HAL
open science

Transcripts' Evolutionary History and Structural Dynamics Give Mechanistic Insights into the Functional Diversity of the JNK Family

Adel Ait-Hamlat, Diego Javier Zea, Antoine Labeeuw, Lélia Polit, Hugues Richard, Elodie Laine

To cite this version:

Adel Ait-Hamlat, Diego Javier Zea, Antoine Labeeuw, Lélia Polit, Hugues Richard, et al.. Transcripts' Evolutionary History and Structural Dynamics Give Mechanistic Insights into the Functional Diversity of the JNK Family. Journal of Molecular Biology, 2020, 432 (7), pp.2121-2140. 10.1016/j.jmb.2020.01.032 . hal-02884145

HAL Id: hal-02884145

https://hal.sorbonne-universite.fr/hal-02884145v1

Submitted on 29 Jun 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Transcripts' evolutionary history and structural dynamics give mechanistic insights into the functional diversity of the JNK family.

Adel Ait-hamlat[1], Diego Javier Zea[1], Antoine Labeeuw[1], Lélia Polit[1], Hugues Richard [1*] and Elodie Laine [1*]

[1] Sorbonne Université, CNRS, IBPS, Laboratoire de Biologie Computationnelle et Quantitative (LCQB), 75005 Paris, France.

[*] corresponding authors: hugues.richard@upmc.fr, elodie.laine@upmc.fr

Abstract

Alternative splicing and alternative initiation/termination transcription sites, have the potential to greatly expand the proteome in eukaryotes by producing several transcript isoforms from the same gene. Although these mechanisms are well described at the genomic level, little is known about their contribution to protein evolution and their impact at the protein structure level. Here, we address both issues by reconstructing the evolutionary history of transcripts and by modeling the tertiary structures of the corresponding protein isoforms. We reconstruct phylogenetic forests relating 60 protein coding transcripts from the c-Jun N-terminal kinase (JNK) family observed in 7 species. We identify two alternative splicing events of ancient origin and show that they induce subtle changes on the protein's structural dynamics. We highlight a previously uncharacterized transcript whose predicted structure seems stable in solution. We further demonstrate that orphan transcripts, for which no phylogeny could be reconstructed, display peculiar sequence and structural properties. Our approach is implemented in PhyloSofS

(Phylogenies of Splicing Isoforms Structures), a fully automated computational tool freely available at `https://github.com/PhyloSofS-Team/PhyloSofS`.

Keywords:

Alternative splicing, Molecular modeling, Evolution, Transcript phylogeny, Kinase

Abbreviations [1]

Introduction

Alternative splicing (AS) of pre-mRNA transcripts and alternative transcription initiation/termination are essential eukaryotic regulatory processes. They can impact the regulation of gene expression, for instance by introducing changes in the three prime untranslated region[61]. Or they can directly modify the content of the coding sequence (CDS)[26], leading to different protein isoforms. Virtually all multi-exons genes in vertebrates are subject to AS[68] and about 25% of the AS events (ASEs) common to human and mouse are conserved in other vertebrates[3;49;51]. This suggests an important role for AS in expanding the protein repertoire through evolution. AS has also gained interest for medicinal purpose, as the ratio of alternatively spliced isoforms is imbalanced in several cancers[43;70].

The extent to which the ASEs detected at the gene level actually result in functional protein isoforms in the cell remains largely unknown. Transcriptomics and proteomics studies suggested that most highly expressed human genes have only one single dominant isoform[22;25], but the detection rate of these experiments is very difficult to assess[37] and likely suffer from strong experimental detection bias[69]. A recent analysis of ribosome profiling data suggested that a major fraction of splice variants is translated, with direct implications on specific cellular functions[72]. Moreover, a large scale assessment of isoforms present in the cell revealed that the majority of isoform pairs share less than 50% of their interactions[74]. From a structural perspective, very few alternatively spliced isoforms have been characterized and are available in the Protein Data Bank (PDB)[7;27]. It was

[1] AS: alternative splicing, ASE: alternative splicing event, JNK: c-jun N-terminal kinase

shown that the boundaries of single constitutive exons or of co-occurring exon pairs tend to overlap those of compact structural units, called protein units[24]. Moreover, tissue-specific alternatively spliced exons are enriched in disordered regions containing binding motifs[13]. It was also suggested that splicing events may induce major fold changes[9;10], and a few cases of isoforms displaying domain atrophy while retaining some activity have been reported[54].

The elusiveness of the significance of AS for protein function and fold diversification though evolution has stimulated the development of knowledge bases, such as APPRIS[57] and Exon Ontology[65]. They provide functional and sequence-based information at the level of the transcript or the exon. A method reconstructing transcripts' phylogenies was also proposed and proved useful for enhancing transcriptome reconstruction from ESTs and investigating proteins functional features (domains, sites)[16;17].

In this work, we combine sequence- and structure-based information to shed led on the evolution of AS. We have developed PhyloSofS (Phylogenies of Splicing Isoforms Structures), an automated tool that infers plausible evolutionary scenarios explaining an ensemble of protein coding transcripts observed in a set of species and predicts the tertiary structures of the protein isoforms. We show how PhyloSofS can be used to identify and date ASEs, and also shed light on the molecular mechanisms underlying their functional outcome in the c-Jun N-terminal kinase (JNK) family. This choice was motivated by the fact that JNKs are among the few tens of families for which alternative transcripts performing different functional tasks have been experimentally identified and characterized[8;36;64]. Moreover, they play essential regulatory roles by targeting specific transcription factors (c-Jun, ATF2...) in response to cellular stimuli. The deregulation of their activity is associated with various diseases (cancer, inflammatory diseases, neuronal disorder...) which makes them important therapeutic targets[46]. About ten JNK splicing isoforms have been documented in the literature[39]. They were shown to perform different context-specific tasks[12;30;32;66] and to have different affinities for their substrates[11;67]. By reconstructing the phylogeny of JNK transcripts across seven species, we identify two ASEs of ancient origin. We further identify key residues that may be responsible for

the selective recognition of JNK substrates by different isoforms and characterize the behaviour of these isoforms in solution by biomolecular simulations. One of the ASEs involves a 80-residue deletion and has never been documented before. We find that its predicted structure is stable in solution. Both ASEs are supported by sequencing evidence from transcriptomics studies.

Our work allows to put together, for the first time, two types of information, one coming from the reconstructed phylogeny of transcripts and the other from the structural modeling of the produced isoforms, and this to shed light on the molecular mechanisms underlying the evolution of protein function. It goes beyond simple conservation analysis, by dating the appearance of ASEs in evolution, and beyond general structural considerations regarding AS, by characterizing in details the isoforms' shapes and motions. We find that the effect of functional ASEs on the structural dynamics of the isoforms may be subtle and require such a detailed investigation. Our results also open the way to the identification and characterization of new isoforms that may be targeted in the future for medicinal purpose.

Computational method

PhyloSofS can be applied to single genes or to gene families. Given a gene tree and the observed protein coding transcripts at the leaves (**Fig. 1a**, on the left), it reconstructs a phylogenetic forest embedded in the gene tree (**Fig. 1a**, on the right) representing plausible evolutionary scenarios explaining the transcripts. Each tree in the forest (in orange, green or purple) represents the phylogeny of one transcript. In other words, each root indicates the appearance of a new transcript and its corresponding ASE(s). Transcript losses are possible (triangles in **Fig. 1a**), and the exon usage of a transcript can change along the branches upon the inclusion or exclusion of one or several exons (which we refer to as "mutations"). The underlying evolutionary model is comprised of two levels, following[16]. At the level of the gene, exons can be absent, constitutive, or alternative (*i.e.* involved in at least one ASE), whereas at the level of the transcript, exons are either present or absent. The cost associated to the mutation of an exon naturally

depends on its impact on the status of the exon at the gene level. For instance when the gain of an exon at the gene level shifts its status from absent to constitutive, the mutation will not be penalized (see *Methods* and **Table II**).

PhyloSofS algorithm seeks to determine the scenario with the smallest number of evolutionary events, following the maximum parsimony principle. It is inspired from that reported in[16]. Our main contribution was to develop heuristics in order to treat complex cases in a computationally tractable way. Specifically, we have implemented a multi-start iterative strategy combined with a systematic local exploration around the best current solution to efficiently search the space of phylogenetic forests (see *Methods* and *Supplementary Fig. S1*). Moreover, we have designed a branch-and-bound algorithm adapted to the problem of assigning transcripts between parent and child nodes (see *Methods* and *Supplementary Text S1*). The reconstructed forests are provided with a user-friendly visualization (**Fig. 1b-c**). In addition to phylogenetic reconstruction, PhyloSofS predicts the 3D structures of the protein isoforms. The predictions are performed based on comparative modeling using the HH-suite[29]. Furthermore, PhyloSofS annotates the generated models with sequence (exon boundaries) and structure (secondary structure, solvent accessibility, model quality) information. For example, it is very easy to visualize the location of each exon on the modeled structure. Here, we present the application of PhyloSofS to the c-Jun N-terminal kinase (JNK) family across 7 species (*H. sapiens*, *M. musculus*, *X. tropicalis*, *T. rubripes*, *D. rerio*, *D. melanogaster* and *C. elegans*). This case represents a high degree of complexity with 60 observed transcripts assembled from a total of 19 different exons. Most of these transcripts comprise more than 10 exons and the number of transcripts per gene per species varies from 1 to 8 (**Fig. 1b-c**).

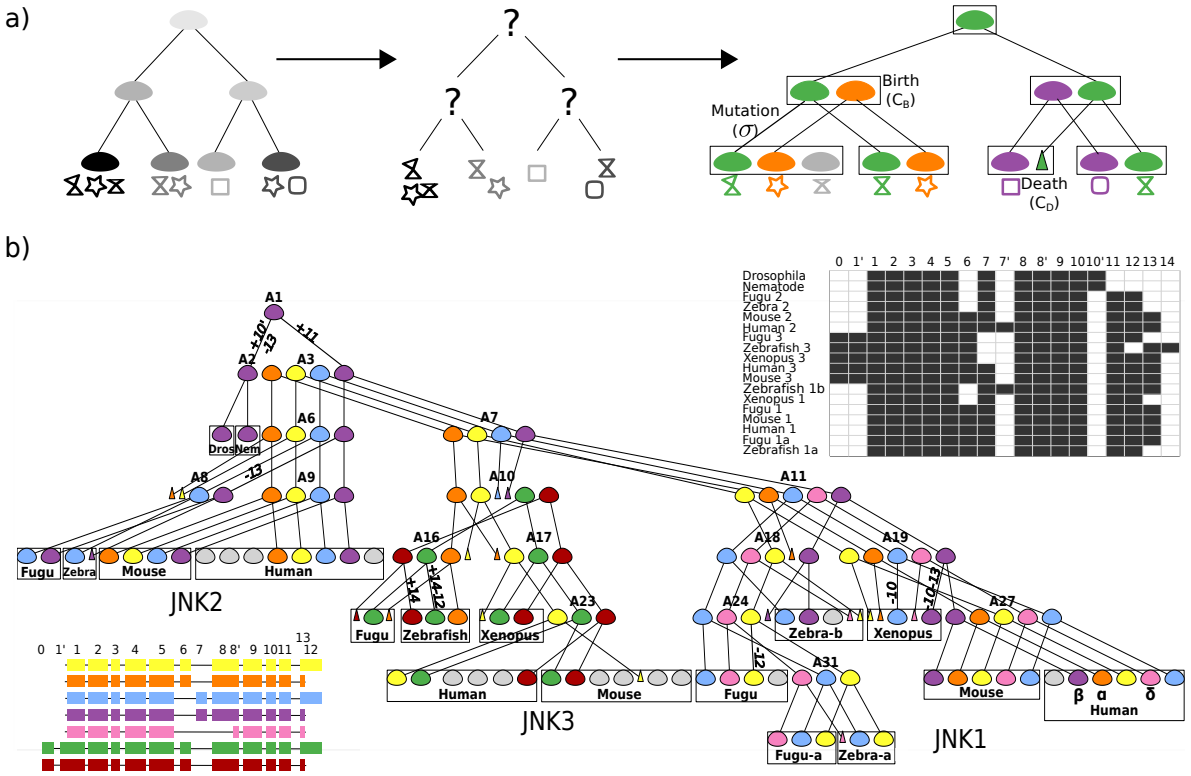

Figure 1: **Transcripts' phylogenies reconstructed by PhyloSofS. (a)** On the left, example of a phylogenetic gene tree where 8 transcripts (represented by geometrical symbols) are observed in 4 current species (leaves of the tree, colored in different grey tones). These data are given as input to PhyloSofS. In the middle, the problem addressed by PhyloSofS is that of a partial assignment: how to pair transcripts so as to maximize their similarity? On the right, example of a solution determined by PhyloSofS. The transcripts' phylogeny is a forest comprised of 3 trees (colored differently). The nodes of the input gene tree are subdivided into subnodes corresponding to observed (current) or reconstructed (ancestral) transcripts. The root of a tree stands for the creation of a new transcript and is associated to a cost C_B. Triangles indicate transcript deaths and are associated to a cost C_D. Mutation events occur along branches and are associated to a cost σ. The grey node corresponds to an orphan transcript for which no phylogeny could be reconstructed. **(b)** Transcripts' phylogeny reconstructed by PhyloSofS for the JNK family. The forest is comprised of 7 trees, 19 deaths (triangles) and 14 orphan transcripts (in grey). Mutation events are indicated on branches by the symbol + or - followed by the number of the exon being included or excluded (*e.g. +11*). The cost of the phylogeny is 69 (with $C_B = 3$, $C_D = 0$ and $\sigma = 2$). On the top right corner are displayed the exons present in each current species (in black). On the bottom left corner are displayed the exon compositions of the human isoforms for which a phylogeny could be reconstructed.

