bioRxiv preprint doi: https://doi.org/10.1101/119891; this version posted March 23, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Transcripts evolutionary conservation and structural dynamics give insights into the role of alternative splicing for the JNK family.

Adel Ait-hamlat[1], Lélia Polit[1], Hugues Richard[1,⋆] and Elodie Laine[1,⋆]

[1] Sorbonne Universités, UPMC University Paris 06, CNRS, IBPS, UMR 7238,
Laboratoire de Biologie Computationnelle et Quantitative (LCQB), 75005 Paris, France.

⋆ corresponding authors: elodie.laine@upmc.fr, hugues.richard@upmc.fr

bioRxiv preprint doi: https://doi.org/10.1101/119891; this version posted March 23, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Abstract

Alternative splicing (AS), by producing several transcript isoforms from the same gene, has the potential to greatly expand the proteome in eukaryotes. Its deregulation has been associated to the development of various diseases, including cancer. Although the AS mechanisms are well described at the genomic level, little is known about the contribution of AS to protein evolution and the impact of AS at the level of the protein structure. Here, we address both issues by reconstructing the evolutionary history of the c-Jun N-terminal kinase (JNK) family, and by describing the tertiary structures and dynamical behavior of several JNK isoforms. JNKs bear a great interest for medicinal research as they are involved in crucial signaling pathways. We reconstruct the phylogenetic forest relating 60 JNK transcripts observed in 7 species. We use it to estimate the evolutionary conservation of transcripts and to identify ASEs likely to be functionally important. We show that ASEs of ancient origin and having significant functional outcome may induce very subtle changes on the protein's structural dynamics. We also propose that phylogenetic reconstruction, combined with structural modeling, can help identify new potential therapeutic targets. Finally, we show that transcripts likely non-functional (*i.e.* not conserved) display peculiar sequence and structural properties. Our approach is implemented in PhyloSofS (Phylogenies of Splicing Isoforms Structures), a fully automated computational tool that infers plausible evolutionary scenarios explaining a set of transcripts observed in several species and models the three-dimensional structures of the protein isoforms. PhyloSofS has broad applicability and can be used, for example, to study transcripts diversity between different individuals (*e.g.* patients affected by a particular disease). It is freely available at `www.lcqb.upmc.fr/PhyloSofS`.

bioRxiv preprint doi: https://doi.org/10.1101/119891; this version posted March 23, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Author Summary

Alternative splicing (AS) is a eukaryotic regulatory process by which multiple proteins are produced from the same gene. Although the mechanisms of AS have been extensively described at the level of the gene, little is known about its contribution to protein evolution and its impact on the shape and motions of the produced isoforms. Here, we address both issues computationally, focusing our study on the c-Jun N-terminal kinases (JNKs) family. JNKs are essential regulators that target specific transcription factors and are thus important therapeutic targets. We reconstruct a phylogenetic forest linking 60 JNK transcript isoforms observed in 7 species and we predict and analyze their 3D structures. We show that an ancient ASE having significant functional outcome induces very subtle changes on the structural dynamics of the protein and we identify the residues likely responsible for the functional change. We highlight a new isoform, not previously documented, and explore its motions in solution. We propose that it may play a role in the cell and serve as a therapeutic target. Finally, we link the evolutionary conservation of transcripts to sequence and structural properties.

Key words: alternative splicing, molecular modeling, evolution, transcript phylogeny, kinase, isoform

bioRxiv preprint doi: https://doi.org/10.1101/119891; this version posted March 23, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Introduction

Alternative Splicing (AS) of pre-mRNA transcripts is an essential eukaryotic regulatory process by which multiple isoforms are produced from the same gene. AS-induced changes in the transcribed sequences can impact the regulation of gene expression or directly modify the content of the coding sequence (CDS). Large-scale studies revealed that virtually all multi-exons genes in vertebrates are subject to AS [1]. Consequently, AS has the potential to greatly contribute to functional diversity in eukaryotes. AS has also gained considerable interest for drug development. It is estimated that 50% of disease causing mutations affect splicing and the ratio of alternatively spliced isoforms is imbalanced in several cancers [2, 3].

About 25% of the AS events (ASEs) common to human and mouse are also conserved in vertebrates [4, 5, 6]. This high degree of conservation supports an important role of AS in expanding the protein repertoire through evolution. However, it is difficult to estimate to what extent the ASEs identified at the gene level actually result in functional protein isoforms in the cell. Transcriptomics and proteomics studies suggested that most highly expressed human genes have only one single dominant isoform [7, 8], but the detection rate of these experiments is very difficult to assess [9]. Larger estimates of the number of functional isoforms in human were reported by machine learning studies [10]. Moreover, a recent analysis of ribosome profiling data suggested that a major fraction of splice variants is translated and that the AS-dependent modulation of the translation output regulates specific cellular functions [11]. At the level of protein structures, it was suggested that splicing events may induce major fold changes [12, 13]. The elusiveness of the significance of AS for protein function through evolution calls for the development of efficient and accurate computational methods that combine protein sequence and structure information.

bioRxiv preprint doi: https://doi.org/10.1101/119891; this version posted March 23, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

To address this issue, we have developed an automated tool, PhyloSofS (Phylogenies of Splicing Isoforms Structures), that infers plausible evolutionary scenarios explaining an ensemble of transcripts observed in a set of species and predicts the tertiary structures of the protein isoforms. Given a gene tree and the observed transcripts at the leaves (**Fig. 1a**, on the left), PhyloSofS reconstructs a phylogenetic forest that is embedded in the gene tree (**Fig. 1a**, on the right), where each tree of the forest (in orange, green or purple) represents the phylogeny of one transcript. The algorithm relies on a combinatorial approach and the maximum parsimony principle. The underlying evolutionary model is inspired from [14]. In parallel, the isoforms' 3D structures are generated using comparative modeling and annotated. Here, we present the application of PhyloSofS to the c-Jun N-terminal kinase (JNK) family across 7 species (human, mouse, xenope, fugu fish, zebrafish, drosophila and nematode). This case represents a high degree of complexity with 60 observed transcripts composed with a total of 19 different exons, most of the transcripts comprising more than 10 exons, and high disparities between species, from 1 to 8 transcripts per gene per species (**Fig. 1b-c**).

In Human, JNKs are essential regulators that target specific transcription factors (c-Jun, ATF2...) in response to cellular stimuli. They are involved in signaling pathways controlling cellular proliferation, differentiation and apoptosis. The deregulation of their activity is associated with various diseases (cancer, inflammatory diseases, neuronal disorder...) which makes them important therapeutic targets [15]. The family comprises three paralogues: JNK1 (MAPK8) and JNK2 (MAPK9) are ubiquitously expressed, while JNK3 (MAPK10) is present primarily in the heart, brain and testes [16]. About 10 JNK splicing isoforms have been documented in the literature, and gene-disruption and functional-interference studies showed that they perform different context-specific tasks [17, 18, 19, 20]. Specifically, isoforms differing by the presence/absence of two mutually exclusive exons (numbered 6 and 7 on **Fig. 1b**) display different affinities for JNK substrates, so that the target genes are in turn differentially regulated [21, 22]. In the context of drug development, the

bioRxiv preprint doi: https://doi.org/10.1101/119891; this version posted March 23, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

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 exon compositions of the human isoforms for which a phylogeny could be reconstructed. They represent a subset of all the exons composing the 60 transcripts observed in the 7 current species. **(c)** Representation of the transcripts' phylogeny embedded in the species tree. In this forest, the duplication events are not explicitly indicated, as the different paralogous genes are not linked. There are 2 duplication events giving rise to JNK2 and JNK3 and 2 additional ones for JNK1a and JNK1b (indicated by stars). A given species may contain several paralogous genes. To differentiate the transcripts belonging to the same tree in (b) but to different paralogous genes, different shades of the same colors are used. For example, in human, the transcripts colored in light purple and in purple are issued from JNK2 and JNK1 and belong to the purple tree in (b). For the sake of clarity, mutations are not indicated along the branches. The lists of transcripts appearing in the different genes are displayed in the top right corner.

identification of the JNK isoforms and the characterization of the structural determinants of their different activities is of paramount importance.

Here, we show how PhyloSofS can be used to provide insight on the contribution of AS to the evolution of the JNK family and we describe the structural determinants of the JNK isoforms' functional differences. By reconstructing the phylogeny of JNK transcripts, we show that the ASE associated to substrate binding affinity modulation appeared in the ancestor common to mammals, amphibians and fishes, before gene duplication. By using molecular modeling techniques, we demonstrate that, despite its important functional outcome, this ASE induces very subtle changes on the protein's internal dynamics. Moreover, our results highlight a set of positively charged and polar residues that may be responsible for substrate molecular recognition specificity. Importantly, we highlight a

bioRxiv preprint doi: https://doi.org/10.1101/119891; this version posted March 23, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

JNK1-specific ASE that has not been documented in the literature. This ASE is of ancient origin, spread across several species, and it induces a large deletion in the protein (about 80 residues). By simulating the dynamical behavior of the resulting isoform in solution, we show that its overall shape and secondary structures remain stable on the time scale of a few hundreds of nanoseconds. We propose that this isoform might be catalytically competent and play a role in the cell.

By crossing sequence-based and structure-based analyses, we show that the 3D structure of the protein and the important regions defined in the litterature are preserved by the 1D structure of the gene (borders of the exons). We also show that the transcripts for which no phylogeny could be reconstructed (orphans) display peculiar properties, indicative of a low stability. They tend to be smaller than the parented ones, the 3D models generated for them are of poorer quality and they have a higher proportion of hydrophobic residues being exposed to the solvent. This result suggests that sequence and structure descriptors can be used to flag transcripts likely non-functional and filter them out early in the phylogenetic reconstruction. These two observations are likely generalizable to other systems.

