eukaryotic signalling domain homologues in archaea and bacteria. Ancient ancestry and horizontal gene transfer. *J Mol Biol* **289:** 729-745.

Pushker, R., A. Mira, and F. Rodriguez-Valera. 2004. Comparative genomics of gene-family size in closely related bacteria. *Genome Biol* **5:** R27.

Qian, J., N.M. Luscombe, and M. Gerstein. 2001. Protein family and fold occurrence in genomes: power-law behaviour and evolutionary model. *J Mol Biol* **313:** 673-681.

Rappe, M.S. and S.J. Giovannoni. 2003. The uncultured microbial majority. *Annu Rev Microbiol* **57:** 369-394.

Ravanat, J.L., T. Douki, and J. Cadet. 2001. Direct and indirect effects of UV radiation on DNA and its components. *J Photochem Photobiol B* **63:** 88-102.

Raven, J.A. 1994. Why are there no picoplanktonic O2 evolvers with volumes less than 10-19 m3? *J Plank Res* **16:** 565-580.

Rippka, R., T. Coursin, W. Hess, C. Lichtle, D.J. Scanlan, K.A. Palinska, I. Iteman, F. Partensky, J. Houmard, and M. Herdman. 2000. *Prochlorococcus marinus* Chisholm et al. 1992 subsp. pastoris subsp. nov. strain PCC 9511, the first axenic chlorophyll a2/b2-containing cyanobacterium (Oxyphotobacteria). *Int J Syst Evol Microbiol* **50 Pt 5:** 1833-1847.

Rivera, M.C., R. Jain, J.E. Moore, and J.A. Lake. 1998. Genomic evidence for two functionally distinct gene classes. *Proc Natl Acad Sci U S A* **95:** 6239-6244.

Rocap, G., D.L. Distel, J.B. Waterbury, and S.W. Chisholm. 2002. Resolution of *Prochlorococcus* and *Synechococcus* ecotypes by using 16S-23S ribosomal DNA internal transcribed spacer sequences. *Appl Environ Microbiol* **68:** 1180-1191.

Rocap, G., F.W. Larimer, J. Lamerdin, S. Malfatti, P. Chain, N.A. Ahlgren, A. Arellano, M. Coleman, L. Hauser, W.R. Hess. *et al.* 2003. Genome divergence in two *Prochlorococcus* ecotypes reflects oceanic niche differentiation. *Nature* **424:** 1042-1047.

Rubin, G.M. 2001. The draft sequences. Comparing species. *Nature* **409:** 820-821.

Sancar, A. 1994. Structure and function of DNA photolyase. *Biochemistry* **33:** 2-9.

Sanderson, M.J., M.F. Wojciechowski, J.M. Hu, T.S. Khan, and S.G. Brady. 2000. Error, bias, and long-branch attraction in data for two chloroplast photosystem genes in seed plants. *Mol Biol Evol* **17:** 782-797.

Shigenobu, S., H. Watanabe, M. Hattori, Y. Sakaki, and H. Ishikawa. 2000. Genome sequence of the endocellular bacterial symbiont of aphids *Buchnera* sp. APS. *Nature* **407:** 81-86.

Shirai, M., H. Hirakawa, M. Kimoto, M. Tabuchi, F. Kishi, K. Ouchi, T. Shiba, K. Ishii, M. Hattori, S. Kuhara. et al. 2000. Comparison of whole genome sequences of *Chlamydia pneumoniae* J138 from Japan and CWL029 from USA. *Nucleic Acids Res* **28:** 2311-2314.

Sieracki, M.E., E.M. Haugen, and T.L. Cucci. 1995. Overestimation of bacteria in the Sargasso Sea: direct evidence by flow and imaging cytometry. *Deep-Sea Research I* **42:** 1399-1409.

Six, C., J.C. Thomas, B. Brahamsha, Y. Lemoine, and F. Partensky. 2004. Photophysiology of the marine cyanobacterium *Synechococcus* sp. WH8102, a new model organism. *Aqu Microb Ecol* **35:** 17-29.

Snel, B., P. Bork, and M.A. Huynen. 2002. Genomes in flux: the evolution of archaeal and proteobacterial gene content. *Genome Res* **12:** 17-25.

Steglich, C. 2003. Biochemical, spectroscopic and molecular genetic characterisation of novel phycoerythrin species from *Prochlorococcus* sp., PhD Thesis. Universität zu Berlin, Berlin. 112 pp.

Storf, M., A. Parbel, M. Meyer, B. Strohmann, H. Scheer, M.G. Deng, M. Zheng, M. Zhou, and K.H. Zhao. 2001. Chromophore attachment to biliproteins: Specificity of PecE/PecF, a lyase-isomerase for the photoactive 3(1)-Cys-alpha 84-phycoviolobilin chromophore of phycoerythrocyanin. *Biochemistry* **40:** 12444-12456.

Ting, C.S., G. Rocap, J. King, and S.W. Chisholm. 2001. Phycobiliprotein genes of the marine photosynthetic prokaryote *Prochlorococcus*: evidence for rapid evolution of genetic heterogeneity. *Microbiology* **147:** 3171-3182.

Ting, C.S., G. Rocap, J. King, and S.W. Chisholm. 2002. Cyanobacterial photosynthesis in the oceans: the origins and significance of divergent light-harvesting strategies. *Trends Microbiol* **10:** 134-142.