Results

Transcripts' phylogeny for the JNK family.

The observed transcripts were collected from the Ensembl[75] database (see *Methods*). PhyloSofS algorithm was run for 10^6 iterations on the JNK family gene tree and we retained the most parsimonious evolutionary scenario (cost = 69, see *Methods* for a detailed description of the parameters). The reconstructed forest is comprised of 7 transcript trees (**Fig. 1b**, each tree is colored differently). Each transcript is described as a collection of exons, numbered from *0* to *14* (**Fig. 1b**, top right and bottom left corners, and see *Methods* for more details on the numbering). We could reconstruct a phylogeny for 46 out of the 60 observed transcripts. The 14 "orphan" transcripts (leaves in grey) are not conserved across the studied species, and thus likely result in non-functional protein products. Mutations occurring along the branches of the trees are labelled (**Fig. 1b**, see +/− symbols followed by the number of the included/excluded exon). In total, JNK transcripts' phylogeny comprises 11 mutations.

The sequences of the JNK genes are highly conserved through evolution (**Table I**). While *Drosophila melanogaster* and nematode are the most distant species to human, their unique JNK genes share as much as 78% and 56% sequence identity, respectively, with human JNK1 (**Table I**). The sequence identities with human JNK2 and JNK3 are slightly lower (**Table I**, in grey). This suggests that the most recent common ancestor of the 7 studied species contained one copy of an ancestral JNK1 gene. Under this assumption, we propose an evolutionary scenario to reconcile the JNK family gene tree (*Supplementary Fig. S2a*) and the species tree (*Supplementary Fig. S2b*). In this scenario, early duplication events led to the creation of JNK2 and JNK3 in the ancestor common to mammals, amphibians and fishes (*Supplementary Fig. S2b*). JNK1 was then further duplicated in fishes while JNK2 was lost in *Xenopus tropicalis*. One can see that there is no conflict between the gene and species trees under these hypotheses (*Supplementary Fig. S2a-b*).

The 7 reconstructed trees relate 12 transcripts observed in human across the three

Table I: **Percentages of sequence identity of JNK genes to human.**

	JNK1	JNK2	JNK3
Mouse	99	97	100
Xenopus tropicalis	89	-	98
Fugu	79 82 (a)	81	96
Zebrafish	87 (a) 87 (b)	85	93
Drosophila melanogaster	78	73	77
Nematode	56	54	56

Each gene of each species was aligned to its orthologous gene in human. Human and mouse genomes contain 3 paralogues: JNK1, JNK2 and JNK3. *Xenopus tropicalis* contains only JNK1 and JNK3. The fishes contain 4 paralogues: JNK1, JNK1a, JNK2 and JNK3 in fugu, JNK1a, JNK1b, JNK2 and JNK3 in zebrafish. *Drosophila melanogaster* and nematode contain only one gene, whose sequence identities with human JNK1, JNK2 and JNK3 are displayed in black, grey and grey, respectively. In addition to the values reported in the table, here are some sequence identities computed between paralogues: (*i*) 83% between human JNK1 and JNK2, and between human JNK1 and JNK3; (*ii*) 86% between fugu JNK1 and JNK1a; (*iii*) 92% between zebrafish JNK1a and JNK1b.

genes (**Fig. 1b**). The transcripts of the same color belong to the same tree and share the same exon composition, even if they come from different gene loci and hence have different amino acid sequences. For instance, the transcript structure including exons *6*, *8* and *12* and excluding exons *0*, *1'*, *7* and *13* (in yellow) is shared by 3 human transcripts present in JNK1, JNK2 and JNK3 (*Supplementary Fig. S2c*). Note that this may not be the case in general, for any protein family: the leaves of a tree may have different exon compositions if mutations occur along the branches.

Two pairs of exons, namely *6-7* and *12-13*, are mutually exclusive (*Supplementary Fig. S2c*). The associated ASEs can be dated early in the phylogeny (**Fig. 1b**), before the gene duplication (*Supplementary Fig. S2b*). Neither *Drosophila melanogaster* nor nematode contain any of exons *12-13*. Hence, it is equivalent to consider that exon *12* or exon *13* appeared first (compare **Fig. 1b** and *Supplementary Fig. S3*). By contrast, exon *7* is clearly predicted as appearing before exon *6* (**Fig. 1b**, compare purple tree with yellow and orange trees). Noticeably, The two transcripts expressing exons *6* (in orange and yellow) are consistently absent from *Zebrafish* JNK1b and *Xenopus tropicalis* JNK1. Although this can correspond to a real loss of transcripts in those species, a more

parsimonious explanation would be that the gene annotation in the Ensembl database is incomplete. We searched for direct experimental evidence of the expression of the transcripts in these two organisms using transcriptome sequencing data from hundreds of RNA-Seq libraries (see *Methods*). In *Xenopus tropicalis*, the analysis of exon-exon junctions revealed the expression of transcripts containing a 72bp-long exon with a translated sequence very similar to that of exon *6* in other species. The sequencing support for this exon is strong as it is present in more than two third of the *Xenopus tropicalis* RNA-seq libraries we studied. Additionally, a transcript containing this exon is predicted in Refseq (Refseq ID: XM_012966153.2) and the corresponding genomic region is strongly conserved. In *D. rerio*, the analysis of exon-exon junctions in JNK1b also identified one new 72bp-long exon with a translated sequence very similar to exon *6* in other species. Hence, there is significant evidence of the expression of transcripts containing exon *6* in both *X. tropicalis* JNK1 and *D. rerio* JNK1b, although they are not annotated in Ensembl. This observation gives support to our choice of not penalizing death ($C_D = 0$) when we reconstruct the transcripts' phylogeny as a way to account for the incomplete transcript data.

Among the three transcripts appearing after the gene duplication events (**Fig. 1b**, in pink, green, and red), one transcript features a large deletion encompassing exons *6*, *7* and *8* (JNK1 sub-forest, internal node A11, in pink). Its exon composition is perfectly conserved along the phylogeny (no mutation). We looked for additional RNA-seq support for this transcript and found evidence in 3 Human RNA-seq libraries (out of 166) for reads aligning to the exon-exon junction between exon 7 and 8'. There was no evidence in the RNA-Seq mouse libraries. The two other transcripts are created at the root of the JNK3 sub-forest (ancestor node 10, in green and red). They are characterized by the presence of exons *0* and *1'*, not found in the other paralogues. Another characteristic feature can be observed for the JNK3 gene, namely exon *7* is completely absent from the associated sub-forest. The genomic sequence of exon *7* is present at the JNK3 locus in all species, but it diverged far more in this gene compared to JNK1 and JNK2.

Mapping of the gene 1D structure onto the protein 3D structure.

About eighty structures of human JNKs are available in the PDB (*Supplementary Table S1*). This abundance of structural data can be explained by the fact that JNKs are important therapeutic targets and they were crystallized with different inhibitors. The three paralogues share the same fold, which is highly conserved among protein kinases. The structures are highly redundant, with an average root mean square deviation (RMSD) of 1.96 ± 0.71 Å, computed over more than 80% of the protein residues. In order to visualize the correspondence between the gene structure and the protein secondary and tertiary structures, the exons were mapped onto a high-resolution PDB structure (3ELJ[15]) of human JNK1 (**Fig. 2**, each exon is colored differently). One can observe that the organization of the protein 3D structure is preserved by the 1D structure of the gene. Most of the secondary structures (10 over 12 α-helices and 7 over 9 β-strands) are completely included in single exons. Moreover, each one of the regions important for the structural stability and/or function of protein kinases (**Fig. 2**, labelled in black) is included in one single exon (see also *Supplementary Table S2*). So are the N-terminal hairpin and the MAPK insert (labelled in grey), two structural motifs specific to the mitogen-activated protein kinase (MAPK) type, to which the JNKs belong. By contrast, binding sites for cofactors and substrates (green circles, see also *Supplementary Table S2*) are comprised of residues belonging to different exons. This is expected as binding sites are comprised of segments that can be very far from each other along the protein sequence. Of note, the block formed by exons *1* to *5*, comprising the N-terminal lobe and the A(ctivation)-loop (**Fig. 2**, from blue to white), is constitutively present in all transcripts belonging to the colored trees on **Fig. 1b**. The correspondence was also analyzed for the JNK protein from *Drosophila melanogaster*, whose 3D structure is very similar to that of human JNK1 (*Supplementary Fig. S4*, RMSD of 0.68 Å).The JNK gene from the *Drosophila melanogaster* genome comprises much fewer exons than the human gene. The match between the borders of these exons and the borders of the secondary structures and known important regions is even better in that species. Considering the high degree of conservation of JNK sequences, one may hypothesize that a good match also exists in all studied

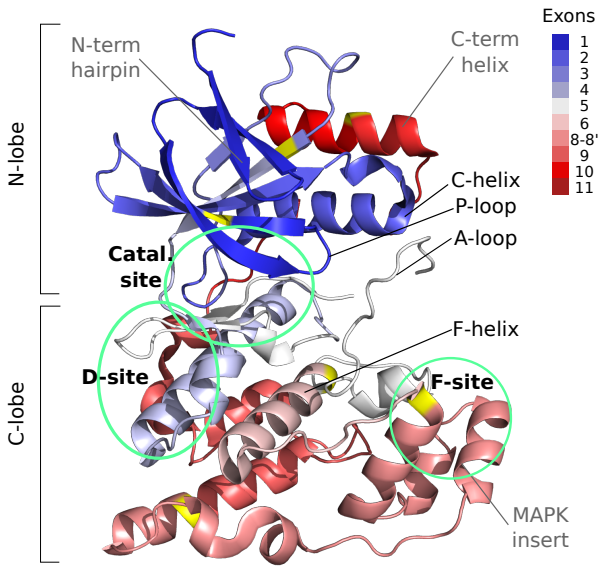

Figure 2: **Exons mapped onto the tertiary structure of human JNK1.** The protein (residues 7 to 364) is represented as a cartoon and the different exons are colored from blue through white to red. The residues in yellow are at the junction of 2 exons. It should be noted that exons *8* and *8'* used in PhyloSofS actually correspond to only one genomic exon (see *Methods*). The regions labelled in black are common to kinases and were reported in the literature (see[33]) for playing important roles in their structural stability and/or function. The regions labelled in grey are specific to MAP kinases. The green circles indicate the catalytic site and binding sites for JNK cellular partners[28;44]. The structure was solved by X-ray crystallography at 1.80 Å resolution (PDB code: 3ELJ[15]).

species. Our observation is in agreement with a previous study establishing a relationship between exon boundaries and structurally consistent protein regions[24].