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 demonstrate that such deep characterization is mandatory in certain cases, in order to unveil the mechanisms underlying AS functional outcome. Our results also open the way to the identification and characterization of new isoforms that may be targeted in the future for medicinal purpose.

bioRxiv preprint doi: https://doi.org/10.1101/119891; this version posted March 23, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Results

PhyloSofS was used to reconstruct the phylogenetic forest relating 60 transcripts from the JNK family observed in 7 species, human, mouse, xenope, fugu, zebrafish, drosophila and nematode. The input data were collected from the Ensembl [23] database (see *Methods*). The algorithm was run for 10^6 iterations and we retained the most parsimonious evolutionary scenario (cost $= 69$, see *Methods* for a detailed description of the parameters). PhyloSofS also generated 3D molecular models for the corresponding protein isoforms, by using homology modeling (see *Methods*). We subsequently performed molecular dynamics simulations of 3 human isoforms, starting from the predicted 3D models. In the following, we describe the analysis of the transcripts' phylogeny, structures and dynamical behavior.

Transcripts' phylogeny for the JNK family.

The forest reconstructed by PhyloSofS is comprised of 7 transcript trees (**Fig. 1b**, each tree is colored differently). The root of a tree corresponds to the appearance of a new transcript in evolution. It indicates the level in the phylogeny where a new ASE occurred, that resulted in the transcripts observed at the leaves of the tree. Dead ends (indicated by triangles) correspond to transcript losses. Each transcript is described as a collection of exons, numbered from *0* to *14* (**Fig. 1b**, top right corner, and see *Methods* for more details on the numbering). Mutations, *i.e.* inclusions or exclusions of exons, occurring along the branches of the trees are labelled (**Fig. 1b**, see $+/-$ symbols followed by the number of the added/removed exon). In total, there are 11 mutations along the JNK transcripts' phylogeny. We also observe 14 orphan leaves (in grey) that correspond to transcripts for which no phylogeny could be reconstructed. These transcripts are not conserved across the studied species, and thus are likely non-functional.

The transcripts' forest is embedded in the gene tree, where each internal node represents an ances-

bioRxiv preprint doi: https://doi.org/10.1101/119891; this version posted March 23, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

tral gene in an ancestral species (**S1 Fig, a**). The sequences of the JNK genes are highly conserved through evolution (**Table I**). The genomes of the two most distant species, namely drosophila and nematode, contain each only one JNK gene. This gene shares a high degree of nucleotidic sequence identity (78% for drosophila, 56% for nematode) 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, the JNK family gene tree (**S1 Fig, a**) can be reconciled with the species tree (**S1 Fig, b**) by hypothesizing that early duplication events led to the creation of JNK2 and JNK3 in the ancestor common to mammals, amphibians and fishes. JNK1 was then further duplicated in fishes while JNK2 was lost in xenope. A representation of the reconstructed transcripts' phylogeny embedded in the species tree is displayed on **Figure 1c**. It permits to appreciate the diversity of transcripts in each species.

Table I: **Percentages of sequence identity between JNK genes.**

	JNK1	JNK2	JNK3
Human	100	100	100
Mouse	99	97	100
Xenope	89	-	98
Fugu	79 82 (a)	81	96
Zebrafish	87 (a) 87 (b)	85	93
Drosophila	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. Xenope 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 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.

bioRxiv preprint doi: https://doi.org/10.1101/119891; this version posted March 23, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

The 7 reconstructed trees relate 12 transcripts observed in human (**Fig. 1b**). Among those, the transcripts colored the same belong to the same tree and share the same exon composition, even if they are issued from different paralogues 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, issued from JNK1, JNK2 and JNK3 (**S1 Fig, c**). 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.

Among the exons composing the JNK transcripts, two pairs, namely *6* and *7*, and *12* and *13*, are mutually exclusive (**S1 Fig, c**). The associated ASEs can be dated early in the phylogeny (**Fig. 1b**), before the gene duplication (**S1 Fig, b**). Exon *7* is already expressed at the root of the forest (**Fig. 1b**, purple tree), while transcripts including exon *6* appear in the ancestor common to mammals, amphibians and fishes (internal node A3, yellow and orange trees). On **Figure 1b**, exon *13* (purple and brown trees) appears before exon *12* (yellow and orange trees). However, the scenario where exon *12* appears before *13* is strictly equivalent (**S8 Fig**, same forest cost = 69). This is explained by the fact that neither drosophila nor nematode contain any of these exons (**Figure 1b**, see mutation of the purple transcript between the root and internal node A2).

New transcripts appear further down the phylogeny (**Fig. 1b**, in pink, green, and red), after the JNK1 to JNK2 and JNK1 to JNK3 gene duplication events (**S1 Fig, b**). They are created in the ancestor common to mammals, amphibians and fishes (**Fig. 1b**). One of them appears in the sub-forest associated to JNK1 (internal node A11, in pink). It features a large deletion (exclusion of exons *6*, *7* and *8*) and its exon composition is perfectly conserved along the phylogeny (no mutation). The two other transcripts are created at the root of the sub-forest associated to JNK3 (ancestor node 10, in green and red). They are characterized by the presence of exons *0* and *1'*, not found in the other paralogues, and they both include exon *6*. Interestingly, all the transcripts containing exon *7*

bioRxiv preprint doi: https://doi.org/10.1101/119891; this version posted March 23, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

(purple and brown trees) die in the same node. Consequently, exon *7* is completely absent from the sub-forest associated to JNK3.

In summary, the analysis of the transcripts' phylogeny inferred by PhyloSofS for JNKs emphasized several characteristics of the evolution of this protein family. First, it revealed a rather low number of mutations, illustrating the fact that the sequences of the JNK genes and their exon sites are highly conserved through evolution. Second, it enabled to date ASEs associated to two pairs of mutually exclusive exons, namely *6* and *7*, and *12* and *13*. Of particular interest is the *6*/*7* pair: the two exons are homologous and were shown to modulate the affinity of JNKs to their cellular substrates [21]. Our phylogenetic reconstruction revealed that the most recent common ancestor of all 7 species contained only one transcript with exon *7*, and that transcripts containing exon *6* appeared in the ancestor common to mammals, amphibians and fishes. Moreover, by analyzing the genomes of drosophila and nematode, we found that exon *6* is absent from them. These observations suggest that exon *6* is issued from the duplication of exon *7* and that this duplication occurred in the ancestor common to mammals, amphibians and fishes, before the duplication of the ancestral JNK gene. Our analysis also highlighted 2 transcripts specific to JNK3 across several species and showed that exon *7* is not expressed in the JNK3 sub-forest. This may suggest a subfunctionalization for JNK3, which is the only paralogue being specifically expressed in certain tissues, namely the heart, brain and testes [16]. Finally it highlighted a transcript lacking exons *6*, *7* and *8* and being specific to JNK1 and its paralogue in Fugu, JNK1a.

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

Eighty structures of human JNKs are available in the Protein Data Bank (PDB) [24], among which 30 for JNK1, 2 for JNK2 and 48 for JNK3 (**S1 Table**). This abundance of structural data can be explained by the fact that JNKs are important therapeutic targets and they were crystallized with

bioRxiv preprint doi: https://doi.org/10.1101/119891; this version posted March 23, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

different inhibitors. The three paralogues share the same fold, which is highly conserved among protein kinases (**Fig. 2**). 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. The activation loop (A-loop on **Fig. 2**, residues 169-195 in JNK1 and JNK2, residues 207-233 in JNK3) displays the highest deviations and comprises residues often unresolved in the PDB structures. The A-loop is found in all kinases and is involved in the control of their activation [25]. The glycine-rich loop (P-loop), the C-helix and the F-helix (labelled in black on **Fig. 2**) are also ubiquitously found in protein kinases and play important roles for their structural stability and/or function [25]. The N-terminal hairpin, the MAPK insert and the C-terminal helix (labelled in grey) are specific to the mitogen-activated protein kinase (MAPK) type, to which the JNKs belong. The catalytic site (green circle), where ATP binds, is located at the junction between the N- and C-terminal lobes. Two regions at the surface of JNKs (indicated by green circles) are known to interact with cellular partners, namely the D-site binding the scaffolding protein JIP-1 [26] and the F-site binding the phosphatase MKP7 [27].

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. The regions labelled in black are common to kinases and were reported in the literature 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 binding sites for JNK cellular partners. The structure was solved by X-ray crystallography at 1.80 Å resolution (PDB code: 3ELJ [28]).

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 [28]) 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 one exon. It

bioRxiv preprint doi: https://doi.org/10.1101/119891; this version posted March 23, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

should be noted that exons *8* and *8'* used in PhyloSofS actually correspond to only one genomic exon (see *Methods*). All the regions important for kinases are preserved (**Fig. 2**, labelled in black), as well as the N-terminal hairpin and the MAPK insert (labelled in grey). By contrast, the catalytic site, the D-site and the F-site (green circles) are comprised of residues belonging to different exons. The precise borders of the exons and the known regions/sites are given in **S2 Table**. Of note, the block formed by exons *1* to *5*, comprising the N-terminal lobe and the A-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 (**S2 Fig, b**). The 3D structures of human JNK1 (**S2 Fig, a**) and drosophila JNK (**S2 Fig, b**) are very similar, with a RMSD of 0.68 Å on 251 over 314 (80%) residues. The JNK gene from the drosophila genome comprises much fewer exons than the human gene, which leads to an even better preservation of the secondary structures and of the known important regions in that species.

This analysis showed that the 1D structure of the JNK genes preserves most of the protein secondary structure elements and most of the regions playing important roles for kinases structural stability and/or function. This is true for human and also for one of the most distant species, namely drosophila. Considering the high degree of conservation of JNK sequences, one may hypothesize that this is a general property across all the studied species. By contrast, the functional binding sites of the protein contain 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.

Previous studies have related the 1D structure of the gene and the 3D structure of the protein. It was shown that compact units in protein structures, namely protein units, tend to overlap the boundaries of single constitutive exons or of co-occurring exon pairs in human [29].

bioRxiv preprint doi: https://doi.org/10.1101/119891; this version posted March 23, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

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). 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's 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 [30] G-factor and Modeller [31] 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).

bioRxiv preprint doi: https://doi.org/10.1101/119891; this version posted March 23, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

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 [30]. **(d)** Normalized DOPE score computed by Modeller [31]. **(e)** Fraction of protein residues being exposed to the solvent ($rsa > 0.25$). **(f)** Fraction of hydrophobic protein residues being exposed to the solvent ($rsa > 0.25$).