Todo, T., H. Ryo, K. Yamamoto, H. Toh, T. Inui, H. Ayaki, T. Nomura, and M. Ikenaga. 1996. Similarity among the Drosophila (6-4)photolyase, a human photolyase homolog, and the DNA photolyase-blue-light photoreceptor family. *Science* **272:** 109-112.

Urbach, E., D.L. Robertson, and S.W. Chisholm. 1992. Multiple evolutionary origins of prochlorophytes within the cyanobacterial radiation. *Nature* **355:** 267-270.

Urbach, E., D.J. Scanlan, D.L. Distel, J.B. Waterbury, and S.W. Chisholm. 1998. Rapid diversification of marine picophytoplankton with dissimilar light-harvesting structures inferred from sequences of *Prochlorococcus* and *Synechococcus* (Cyanobacteria). *J Mol Evol* **46:** 188-201.

Walsh, M.A., Z. Otwinowski, A. Perrakis, P.M. Anderson, and A. Joachimiak. 2000. Structure of cyanase reveals that a novel dimeric and decameric arrangement of subunits is required for formation of the enzyme active site. *Structure Fold Des* **8:** 505-514.

Waterbury, J.B., S.W. Watson, F.W. Valois, and D.G. Franks. 1986. Biological and ecological characterization of the marine unicellular cyanobacterium *Synechococcus*. In *Photosynthetic Picoplankton* eds T. Platt and W.K.W. Li), pp. 71-120.

Wernegreen, J.J. 2002. Genome evolution in bacterial endosymbionts of insects. *Nat Rev Genet* **3:** 850-861.

West, N.J. and D.J. Scanlan. 1999. Niche-partitioning of *Prochlorococcus* populations in a stratified water column in the eastern North Atlantic Ocean. *Appl Environ Microbiol* **65:** 2585-2591.

West, N.J., W.A. Schonhuber, N.J. Fuller, R.I. Amann, R. Rippka, A.F. Post, and D.J. Scanlan. 2001. Closely related *Prochlorococcus* genotypes show remarkably different depth distributions in two oceanic regions as revealed by in situ hybridization using 16S rRNA-targeted oligonucleotides. *Microbiology* **147:** 1731-1744.

Wilbanks, S.M. and A.N. Glazer. 1993. Rod structure of a phycoerythrin II-containing phycobilisome. I. Organization and sequence of the gene cluster encoding the major phycobiliprotein rod components in the genome of marine *Synechococcus* sp. WH8020. *J Biol Chem* **268:** 1226-1235.

Yamada, Y., T. Fujiwara, T. Sato, N. Igarashi, and N. Tanaka. 2002. The 2.0 A crystal structure of catalase-peroxidase from *Haloarcula marismortui*. *Nat Struct Biol* **9:** 691-695.

RÉSUMÉ

Une grande part de la production primaire océanique est réalisée par deux genres phylogénétiquement très proches de picocyanobactéries (procaryotes oxyphototrophes de taille < 2 µm): *Prochlorococcus* et *Synechococcus*. En plus d'être le plus petit organisme photosynthétique connu à ce jour, *Prochlorococcus* est également le plus abondant. Il est présent essentiellement dans les régions oligotrophes intertropicales alors que *Synechococcus* possède une répartition géographique beaucoup plus large et domine dans les environnements plus riches en sels nutritifs. De plus, *Prochlorococcus* est caractérisé par une distribution verticale plus grande que *Synechococcus*.

L'extraordinaire succès écologique de ces cyanobactéries semble être lié à la mise en place de stratégies différentes leur permettant de s'adapter à des niches écologiques distinctes. Cependant, les bases moléculaires expliquant l'adaptation de ces cyanobactéries à leurs environnements respectifs restent mystérieuses. De manière plus générale, on sait peu de choses des mécanismes évolutifs à l'origine de la diversification au sein de ces deux genres qui possèdent un ancêtre commun relativement récent.

Récemment, les génomes de trois souches de *Prochlorococcus* (MED4, SS120 et MIT9313) et d'une souche de *Synechococcus* (WH8102) ont été entièrement séquencés et annotés. Le génome d'une seconde souche de *Synechococcus* (WH7803) est en cours d'annotation. Parmi ces génomes, ceux de *Prochlorococcus* MED4 et SS120 ont une taille inférieure à 2 Mb. Ils représentent les deux plus petits génomes d'organismes oxyphototrophes identifiés jusqu'ici et contiennent un jeu de gènes quasiment minimal pour une cyanobactérie.

La comparaison des cinq génomes de picocyanobactéries révèle un faible niveau de différenciation des répertoires de gènes. Néanmoins, la présence de certains gènes (HLIPs, photolyases, systèmes à deux composants, transporteurs ABC...) paraît pouvoir être reliée aux caractéristiques des niches écologiques de ces organismes.

Finalement, il apparaît que la petite taille des génomes de MED4 et SS120 est le résultat d'une perte massive de gènes associée à une accélération de l'évolution des protéines. Il s'agit du premier cas d'évolution réductive identifié chez un organisme à cycle de vie complètement libre. Ce processus évolutif pourrait avoir joué un rôle déterminant dans l'adaptation de *Prochlorococcus* au milieu très stable et très oligotrophe au sein duquel celui-ci prospère.

Mots clés : cyanobactéries, *Prochlorococcus*, *Synechococcus*, génome, annotation, adaptation, évolution réductive