Properties of the orphan transcripts.

We investigated whether the orphan transcripts, for which no phylogeny could be reconstructed (**Fig. 1b**, grey leaves), displayed peculiar sequence and structural properties compared to the "parented" transcripts (**Fig. 1b**, colored leaves). Our assumption is that an orphan transcript is less likely to have functional importance. First, the orphan transcripts are significantly smaller than the parented ones (**Fig. 3a**). While the minimum length for parented transcripts is 308 residues, with an average of 406 ± 40 residues (**Fig. 3a**, in white), the orphan transcripts can be as small as 124 residues, with an average of 280 ± 88 residues (**Fig. 3a**, in grey). Second, regarding secondary structure

content, both types of transcripts contain about 40% of residues predicted in α-helices or β-sheets (**Fig. 3b**). Third, the 3D models generated by PhyloSofS molecular modeling routine for the orphan transcript isoforms are of poorer quality than those for the transcripts belonging to a phylogeny (**Fig. 3c-d**). The quality of the models was assessed by computing Procheck[42] G-factor and Modeller[47] normalized DOPE score (**Fig. 3c-d**). A model resembling experimental structures deposited in the PDB should have a G-factor greater than -0.5 (the higher the better) and a normalized DOPE score lower than -1 (the lower the better). The distributions obtained for the parented isoforms are clearly shifted toward better values and are more narrow than those for the orphan transcripts. Finally, the proportion of protein residues being exposed to the solvent (relative accessible surface area $rsa > 25\%$) is significantly higher for the orphan isoforms (**Fig. 3e**), as is the proportion of hydrophobic residues being exposed to the solvent (**Fig. 3f**). Overall, these observations suggest that simple sequence and structure descriptors enable to distinguish the orphan transcripts from the ones within a phylogeny and that the formers display properties likely reflecting structural instability (large truncations, poorer quality, larger and more hydrophobic surfaces).

Subtle changes in the protein's internal dynamics linked to substrate differential affinity.

The two mutually exclusive exons *6* and *7* are particularly important for JNK cellular functions, as they confer substrate specificity. The inclusion or exclusion of one or the other results in different substrate-binding affinities[11;67]. From a sequence perspective, the two exons are homologous, highly conserved through evolution, and differ only by a few positions (*Supplementary Fig. S5*). From a structural perspective, they both fold into an α-helix, known as the F-helix, followed by a loop (**Fig. 2**, in light pink). The F-helix was shown to play a central role in the structural stability and catalytic activity of protein kinases[38;53]. It serves as an anchor for two clusters of hydrophobic residues, namely the catalytic and regulatory spines (see illustration on the PKA kinase on *Supplementary Fig. S6a*), and for the HDR motif of the catalytic loop (see illustration on the CDK-substrate

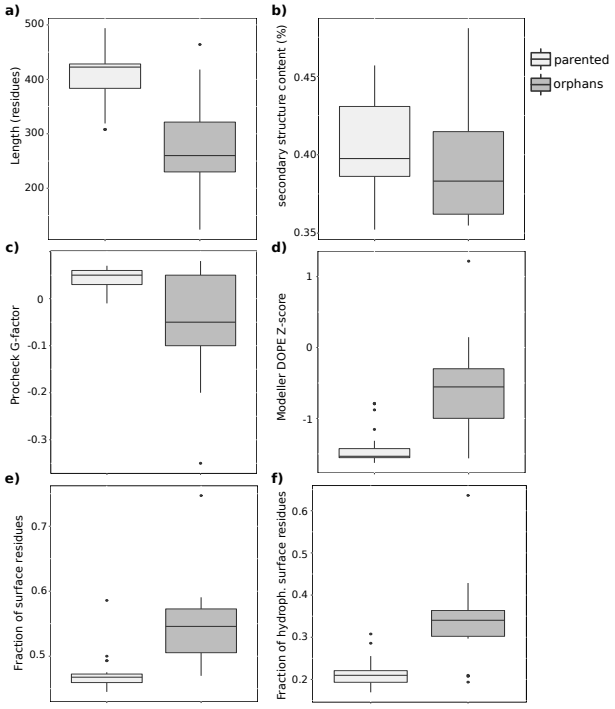

Figure 3: **Structural features of the transcript isoforms.** Distributions are reported for the parented transcripts (in light gray) and the orphan transcripts (in dark grey) in the transcripts' phylogeny (see **Fig. 1b**). **(a)** Length of the transcript (in residues). **(b)** Predicted secondary structure content (in percentages of residues). **(c)** Overall G-factor computed by Procheck[42]. **(d)** Normalized DOPE score computed by Modeller[47]. **(e)** Fraction of protein residues being exposed to the solvent ($rsa > 0.25$). **(f)** Fraction of hydrophobic protein residues being exposed to the solvent ($rsa > 0.25$).

complex on *Supplementary Fig. S7a*). In the following, we will use these known structural features as proxies for the stability and catalytic competence of the studied isoforms.

The available JNK crystallographic structures and the 3D models generated by PhyloSofS do not display any significant structural change upon exchanging exons *6* and *7*. The catalytic and regulatory spines, together with their anchors in the F-helix, are present in both types of isoforms (*Supplementary Fig. S6b-c*). The N-terminal aspartate (D207) of the F-helix, which serves as an anchor for the spines, is 100% conserved in both exons *6* and *7* in the 7 studied species (*Supplementary Fig. S5*, indicated by an arrow). The two other anchor points are also present, namely I214 and L/M218 (*Supplementary Fig. S5*, indicated by arrows). Moreover, the characteristic H-bond pattern with the HRD motif and the associated strained backbone conformation are also observed in both types of isoforms (*Supplementary Fig. S7b-c*). Consequently, both exons *6* and *7*, and

thus the isoforms containing them, possess the structural features known to be important for kinase catalytic activity and/or regulation.

To further investigate the potential impact of the inclusion/exclusion of exon *6* or *7* on the dynamical behavior of the protein, we performed all-atom molecular dynamics (MD) simulations of the human isoforms colored in orange and purple on **Figure 1b**. We shall refer to these isoforms as JNK1α (with exon *6*) and JNK1β (with exon *7*), in agreement with the nomenclature found in the literature[67]. JNK1α and JNK1β were simulated in explicit solvent for 250 ns (5 replicates of 50 ns, see *Methods*). The backbone atomic fluctuation profiles of the two isoforms are very similar (**Fig. 4a**, orange and purple curves), except for the A-loop which is significantly more flexible in JNK1α: the region from residue 176 to 188 displays averaged Cα fluctuations of 1.55 $\pm$ 0.28 Å in JNK1α and of 0.98 $\pm$ 0.16 Å in JNK1β (**Fig. 4a**). We should stress that this loop displays the highest deviations among the JNK structures available in the PDB and often comprises unresolved residues. The two exons, *6* and *7*, have similar backbone flexibility. In the F-helix, the anchor residues for the spines, D207, I214 and M218 adopt stable and very similar conformations (**Fig. 4b**). Moreover, the HRD backbone strain and the associated H-bond pattern are maintained along the simulations of both systems (*Supplementary Fig. S8a-b*). Consequently, the observations realized on the static 3D models hold true when simulating their dynamical behavior: the *6*/*7* variation does not induce any drastic change on the protein's overall shape and behaviour.

Nevertheless, we observe differences in the side-chain flexibilities of a few residues lying in the loop following the F-helix between the two isoforms (**Fig. 4b**). On the one hand, in exon *6* (in orange), the polar and positively charged residues H221, K222 and R228 are exposed to the solvent and display large amplitude side-chain motions. These amino acids are 100% conserved in exon *6* across all species (*Supplementary Fig. S5*). On the other hand, in exon *7* (**Fig. 4b**, in purple), G221, G222 and T228 have small side chains with much reduced motions. While G221 is conserved across all species, position 222 is variable and position 228 features G, T or S (*Supplementary Fig. S5*). This region of the protein is involved in the binding of substrates (see **Fig. 2**, F-site). Moreover, in

both isoforms, we predicted residues 223-230 as directly interacting with cellular partners (see *Methods*). Consequently, one may hypothesize that the differences highlighted here may be crucial for substrate molecular recognition specificity. The positive charges, high fluctuations, high solvent accessibility and high conservation of residues H221, K222 and R228 in JNK1α support a determinant role for these residues in selectively recognizing specific substrates.

Structural dynamics of a newly identified isoform.

Our reconstruction of the JNK transcripts' phylogeny highlighted a JNK1 isoform (**Figure 1b**, in pink) that has not been documented in the literature so far. It is expressed in human, mouse and fugu fish (**Figure 1b**), suggesting that it could play a functional role in the cell. To investigate this hypothesis, we analyzed the 3D structure and dynamical behavior of this isoform in human. We refer to it as JNK1δ.

JNK1δ displays a large deletion (of about 80 residues), lacking exons *6*, *7* and *8*. It does not contain the F-helix, shown to be crucial for kinases structural stability[38], nor the MAPK insert, involved in the binding of the phosphatase MKP7[44] (**Fig. 2**). The 3D model generated by PhyloSofS superimposes well to those of JNK1α and JNK1β, with a RMSD lower than 0.5 Å on 245 residues. This is somewhat expected as we use homology modeling. Nevertheless, cases were reported in the literature where homology modeling detected big changes in protein structures induced by exon skipping[52]. In the model of JNK1δ, the F-helix present in JNK1α and JNK1β (residues 207 to 220) is replaced by a loop (residues 282 to 288) corresponding to exon *8'* (**Fig. 4c**, indicated by the two stars). The sequence of this loop (exon *8'*) does not share any significant identity with the F-helix (N-terminal parts of exons *6* and *7*), except for the N-terminal residue which is an aspartate, namely D282 (D207 in JNK1α and JNK1β). This replacement results in the regulatory spine being intact in JNK1δ (*Supplementary Fig. S6d*, in red). Moreover, the HRD motif's strained backbone conformation and the associated H-bond pattern, which are stabilized by the aspartate, are maintained (*Supplementary Fig. S7d*). By contrast, the catalytic spine lacks its two anchors (*Supplementary Fig. S6d*, in yellow).

JNK1δ was simulated in explicit solvent for 250 ns (5 replicates of 50 ns). The isoform displays stable secondary structures (*Supplementary Fig. S9*, at the bottom) and atomic fluctuations comparable to those of JNK1α and JNK1β (**Fig. 4a**, pink curve to be compared with the purple and orange curves). The Cα atomic fluctuations averaged over the loop replacing the F-helix are of 0.88 $\pm$ 0.18 Å. This is higher than the values computed for the F-helix in JNK1α and JNK1β (0.57 $\pm$ 0.10 Å and 0.53 $\pm$ 0.09 Å), but it still indicates a limited flexibility. Moreover, the N-terminal aspartate D282 establishes stable H-bonds with the HRD motif along all but one of the replicates (*Supplementary Fig. S8a*, on the right) and the HRD motif's backbone remains in a strained conformation (*Supplementary Fig. S8b*, on the right), as was observed for JNK1α and JNK1β. Consequently, JNK1δ seems stable in solution, and, as observed on the static 3D model, the absence of the F-helix in this isoform is partially compensated by the presence of D282, which is sufficient to maintain H-bonds with the HRD motif and a resulting backbone strain of the motif, important for kinase structural stability.

The main difference between JNK1δ and the two other isoforms lies in the amplitude of the motions of the A-loop. In JNK1δ, the C-terminal part of the A-loop can detach from the rest of the protein along the simulations (**Fig. 4c**). The amplitude of the angle computed between the most retracted conformation (in grey) and the most extended one (in black) is 107°. By contrast, in JNK1α and JNK1β, the A-loop always stays close to the rest of the protein, with amplitude angles of 18° and 19°, respectively. The A-loop contains two residues, T183 and Y185 (**Fig. 4c**, highlighted in sticks), whose phosphorylation is required for JNK activation. We hypothesize that the large amplitude motion in JNK1δ might favor their accessibility and, in turn, the activation of the protein.