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 [21, 22]. From a sequence perspective, the two exons are homologous, highly conserved through evolution, and differ only by a few positions (**S3 Fig**). 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 of protein kinases [32]. In particular, it contains a N-terminal aspartate and 2 hydrophobic residues highly conserved across the whole kinase family. These 3 residues were shown to serve as anchor points for two clusters of hydrophobic residues, namely the catalytic and regulatory spines, essential for kinase activity and regulation [32] (see illustration on the PKA kinase on **S4 Fig, a**). Moreover, the N-terminal aspartate was shown to form hydrogen bonds (H-bonds) with the HRD motif in the catalytic loop and to consequently stabilize the backbone of this motif in a strained conformation characteristic of protein kinase structures and important for their catalytic activity [33] (see illustration on the CDK-substrate complex on **S5 Fig, a**). To sum up, the F-helix is essential for kinase structural stability and some particular residues in this helix are involved in structural features important for kinase catalytic activity and/or regulation. In the following, we will use these known structural features as

bioRxiv preprint doi: https://doi.org/10.1101/119891; this version posted March 23, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

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 between the isoforms including exon *6* and those including exon *7*. The catalytic and regulatory spines, together with their anchors in the F-helix, are present in both types of isoforms (**S4 Fig, b-c**). The HRD motif's strained backbone conformation and the associated H-bond pattern are also observed in both types of isoforms (**S5 Fig, b-c**). The N-terminal aspartate (D207) of the F-helix is 100% conserved in both exons *6* and *7* in the 7 studied species (**S3 Fig**, indicated by an arrow). The two other anchor points are also present, namely I214 and L/M218 (**S3 Fig**, indicated by arrows). 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 [21]. 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 ± 0.28 Å in JNK1α and of 0.98 ± 0.16 Å in JNK1β (**Fig. 4a**). 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 (**S7 Fig, a-b**). Consequently, the observa-

bioRxiv preprint doi: https://doi.org/10.1101/119891; this version posted March 23, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

tions realized on the static 3D models hold true when simulating their dynamical behavior: the *6*/*7* variation does not induce any drastic change.

Nevertheless, an interesting observation can be made regarding the loop following the F-helix: a few residues lying in this loop display very different side-chain flexibilities 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 (**S3 Fig**). 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 (**S3 Fig**). 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.

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 **S6 Fig**). 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.

bioRxiv preprint doi: https://doi.org/10.1101/119891; this version posted March 23, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

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 [32], nor the MAPK insert, involved in the binding of the phosphatase MKP7 [27] (**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 [34]. 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 (**S4 Fig, d**, in red). Moreover, the HRD motif's strained backbone conformation and the associated H-bond pattern, which are stabilized by the aspartate, are maintained (**S5 Fig, d**). By contrast, the catalytic spine lacks its two anchors (**S4 Fig, d**, in yellow). Consequently, despite its lacking of an important and large part of the protein, JNK1δ still possesses some structural features important for kinase catalytic activity and/or regulation.

bioRxiv preprint doi: https://doi.org/10.1101/119891; this version posted March 23, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

JNK1δ was simulated in explicit solvent for 250 ns (5 replicates of 50 ns). The isoform displays stable secondary structures (**S6 Fig**, 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 values 0.88 ± 0.18 Å. This is higher than the values computed for the F-helix in JNK1α and JNK1β (0.57 ± 0.10 Å and 0.53 ± 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 (**S7 Fig, a**, on the right) and the HRD motif's backbone remains in a strained conformation (**S7 Fig, b**, 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.

Alternative transcripts' phylogenies.

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

bioRxiv preprint doi: https://doi.org/10.1101/119891; this version posted March 23, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

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*). Depending on the input data and the set of parameters, it may find several solutions with equivalent costs. Over 10^6 iterations of the program, the forest described above (**Fig. 1b**, or **S8 Fig** with branch swapping), comprising 7 trees, 19 deaths and 14 orphans, was visited 1 219 times. An alternative phylogeny was visited 310 times, that comprises the same number of trees and orphans, but 2 more deaths (**S9 Fig**). The difference between the two forests lies among the fugu JNK1 transcripts, where one transcript belongs to the orange tree (**S9 Fig**) 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 (**S9 Fig**), 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 (**S10 Fig**). 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. Both transcripts are truncated at the C-terminus and 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.

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's homology modeling routine (see *Methods*) did not enable to improve their reconstruction. In fact, the models generated by the threading algorithm are

bioRxiv preprint doi: https://doi.org/10.1101/119891; this version posted March 23, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

very similar to those generated by PhyloSofS.

All the missing exons are predicted to contain some intrinsically disordered regions (**S11 Fig**). At the N-terminus, exons *0* and *1'* contain two segments of about 10 residues predicted as disordered protein-binding regions (**S11 Fig, b**, 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 [35]. These exons are present in only two JNK3 transcript isoforms (**Fig. 1b**, colored in red and green). Considering that JNK3 isoforms are specifically expressed in the heart, brain and testes [21], one can hypothesize that the two exons are involved in interactions with specific cellular partners in these tissues. At the C-terminus, exons *12* or *13* are completely predicted as intrinsically disordered (**S11 Fig, a** and **S11 Fig, b**, blue curve). The functional implication of the inclusion or exclusion of *12*/*13* has not been assessed experimentally [21].

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 [36, 37]. 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.

Importantly, our approach enables to go beyond a mere 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 and we date the ASEs. This is important when

bioRxiv preprint doi: https://doi.org/10.1101/119891; this version posted March 23, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

one wants to study the sequence of ASEs and how it translates in terms of protein structure evolution. Another important aspect is that, in this study, we have inferred the phylogeny of all transcripts observed for the whole JNK family at once. This means that we have directly addressed the issue of pairing transcripts across homologous and paralogous genes between different species, starting from a given reconciled gene tree. This general problem is much more complex than that of inferring the transcripts' phylogeny of each gene separately. We can thus perform an integrated phylogenetic reconstruction that combines creation/loss events at both gene and transcript levels.

The reconstructed phylogenies enable to rapidly and easily identify transcript isoforms conserved during long evolutionary times and thus likely to be functionally important, and/or ASEs specific to one gene of the family. One can then investigate the structural impact of the AS-induced sequence variations on these isoforms by molecular modeling. Characterizing in details their dynamical behavior further permits to get insight into the molecular mechanisms underlying AS-induced functional changes. Such *in silico* analyses provide a way to complement findings from large-scale proteomics and ribosome profiling studies [11, 7, 8] with a mechanistic explanation.

We summarize below our main findings on the JNK family, some of which likely have general applicability.

First, we dated an ASE consisting of two mutually exclusive homologous exons (*6* and *7*) in the ancestor common to mammals, amphibians and fishes. 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. Alternatively spliced homologous exons were recently shown to be highly expressed at the protein level and to have ancient origin, supporting an important cellular role [38].

Second, our analysis highlighted an isoform, JNK1δ, conserved across several species, displaying

bioRxiv preprint doi: https://doi.org/10.1101/119891; this version posted March 23, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

a large deletion (about 80 residues), and not previously described in the literature. It is recorded in the UniProt database [39] (accession id: P45983-5). The APPRIS database v20 [40] annotates it as *minor* and indicates that there are 4 peptides matching the isoform in publicly available proteomics data. By comparison, the human JNK1 isoforms identifed as orphans by our phylogenetic analysis are also annotated as *minor* in the APPRIS database and have between zero and 2 matching peptides. The other human JNK1 isoforms, which possess a phylogeny and are described in the literature [21], are annotated as *alternative* or *principal* and have between 5 and 7 matching peptides. Our analysis showed that JNK1δ remains stable in solution and that its catalytic site is intact. We propose that JNK1δ 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.

Third, we found characteristics specific to the JNK3 isoforms, expressed in the heart, brain and testes. In the phylogeny, we observed that exon *7* is absent from the JNK3 sub-forest. One may wonder whether this could be due to under-annotation of the transcripts. In fact, the genomic sequence of exon *7* is present at the JNK3 locus in all species. Nevertheless, this sequence (exon *7*, JNK3) diverged far more than the other ones (exon *6*, JNK3, and exons *6/7*, JNK1 and JNK2). This observation supports the transcriptomic data used as input and our results. Studies investigating the gain/loss of alternative splice forms associated to gene duplication at large scale [41, 42] have highlighted a wide diversity of cases and have suggested that it depends on the specific cellular context of each gene. By analyzing the structural models, we also observed that two exons (*0* and *1'*) contain regions predicted to be disordered protein-binding regions. This is in agreement with a study linking

bioRxiv preprint doi: https://doi.org/10.1101/119891; this version posted March 23, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

protein-protein interaction networks remodeling with tissue specific AS [43]. The authors showed that tissue-specifically included exons are frequently enriched in intrinsically disordered regions likely to influence protein interactions. These observations call for the development of molecular modeling methods able to correctly handle these regions and predict their partner(s) and their stabilized-upon-binding fold(s).

Under-annotation of transcripts is a potential source of error coming from the input data. It can impact the phylogenetic reconstruction by 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 supposedly non-functional. We expect that these transcripts 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. The present work opens the way to such assessment at large-scale.

To efficiently search the space of possible phylogenies, the algorithm implemented in PhyloSofS relies on a multi-start iterative procedure and on the computation of a lower bound that enables to early eliminate unsuitable candidate solutions (see *Methods*). 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

bioRxiv preprint doi: https://doi.org/10.1101/119891; this version posted March 23, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

gene being estimated at 8.8 in the human genome [44]) 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 [40]). To reduce the computing time, the user can easily parallelize the multi-start iterative search on multiple cores and he/she has the possibility to give as input a previously computed value for the lower bound (to increase the efficiency of the cut). This implementation makes feasible, for the first time, the reconstruction of transcripts' phylogenies for any gene family.

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 medicinal research.