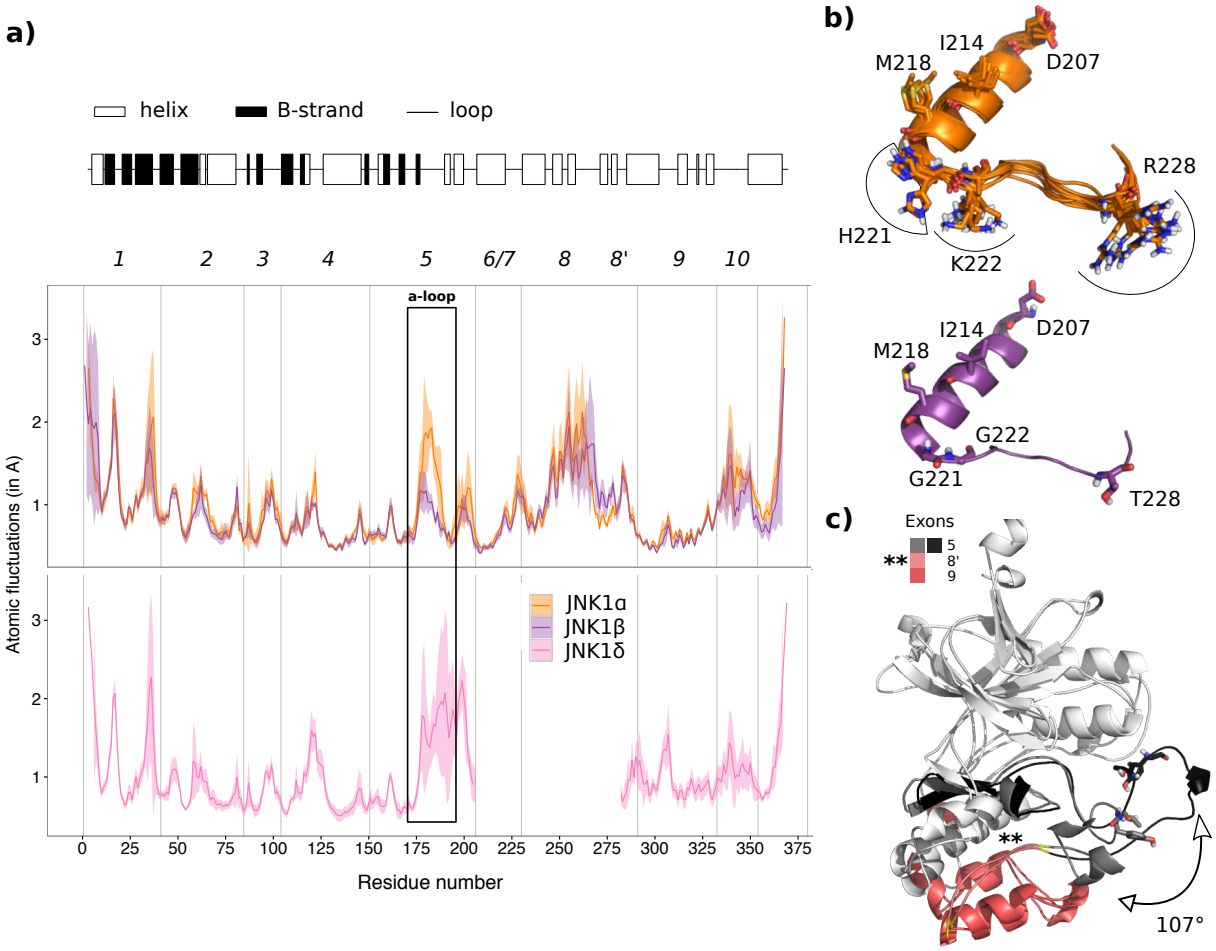

Figure 4: **Dynamical behavior of the human JNK1 isoforms in solution. (a)** The secondary structures for JNK1α (with exon *6*) are depicted on top (the profiles for the 2 other isoforms are very similar, see *Supplementary Fig. S9*). The atomic fluctuations (computed on the Cα) averaged over 5 50-ns MD replicates are reported for JNK1α in orange, JNK1β in purple and JNK1δ in pink. The envelopes around the curves indicate the standard deviation. **(b)** Representative MD conformations obtained by clustering based on position 228 (RMSD cutoff of 1.5 Å). There are 8 conformations for JNK1α (in orange) and only 1 for JNK1β (in purple). **(c)** Superimposed pair of MD conformations illustrating the amplitude of the A-loop motion in JNK1δ (see *Materials ad Methods* for details on the calculation of the angle). Exons *5*, *8'* and *9* are indicated by colors and labels. For clarity, *8'* is also indicated by two stars on the structure.

Unresolved residues in the 3D models.

In the 3D models generated by PhyloSofS, the N-terminal exons *0* and *1'* and the C-terminal exons *12* and *13* are systematically missing. This is due to the lack of structural templates for these regions. Using a threading approach instead of PhyloSofS homology modeling routine (see *Methods*) did not enable to improve their reconstruction. In fact, the models generated by the threading algorithm are very similar to those generated by PhyloSofS.

At the C-terminus, exons *12* and *13* are completely predicted as intrinsically disordered (*Supplementary Fig. S12a* and *Supplementary Fig. S12b*, blue curve). At the N-terminus, exons *0* and *1'* contain two segments of about 10 residues predicted as disordered protein-binding regions (*Supplementary Fig. S12b*, orange curve), *i.e* regions unable to form enough favorable intra-chain interactions to fold on their own and likely stabilized upon interaction with a globular protein partner[50]. These exons are present in only two JNK3 transcript isoforms (**Fig. 1b**, colored in dark red and green).

Alternative phylogenies and robustness to parameter changes

PhyloSofS phylogenetic reconstruction's algorithm may find several solutions with equivalent costs, depending on the input data and the set of parameters. The forest described above (**Fig. 1b**, and *Supplementary Fig. S3* with branch swapping), comprising 7 trees, 19 deaths and 14 orphans, was visited 1 219 times over 10^6 iterations of the program. An alternative phylogeny was visited 310 times, that comprises the same number of trees and orphans, but 2 more deaths (*Supplementary Fig. S10*). The difference between the two forests lies among the fugu JNK1 transcripts, where one transcript belongs to the orange tree (*Supplementary Fig. S10*) instead of the yellow one (**Fig. 1b**). The two trees differ by the inclusion or exclusion of exon *12* or *13*, and the re-assigned transcript lacks both exons. Consequently, the new branching results in the loss of exon *13* between the internal nodes A11 and A18 (*Supplementary Fig. S10*), instead of the loss of exon *12* between A24 and fugu JNK1 (**Fig. 1b**). Another forest with the same cost comprising 8 trees, 23 deaths and 13 orphans was visited 190 times (*Supplementary Fig. S11*). The additional

tree is created in the internal node A10 and links two observed JNK3 transcripts: one from the mouse that was previously orphan (**Fig. 1b**) and one from zebrafish that previously belonged to the green tree. These two transcripts are very similar to the green and dark red transcripts, in terms of exon composition and of structural properties. The only difference is that they lack exons *12* and *13*. Consequently, this new branching avoids the loss of exon *12* between A16 and zebrafish JNK3. Overall the differences between the three solutions are minor and these ambiguities do not impact our interpretation of the results.

To further assess the robustness of our results, we ran PhyloSofS algorithm with different parameters and analyzed the output phylogenies. The three main parameters of the algorithm are the costs C_B, C_D and σ, associated to the creation (birth), the loss (death) and the mutation of a transcript, respectively. The forests described above were obtained by setting the death cost to zero ($C_D = 0$) and the ratio between the birth and mutation costs below 2 ($C_B/\sigma = 3/2 = 1.5$). The choice of not penalizing death was motivated by the fact that the transcriptome data and annotations we are working with may be incomplete. Indeed, the different genomes available in Ensembl are not annotated with the same accuracy. The choice of tolerating few mutations within each tree was motivated by the fact that several pairs of transcripts from the same species differ by only two exons (**Fig. 1b**, compare exon compositions in the bottom left corner). Varying slightly the birth-to-mutation ratio while maintaining the death cost to zero did not impact the results (**Fig. 5** and *Supplementary Table S3*, compare combinations 5.0.3 and 5.0.4 with 3.0.2). A bigger parameter change, resulting in a birth-to-mutation ratio larger than 2, had a minor impact on the results (**Fig. 5**, combination 5.0.2). The numbers of orphans and deaths were slightly modified, but the main information contained in the phylogenies remained the same (*Supplementary Table S3*). Penalizing death reduced the number of deaths, as expected, and increased the number of orphans (**Fig. 5**, combinations 5.3.2, 3.3.2 and 2.2.2), while the resulting scenarios were less parsimonious (*Supplementary Table S3*). For example, the JNK1δ transcript was created twice, at the internal nodes A24 and A27 (*Supplementary Table S3*). The two mutually exclusive events involving

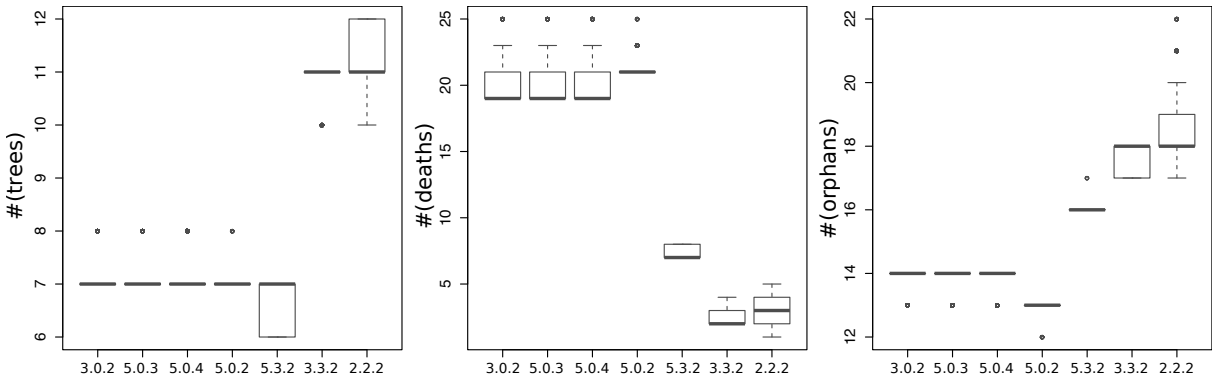

Figure 5: **Statistical analysis of the phylogenies obtained with different parameters.** The numbers of trees, deaths and orphans are given in y-axis for seven combinations of parameters. The parameter values are indicated on the x-axis, in the following order: $C_B.C_D.\sigma$. The distributions are computed over 1756, 4204, 2880, 576, 47, 326 and 1100 solutions of equal costs, found during 10^6 iterations of the algorithm.

exons *6/7* and *12/13* also appeared multiple times in independent sub-forests (*Supplementary Table S3*), except when the number of trees was constrained by a high birth cost (combination 5.3.2, see also **Fig. 5**). These observations illustrate well how the presence of under-annotated species in the input data may artificially scatter very similar or identical transcripts in different trees. We also investigated the influence of gene-level changes on the phylogenetic reconstruction. To do this, we changed the priority rule for the determination of the exon states at the gene level and we changed the cost associated to gene-level induced transcript mutations from zero to σ. The former did not have any impact on the results (*Supplementary Table S3*, combination $3.0.2^a$). The latter resulted in more trees and more deaths, which can be explained by the fact that all mutations being penalized it may become more advantageous to create a new transcript than to pair up two transcripts. However, the dating of the main ASEs did not change (*Supplementary Table S3*, combination $3.0.2^b$). We should stress that in all our simulations, exon *7* is present at the root of the forest, while exon *6* appears afterwards (*Supplementary Table S3*). Altogether, these results validate our choice of parameters and show that our JNK phylogeny is robust to small parameter changes.