Methods

PhyloSofS workflow

PhyloSofS takes as input 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

bioRxiv preprint doi: https://doi.org/10.1101/119891; this version posted March 23, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

species (symbols at the leaves). PhyloSofS comprises two main steps:

a. It reconstructs a forest of phylogenetic trees describing plausible evolutionary scenarios that can explain the observed transcripts by using the maximum parsimony principle (**Fig. 1a**, on the right). The forest is embedded in the input gene tree. The leaves of each tree correspond to a subset of the observed transcripts (one transcript at every leaf of every tree). The root of a tree corresponds to the creation of a new transcript while dead ends (indicated by triangles on **Fig. 1a**, on the right) correspond to transcript losses. Transcripts can mutate along the branches of the trees.

b. It predicts the three-dimensional structures of the protein isoforms corresponding to the observed transcripts by using homology modeling. The molecular models are then annotated with quality measures. For each isoform, the exons composing it are mapped onto its 3D structural model.

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.

Step a. Transcripts' phylogenies reconstruction

For simplicity, we describe here the case where only one gene of interest is studied across several species. Nevertheless, PhyloSofS can reconstruct phylogenies for several genes from the same family, as exemplified by its application to the JNK family.

Evolution model. PhyloSofS models transcript evolution as a two-level process. The first level corresponds to the gene structure, where the status (absent, alternative or constitutive) of each exon is determined, while the second level corresponds to the transcripts, where the presence or absence of each exon is determined for each transcript. Modification of the gene structure affects the set

bioRxiv preprint doi: https://doi.org/10.1101/119891; this version posted March 23, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

of transcripts that can be expressed, but modification of the transcripts does not affect the gene structure. Three evolutionary events are considered, namely creation of a transcript, death of a transcript and mutation of a transcript, and three associated costs are defined, C_B, C_D and σ (Table II). This model is inspired by a previous work [14].

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

child/parent	(0,0)	(0,1)	(1,1)	(1,2)	state of the exon e in the child transcript t_i^s of species s
(0,0)	0	0	0	0	excluded from t_i^s and from all transcripts of s
(0,1)	0	0	σ	σ	excluded from t_i^s but included in some transcript(s) of s
(1,1)	σ	σ	0	0	included in t_i^s but excluded from some transcript(s) of s
(1,2)	0	σ	0	0	included in t_j^s and in all transcripts of s

Only evolutionary changes taking place at the level of the transcript are taken into consideration.

Input data. The input consists in a gene tree with the observed transcripts at the leaves (**Fig. 5a**). 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 (colored squares on **Fig. 5a**), 0 if it is excluded (white squares).

bioRxiv preprint doi: https://doi.org/10.1101/119891; this version posted March 23, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Figure 5: **Workflow of the transcripts' phylogeny reconstruction algorithm.** **(a)** A binary tree representing the phylogeny of the gene(s) of interest is given as input, along with the transcripts observed at the leaves (symbols). Each transcript is described as a collection of exons, each exon being colored differently (white means that the exon is absent from the transcript). **(b)** The first step consists in determining the states of the exons at the level of the gene, either absent (white square), alternative (black/white square) or constitutive (black square). To determine the exon states at the internal nodes, Sankoff's algorithm and Dollo's parsimony principle are used. **(c-d)** The algorithm then proceeds iteratively by searching the space of possible forest structures (c) and evaluating the phylogeny of minimum cost for each chosen structure (d). **(c)** A forest structure S_i is fixed by setting the number of binary (with two children), left (with one left child) and right (with one right child) subnodes at each internal node. **(d)** The phylogeny φ_i associated to the forest S_i is computed only if the cost associated to S_i, which depends on the number of transcript births and deaths, is lower than the cost C_{min} of the best phylogeny found so far. At this stage, each transcript is represented by a table of costs, where each line corresponds to an exon and each column corresponds to an exon state. There are four possible states: absent (white square), alternative absent (grey/white square), alternative present (black/grey square) and present (black square). Only the cells permitted by the exon states at the gene level (determined in a) are considered. Sankoff's algorithm is used bottom up to compute the minimal pairing costs (see Table II for the list of elementary mutation costs). At each internal node, the problem of pairing the children transcripts is that of a partial assignment and is solved by using a branch-and-bound algorithm (see inserted table on the left: the chosen pairs are those with minimum costs and compatible, and *Supplementary text S1*). The total cost associated to mutations along the branches is obtained by summing the costs over all tables, where the cost of each table is the sum of the minimum costs determined for each line (exon). The cost associated to each observed transcript (leaf) is obviously zero.

Exon states at the gene level. 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 (**Fig. 5b**, white, black/white and black squares). At the leaves (current species), the components of g^s are calculated as:

$$g_i^s = \prod_{j=1}^{n_s} T_{i,j}^s + 1 - \prod_{j=1}^{n_s} (1 - T_{i,j}^s) \tag{1}$$

The g^s vectors for internal nodes (ancestral species) are determined by using Sankoff's algorithm [45]. Dollo's parsimony principle is also respected, such that an exon cannot be created twice [46]. If different exon states have equal cost, we follow the priority rule $2 > 0 > 1$.

Forest structure. 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 (**Fig.**

bioRxiv preprint doi: https://doi.org/10.1101/119891; this version posted March 23, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

5c). There exist three types of subnodes: binary (two transcript 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:

$$C_{birth}(S) = C_B \times |S| \text{ with } |S| \text{ the number of trees in the forest,} \tag{2}$$

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

Transcripts' phylogeny. A transcripts' phylogeny determines the pairings of transcripts at each level of the forest structure (**Fig. 5d**). The cost of the phylogeny φ complying with the 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 x^{th} transcript of node y. The evolution costs σ are given in Table II.

Detailed algorithm. PhyloSofS's algorithm seeks to determine the scenario with the smallest number of evolutionary events, *i.e.* the transcripts' phylogeny with the minimum cost (**Fig. 5c-d**). It proceeds as follows:

bioRxiv preprint doi: https://doi.org/10.1101/119891; this version posted March 23, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

```
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 $\varphi_i$ 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 (**Fig. 5c**), 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 (**Fig. 5d**, 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 [47] (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

bioRxiv preprint doi: https://doi.org/10.1101/119891; this version posted March 23, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

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 (**Fig. 5d**). 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 (**S12 Fig**). 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).

Visualization. PhyloSofS generates PDF files displaying the computed transcripts’ phylogenies using a Python driver to the Graphviz [48] DOT format.

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 Protein Data Bank (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.

bioRxiv preprint doi: https://doi.org/10.1101/119891; this version posted March 23, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Search for templates. Step 1 makes extensive use of the HH-suite [49] and can be decomposed in: (a) search for homologous sequences and building of a multiple sequence alignment (MSA), by using HHblits [50], (b) addition of secondary structure predictions, obtained by PSIPRED [51], 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 [52].

3D model building. Step 3 is performed by Modeller [31] with default options.

Annotation of the models. 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 [51], (c) assessing the quality of the model with Procheck [30] and with the normalized DOPE score from Modeller, (d) determining the by-residue solvent accessible surface areas with Naccess [53] and computing the proportions of surface residues and of hydrophobic surface residues.

Application of PhyloSofS to the JNK family

Retrieval and pre-processing of transcriptome data. The peptide sequences of all splice variants from the JNK family observed in human, mouse, xenope, zebrafish, fugu, drosophila and nematode were retrieved from Ensembl [23] 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 homologous exons between the different genes in the different species were identified by aligning the sequences with MAFFT [54], and projecting the alignment on the human annotation. The isoforms resulting in the same amino acid sequence were merged. In total, 64 transcripts comprised of 38 exons were given as input to PhyloSofS.

bioRxiv preprint doi: https://doi.org/10.1101/119891; this version posted March 23, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Exon numbering. The set of homologous exons used in PhyloSofS were defined so as to account for all the variations occurring between the observed transcripts in any species. They do not necessarily represent exons definition based on the genomic sequence, for two reasons. First, the structure of the genes may be different from one species to another. For instance, the third and fourth exons of human JNK1 genes are completely covered by only one exon in the drosophila JNK gene (**S2 Fig**). In that case, we keep the highest level of resolution and define two exons (*3* and *4*). Second, 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'*.

Reconstruction of the transcripts' phylogeny. 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.

Generation of the 3D models. The 3D models of all observed isoforms were generated by PhyloSofS's molecular modeling routine by setting the number of retained best templates to 5 (default

bioRxiv preprint doi: https://doi.org/10.1101/119891; this version posted March 23, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

parameter) for every isoform.

Analysis of JNK tertiary structures.

The list of experimental structures deposited in the PDB for the human JNKs was retrieved from UniProt [39]. The structures were aligned with PyMOL [55] 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 [26]) and the catalytic domain of MKP7 (PDB code: 4YR8, resolution: 2.4 Å [27]), respectively. They were detected as displaying a change in relative solvent accessibility >1 Å^2 upon binding.

The I-TASSER webserver [56, 57, 58] was used to try and model the regions for which no structural templates could be found. DISOPRED [59] and IUPred [60] were used to predict intrinsic disorder. JET2 [61] was used to predict binding sites at the surface of the isoforms.

Molecular dynamics simulations of human isoforms.

Set up of the systems. 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 [62], 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.

bioRxiv preprint doi: https://doi.org/10.1101/119891; this version posted March 23, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Minimization, heating and equilibration. 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 [63] 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) [64] to maintain the pressure. Finally the restraints were removed and the system was equilibrated for a final 100 *ps* run.

Production of the trajectories. 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 [65] was employed to treat long-range electrostatics. The coordinates of the system were written every ps.

Analysis of the trajectories 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.

bioRxiv preprint doi: https://doi.org/10.1101/119891; this version posted March 23, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

For each residue, the most frequent secondary structure type was retained (see **Fig. 4a** and **S6 Fig**). 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.

Acknowledgments We thank Y. Christinat for providing information on the algorithm he developed for the reconstruction of transcript phylogenies.

bioRxiv preprint doi: https://doi.org/10.1101/119891; this version posted March 23, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

References

1. Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, et al. Alternative isoform regulation in human tissue transcriptomes. Nature. 2008 Nov;456(7221):470–476.

2. Ward AJ, Cooper TA. The pathobiology of splicing. J Pathol. 2010 Jan;220(2):152–163.

3. Lim KH, Ferraris L, Filloux ME, Raphael BJ, Fairbrother WG. Using positional distribution to identify splicing elements and predict pre-mRNA processing defects in human genes. Proc Natl Acad Sci USA. 2011 Jul;108(27):11093–11098.

4. Mudge JM, Frankish A, Fernandez-Banet J, Alioto T, Derrien T, Howald C, et al. The origins, evolution, and functional potential of alternative splicing in vertebrates. Mol Biol Evol. 2011 Oct;28(10):2949–2959.