Complexity and comparison with other methods

The size of the search space for the transcripts' phylogeny reconstruction grows exponentially with the number of observed transcripts (leaves). To explore that space, the heuristic algorithm implemented in PhyloSofS relies on a multi-start iterative procedure and on the computation of a lower bound to early filter out unlikely scenarios (see *Methods*). As the algorithm finds better and better solutions, the filtering procedure becomes more and more efficient. Transcripts assignment (**Fig. 1a**, in the middle) is performed only when the computed lower bound passes the filter. At each internal node, k (among n_l) transcripts from the left child must be paired with k (among n_r) transcripts from the right child. We solve this problem with a branch-and-bound algorithm whose complexity is of $\mathcal{O}(n^3)$ for $n_l = n_r = n$ (see details in *Supplementary Text S1*). Hence, it requires about $(s-1)n^3$ operations to pair up transcripts from the leaves to the root, with s the number of current species ($s-1$ being the number of ancestral species). Another heuristic algorithm, making use of a neighbor-joining operation, has been proposed in the literature[17]. This algorithm requires to look at up to $\binom{(s-1)n}{2}+\binom{(s-1)(n-1)}{2}+\binom{(s-1)(n-1)(n-2)}{2}+....+1 \simeq \mathcal{O}(s^3n^3)$ possible pairs of transcripts at the level just below the root, and recursively applies a neighbor-joining procedure to each one of these pairs down to the leaves. One advantage of this algorithm is that it constrains the space of considered phylogenies and thus may in practice be more efficient than PhyloSofS. However, it does not consider transcript mutations and hence cannot reconstruct ancestral transcripts (see Fig. 1A in[17]). Moreover, the information about transcript ancestry is virtually lost as trees can be merged by the neighbor-joining operation (see Fig. 1B in[17]). As a consequence, this algorithm is not suitable for inferring evolutionary scenarios explaining observed transcripts nor for dating AS events.

Discussion

To what extent the transcript diversity generated by AS translates at the protein level and has functional implications in the cell remains a very challenging question and has

been subject to much debate[48;56]. The present work contributes to elaborating strategies to answer it, by crossing sequence analysis and phylogenetic inference with molecular modeling. We report the first joint analysis of the evolution of alternative splicing across several species and of its structural impact on the produced isoforms. The analysis was performed on the JNK family, which represents a high interest for medicinal research and for which a number of human isoforms have been described and biochemically characterized.

Firstly, our results allowed dating an ASE consisting of two mutually exclusive homologous exons (*6* and *7*) in the ancestor common to mammals, amphibians and fishes. We find that the most ancient of these two exons is exon *7*. By characterizing in details the structural dynamics of two human isoforms, JNK1α and JNK1β, bearing one or the other exon, we could emphasize subtle changes associated to this ASE and identify residues that may be responsible for the selectivity of the JNK isoforms toward their substrates. Secondly, we highlighted an isoform that was not previously described in the literature, namely JNK1δ. Despite displaying a large deletion (about 80 residues), it is conserved across several species and short MD simulations suggest that it is stable in solution. According to the APPRIS database v20[57], there are 4 peptides matching this isoform in publicly available proteomics data. By comparison, the other human JNK1 isoforms with a phylogeny have between 5 and 7 matching peptides, while the orphan transcripts identified by our analysis have between zero and 2 matching peptides, suggesting that JNK1δ is indeed translated and stable in solution. Hence, considering that the catalytic site is intact in JNK1δ, we propose that this isoform might be catalytically competent and that the large amplitude motion of the A-loop observed in the simulations might facilitate the activation of the protein by exposing a couple of tyrosine and threonine residues that are targeted by MAPK kinases. The validation of this hypothesis would require further calculations and experiments that fall beyond the scope of this study. Already, this interesting result suggests that our approach could be used to identify and characterize new isoforms, that may play a role in the cell and thus serve as therapeutic targets. Thirdly, we found characteristics specific to the JNK3 isoforms, namely the ab-

sence of exon *7* and the presence of twos exons (*0* and *1'*) containing regions predicted to be disordered and involved in interactions. These observations suggest specific competences or functions for this gene. Studies investigating the gain/loss of alternative splice forms associated to gene duplication at large scale[1;58] have highlighted a wide diversity of cases and have suggested that it depends on the specific cellular context of each gene. Although we did not have a sufficient sample resolution to confirm it with RNA-Seq data, JNK3 is reported to be specifically expressed in the heart brain and testes[67]

Our approach enables to go beyond a description of transcript variability across species and/or across genes. Indeed, by reconstructing phylogenies, we do not only cluster transcripts but we also add a temporal dimension to the analysis. Previous methods reported in the literature were only applied on simple cases[16] and/or largely simplified the evolution model to increase computational efficiency[17], such that they could not be used for ancestral transcripts' reconstruction. PhyloSofS algorithm makes the reconstruction of transcripts' phylogenies feasible for any gene family. For the JNK family, the execution of 1 million iterations took about two weeks on a single CPU. This case represents a high level of complexity as most of the transcripts contain more than 10 exons (the average number of exons per gene being estimated at 8.8 in the human genome[60]) and up to 8 transcripts are observed within each species (it is estimated that about 4 distinct-coding transcripts per gene are expressed in human[57]). To reduce the computing time, the user can easily parallelize the multi-start iterative search on multiple cores and s/he has the possibility to give as input a previously computed value for the lower bound (to increase the efficiency of the cut). We should stress that the problem of pairing transcripts across homologous and paralogous genes between different species, addressed here, is much more complex than that of inferring the transcripts' phylogeny of each gene separately. Indeed, in the former case, the problem size is bigger, one needs to reconcile the gene tree with the species tree, and the sequences are more divergent.

Our phylogenies may be impacted by two main sources of error coming from the input data. Specifically, under-annotation of transcripts can lead to missing distant evolutionary relationships. To deal with this issue, we set the cost associated to transcript death

to zero. This enables to construct trees that can relate transcripts possibly very far from each other in the phylogeny (*i.e.* expressed in very distant species, because some species in between are under-annotated). This parameter may be tuned by the user depending on the quality and reliability of the input data. A second source of error comes from annotated transcripts that are not translated or not functional at the protein level. However, we do not expect that these transcripts will significantly pollute the phylogenetic reconstruction. Indeed, they are likely not conserved across species and thus will be attributed the status of orphans in the phylogenetic reconstruction. Moreover, we have emphasized an independent source of evidence coming from their structural characterization which can help us flag them. The reliability of the transcript expression data clearly constitutes a present limitation of the method. However, as experimental evidence accumulate and precise quantitative data become available, computational methods such as PhyloSofS will become instrumental in assessing the contribution of AS in protein evolution.

Although PhyloSofS was applied here to study the evolution of transcripts in different species, it has broad applicability and can be used to study transcript diversity and conservation among diverse biological entities. The entities could be at the scale of (*i*) one individual/species (tissue/cell differentiation), (*ii*) different species (matching cell types), (*iii*) population of individuals affected or not by a multifactorial disorder. In the first case, the tree given as input should describe checkpoints during cell differentiation and PhyloSofS will provide insights on the ASEs occurring along this process. In the second case, PhyloSofS can be applied to study one particular tissue across several species in a straightforward manner (explicitly dealing with the dimension of different tissues requires further development). In the third case, the tree given as input may be constructed based on genome comparison, a biological trait or disease symptoms. PhyloSofS can be used to evaluate the pertinence of such criteria to relate the patients, with regards to the likelihood (parsimony) of the associated transcripts scenarios. This case is particularly relevant in the context of medical research.

1 Materials and Methods

1.1 PhyloSofS workflow

PhyloSofS can be applied to single genes or to gene families. The input is a binary tree (called a gene tree) describing the phylogeny of the gene(s) of interest for a set of species (**Fig. 1a**, on the left), and the ensemble of transcripts observed in these species (symbols at the leaves). Only transcripts annotated as coding for a protein are considered. PhyloSofS comprises two main steps:

- It reconstructs a forest of phylogenetic trees describing plausible evolutionary scenarios that can explain the observed transcripts (**Fig. 1a**, on the right). The forest is embedded in the input gene tree and is reconstructed by using the maximum parsimony principle. The root of a tree corresponds to the creation of a new transcript, each leaf stands for an observed transcript and a dead end (indicated by a triangle on **Fig. 1a**, on the right) indicates a transcript loss. Transcripts can mutate along the branches of the trees.

- It predicts the three-dimensional structures of the protein isoforms corresponding to the observed transcripts by using homology modeling. The 3D models are then annotated with quality measures and with exon labels.

PhyloSofS comes with helper functions for the visualization of the output transcripts' phylogeny(ies) and of the isoforms' molecular models. The program is implemented in Python 3 and freely available at GitHub under MIT license: https://github.com/PhyloSofS-Team/PhyloSofS.

1.1.1 Step a. Transcripts' phylogenies reconstruction

For simplicity and without loss of generality, we describe here the case of one gene of interest studied across several species. The gene is represented by an ensemble E of n_e exons. The identification and alignment of the n_e homologous exons between the different transcripts must be performed prior to the application of the method (see below for details

on data preprocessing for the JNK family). The n_s transcripts of species s are described by a binary table T^s of $n_e \times n_s$ elements, where $T^s_{i,j} = 1$ if exon i is included in transcript j (*Supplementary Fig. S1a*, see colored squares), 0 if it is excluded (white squares).

We model transcripts evolution as a two-level process, at the gene and transcript levels, as described by Christinat and Moret[16]. At the level of the gene, each exon can be either absent, alternative or constitutive. This status is inferred from the occurrence of the exon in the transcripts. Hence, for a given species s, a vector g^s of length n_e encodes the state of each exon by the values $\{0,\ 1,\ 2\}$ for absent, alternative and constitutive, respectively (*Supplementary Fig. S1b*, white, black/white and black squares). At the leaves (current species), the components of g^s are calculated as:

$$g^s_i = \prod_{j=1}^{n_s} T^s_{i,j} + 1 - \prod_{j=1}^{n_s} (1 - T^s_{i,j}) \tag{1}$$

As in[16], the g^s vectors for internal nodes (ancestral species) are determined by using Sankoff's algorithm[62]. Dollo's parsimony principle is also respected, such that an exon cannot be created twice[2]. If different exon states have equal cost, we follow the priority rule $1 > 0 > 2$.

Three evolutionary events are considered, namely creation, death and mutation of a transcript with costs C_B, C_D and σ, respectively. The mutation cost σ is accounted for only when the associated evolutionary change occur at the level of the transcript (Table II). This reflects the fact that changes at the level of the gene affects the expression of exons in the transcripts but changes at the level of the transcripts do not affect the gene structure. For instance, if an exon is absent in a parent and becomes present in the child, then this change of status at the transcript level will be penalized by σ only if the exon could be absent in the child, *i.e.* its status at the gene level is "alternative". If the "constitutive" exon is the child, then the mutation is not penalized (Table II, compare the cells (0,0)→(1,1) and (0,0)→(1,2).

Each internal node of the gene tree, representing an ancestral species, is expanded in several subnodes, representing the transcripts of the gene in this ancestral species (*Supplementary Fig. S1c*). There exist three types of subnodes: binary (two transcript

Table II: **Exon states and associated costs σ.**

child / parent	(0,0)	(0,1)	(1,1)	(1,2)
(0,0)	0	0	**0**	**0**
(0,1)	0	0	σ	σ
(1,1)	σ	σ	0	0
(1,2)	**0**	σ	0	0

Each cell is associated to the evolution of the state of the exon e from a parent transcript to a child transcript. The first and second terms in parenthesis correspond to the status of the exon at the transcript and gene levels, respectively. Only evolutionary changes taking place at the transcript level, without being directly induced by a gene-level change, are penalized (σ). Zeros highlighted in bold indicate transcript-level changes being a direct consequence of a gene-level change. This table of costs was taken from[16].

children), left (one transcript child in the node's left child) and right (one transcript child in the node's right child). Left and right subnodes imply that a transcript death occurred along the branch. A *forest structure* S is fixed by setting n_b, n_l and n_r the respective numbers of binary, left and right subnodes for every internal node of the gene tree. The cost associated to structure S is calculated as $C_S = C_{birth}(S) + C_{death}(S)$, where $C_{birth}(S)$ and $C_{death}(S)$ are the total costs of creation and loss of transcripts, expressed as

$$C_{birth}(S) = C_B \times |S| \tag{2}$$

$$C_{death}(S) = C_D \times \sum_{\text{nodes } N} n_l(N) + n_r(N), \tag{3}$$

where $|S|$ is the number of trees in the forest.

Given a forest structure, a *transcripts' phylogeny* determines the pairings of transcripts at each internal node (*Supplementary Fig. S1d*). The cost of the transcripts' phylogeny φ complying with the forest structure S is calculated as:

$$C_\varphi = C_S + \sum_{A \text{ tree of } \varphi} \Gamma(A) \tag{4}$$

where $\Gamma(A)$ is computed for each tree A of φ by evaluating the changes of exon states

along the branches of φ:

$$\Gamma(A) = \sum_{t_i^k \to t_j^l \text{ branch of } A} \Gamma(t_i^k \to t_j^l) \tag{5}$$

where t_i^k is the parent transcript, i^{th} subnode of node k, t_j^l is the child transcript, j^{th} subnode of node l and $\Gamma(t_i^k \to t_j^l) = \sum_{e \in E} \sigma((T_{e,i}^k; g_e^k), (T_{e,j}^l; g_e^l))$, with $g_e^y \in \{0, 1, 2\}$ the state of exon e at the level of the gene at node y and $T_{e,x}^y \in \{0, 1\}$ the state of exon e at the level of the xth transcript of node y. The evolution costs σ are given in Table II.

PhyloSofS algorithm seeks to determine the scenario with the smallest number of evolutionary events, *i.e.* the transcripts' phylogeny with the minimum cost (*Supplementary Fig. S1c-d*). It proceeds as follows:

Initialization:
$C_{min} \leftarrow \infty$
Choose the forest structure S_0 that maximizes the n_b values
Iteration:
for $i = 0$ to $t_{max} - 1$ **do**
  if $C_{S_i} < C_{min}$ **then**
    Find the most parsimonious phylogeny φ_i given structure S_i
    if $C_{\varphi_i} < C_{min}$ **then**
      $C_{min} \leftarrow C_{\varphi_i}$
    end if
  end if
  Choose forest structure S_{i+1} by setting n_b, n_l and n_r at every internal node
end for

To efficiently search the space of all possible forest structures (*Supplementary Fig. S1c*), PhyloSofS relies on a multi-start iterative procedure. Random jumps in the search space are performed until a suitable forest structure S_i (with $C_{S_i} < C_{min}$) is found. The cost C_{S_i} of the forest structure S_i serves as a lower bound for the cost C_{φ_i} of the phylogeny

φ_i. Forest structures that are too costly are simply discarded, without calculating the corresponding phylogenies. As the algorithm finds better and better solutions, the cut becomes more and more efficient. The phylogeny φ_i is reconstructed by using dynamic programming. Sankoff's algorithm is applied bottom up to compute the minimum pairing costs between transcripts (*Supplementary Fig. S1d*, each transcript is represented by a matrix of costs). At each internal node, the pairings are determined by using a specific version of the branch-and-bound algorithm[41] (see *Supplementary Text S1*). If the reconstructed phylogeny is more parsimonious than those previously visited ($C_{\varphi_i} < C_{min}$), then the minimum cost C_{min} is updated. There may be more than one phylogeny with minimum cost that comply with a given structure S_i. The next forest structure S_j will be randomly chosen among the immediate neighbors of S_i (*Supplementary Fig. S1d*). Two structures are immediate neighbors if each one of them can be obtained by an elementary operation applied to only one node of the other one (*Supplementary Fig. S13*). If the phylogeny φ_j is such that $C_{\varphi_j} < C_{min}$, then the next forest structure will be chosen among the neighbors of S_j, which serves as a new "base" for the search. Otherwise, the algorithm continues to sample the neighborhood of S_i. This step-by-step search is applied until no better solution can be found. At this point, a new random jump is performed. The total number of iterations t_{max} is given as input by the user (1 by default).

We should stress that PhyloSofS algorithm is designed to deal with much more complex cases than those reported in[16] in a computationally tractable way. Hence, it differs from the algorithm reported in[16] in several respects. First, our multi-start iterative strategy relies on random jumps in the forest structure space combined with systematic local exploration around the best current solution, while Christinat and Moret[16] proposed an exhaustive generation and evaluation of forest structures. Secondly, we have designed a branch-and-bound algorithm specifically adapted to the problem of determining the best phylogeny complying with a given forest structure (see *Supplementary Text S1*). Both aspects contribute to PhyloSofS efficiency in reconstructing transcripts' phylogenies.

PhyloSofS generates PDF files displaying the computed transcripts' phylogenies using a Python driver to the Graphviz[23] DOT format.

1.1.2 Step b. Isoforms structures prediction

The molecular modeling routine implemented in PhyloSofS relies on homology modeling. It takes as input an ensemble of multi-fasta files (one per species) containing the sequences of the splicing isoforms. For each isoform, it proceeds as follows:

1. search for homologous sequences whose 3D structures are available in the PDB (templates) and align them to the query sequence;
2. select the n (5 by default, adjustable by the user) best templates;
3. build the 3D model of the query;
4. remove the N- and C-terminal residues unresolved in the model (no structural template);
5. annotate the model with sequence and structure information.

Step 1 makes extensive use of the HH-suite[29] and can be decomposed in: (a) search for homologous sequences and building of a multiple sequence alignment (MSA), by using HHblits[55], (b) addition of secondary structure predictions, obtained by PSIPRED[34], to the MSA, (c) generation of a profile hidden markov model (HMM) from the MSA, (d) search of a database of profile HMMs for homologous proteins, using HHsearch[63]. Step 3 is performed by Modeller[47] with default options. Step 5 consists in: (a) inserting the numbers of the exons in the β-factor column of the PDB file of the 3D model, (b) computing the proportion of residues predicted in well-defined secondary structures by PSIPRED[34], (c) assessing the quality of the model with Procheck[42] and with the normalized DOPE score from Modeller, (d) determining the by-residue solvent accessible surface areas with Naccess[31] and computing the proportions of surface residues and of hydrophobic surface residues.

1.2 Retrieval and pre-processing of JNK annotated transcriptome data

The peptide sequences of all splice variants from the JNK family observed in human, mouse, *Xenopus tropicalis*, zebrafish, fugu, *Drosophila melanogaster* and nematode were retrieved from Ensembl[75] release 84 (March 2016) along with the phylogenetic gene tree. Only the transcripts containing an open reading frame and not annotated as undergoing nonsense mediated decay or lacking 3' or 5' truncation were retained.The isoforms sharing the same amino acid sequence were merged. The homologous exons between the different genes in the different species were identified by aligning the sequences with MAFFT[35], and projecting the alignment on the human annotation. They do not necessarily represent exons definition based on the genomic sequence and this can be explained by two reasons. First, the gene structure may be different from one species to another. For instance, the third and fourth exons of human JNK1 genes are completely covered by a single exon in the *Drosophila melanogaster* JNK gene (*Supplementary Fig. S4*). In cases like this, we keep the highest level of resolution and define two exons (*e.g.* numbered *3* and *4*). Secondly, it may happen that a transcript contains only a part of an exon in a given species translated in another frame. In that case, we define two exons sharing the same number but distinguished by the prime symbol (*e.g.* exons *8* and *8'*). In total, 64 transcripts comprised of 38 exons were given as input to PhyloSofS.

1.3 PhyloSofS' parameter setting

To set the parameters, two criteria were taken into consideration. First, the different genomes available in Ensembl are not annotated with the same accuracy and the transcriptome data and annotations may be incomplete. This may challenge the reconstruction of transcripts' phylogenies across species. To cope with this issue, we chose not to penalize transcript death (C_D=0). Second, the JNK genes are highly conserved across the seven studied species (**Table I**), indicating that this family has not diverged much through evolution. Consequently, we set the transcript mutation and birth costs to $\sigma = 2$ and $C_B = 3$ ($C_B < \sigma \times 2$). This implies that few mutations will be tolerated along a

phylogeny. Prior to the phylogenetic reconstruction, PhyloSofS removed 19 exons that appeared in only one transcript (default option), reducing the number of transcripts to 60. This pruning enables to limit the noise contained in the input data and to more efficiently reconstruct phylogenies. PhyloSofS algorithm was then run for 10^6 iterations.

The 3D models of all observed isoforms were generated by PhyloSofS molecular modeling routine by setting the number of retained best templates to 5 (default parameter) for every isoform.

1.4 Analysis of JNK tertiary structures.

The list of experimental structures deposited in the PDB for the human JNKs was retrieved from UniProt[5]. The structures were aligned with PyMOL[19] and the RMSD between each pair was computed. Residues comprising the catalytic site were defined from the complex between human JNK3 and adenosine mono-phosphate (PDB code: 4KKE, resolution: 2.2 Å), as those located less than 6 Å away from the ligand. Residues comprising the D-site and the F-site were defined from the complexes between human JNK1 and the scaffolding protein JIP-1 (PDB code: 1UKH, resolution: 2.35 Å[28]) and the catalytic domain of MKP7 (PDB code: 4YR8, resolution: 2.4 Å[44]), respectively. They were detected as displaying a change in relative solvent accessibility >1 $Å^2$ upon binding.

The I-TASSER webserver[59;73;76] was used to try and model the regions for which no structural templates could be found. DISOPRED[71] and IUPred[21] were used to predict intrinsic disorder. JET2[40] was used to predict binding sites at the surface of the isoforms.

1.5 Molecular dynamics simulations of human isoforms.

The 3D coordinates of the human JNK1 isoforms JNK1α (369 res., containing exon *6*), JNK1β (369 res., containing exon *7*) and JNK1δ (304 res., containing neither exon *6* nor exon *7*) were predicted by PhyloSofS pipeline. The 3 systems were prepared with the LEAP module of AMBER 12[14], using the ff12SB forcefield parameter set: (*i*) hydrogen atoms were added, (*ii*) the protein was hydrated with a cuboid box of explicit TIP3P water molecules with a buffering distance up to 10Å, (*iii*) Na^+ and Cl^- counter-ions were

added to neutralize the protein.

The systems were minimized, thermalized and equilibrated using the SANDER module of AMBER 12. The following minimization procedure was applied: (*i*) 10,000 steps of minimization of the water molecules keeping protein atoms fixed, (*ii*) 10,000 steps of minimization keeping only protein backbone fixed to allow protein side chains to relax, (*iii*) 10,000 steps of minimization without any constraint on the system. Heating of the system to the target temperature of 310 K was performed at constant volume using the Berendsen thermostat[6] and while restraining the solute C_α atoms with a force constant of 10 $kcal/mol/\AA^2$. Thereafter, the system was equilibrated for 100 ps at constant volume (NVT) and for further 100 ps using a Langevin piston (NPT)[45] to maintain the pressure. Finally the restraints were removed and the system was equilibrated for a final 100 ps run.

Each system was simulated during 250 ns (5 replicates of 50 ns, starting from different initial velocities) in the NPT ensemble using the PMEMD module of AMBER 12. The temperature was kept at 310 K and pressure at 1 bar using the Langevin piston coupling algorithm. The SHAKE algorithm was used to freeze bonds involving hydrogen atoms, allowing for an integration time step of 2.0 fs. The Particle Mesh Ewald (PME) method[18] was employed to treat long-range electrostatics. The coordinates of the system were written every ps.

Standard analyses of the MD trajectories were performed with the *ptraj* module of AMBER 12. The calculation of the root mean square deviation (RMSD) over all atoms indicated that it took between 5 and 20 ns for the systems to relax. Consequently, the last 30 ns of each replicate were retained for further analysis, totaling 150 000 snapshots for each system. The fluctuations of the C-α atoms were recorded along each replicate. For each residue or each system, we report the value averaged over the 5 replicates and the standard deviation (see **Fig. 4a**). The secondary structures were assigned by DSSP algorithm over the whole conformational ensembles. For each residue, the most frequent secondary structure type was retained (see **Fig. 4a** and *Supplementary Fig. S9*). If no secondary structure was present in more than 50% of the MD conformations, then the

residue was assigned to a loop. The amplitude of the motion of the A-loop compared to the rest of the protein was estimated by computing the angle between the geometric center of residues 189-192, residue 205 and either residue 211 in the isoforms JNK1α and JNK1β or residue 209 in the isoform JNK1δ. Only C-α atoms were considered.

1.6 RNA-Seq Data integration

To obtain additional support for transcript isoform expression, we queried the Bgee database (v.14)[4] for a list of all RNA-Seq experiments related to the selected species. Using SRA tools, we downloaded raw sequences from *H. sapiens* (224 samples), *M. musculus* (155 samples), *Xenopus tropicalis* (69 samples) and *D. rerio* (67 samples) and then aligned the reads using STAR v.2.5.3a[20] with default parameters. *T. rubripes* is not annotated in Bgee and was not integrated in this part of the analysis. Reads overlapping exon-exon boundaries (*e.g.* splice-junction reads) next to alternative splicing events provide direct evidence for the expression of specific transcripts isoforms. Combined with sample annotation, they could also inform on tissue specific isoform expression. We thus considered all reads included within one of the JNK genes and monitored the alignment of splice junctions between different exons as support for the transcripts isoforms: exons 5-6 and 6-8 for JN1α, exons 5-7 and 7-8 for JNK1β, and exons 5-8' for JNK1δ.

Acknowledgments A grant of the French national research agency (MASSIV project, ANR-17-CE12-0009) provided a salary to D.J.Z. and A.L.. We thank Y. Christinat for providing information on the algorithm he developed for the reconstruction of transcript phylogenies.

References

[1] Abascal, F., Tress, M. L., and Valencia, A. (2015). The evolutionary fate of alternatively spliced homologous exons after gene duplication. *Genome Biol Evol*, 7(6):1392–1403.

[2] Alekseyenko, A. V., Lee, C. J., and Suchard, M. A. (2008). Wagner and Dollo: a stochastic duet by composing two parsimonious solos. *Syst. Biol.*, 57(5):772–784.

[3] Barbosa-Morais, N. L., Irimia, M., Pan, Q., Xiong, H. Y., Gueroussov, S., Lee, L. J., Slobodeniuc, V., Kutter, C., Watt, S., Colak, R., Kim, T., Misquitta-Ali, C. M., Wilson, M. D., Kim, P. M., Odom, D. T., Frey, B. J., and Blencowe, B. J. (2012). The evolutionary landscape of alternative splicing in vertebrate species. *Science*, 338(6114):1587–1593.

[4] Bastian, F., Parmentier, G., Roux, J., Moretti, S., Laudet, V., and Robinson-Rechavi, M. (2008). chapter Bgee: Integrating and Comparing Heterogeneous Transcriptome Data Among Species, pages 124–131.

[5] Bateman, A., Martin, M. J., O'Donovan, C., Magrane, M., Apweiler, R., Alpi, E., Antunes, R., Arganiska, J., Bely, B., Bingley, M., Bonilla, C., Britto, R., Bursteinas, B., Chavali, G., Cibrian-Uhalte, E., Silva, A. D., De Giorgi, M., Dogan, T., Fazzini, F., Gane, P., Castro, L. G., Garmiri, P., Hatton-Ellis, E., Hieta, R., Huntley, R., Legge, D., Liu, W., Luo, J., MacDougall, A., Mutowo, P., Nightingale, A., Orchard, S., Pichler, K., Poggioli, D., Pundir, S., Pureza, L., Qi, G., Rosanoff, S., Saidi, R., Sawford, T., Shypitsyna, A., Turner, E., Volynkin, V., Wardell, T., Watkins, X., Zellner, H., Cowley, A., Figueira, L., Li, W., McWilliam, H., Lopez, R., Xenarios, I., Bougueleret, L., Bridge, A., Poux, S., Redaschi, N., Aimo, L., Argoud-Puy, G., Auchincloss, A., Axelsen, K., Bansal, P., Baratin, D., Blatter, M. C., Boeckmann, B., Bolleman, J., Boutet, E., Breuza, L., Casal-Casas, C., de Castro, E., Coudert, E., Cuche, B., Doche, M., Dornevil, D., Duvaud, S., Estreicher, A., Famiglietti, L., Feuermann, M., Gasteiger, E., Gehant, S., Gerritsen, V., Gos, A., Gruaz-Gumowski, N., Hinz, U., Hulo, C., Jungo, F., Keller, G., Lara, V., Lemercier, P., Lieberherr, D., Lombardot, T., Martin, X., Masson, P., Morgat, A., Neto, T., Nouspikel, N., Paesano, S., Pedruzzi, I., Pilbout, S., Pozzato, M., Pruess, M., Rivoire, C., Roechert, B., Schneider, M., Sigrist, C., Sonesson, K., Staehli, S., Stutz, A., Sundaram, S., Tognolli, M., Verbregue, L., Veuthey, A. L., Wu, C. H., Arighi, C. N., Arminski, L., Chen, C., Chen, Y., Garavelli, J. S., Huang,

826 H., Laiho, K., McGarvey, P., Natale, D. A., Suzek, B. E., Vinayaka, C., Wang, Q.,
827 Wang, Y., Yeh, L. S., Yerramalla, M. S., and Zhang, J. (2015). UniProt: a hub for
828 protein information. *Nucleic Acids Res.*, 43(Database issue):D204–212.

829 [6] Berendsen, H., Postma, J., van Gunsteren, W., DiNola, A., and Haak, J. (1984).
830 Molecular dynamics with coupling to an external bath. *The Journal of chemical physics*,
831 81(8):3684–3690.

832 [7] Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H.,
833 Shindyalov, I. N., and Bourne, P. E. (2000). The Protein Data Bank. *Nucleic Acids*
834 *Res.*, 28(1):235–242.

835 [8] Bhuiyan, S. A., Ly, S., Phan, M., Huntington, B., Hogan, E., Liu, C. C., Liu, J., and
836 Pavlidis, P. (2018). Systematic evaluation of isoform function in literature reports of
837 alternative splicing. *BMC Genomics*, 19(1):637.

838 [9] Birzele, F., Csaba, G., and Zimmer, R. (2008a). Alternative splicing and protein
839 structure evolution. *Nucleic Acids Res.*, 36(2):550–558.

840 [10] Birzele, F., Kuffner, R., Meier, F., Oefinger, F., Potthast, C., and Zimmer, R.
841 (2008b). ProSAS: a database for analyzing alternative splicing in the context of protein
842 structures. *Nucleic Acids Res.*, 36(Database issue):D63–68.

843 [11] Bogoyevitch, M. A. and Kobe, B. (2006). Uses for JNK: the many and varied
844 substrates of the c-Jun N-terminal kinases. *Microbiol. Mol. Biol. Rev.*, 70(4):1061–
845 1095.

846 [12] Brecht, S., Kirchhof, R., Chromik, A., Willesen, M., Nicolaus, T., Raivich, G.,
847 Wessig, J., Waetzig, V., Goetz, M., Claussen, M., Pearse, D., Kuan, C. Y., Vaudano,
848 E., Behrens, A., Wagner, E., Flavell, R. A., Davis, R. J., and Herdegen, T. (2005).
849 Specific pathophysiological functions of JNK isoforms in the brain. *Eur. J. Neurosci.*,
850 21(2):363–377.

851 [13] Buljan, M., Chalancon, G., Eustermann, S., Wagner, G. P., Fuxreiter, M., Bateman,

A., and Babu, M. M. (2012). Tissue-specific splicing of disordered segments that embed binding motifs rewires protein interaction networks. *Mol. Cell*, 46(6):871–883.

[14] Case, D., Darden, T., Cheatham III, T., Simmerling, C., Wang, J., Duke, R., Luo, R., Walker, R., Zhang, W., Merz, K., et al. (2012). Amber 12. *University of California, San Francisco*, 1(2):3.

[15] Chamberlain, S. D., Redman, A. M., Wilson, J. W., Deanda, F., Shotwell, J. B., Gerding, R., Lei, H., Yang, B., Stevens, K. L., Hassell, A. M., Shewchuk, L. M., Leesnitzer, M. A., Smith, J. L., Sabbatini, P., Atkins, C., Groy, A., Rowand, J. L., Kumar, R., Mook, R. A., Moorthy, G., and Patnaik, S. (2009). Optimization of 4,6-bis-anilino-1H-pyrrolo[2,3-d]pyrimidine IGF-1R tyrosine kinase inhibitors towards JNK selectivity. *Bioorg. Med. Chem. Lett.*, 19(2):360–364.

[16] Christinat, Y. and Moret, B. M. (2012). Inferring transcript phylogenies. *BMC Bioinformatics*, 13 Suppl 9:S1.

[17] Christinat, Y. and Moret, B. M. (2013). A transcript perspective on evolution. *IEEE/ACM Trans Comput Biol Bioinform*, 10(6):1403–1411.

[18] Darden, T., York, D., and Pedersen, L. (1993). Particle mesh ewald: An nlog(n) method for ewald sums in large systems. *The Journal of Chemical Physics*, 98:10089–10092.

[19] DeLano, W. (2002). The PyMOL Molecular Graphics System. http://www.pymol.org.

[20] Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T. R. (2013). Star: ultrafast universal rna-seq aligner. *Bioinformatics*, 29(1):15–21.

[21] Dosztanyi, Z., Csizmok, V., Tompa, P., and Simon, I. (2005). IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics*, 21(16):3433–3434.

[22] Ezkurdia, I., Rodriguez, J. M., Carrillo-de Santa Pau, E., Vazquez, J., Valencia, A., and Tress, M. L. (2015). Most highly expressed protein-coding genes have a single dominant isoform. *J. Proteome Res.*, 14(4):1880–1887.

[23] Gansner, E. R. and North, S. C. (2000). An open graph visualization system and its applications to software engineering. *SOFTWARE - PRACTICE AND EXPERIENCE*, 30(11):1203–1233.

[24] Gelly, J. C., Lin, H. Y., de Brevern, A. G., Chuang, T. J., and Chen, F. C. (2012). Selective constraint on human pre-mRNA splicing by protein structural properties. *Genome Biol Evol*, 4(9):966–975.

[25] Gonzalez-Porta, M., Frankish, A., Rung, J., Harrow, J., and Brazma, A. (2013). Transcriptome analysis of human tissues and cell lines reveals one dominant transcript per gene. *Genome Biol.*, 14(7):R70.

[26] Graveley, B. R. (2001). Alternative splicing: increasing diversity in the proteomic world. *Trends Genet*, 17(2):100–107.

[27] Hegyi, H., Kalmar, L., Horvath, T., and Tompa, P. (2011). Verification of alternative splicing variants based on domain integrity, truncation length and intrinsic protein disorder. *Nucleic Acids Res.*, 39(4):1208–1219.

[28] Heo, Y. S., Kim, S. K., Seo, C. I., Kim, Y. K., Sung, B. J., Lee, H. S., Lee, J. I., Park, S. Y., Kim, J. H., Hwang, K. Y., Hyun, Y. L., Jeon, Y. H., Ro, S., Cho, J. M., Lee, T. G., and Yang, C. H. (2004). Structural basis for the selective inhibition of JNK1 by the scaffolding protein JIP1 and SP600125. *EMBO J.*, 23(11):2185–2195.

[29] Hildebrand, A., Remmert, M., Biegert, A., and Soding, J. (2009). Fast and accurate automatic structure prediction with HHpred. *Proteins*, 77 Suppl 9:128–132.

[30] Hirosumi, J., Tuncman, G., Chang, L., Gorgun, C. Z., Uysal, K. T., Maeda, K., Karin, M., and Hotamisligil, G. S. (2002). A central role for JNK in obesity and insulin resistance. *Nature*, 420(6913):333–336.

[31] Hubbard, S. and Thornton, J. (1992-6). `http://www.bioinf.manchester.ac.uk/naccess/`.

[32] Hunot, S., Vila, M., Teismann, P., Davis, R. J., Hirsch, E. C., Przedborski, S., Rakic, P., and Flavell, R. A. (2004). JNK-mediated induction of cyclooxygenase 2 is required for neurodegeneration in a mouse model of Parkinson's disease. *Proc. Natl. Acad. Sci. U.S.A.*, 101(2):665–670.

[33] Huse, M. and Kuriyan, J. (2002). The conformational plasticity of protein kinases. *Cell*, 109(3):275–282.

[34] Jones, D. T. (1999). Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.*, 292(2):195–202.

[35] Katoh, K. and Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.*, 30(4):772–780.

[36] Kelemen, O., Convertini, P., Zhang, Z., Wen, Y., Shen, M., Falaleeva, M., and Stamm, S. (2013). Function of alternative splicing. *Gene*, 514(1):1–30.

[37] Kim, M. S., Pinto, S. M., Getnet, D., Nirujogi, R. S., Manda, S. S., Chaerkady, R., Madugundu, A. K., Kelkar, D. S., Isserlin, R., Jain, S., Thomas, J. K., Muthusamy, B., Leal-Rojas, P., Kumar, P., Sahasrabuddhe, N. A., Balakrishnan, L., Advani, J., George, B., Renuse, S., Selvan, L. D., Patil, A. H., Nanjappa, V., Radhakrishnan, A., Prasad, S., Subbannayya, T., Raju, R., Kumar, M., Sreenivasamurthy, S. K., Marimuthu, A., Sathe, G. J., Chavan, S., Datta, K. K., Subbannayya, Y., Sahu, A., Yelamanchi, S. D., Jayaram, S., Rajagopalan, P., Sharma, J., Murthy, K. R., Syed, N., Goel, R., Khan, A. A., Ahmad, S., Dey, G., Mudgal, K., Chatterjee, A., Huang, T. C., Zhong, J., Wu, X., Shaw, P. G., Freed, D., Zahari, M. S., Mukherjee, K. K., Shankar, S., Mahadevan, A., Lam, H., Mitchell, C. J., Shankar, S. K., Satishchandra, P., Schroeder, J. T., Sirdeshmukh, R., Maitra, A., Leach, S. D., Drake, C. G., Halushka, M. K., Prasad, T. S., Hruban, R. H., Kerr, C. L., Bader, G. D., Iacobuzio-Donahue, C. A., Gowda, H.,

and Pandey, A. (2014). A draft map of the human proteome. *Nature*, 509(7502):575–581.

[38] Kornev, A. P., Haste, N. M., Taylor, S. S., and Eyck, L. F. (2006). Surface comparison of active and inactive protein kinases identifies a conserved activation mechanism. *Proc. Natl. Acad. Sci. U.S.A.*, 103(47):17783–17788.

[39] Kyriakis, J. M. and Avruch, J. (2012). Mammalian MAPK signal transduction pathways activated by stress and inflammation: a 10-year update. *Physiol. Rev.*, 92(2):689–737.

[40] Laine, E. and Carbone, A. (2015). Local Geometry and Evolutionary Conservation of Protein Surfaces Reveal the Multiple Recognition Patches in Protein-Protein Interactions. *PLoS Comput. Biol.*, 11(12):e1004580.

[41] Land, A. H. and Doig, A. G. (1960). An automatic method of solving discrete programming problems. *Econometrica*, 28:497–520.

[42] Laskowski, R. A., MacArthur, M. W., Moss, D. S., and Thornton, J. M. (1993). PROCHECK: a program to check the stereochemical quality of protein structures. *Journal of Applied Crystallography*, 26(2):283–291.

[43] Lim, K. H., Ferraris, L., Filloux, M. E., Raphael, B. J., and Fairbrother, W. G. (2011). Using positional distribution to identify splicing elements and predict pre-mRNA processing defects in human genes. *Proc. Natl. Acad. Sci. U.S.A.*, 108(27):11093–11098.

[44] Liu, X., Zhang, C. S., Lu, C., Lin, S. C., Wu, J. W., and Wang, Z. X. (2016). A conserved motif in JNK/p38-specific MAPK phosphatases as a determinant for JNK1 recognition and inactivation. *Nat Commun*, 7:10879.

[45] Loncharich, R., Brooks, B., and Pastor, R. (1992). Langevin dynamics of peptides: The frictional dependence of isomerization rates of n-acetylalanyl-N'-methylamide. *Biopolymers*, 32(5):523–535.

[46] Manning, A. M. and Davis, R. J. (2003). Targeting JNK for therapeutic benefit: from junk to gold? *Nat Rev Drug Discov*, 2(7):554–565.

[47] Marti-Renom, M. A., Stuart, A. C., Fiser, A., Sanchez, R., Melo, F., and Sali, A. (2000). Comparative protein structure modeling of genes and genomes. *Annu Rev Biophys Biomol Struct*, 29:291–325.

[48] Melamud, E. and Moult, J. (2009). Stochastic noise in splicing machinery. *Nucleic Acids Res.*, 37(14):4873–4886.

[49] Merkin, J., Russell, C., Chen, P., and Burge, C. B. (2012). Evolutionary dynamics of gene and isoform regulation in Mammalian tissues. *Science*, 338(6114):1593–1599.

[50] Meszaros, B., Simon, I., and Dosztanyi, Z. (2009). Prediction of protein binding regions in disordered proteins. *PLoS Comput. Biol.*, 5(5):e1000376.

[51] Mudge, J. M., Frankish, A., Fernandez-Banet, J., Alioto, T., Derrien, T., Howald, C., Reymond, A., Guigo, R., Hubbard, T., and Harrow, J. (2011). The origins, evolution, and functional potential of alternative splicing in vertebrates. *Mol. Biol. Evol.*, 28(10):2949–2959.

[52] Nicolas, A., Raguenes-Nicol, C., Ben Yaou, R., Ameziane-Le Hir, S., Cheron, A., Vie, V., Claustres, M., Leturcq, F., Delalande, O., Hubert, J. F., Tuffery-Giraud, S., Giudice, E., and Le Rumeur, E. (2015). Becker muscular dystrophy severity is linked to the structure of dystrophin. *Hum. Mol. Genet.*, 24(5):1267–1279.

[53] Oruganty, K., Talathi, N. S., Wood, Z. A., and Kannan, N. (2013). Identification of a hidden strain switch provides clues to an ancient structural mechanism in protein kinases. *Proc. Natl. Acad. Sci. U.S.A.*, 110(3):924–929.

[54] Prakash, A. and Bateman, A. (2015). Domain atrophy creates rare cases of functional partial protein domains. *Genome Biol.*, 16:88.

[55] Remmert, M., Biegert, A., Hauser, A., and Soding, J. (2011). HHblits: lightning-

fast iterative protein sequence searching by HMM-HMM alignment. *Nat. Methods*, 9(2):173–175.

[56] Reyes, A., Anders, S., Weatheritt, R. J., Gibson, T. J., Steinmetz, L. M., and Huber, W. (2013). Drift and conservation of differential exon usage across tissues in primate species. *Proc. Natl. Acad. Sci. U.S.A.*, 110(38):15377–15382.

[57] Rodriguez, J. M., Maietta, P., Ezkurdia, I., Pietrelli, A., Wesselink, J. J., Lopez, G., Valencia, A., and Tress, M. L. (2013). APPRIS: annotation of principal and alternative splice isoforms. *Nucleic Acids Res.*, 41(Database issue):D110–117.

[58] Roux, J. and Robinson-Rechavi, M. (2011). Age-dependent gain of alternative splice forms and biased duplication explain the relation between splicing and duplication. *Genome Res.*, 21(3):357–363.

[59] Roy, A., Kucukural, A., and Zhang, Y. (2010). I-TASSER: a unified platform for automated protein structure and function prediction. *Nat Protoc*, 5(4):725–738.

[60] Sakharkar, M. K., Chow, V. T., and Kangueane, P. (2004). Distributions of exons and introns in the human genome. *In Silico Biol. (Gedrukt)*, 4(4):387–393.

[61] Sandberg, R., Neilson, J. R., Sarma, A., Sharp, P. A., and Burge, C. B. (2008). Proliferating cells express mrnas with shortened 3' untranslated regions and fewer microrna target sites. *Science*, 320(5883):1643–1647.

[62] Sankoff, D. (1975). Minimal mutation trees of sequences. *SIAM Journal on Applied Mathematics*, 28(1):35–42.

[63] Soding, J. (2005). Protein homology detection by HMM-HMM comparison. *Bioinformatics*, 21(7):951–960.

[64] Stamm, S., Ben-Ari, S., Rafalska, I., Tang, Y., Zhang, Z., Toiber, D., Thanaraj, T. A., and Soreq, H. (2005). Function of alternative splicing. *Gene*, 344:1–20.

[65] Tranchevent, L. C., Aube, F., Dulaurier, L., Benoit-Pilven, C., Rey, A., Poret, A., Chautard, E., Mortada, H., Desmet, F. O., Chakrama, F. Z., Moreno-Garcia, M. A.,

Goillot, E., Janczarski, S., Mortreux, F., Bourgeois, C. F., and Auboeuf, D. (2017). Identification of protein features encoded by alternative exons using Exon Ontology. *Genome Res.*, 27(6):1087–1097.

[66] Tuncman, G., Hirosumi, J., Solinas, G., Chang, L., Karin, M., and Hotamisligil, G. S. (2006). Functional in vivo interactions between JNK1 and JNK2 isoforms in obesity and insulin resistance. *Proc. Natl. Acad. Sci. U.S.A.*, 103(28):10741–10746.

[67] Waetzig, V. and Herdegen, T. (2005). Context-specific inhibition of JNKs: overcoming the dilemma of protection and damage. *Trends Pharmacol. Sci.*, 26(9):455–461.

[68] Wang, E. T., Sandberg, R., Luo, S., Khrebtukova, I., Zhang, L., Mayr, C., Kingsmore, S. F., Schroth, G. P., and Burge, C. B. (2008). Alternative isoform regulation in human tissue transcriptomes. *Nature*, 456(7221):470–476.

[69] Wang, X., Codreanu, S. G., Wen, B., Li, K., Chambers, M. C., Liebler, D. C., and Zhang, B. (2018). Detection of proteome diversity resulted from alternative splicing is limited by trypsin cleavage specificity. *Mol Cell Proteomics*, 17(3):422–430.

[70] Ward, A. J. and Cooper, T. A. (2010). The pathobiology of splicing. *J. Pathol.*, 220(2):152–163.

[71] Ward, J. J., McGuffin, L. J., Bryson, K., Buxton, B. F., and Jones, D. T. (2004). The DISOPRED server for the prediction of protein disorder. *Bioinformatics*, 20(13):2138–2139.

[72] Weatheritt, R. J., Sterne-Weiler, T., and Blencowe, B. J. (2016). The ribosome-engaged landscape of alternative splicing. *Nat. Struct. Mol. Biol.*, 23(12):1117–1123.

[73] Yang, J., Yan, R., Roy, A., Xu, D., Poisson, J., and Zhang, Y. (2015). The I-TASSER Suite: protein structure and function prediction. *Nat. Methods*, 12(1):7–8.

[74] Yang, X., Coulombe-Huntington, J., Kang, S., Sheynkman, G. M., Hao, T., Richardson, A., Sun, S., Yang, F., Shen, Y. A., Murray, R. R., Spirohn, K., Begg, B. E., Duran-Frigola, M., MacWilliams, A., Pevzner, S. J., Zhong, Q., Trigg, S. A., Tam,

S., Ghamsari, L., Sahni, N., Yi, S., Rodriguez, M. D., Balcha, D., Tan, G., Costanzo, M., Andrews, B., Boone, C., Zhou, X. J., Salehi-Ashtiani, K., Charloteaux, B., Chen, A. A., Calderwood, M. A., Aloy, P., Roth, F. P., Hill, D. E., Iakoucheva, L. M., Xia, Y., and Vidal, M. (2016). Widespread expansion of protein interaction capabilities by alternative splicing. *Cell*, 164(4):805–817.

[75] Yates, A., Akanni, W., Amode, M. R., Barrell, D., Billis, K., Carvalho-Silva, D., Cummins, C., Clapham, P., Fitzgerald, S., Gil, L., Giron, C. G., Gordon, L., Hourlier, T., Hunt, S. E., Janacek, S. H., Johnson, N., Juettemann, T., Keenan, S., Lavidas, I., Martin, F. J., Maurel, T., McLaren, W., Murphy, D. N., Nag, R., Nuhn, M., Parker, A., Patricio, M., Pignatelli, M., Rahtz, M., Riat, H. S., Sheppard, D., Taylor, K., Thormann, A., Vullo, A., Wilder, S. P., Zadissa, A., Birney, E., Harrow, J., Muffato, M., Perry, E., Ruffier, M., Spudich, G., Trevanion, S. J., Cunningham, F., Aken, B. L., Zerbino, D. R., and Flicek, P. (2016). Ensembl 2016. *Nucleic Acids Res.*, 44(D1):D710–716.

[76] Zhang, Y. (2008). I-TASSER server for protein 3D structure prediction. *BMC Bioinformatics*, 9:40.