5. Barbosa-Morais NL, Irimia M, Pan Q, Xiong HY, Gueroussov S, Lee LJ, et al. The evolutionary landscape of alternative splicing in vertebrate species. Science. 2012 Dec;338(6114):1587–1593.

6. Merkin J, Russell C, Chen P, Burge CB. Evolutionary dynamics of gene and isoform regulation in Mammalian tissues. Science. 2012 Dec;338(6114):1593–1599.

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

8. Ezkurdia I, Rodriguez JM, Carrillo-de Santa Pau E, Vazquez J, Valencia A, Tress ML. Most highly expressed protein-coding genes have a single dominant isoform. J Proteome Res. 2015 Apr;14(4):1880–1887.

9. Kim MS, Pinto SM, Getnet D, Nirujogi RS, Manda SS, Chaerkady R, et al. A draft map of the human proteome. Nature. 2014 May;509(7502):575–581.

bioRxiv preprint doi: https://doi.org/10.1101/119891; this version posted March 23, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

10. Hao Y, Colak R, Teyra J, Corbi-Verge C, Ignatchenko A, Hahne H, et al. Semi-supervised Learning Predicts Approximately One Third of the Alternative Splicing Isoforms as Functional Proteins. Cell Rep. 2015 Jul;12(2):183–189.

11. Weatheritt RJ, Sterne-Weiler T, Blencowe BJ. The ribosome-engaged landscape of alternative splicing. Nat Struct Mol Biol. 2016 Dec;23(12):1117–1123.

12. Birzele F, Kuffner R, Meier F, Oefinger F, Potthast C, Zimmer R. ProSAS: a database for analyzing alternative splicing in the context of protein structures. Nucleic Acids Res. 2008 Jan;36(Database issue):D63–68.

13. Birzele F, Csaba G, Zimmer R. Alternative splicing and protein structure evolution. Nucleic Acids Res. 2008 Feb;36(2):550–558.

14. Christinat Y, Moret BM. Inferring transcript phylogenies. BMC Bioinformatics. 2012 Jun;13 Suppl 9:S1.

15. Manning AM, Davis RJ. Targeting JNK for therapeutic benefit: from junk to gold? Nat Rev Drug Discov. 2003 Jul;2(7):554–565.

16. Kyriakis JM, Avruch J. Mammalian MAPK signal transduction pathways activated by stress and inflammation: a 10-year update. Physiol Rev. 2012 Apr;92(2):689–737.

17. Hirosumi J, Tuncman G, Chang L, Gorgun CZ, Uysal KT, Maeda K, et al. A central role for JNK in obesity and insulin resistance. Nature. 2002 Nov;420(6913):333–336.

18. Hunot S, Vila M, Teismann P, Davis RJ, Hirsch EC, Przedborski S, et al. JNK-mediated induction of cyclooxygenase 2 is required for neurodegeneration in a mouse model of Parkinson's disease. Proc Natl Acad Sci USA. 2004 Jan;101(2):665–670.

bioRxiv preprint doi: https://doi.org/10.1101/119891; this version posted March 23, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

19. Brecht S, Kirchhof R, Chromik A, Willesen M, Nicolaus T, Raivich G, et al. Specific pathophysiological functions of JNK isoforms in the brain. Eur J Neurosci. 2005 Jan;21(2):363–377.

20. Tuncman G, Hirosumi J, Solinas G, Chang L, Karin M, Hotamisligil GS. Functional in vivo interactions between JNK1 and JNK2 isoforms in obesity and insulin resistance. Proc Natl Acad Sci USA. 2006 Jul;103(28):10741–10746.

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

22. Bogoyevitch MA, Kobe B. Uses for JNK: the many and varied substrates of the c-Jun N-terminal kinases. Microbiol Mol Biol Rev. 2006 Dec;70(4):1061–1095.

23. Yates A, Akanni W, Amode MR, Barrell D, Billis K, Carvalho-Silva D, et al. Ensembl 2016. Nucleic Acids Res. 2016 Jan;44(D1):D710–716.

24. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, et al. The Protein Data Bank. Nucleic Acids Res. 2000 Jan;28(1):235–242.

25. Huse M, Kuriyan J. The conformational plasticity of protein kinases. Cell. 2002 May;109(3):275–282.

26. Heo YS, Kim SK, Seo CI, Kim YK, Sung BJ, Lee HS, et al. Structural basis for the selective inhibition of JNK1 by the scaffolding protein JIP1 and SP600125. EMBO J. 2004 Jun;23(11):2185–2195.

27. Liu X, Zhang CS, Lu C, Lin SC, Wu JW, Wang ZX. A conserved motif in JNK/p38-specific MAPK phosphatases as a determinant for JNK1 recognition and inactivation. Nat Commun. 2016;7:10879.

bioRxiv preprint doi: https://doi.org/10.1101/119891; this version posted March 23, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

28. Chamberlain SD, Redman AM, Wilson JW, Deanda F, Shotwell JB, Gerding R, et al. Optimization of 4,6-bis-anilino-1H-pyrrolo[2,3-d]pyrimidine IGF-1R tyrosine kinase inhibitors towards JNK selectivity. Bioorg Med Chem Lett. 2009 Jan;19(2):360–364.

29. Gelly JC, Lin HY, de Brevern AG, Chuang TJ, Chen FC. Selective constraint on human pre-mRNA splicing by protein structural properties. Genome Biol Evol. 2012;4(9):966–975.

30. Laskowski RA, MacArthur MW, Moss DS, Thornton JM. PROCHECK: a program to check the stereochemical quality of protein structures. Journal of Applied Crystallography. 1993 Apr;26(2):283–291. Available from: http://dx.doi.org/10.1107/s0021889892009944.

31. Marti-Renom MA, Stuart AC, Fiser A, Sanchez R, Melo F, Sali A. Comparative protein structure modeling of genes and genomes. Annu Rev Biophys Biomol Struct. 2000;29:291–325.

32. Kornev AP, Haste NM, Taylor SS, Eyck LF. Surface comparison of active and inactive protein kinases identifies a conserved activation mechanism. Proc Natl Acad Sci USA. 2006 Nov;103(47):17783–17788.

33. Oruganty K, Talathi NS, Wood ZA, Kannan N. Identification of a hidden strain switch provides clues to an ancient structural mechanism in protein kinases. Proc Natl Acad Sci USA. 2013 Jan;110(3):924–929.

34. Nicolas A, Raguenes-Nicol C, Ben Yaou R, Ameziane-Le Hir S, Cheron A, Vie V, et al. Becker muscular dystrophy severity is linked to the structure of dystrophin. Hum Mol Genet. 2015 Mar;24(5):1267–1279.

35. Meszaros B, Simon I, Dosztanyi Z. Prediction of protein binding regions in disordered proteins. PLoS Comput Biol. 2009 May;5(5):e1000376.

bioRxiv preprint doi: https://doi.org/10.1101/119891; this version posted March 23, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

36. Reyes A, Anders S, Weatheritt RJ, Gibson TJ, Steinmetz LM, Huber W. Drift and conservation of differential exon usage across tissues in primate species. Proc Natl Acad Sci USA. 2013 Sep;110(38):15377–15382.

37. Melamud E, Moult J. Stochastic noise in splicing machinery. Nucleic Acids Res. 2009 Aug;37(14):4873–4886.

38. Abascal F, Ezkurdia I, Rodriguez-Rivas J, Rodriguez JM, del Pozo A, Vazquez J, et al. Alternatively Spliced Homologous Exons Have Ancient Origins and Are Highly Expressed at the Protein Level. PLoS Comput Biol. 2015 Jun;11(6):e1004325.

39. Bateman A, Martin MJ, O'Donovan C, Magrane M, Apweiler R, Alpi E, et al. UniProt: a hub for protein information. Nucleic Acids Res. 2015 Jan;43(Database issue):D204–212.

40. Rodriguez JM, Maietta P, Ezkurdia I, Pietrelli A, Wesselink JJ, Lopez G, et al. APPRIS: annotation of principal and alternative splice isoforms. Nucleic Acids Res. 2013 Jan;41(Database issue):D110–117.

41. Abascal F, Tress ML, Valencia A. The evolutionary fate of alternatively spliced homologous exons after gene duplication. Genome Biol Evol. 2015 Apr;7(6):1392–1403.

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

43. Ellis JD, Barrios-Rodiles M, Colak R, Irimia M, Kim T, Calarco JA, et al. Tissue-specific alternative splicing remodels protein-protein interaction networks. Mol Cell. 2012 Jun;46(6):884–892.

bioRxiv preprint doi: https://doi.org/10.1101/119891; this version posted March 23, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

44. Sakharkar MK, Chow VT, Kangueane P. Distributions of exons and introns in the human genome. In Silico Biol (Gedrukt). 2004;4(4):387–393.

45. Sankoff D. Minimal Mutation Trees of Sequences. SIAM Journal on Applied Mathematics. 1975;28(1):35–42. Available from: `http://dx.doi.org/10.1137/0128004`.

46. Alekseyenko AV, Lee CJ, Suchard MA. Wagner and Dollo: a stochastic duet by composing two parsimonious solos. Syst Biol. 2008 Oct;57(5):772–784.

47. Land AH, Doig AG. An Automatic Method of Solving Discrete Programming Problems. Econometrica. 1960;28:497–520.

48. Gansner ER, North SC. An open graph visualization system and its applications to software engineering. SOFTWARE - PRACTICE AND EXPERIENCE. 2000;30(11):1203–1233.

49. Hildebrand A, Remmert M, Biegert A, Soding J. Fast and accurate automatic structure prediction with HHpred. Proteins. 2009;77 Suppl 9:128–132.

50. Remmert M, Biegert A, Hauser A, Soding J. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. Nat Methods. 2011 Dec;9(2):173–175.

51. Jones DT. Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol. 1999 Sep;292(2):195–202.

52. Soding J. Protein homology detection by HMM-HMM comparison. Bioinformatics. 2005 Apr;21(7):951–960.

53. Hubbard S, Thornton J; 1992-6. `http://www.bioinf.manchester.ac.uk/naccess/`.

54. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 2013 Apr;30(4):772–780.

bioRxiv preprint doi: https://doi.org/10.1101/119891; this version posted March 23, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

55. DeLano WL. The PyMOL Molecular Graphics System; 2002. Http://www.pymol.org.

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

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

58. Zhang Y. I-TASSER server for protein 3D structure prediction. BMC Bioinformatics. 2008 Jan;9:40.

59. Ward JJ, McGuffin LJ, Bryson K, Buxton BF, Jones DT. The DISOPRED server for the prediction of protein disorder. Bioinformatics. 2004 Sep;20(13):2138–2139.

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

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

62. Case D, Darden T, Cheatham III T, Simmerling C, Wang J, Duke R, et al. AMBER 12. University of California, San Francisco. 2012;1(2):3.

63. Berendsen HJC, Postma JPM, van Gunsteren WF, DiNola A, Haak JR. Molecular dynamics with coupling to an external bath. The Journal of chemical physics. 1984;81(8):3684–3690.

64. Loncharich RJ, Brooks BR, Pastor RW. Langevin dynamics of peptides: The frictional dependence of isomerization rates of N-acetylalanyl-N'-methylamide. Biopolymers. 1992;32(5):523–535.

bioRxiv preprint doi: https://doi.org/10.1101/119891; this version posted March 23, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

65. Darden T, York D, Pedersen L. Particle mesh Ewald: An Nlog(N) method for Ewald sums in large systems. The Journal of Chemical Physics. 1993;98:10089–10092.

66. Chimnaronk S, Sitthiroongruang J, Srisucharitpanit K, Srisaisup M, Ketterman AJ, Boonserm P. The crystal structure of JNK from Drosophila melanogaster reveals an evolutionarily conserved topology with that of mammalian JNK proteins. BMC Struct Biol. 2015 Sep;15:17.

67. Knighton DR, Zheng JH, Ten Eyck LF, Ashford VA, Xuong NH, Taylor SS, et al. Crystal structure of the catalytic subunit of cyclic adenosine monophosphate-dependent protein kinase. Science. 1991 Jul;253(5018):407–414.

68. Brown NR, Noble ME, Endicott JA, Johnson LN. The structural basis for specificity of substrate and recruitment peptides for cyclin-dependent kinases. Nat Cell Biol. 1999 Nov;1(7):438–443.

bioRxiv preprint doi: https://doi.org/10.1101/119891; this version posted March 23, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Supporting Information Captions

S1 Text

S1 Table

3D structures of human JNKs deposited in the Protein Data Bank.

S2 Table

Overlap between exons and known regions in human JNK tertiary structure. For each exon, the corresponding residue range in human JNK tertiary structure is indicated, along with the known region(s) overlapping with the exon and the residues from this(ese) region(s) being included in the exon. Exons for which no residue range is indicated (-) map to disordered parts. The horizontal lines separate actual exons (exons *1* and *8* were splitted in *1*, *1'* and *8*, *8'* to model the transcripts in PhyloSofS, see *Methods*).

S1 Fig

Gene tree, species tree and parented human isoforms. (a) Gene tree for the JNK family, comprising JNK1 (in black), JNK2 (in grey) and JNK3 (in light grey), in 7 species: human, mouse, xenope, fugu, zebrafish, drosophila and nematode. Each node represents a gene from a current or ancestral species. The internal nodes are numbered and the leaves are labelled. The fishes contain two paralogs of JNK1 each, named JNK1 and JNK1a in fugu, JNK1a and JNK1b in zebrafish. Drosophila and nematode contain only one JNK gene and are used as outgroups. **(b)** Mapping of the duplication, loss and speciation events for the JNK family onto the phylogeny of the 7 studied species. The genes being duplicated or lost are indicated on the corresponding branches. **(c)** On top, a simplified representation of the JNK genes structure is displayed. Below, the human isoforms for which a phylogeny could be reconstructed (see **Fig. 1b**) are listed and the gene(s) producing each isoform is(are) indicated on the left. Rectangles represent exons and are labelled from 0 to 13.

bioRxiv preprint doi: https://doi.org/10.1101/119891; this version posted March 23, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Exons colored in black in the gene structure (on top) are present in all the listed isoforms (below) and those in gray are present in only a subset of the isoforms. The splicing paths corresponding to the listed isoforms are highlighted in colors.

S2 Fig

Comparison of exon mapping onto JNK tertiary structure between human and drosophila. The structures of JNK1 from human (**a**, PDB code: 3ELJ [28]) and JNK from drosophila (**b**, PDB code: 5AWM [66]) are represented as cartoons. The different exons are colored from blue through white to red. The residues in yellow are at the junction of 2 exons. The exon numbers next to the color strip correspond to those used in PhyloSofS (see *Methods*).

S3 Fig

Multiple sequence alignments of the exons *6* and *7*. The colors indicate the physico-chemical properties of the amino acids: hydrophobic (AFILMPVW) in red, negatively charged (DE) in blue, positively charged (KR) in magenta, polar or special (CGHNQSTY) in green. The symbols at the bottom of each alignment give the conservation degree of the column: $\star$ for completely conserved, **:** for conserved physico-chemical properties, **.** for variable, and void for highly variable. The numbering on top corresponds to JNK1 and JNK2. The arrows indicate anchor residues found in all kinases.

S4 Fig

Catalytic and regulatory spines. **(a)** The PKA kinase (PDB code: 2CPK [67]) is taken as a reference to illustrate the spines identified in all kinases in [32]. **(b-d)** 3D models for the human isoforms JNK1α (b), JNK1β (c) and JNK1δ (d). The residues comprising the spines are displayed as sticks and transparent surfaces and are labelled. The catalytic spine is in yellow. It is anchored by two hydrophobic residues located in the F-helix (in all panels, except for d, JNK1δ). The regulatory spine is in red. It is anchored by the N-terminal aspartate of the F-helix, namely D220 in PKA (a)

bioRxiv preprint doi: https://doi.org/10.1101/119891; this version posted March 23, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

and D207 in JNK (b-c) that forms an H-bond with Y164 (a) or H149 (b-c). In JNK1δ, the aspartate is replaced by D282 and the H-bond with H149 is maintained (d).

S5 Fig

Hydrogen-bond pattern associated to the HRD motif backbone strain. (a) The CDK-substrate complex (PDB code: 1QMZ [68]) is taken as a reference to illustrate the hidden strain identified in kinases structures in [33]. **(b-d)** 3D models for the human isoforms JNK1α (b), JNK1β (c) and JNK1δ (d). Hydrogen-bonds are formed between 3 components: the aspartate in the F-helix, the HRD motif in the catalytic loop and the DFG motif in the activation loop. The 3 component are conserved across the whole protein kinase family.

S6 Fig

Secondary structures for the human JNK1 isoforms. The secondary structures were recorded along MD simulations of human JNK1α (on top), JNK1β (in the middle) and JNK1δ (at the bottom). For each residue, the most frequent secondary structure type is depicted. If a residue does not adopt a well-defined secondary structure, either α-helix or β-sheet, for more than 50% of the MD conformations, then it is assigned to a loop.

S7 Fig

H-bond pattern and backbone strain of the HRD motif. (a) Persistence of the H-bonds formed between H149 and R150 from the HRD motif in the catalytic loop and D207 from the F-helix and D169 from the DFG motif in the catalytic site, recorded along the 5 replicates of MD simulations for JNK1α, JNK1β and JNK1δ. **(b)** Ramachandran plot showing the torsion angle values (phi, psi) of R150 from the HRD motif in the MD conformations of the 3 isoforms.

bioRxiv preprint doi: https://doi.org/10.1101/119891; this version posted March 23, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

S8 Fig

Transcripts' phylogeny reconstructed by PhyloSofS for the JNK family. The forest is comprised of 7 phylogenetic trees, 19 deaths and 14 orphan leaves. The cost of the phylogeny is 69 (with $C_B = 3$, $C_D = 0$ and $\sigma = 2$). The legend is the same as in **Fig. 1b**.

S9 Fig

Transcripts' phylogeny reconstructed by PhyloSofS for the JNK family. The forest is comprised of 7 phylogenetic trees, 21 deaths and 14 orphan leaves. The cost of the phylogeny is 69 (with $C_B = 3$, $C_D = 0$ and $\sigma = 2$). The legend is the same as in **Fig. 1b**. Compared to **Figure 1b**, there is one additional subnode (in orange) in the internal nodes A18 and A24 and one JNK1 transcript in fugu is the leaf of the orange tree instead of the yellow one.

S10 Fig

Transcripts' phylogeny reconstructed by PhyloSofS for the JNK family. The forest is comprised of 8 phylogenetic trees, 23 deaths and 13 orphan leaves. The cost of the phylogeny is 69 (with $C_B = 3$, $C_D = 0$ and $\sigma = 2$). The legend is the same as in **Fig. 1b**. Compared to **Fig. 1b**, there is one additional tree (in blue, exon composition on the right). The murine JNK3 transcript serving as a leaf for this tree was previously an orphan.

S11 Fig

Prediction of intrinsic disorder in JNK isoforms. The predictions were performed with DISOPRED [59] (a) and with IUPred [60] (b) on the 2 human JNK3 isoforms colored in green and red on **Figure 1b**, differing by the inclusion/exclusion of exons *12* or *13*. Exon numbers are indicated on top of the plots. The predictions for exon *13* were added on the right of the 2 plots.

bioRxiv preprint doi: https://doi.org/10.1101/119891; this version posted March 23, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

S12 Fig

State diagram illustrating the 12 possible elementary operations that can be applied to a forest structure. Each elementary operation consists in removing and/or adding subnode(s) to a randomly chosen node in the forest. Each subnode corresponds to a transcript and is represented by a round. It is colored according to the node to which it belongs: the chosen node is in black while its left and right children are colored in orange and green respectively. Deaths are displayed as crosses on the branches. For each transition between 2 states, represented by an arrow, the numbers of binary, left and right subnodes being added or removed are indicated in parenthesis.

bioRxiv preprint doi: https://doi.org/10.1101/119891; this version posted March 23, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Figures

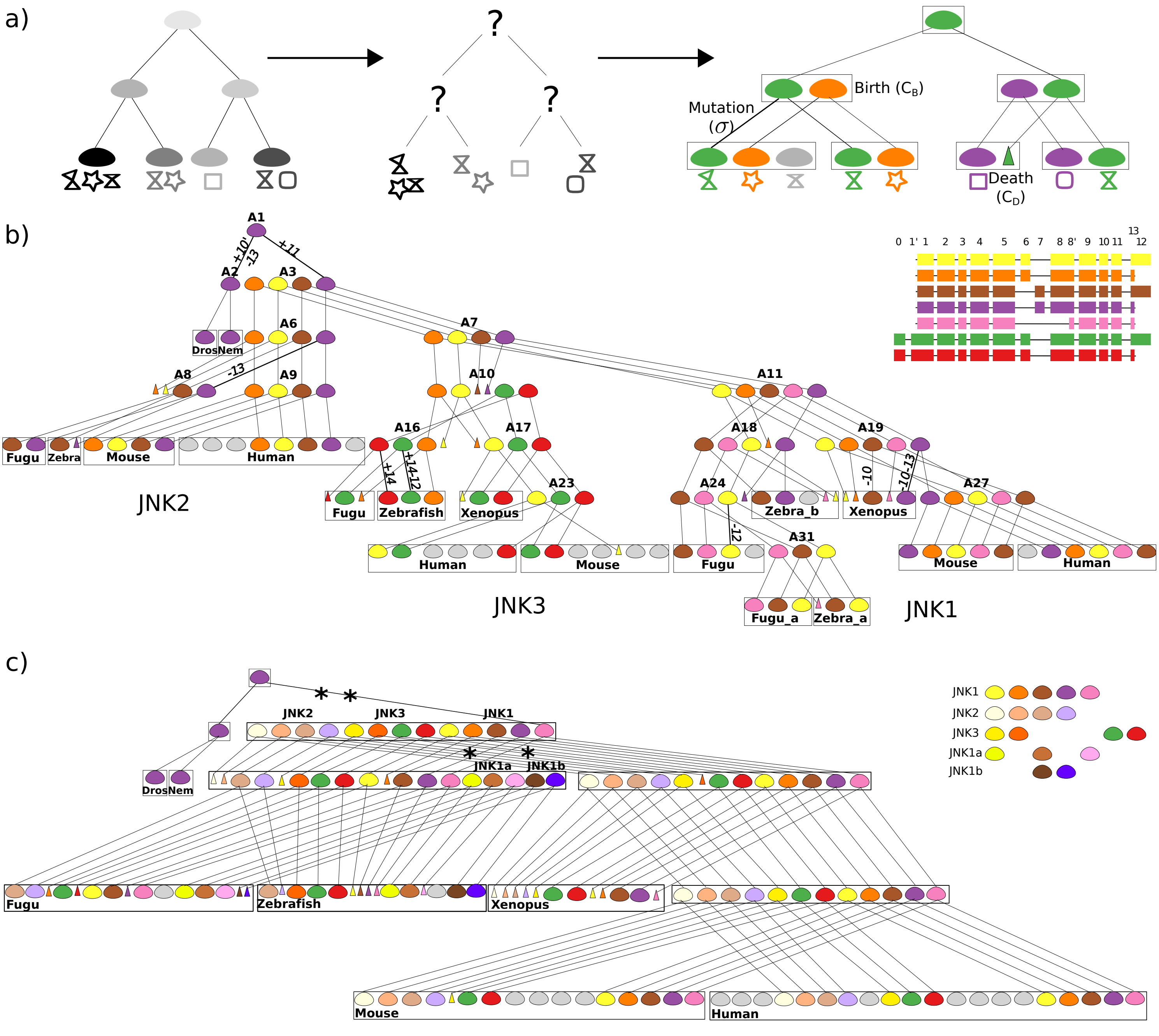

Figure 1:

bioRxiv preprint doi: https://doi.org/10.1101/119891; this version posted March 23, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

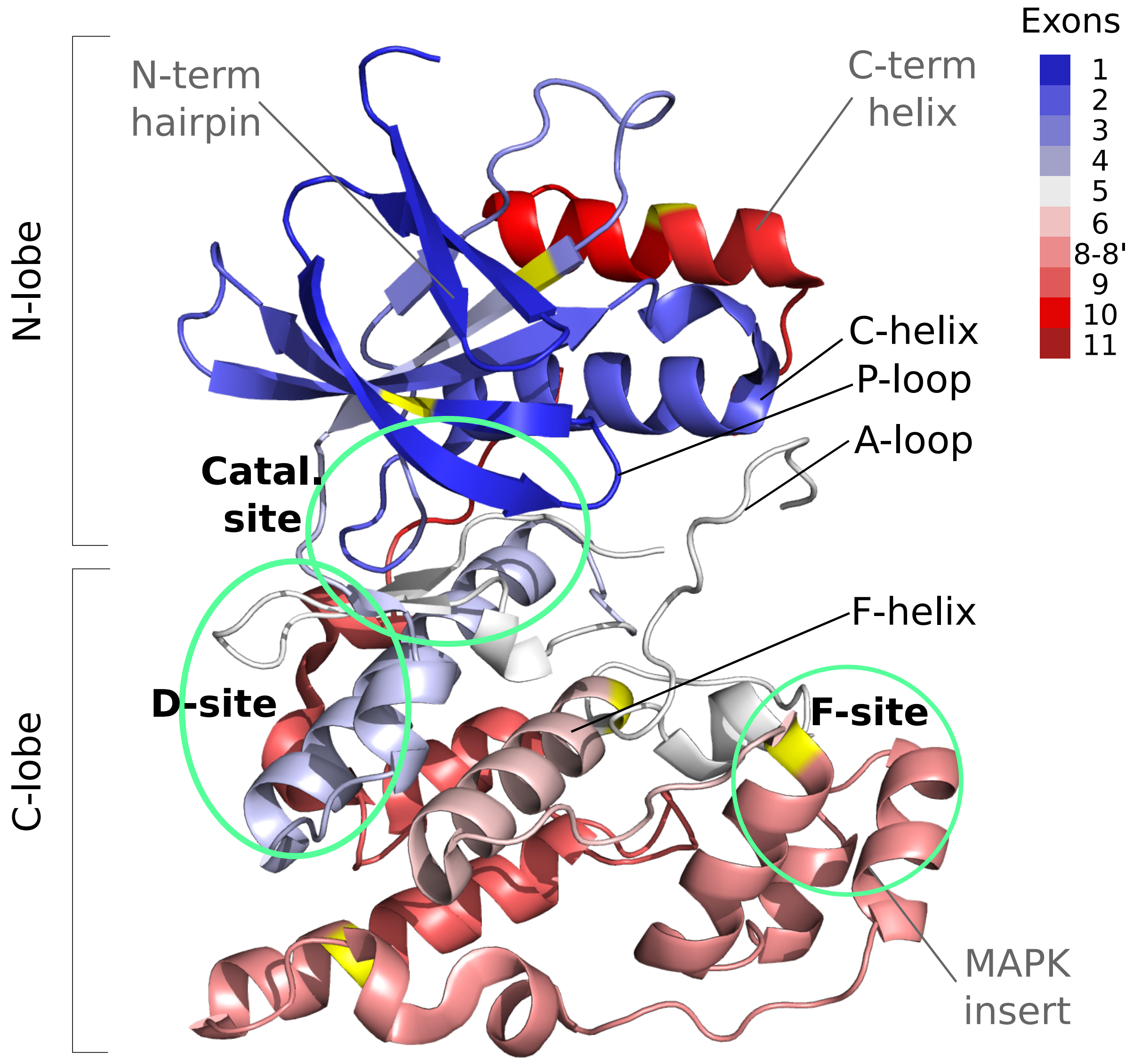

Figure 2:

bioRxiv preprint doi: https://doi.org/10.1101/119891; this version posted March 23, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

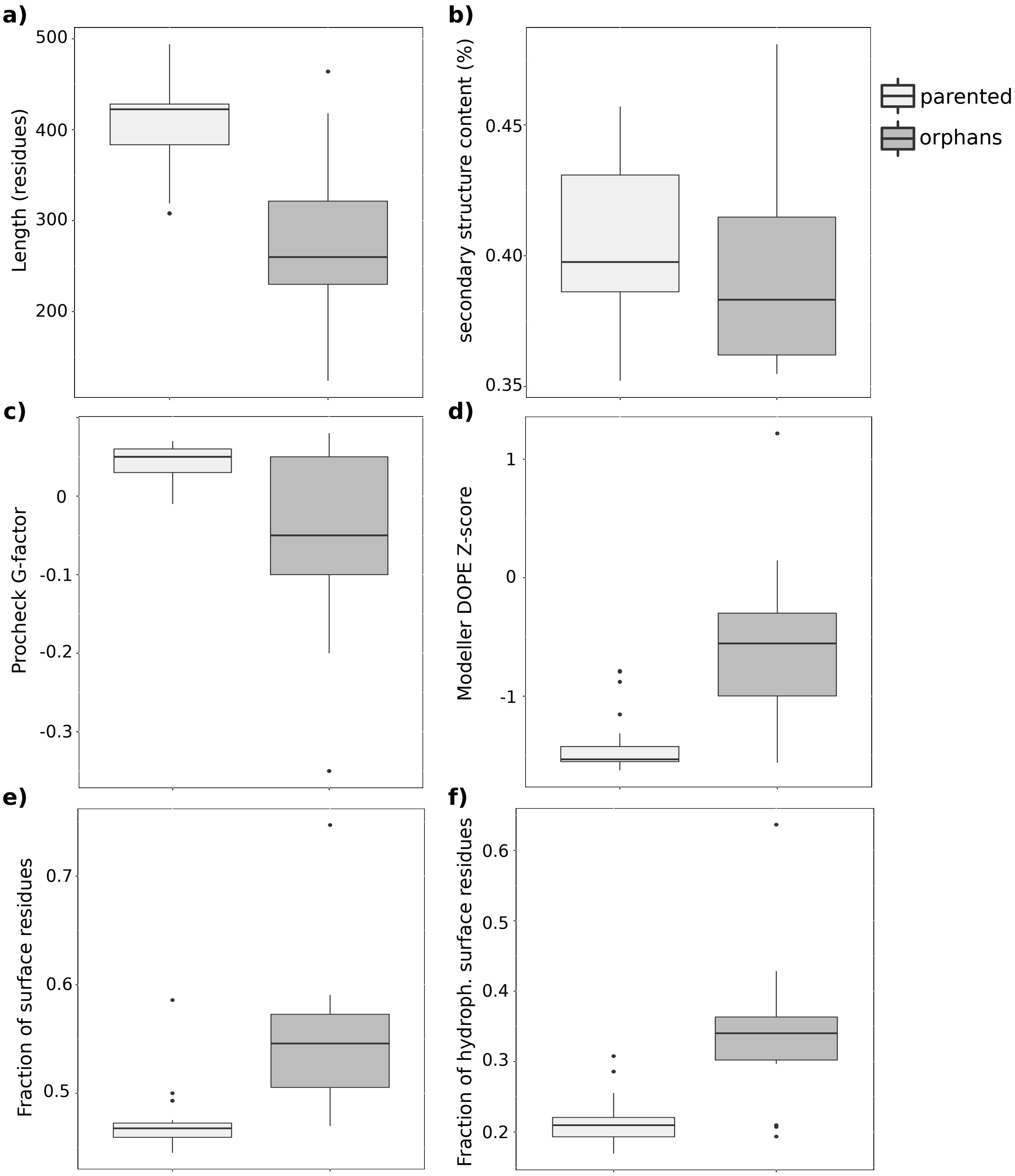

Figure 3:

bioRxiv preprint doi: https://doi.org/10.1101/119891; this version posted March 23, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

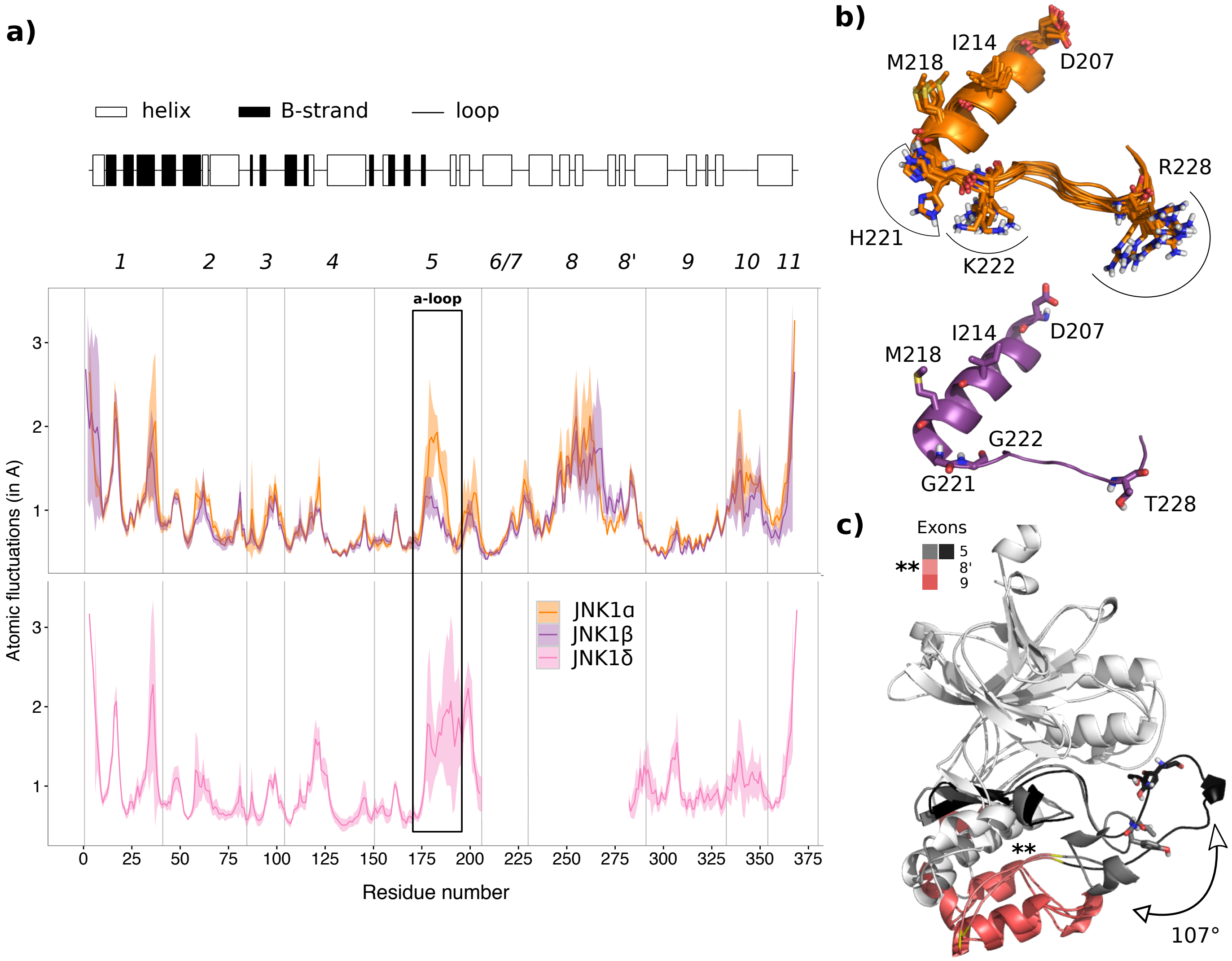

Figure 4:

bioRxiv preprint doi: https://doi.org/10.1101/119891; this version posted March 23, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

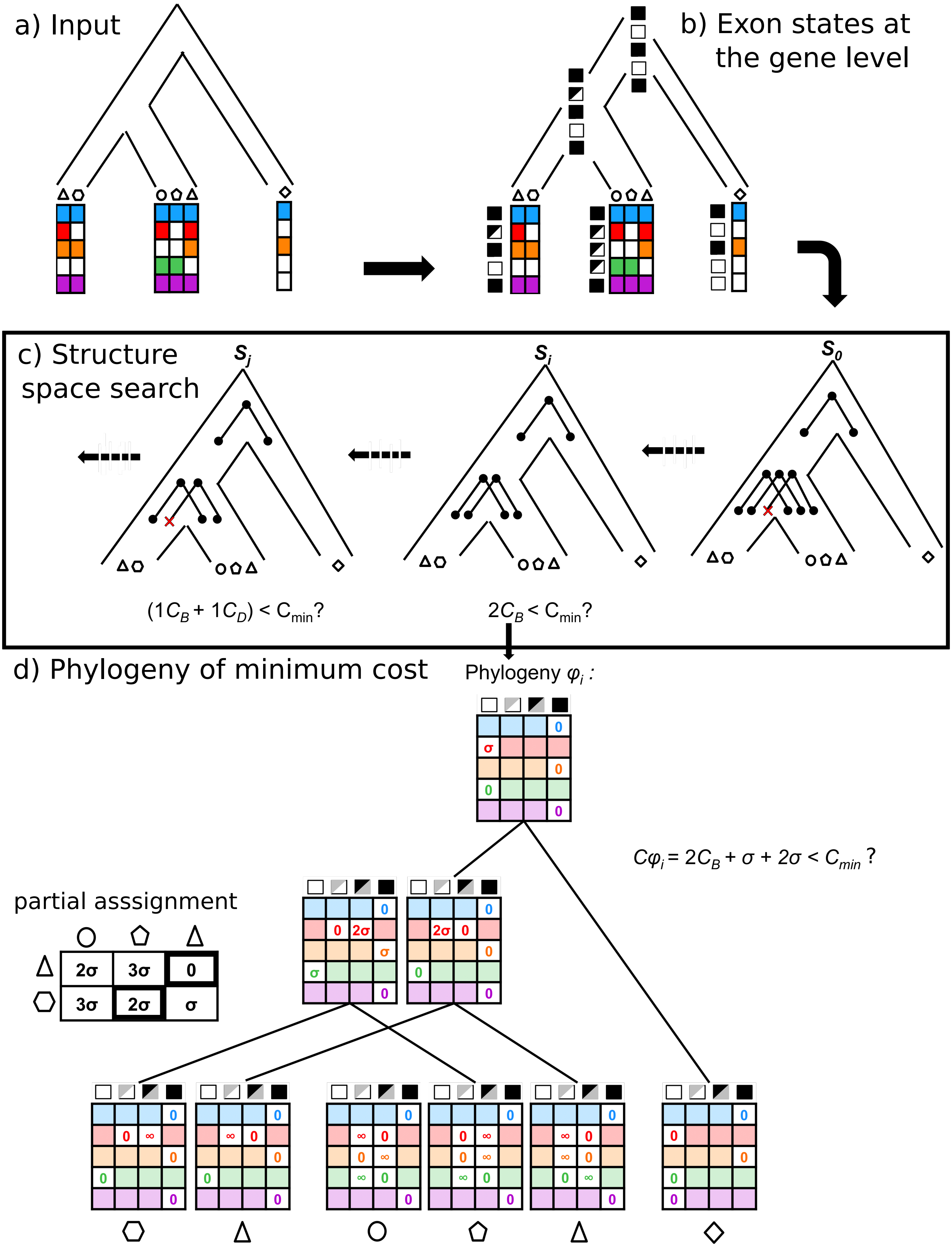
a) Input
b) Exon states at the gene level
c) Structure space search
S_j
S_i
S_0
$(1C_B + 1C_D) < C_{min}$?
$2C_B < C_{min}$?
d) Phylogeny of minimum cost
Phylogeny φ_i :
$C\varphi_i = 2C_B + \sigma + 2\sigma < C_{min}$?
partial asssignment
2σ
3σ
0
3σ
2σ
σ
0
∞

Figure 5: