



## 8 **Abstract**

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

## 29 **Author Summary**

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

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

## 44 **Introduction**

45 Alternative Splicing (AS) of pre-mRNA transcripts is an essential eukaryotic regulatory process  
46 by which multiple isoforms are produced from the same gene. AS-induced changes in the transcribed  
47 sequences can impact the regulation of gene expression or directly modify the content of the coding  
48 sequence (CDS). Large-scale studies revealed that virtually all multi-exons genes in vertebrates are  
49 subject to AS [1]. Consequently, AS has the potential to greatly contribute to functional diversity  
50 in eukaryotes. AS has also gained considerable interest for drug development. It is estimated that  
51 50% of disease causing mutations affect splicing and the ratio of alternatively spliced isoforms is  
52 imbalanced in several cancers [2, 3].

53 About 25% of the AS events (ASEs) common to human and mouse are also conserved in vertebrates  
54 [4, 5, 6]. This high degree of conservation supports an important role of AS in expanding the protein  
55 repertoire through evolution. However, it is difficult to estimate to what extent the ASEs identified  
56 at the gene level actually result in functional protein isoforms in the cell. Transcriptomics and  
57 proteomics studies suggested that most highly expressed human genes have only one single dominant  
58 isoform [7, 8], but the detection rate of these experiments is very difficult to assess [9]. Larger  
59 estimates of the number of functional isoforms in human were reported by machine learning studies  
60 [10]. Moreover, a recent analysis of ribosome profiling data suggested that a major fraction of splice  
61 variants is translated and that the AS-dependent modulation of the translation output regulates  
62 specific cellular functions [11]. At the level of protein structures, it was suggested that splicing  
63 events may induce major fold changes [12, 13]. The elusiveness of the significance of AS for protein  
64 function through evolution calls for the development of efficient and accurate computational methods  
65 that combine protein sequence and structure information.

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

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

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  $\sigma$ . 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.

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

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

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

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

114 Our work allows to put together, for the first time, two types of information, one coming from the  
115 reconstructed phylogeny of transcripts and the other from the structural modeling of the produced  
116 isoforms, and this to shed light on the molecular mechanisms underlying the evolution of protein  
117 function. It goes beyond simple conservation analysis, by dating the appearance of ASEs in evolution,  
118 and beyond general structural considerations regarding AS, by characterizing in details the isoforms'  
119 shapes and motions. We demonstrate that such deep characterization is mandatory in certain cases,  
120 in order to unveil the mechanisms underlying AS functional outcome. Our results also open the way  
121 to the identification and characterization of new isoforms that may be targeted in the future for  
122 medicinal purpose.

## 123 **Results**

124 PhyloSofS was used to reconstruct the phylogenetic forest relating 60 transcripts from the JNK  
125 family observed in 7 species, human, mouse, xenope, fugu, zebrafish, drosophila and nematode. The  
126 input data were collected from the Ensembl [23] database (see *Methods*). The algorithm was run for  
127  $10^6$  iterations and we retained the most parsimonious evolutionary scenario (cost = 69, see *Methods*  
128 for a detailed description of the parameters). PhyloSofS also generated 3D molecular models for  
129 the corresponding protein isoforms, by using homology modeling (see *Methods*). We subsequently  
130 performed molecular dynamics simulations of 3 human isoforms, starting from the predicted 3D  
131 models. In the following, we describe the analysis of the transcripts' phylogeny, structures and  
132 dynamical behavior.

### 133 **Transcripts' phylogeny for the JNK family.**

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

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



146 tral gene in an ancestral species (**S1 Fig, a**). The sequences of the JNK genes are highly conserved  
147 through evolution (**Table I**). The genomes of the two most distant species, namely drosophila and  
148 nematode, contain each only one JNK gene. This gene shares a high degree of nucleotidic sequence  
149 identity (78% for drosophila, 56% for nematode) with human JNK1 (**Table I**). The sequence iden-  
150 tities with human JNK2 and JNK3 are slightly lower (**Table I**, in grey). This suggests that the  
151 most recent common ancestor of the 7 studied species contained one copy of an ancestral JNK1 gene.  
152 Under this assumption, the JNK family gene tree (**S1 Fig, a**) can be reconciled with the species tree  
153 (**S1 Fig, b**) by hypothesizing that early duplication events led to the creation of JNK2 and JNK3  
154 in the ancestor common to mammals, amphibians and fishes. JNK1 was then further duplicated in  
155 fishes while JNK2 was lost in xenope. A representation of the reconstructed transcripts' phylogeny  
156 embedded in the species tree is displayed on **Figure 1c**. It permits to appreciate the diversity of  
157 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.

158 The 7 reconstructed trees relate 12 transcripts observed in human (**Fig. 1b**). Among those, the  
159 transcripts colored the same belong to the same tree and share the same exon composition, even if  
160 they are issued from different paralogues and hence have different amino acid sequences. For instance,  
161 the transcript structure including exons 6, 8 and 12 and excluding exons 0, 1', 7 and 13 (in yellow)  
162 is shared by 3 human transcripts, issued from JNK1, JNK2 and JNK3 (**S1 Fig, c**). Note that this  
163 may not be the case in general, for any protein family: the leaves of a tree may have different exon  
164 compositions if mutations occur along the branches.

165 Among the exons composing the JNK transcripts, two pairs, namely 6 and 7, and 12 and 13, are  
166 mutually exclusive (**S1 Fig, c**). The associated ASEs can be dated early in the phylogeny (**Fig. 1b**),  
167 before the gene duplication (**S1 Fig, b**). Exon 7 is already expressed at the root of the forest (**Fig.**  
168 **1b**, purple tree), while transcripts including exon 6 appear in the ancestor common to mammals,  
169 amphibians and fishes (internal node A3, yellow and orange trees). On **Figure 1b**, exon 13 (purple  
170 and brown trees) appears before exon 12 (yellow and orange trees). However, the scenario where  
171 exon 12 appears before 13 is strictly equivalent (**S8 Fig**, same forest cost = 69). This is explained by  
172 the fact that neither drosophila nor nematode contain any of these exons (**Figure 1b**, see mutation  
173 of the purple transcript between the root and internal node A2).

174 New transcripts appear further down the phylogeny (**Fig. 1b**, in pink, green, and red), after  
175 the JNK1 to JNK2 and JNK1 to JNK3 gene duplication events (**S1 Fig, b**). They are created in  
176 the ancestor common to mammals, amphibians and fishes (**Fig. 1b**). One of them appears in the  
177 sub-forest associated to JNK1 (internal node A11, in pink). It features a large deletion (exclusion of  
178 exons 6, 7 and 8) and its exon composition is perfectly conserved along the phylogeny (no mutation).  
179 The two other transcripts are created at the root of the sub-forest associated to JNK3 (ancestor node  
180 10, in green and red). They are characterized by the presence of exons 0 and 1', not found in the  
181 other paralogues, and they both include exon 6. Interestingly, all the transcripts containing exon 7

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

184 In summary, the analysis of the transcripts' phylogeny inferred by PhyloSofS for JNKs emphasized  
185 several characteristics of the evolution of this protein family. First, it revealed a rather low number  
186 of mutations, illustrating the fact that the sequences of the JNK genes and their exon sites are highly  
187 conserved through evolution. Second, it enabled to date ASEs associated to two pairs of mutually  
188 exclusive exons, namely 6 and 7, and 12 and 13. Of particular interest is the 6/7 pair: the two exons  
189 are homologous and were shown to modulate the affinity of JNKs to their cellular substrates [21]. Our  
190 phylogenetic reconstruction revealed that the most recent common ancestor of all 7 species contained  
191 only one transcript with exon 7, and that transcripts containing exon 6 appeared in the ancestor  
192 common to mammals, amphibians and fishes. Moreover, by analyzing the genomes of drosophila and  
193 nematode, we found that exon 6 is absent from them. These observations suggest that exon 6 is  
194 issued from the duplication of exon 7 and that this duplication occurred in the ancestor common to  
195 mammals, amphibians and fishes, before the duplication of the ancestral JNK gene. Our analysis  
196 also highlighted 2 transcripts specific to JNK3 across several species and showed that exon 7 is not  
197 expressed in the JNK3 sub-forest. This may suggest a subfunctionalization for JNK3, which is the  
198 only paralogue being specifically expressed in certain tissues, namely the heart, brain and testes  
199 [16]. Finally it highlighted a transcript lacking exons 6, 7 and 8 and being specific to JNK1 and its  
200 paralogue in Fugu, JNK1a.

## 201 **Mapping of the gene 1D structure onto the protein 3D structure.**

202 Eighty structures of human JNKs are available in the Protein Data Bank (PDB) [24], among which  
203 30 for JNK1, 2 for JNK2 and 48 for JNK3 (**S1 Table**). This abundance of structural data can be  
204 explained by the fact that JNKs are important therapeutic targets and they were crystallized with

205 different inhibitors. The three paralogues share the same fold, which is highly conserved among  
206 protein kinases (**Fig. 2**). The structures are highly redundant, with an average root mean square  
207 deviation (RMSD) of  $1.96 \pm 0.71$  Å, computed over more than 80% of the protein residues. The  
208 activation loop (A-loop on **Fig. 2**, residues 169-195 in JNK1 and JNK2, residues 207-233 in JNK3)  
209 displays the highest deviations and comprises residues often unresolved in the PDB structures. The  
210 A-loop is found in all kinases and is involved in the control of their activation [25]. The glycine-  
211 rich loop (P-loop), the C-helix and the F-helix (labelled in black on **Fig. 2**) are also ubiquitously  
212 found in protein kinases and play important roles for their structural stability and/or function [25].  
213 The N-terminal hairpin, the MAPK insert and the C-terminal helix (labelled in grey) are specific to  
214 the mitogen-activated protein kinase (MAPK) type, to which the JNKs belong. The catalytic site  
215 (green circle), where ATP binds, is located at the junction between the N- and C-terminal lobes.  
216 Two regions at the surface of JNKs (indicated by green circles) are known to interact with cellular  
217 partners, namely the D-site binding the scaffolding protein JIP-1 [26] and the F-site binding the  
218 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]).

219 In order to visualize the correspondence between the gene structure and the protein secondary  
220 and tertiary structures, the exons were mapped onto a high-resolution PDB structure (3ELJ [28])  
221 of human JNK1 (**Fig. 2**, each exon is colored differently). One can observe that the organization  
222 of the protein 3D structure is preserved by the 1D structure of the gene. Most of the secondary  
223 structures (10 over 12  $\alpha$ -helices and 7 over 9  $\beta$ -strands) are completely included in one exon. It

224 should be noted that exons 8 and 8' used in PhyloSofS actually correspond to only one genomic  
225 exon (see *Methods*). All the regions important for kinases are preserved (**Fig. 2**, labelled in black),  
226 as well as the N-terminal hairpin and the MAPK insert (labelled in grey). By contrast, the catalytic  
227 site, the D-site and the F-site (green circles) are comprised of residues belonging to different exons.  
228 The precise borders of the exons and the known regions/sites are given in **S2 Table**. Of note, the  
229 block formed by exons 1 to 5, comprising the N-terminal lobe and the A-loop (**Fig. 2**, from blue to  
230 white), is constitutively present in all transcripts belonging to the colored trees on **Fig. 1b**.

231 The correspondence was also analyzed for the JNK protein from drosophila (**S2 Fig, b**). The  
232 3D structures of human JNK1 (**S2 Fig, a**) and drosophila JNK (**S2 Fig, b**) are very similar, with  
233 a RMSD of 0.68 Å on 251 over 314 (80%) residues. The JNK gene from the drosophila genome  
234 comprises much fewer exons than the human gene, which leads to an even better preservation of the  
235 secondary structures and of the known important regions in that species.

236 This analysis showed that the 1D structure of the JNK genes preserves most of the protein sec-  
237 ondary structure elements and most of the regions playing important roles for kinases structural  
238 stability and/or function. This is true for human and also for one of the most distant species, namely  
239 drosophila. Considering the high degree of conservation of JNK sequences, one may hypothesize that  
240 this is a general property across all the studied species. By contrast, the functional binding sites  
241 of the protein contain residues belonging to different exons. This is expected as binding sites are  
242 comprised of segments that can be very far from each other along the protein sequence.

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

## 246 **Properties of the orphan transcripts.**

247 We investigated whether the orphan transcripts, for which no phylogeny could be reconstructed  
248 (**Fig. 1b**, grey leaves), displayed peculiar sequence and structural properties compared to the "par-  
249 ented" transcripts (**Fig. 1b**, colored leaves). First, the orphan transcripts are significantly smaller  
250 than the parented ones (**Fig. 3a**). While the minimum length for parented transcripts is 308 residues,  
251 with an average of  $406 \pm 40$  residues (**Fig. 3a**, in white), the orphan transcripts can be as small as  
252 124 residues, with an average of  $280 \pm 88$  residues (**Fig. 3a**, in grey). Second, regarding secondary  
253 structure content, both types of transcripts contain about 40% of residues predicted in  $\alpha$ -helices or  
254  $\beta$ -sheets (**Fig. 3b**). Third, the 3D models generated by PhyloSofS's molecular modeling routine  
255 for the orphan transcript isoforms are of poorer quality than those for the transcripts belonging  
256 to a phylogeny (**Fig. 3c-d**). The quality of the models was assessed by computing Procheck [30]  
257 G-factor and Modeller [31] normalized DOPE score (**Fig. 3c-d**). A model resembling experimental  
258 structures deposited in the PDB should have a G-factor greater than -0.5 (the higher the better) and  
259 a normalized DOPE score lower than -1 (the lower the better). The distributions obtained for the  
260 parented isoforms are clearly shifted toward better values and are more narrow than those for the  
261 orphan transcripts. Finally, the proportion of protein residues being exposed to the solvent (relative  
262 accessible surface area  $rsa > 25\%$ ) is significantly higher for the orphan isoforms (**Fig. 3e**), as  
263 is the proportion of hydrophobic residues being exposed to the solvent (**Fig. 3f**). Overall, these  
264 observations suggest that simple sequence and structure descriptors enable to distinguish the orphan  
265 transcripts from the ones within a phylogeny and that the formers display properties likely reflecting  
266 structural instability (large truncations, poorer quality, larger and more hydrophobic surfaces).

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 ( $rsa > 0.25$ ).

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

269 The two mutually exclusive exons 6 and 7 are particularly important for JNK cellular functions,  
270 as they confer substrate specificity. The inclusion or exclusion of one or the other results in different  
271 substrate-binding affinities [21, 22]. From a sequence perspective, the two exons are homologous,  
272 highly conserved through evolution, and differ only by a few positions (**S3 Fig**). From a structural  
273 perspective, they both fold into an  $\alpha$ -helix, known as the F-helix, followed by a loop (**Fig. 2**, in light  
274 pink).

275 The F-helix was shown to play a central role in the structural stability of protein kinases [32].  
276 In particular, it contains a N-terminal aspartate and 2 hydrophobic residues highly conserved across  
277 the whole kinase family. These 3 residues were shown to serve as anchor points for two clusters  
278 of hydrophobic residues, namely the catalytic and regulatory spines, essential for kinase activity  
279 and regulation [32] (see illustration on the PKA kinase on **S4 Fig, a**). Moreover, the N-terminal  
280 aspartate was shown to form hydrogen bonds (H-bonds) with the HRD motif in the catalytic loop  
281 and to consequently stabilize the backbone of this motif in a strained conformation characteristic of  
282 protein kinase structures and important for their catalytic activity [33] (see illustration on the CDK-  
283 substrate complex on **S5 Fig, a**). To sum up, the F-helix is essential for kinase structural stability  
284 and some particular residues in this helix are involved in structural features important for kinase  
285 catalytic activity and/or regulation. In the following, we will use these known structural features as



286 proxies for the stability and catalytic competence of the studied isoforms.

287 The available JNK crystallographic structures and the 3D models generated by PhyloSofS do not  
288 display any significant structural change between the isoforms including exon 6 and those including  
289 exon 7. The catalytic and regulatory spines, together with their anchors in the F-helix, are present  
290 in both types of isoforms (**S4 Fig, b-c**). The HRD motif's strained backbone conformation and the  
291 associated H-bond pattern are also observed in both types of isoforms (**S5 Fig, b-c**). The N-terminal  
292 aspartate (D207) of the F-helix is 100% conserved in both exons 6 and 7 in the 7 studied species  
293 (**S3 Fig**, indicated by an arrow). The two other anchor points are also present, namely I214 and  
294 L/M218 (**S3 Fig**, indicated by arrows). Consequently, both exons 6 and 7, and thus the isoforms  
295 containing them, possess the structural features known to be important for kinase catalytic activity  
296 and/or regulation.

297 To further investigate the potential impact of the inclusion/exclusion of exon 6 or 7 on the dy-  
298 namical behavior of the protein, we performed all-atom molecular dynamics (MD) simulations of the  
299 human isoforms colored in orange and purple on **Figure 1b**. We shall refer to these isoforms as  
300 JNK1 $\alpha$  (with exon 6) and JNK1 $\beta$  (with exon 7), in agreement with the nomenclature found in the  
301 literature [21]. JNK1 $\alpha$  and JNK1 $\beta$  were simulated in explicit solvent for 250 ns (5 replicates of 50  
302 ns, see *Methods*). The backbone atomic fluctuation profiles of the two isoforms are very similar (**Fig.**  
303 **4a**, orange and purple curves), except for the A-loop which is significantly more flexible in JNK1 $\alpha$ :  
304 the region from residue 176 to 188 displays averaged C $\alpha$  fluctuations of  $1.55 \pm 0.28$  Å in JNK1 $\alpha$  and  
305 of  $0.98 \pm 0.16$  Å in JNK1 $\beta$  (**Fig. 4a**). The two exons, 6 and 7, have similar backbone flexibility. In  
306 the F-helix, the anchor residues for the spines, D207, I214 and M218 adopt stable and very similar  
307 conformations (**Fig. 4b**). Moreover, the HRD backbone strain and the associated H-bond pattern  
308 are maintained along the simulations of both systems (**S7 Fig, a-b**). Consequently, the observa-



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

311 Nevertheless, an interesting observation can be made regarding the loop following the F-helix: a  
312 few residues lying in this loop display very different side-chain flexibilities between the two isoforms  
313 (**Fig. 4b**). On the one hand, in exon 6 (in orange), the polar and positively charged residues H221,  
314 K222 and R228 are exposed to the solvent and display large amplitude side-chain motions. These  
315 amino acids are 100% conserved in exon 6 across all species (**S3 Fig**). On the other hand, in exon  
316 7 (**Fig. 4b**, in purple), G221, G222 and T228 have small side chains with much reduced motions.  
317 While G221 is conserved across all species, position 222 is variable and position 228 features G, T  
318 or S (**S3 Fig**). This region of the protein is involved in the binding of substrates (see **Fig. 2**, F-  
319 site). Moreover, in both isoforms, we predicted residues 223-230 as directly interacting with cellular  
320 partners (see *Methods*). Consequently, one may hypothesize that the differences highlighted here may  
321 be crucial for substrate molecular recognition specificity. The positive charges, high fluctuations, high  
322 solvent accessibility and high conservation of residues H221, K222 and R228 in JNK1 $\alpha$  support a  
323 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 JNK1 $\alpha$  (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 C $\alpha$ ) averaged over 5 50-ns MD replicates are reported for JNK1 $\alpha$  in orange, JNK1 $\beta$  in purple and JNK1 $\delta$  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 JNK1 $\alpha$  (in orange) and only 1 for JNK1 $\beta$  (in purple). (c) Superimposed pair of MD conformations illustrating the amplitude of the A-loop motion in JNK1 $\delta$  (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.

## 324 **Structural dynamics of a newly identified isoform.**

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

330 JNK1 $\delta$  displays a large deletion (of about 80 residues), lacking exons 6, 7 and 8. It does not contain  
331 the F-helix, shown to be crucial for kinases structural stability [32], nor the MAPK insert, involved  
332 in the binding of the phosphatase MKP7 [27] (**Fig. 2**). The 3D model generated by PhyloSofS  
333 superimposes well to those of JNK1 $\alpha$  and JNK1 $\beta$ , with a RMSD lower than 0.5 Å on 245 residues.  
334 This is somewhat expected as we use homology modeling. Nevertheless, cases were reported in  
335 the literature where homology modeling detected big changes in protein structures induced by exon  
336 skipping [34]. In the model of JNK1 $\delta$ , the F-helix present in JNK1 $\alpha$  and JNK1 $\beta$  (residues 207 to 220)  
337 is replaced by a loop (residues 282 to 288) corresponding to exon 8' (**Fig. 4c**, indicated by the two  
338 stars). The sequence of this loop (exon 8') does not share any significant identity with the F-helix  
339 (N-terminal parts of exons 6 and 7), except for the N-terminal residue which is an aspartate, namely  
340 D282 (D207 in JNK1 $\alpha$  and JNK1 $\beta$ ). This replacement results in the regulatory spine being intact  
341 in JNK1 $\delta$  (**S4 Fig, d**, in red). Moreover, the HRD motif's strained backbone conformation and the  
342 associated H-bond pattern, which are stabilized by the aspartate, are maintained (**S5 Fig, d**). By  
343 contrast, the catalytic spine lacks its two anchors (**S4 Fig, d**, in yellow). Consequently, despite its  
344 lacking of an important and large part of the protein, JNK1 $\delta$  still possesses some structural features  
345 important for kinase catalytic activity and/or regulation.

346 JNK1 $\delta$  was simulated in explicit solvent for 250 ns (5 replicates of 50 ns). The isoform displays  
347 stable secondary structures (**S6 Fig**, at the bottom) and atomic fluctuations comparable to those  
348 of JNK1 $\alpha$  and JNK1 $\beta$  (**Fig. 4a**, pink curve to be compared with the purple and orange curves).  
349 The C $\alpha$  atomic fluctuations averaged over the loop replacing the F-helix values  $0.88 \pm 0.18$  Å.  
350 This is higher than the values computed for the F-helix in JNK1 $\alpha$  and JNK1 $\beta$  ( $0.57 \pm 0.10$  Å and  
351  $0.53 \pm 0.09$  Å), but it still indicates a limited flexibility. Moreover, the N-terminal aspartate D282  
352 establishes stable H-bonds with the HRD motif along all but one of the replicates (**S7 Fig, a**, on the  
353 right) and the HRD motif's backbone remains in a strained conformation (**S7 Fig, b**, on the right),  
354 as was observed for JNK1 $\alpha$  and JNK1 $\beta$ . Consequently, JNK1 $\delta$  seems stable in solution, and, as  
355 observed on the static 3D model, the absence of the F-helix in this isoform is partially compensated  
356 by the presence of D282, which is sufficient to maintain H-bonds with the HRD motif and a resulting  
357 backbone strain of the motif, important for kinase structural stability.

358 The main difference between JNK1 $\delta$  and the two other isoforms lies in the amplitude of the  
359 motions of the A-loop. In JNK1 $\delta$ , the C-terminal part of the A-loop can detach from the rest of  
360 the protein along the simulations (**Fig. 4c**). The amplitude of the angle computed between the  
361 most retracted conformation (in grey) and the most extended one (in black) is  $107^\circ$ . By contrast, in  
362 JNK1 $\alpha$  and JNK1 $\beta$ , the A-loop always stays close to the rest of the protein, with amplitude angles  
363 of  $18^\circ$  and  $19^\circ$ , respectively. The A-loop contains two residues, T183 and Y185 (**Fig. 4c**, highlighted  
364 in sticks), whose phosphorylation is required for JNK activation. We hypothesize that the large  
365 amplitude motion in JNK1 $\delta$  might favor their accessibility and, in turn, the activation of the protein.

## 366 **Alternative transcripts' phylogenies.**

367 The size of the search space for the transcripts' phylogeny reconstruction grows exponentially  
368 with the number of observed transcripts (leaves). To explore that space, the heuristic algorithm

369 implemented in PhyloSofS relies on a multi-start iterative procedure and on the computation of a  
370 lower bound to early filter out unlikely scenarios (see *Methods*). Depending on the input data and  
371 the set of parameters, it may find several solutions with equivalent costs. Over  $10^6$  iterations of the  
372 program, the forest described above (**Fig. 1b**, or **S8 Fig** with branch swapping), comprising 7 trees,  
373 19 deaths and 14 orphans, was visited 1 219 times. An alternative phylogeny was visited 310 times,  
374 that comprises the same number of trees and orphans, but 2 more deaths (**S9 Fig**). The difference  
375 between the two forests lies among the fugu JNK1 transcripts, where one transcript belongs to the  
376 orange tree (**S9 Fig**) instead of the yellow one (**Fig. 1b**). The two trees differ by the inclusion or  
377 exclusion of exon *12* or *13*, and the re-assigned transcript lacks both exons. Consequently, the new  
378 branching results in the loss of exon *13* between the internal nodes A11 and A18 (**S9 Fig**), instead  
379 of the loss of exon *12* between A24 and fugu JNK1 (**Fig. 1b**). Another forest with the same cost  
380 comprising 8 trees, 23 deaths and 13 orphans was visited 190 times (**S10 Fig**). The additional tree  
381 is created in the internal node A10 and links two observed JNK3 transcripts: one from the mouse  
382 that was previously orphan (**Fig. 1b**) and one from zebrafish that previously belonged to the green  
383 tree. Both transcripts are truncated at the C-terminus and lack exons *12* and *13*. Consequently, this  
384 new branching avoids the loss of exon *12* between A16 and zebrafish JNK3. Overall the differences  
385 between the three solutions are minor and these ambiguities do not impact our interpretation of the  
386 results.

### 387 **Unresolved residues in the 3D models.**

388 In the 3D models generated by PhyloSofS, the N-terminal exons *0* and *1'* and the C-terminal exons  
389 *12* and *13* are systematically missing. This is due to the lack of structural templates for these regions.  
390 Using a threading approach instead of PhyloSofS's homology modeling routine (see *Methods*) did not  
391 enable to improve their reconstruction. In fact, the models generated by the threading algorithm are

392 very similar to those generated by PhyloSofS.

393 All the missing exons are predicted to contain some intrinsically disordered regions (**S11 Fig**).  
394 At the N-terminus, exons *0* and *1'* contain two segments of about 10 residues predicted as disordered  
395 protein-binding regions (**S11 Fig, b**, orange curve), *i.e* regions unable to form enough favorable  
396 intra-chain interactions to fold on their own and likely stabilized upon interaction with a globular  
397 protein partner [35]. These exons are present in only two JNK3 transcript isoforms (**Fig. 1b**, colored  
398 in red and green). Considering that JNK3 isoforms are specifically expressed in the heart, brain and  
399 testes [21], one can hypothesize that the two exons are involved in interactions with specific cellular  
400 partners in these tissues. At the C-terminus, exons *12* or *13* are completely predicted as intrinsically  
401 disordered (**S11 Fig, a** and **S11 Fig, b**, blue curve). The functional implication of the inclusion or  
402 exclusion of *12/13* has not been assessed experimentally [21].

## 403 Discussion

404 To what extent the transcript diversity generated by AS translates at the protein level and has  
405 functional implications in the cell remains a very challenging question and has been subject to much  
406 debate [36, 37]. The present work contributes to elaborating strategies to answer it, by crossing  
407 sequence analysis and phylogenetic inference with molecular modeling. We report the first joint  
408 analysis of the evolution of alternative splicing across several species and of its structural impact  
409 on the produced isoforms. The analysis was performed on the JNK family, which represents a high  
410 interest for medicinal research and for which a number of human isoforms have been described and  
411 biochemically characterized.

412 Importantly, our approach enables to go beyond a mere description of transcript variability across  
413 species and/or across genes. Indeed, by reconstructing phylogenies, we do not only cluster transcripts  
414 but we also add a temporal dimension to the analysis and we date the ASEs. This is important when

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

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

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

431 First, we dated an ASE consisting of two mutually exclusive homologous exons (6 and 7) in the  
432 ancestor common to mammals, amphibians and fishes. By characterizing in details the structural  
433 dynamics of two human isoforms, JNK1 $\alpha$  and JNK1 $\beta$ , bearing one or the other exon, we could  
434 emphasize subtle changes associated to this ASE and identify residues that may be responsible for  
435 the selectivity of the JNK isoforms toward their substrates. Alternatively spliced homologous exons  
436 were recently shown to be highly expressed at the protein level and to have ancient origin, supporting  
437 an important cellular role [38].

438 Second, our analysis highlighted an isoform, JNK1 $\delta$ , conserved across several species, displaying

439 a large deletion (about 80 residues), and not previously described in the literature. It is recorded in  
440 the UniProt database [39] (accession id: P45983-5). The APPRIS database v20 [40] annotates it as  
441 *minor* and indicates that there are 4 peptides matching the isoform in publicly available proteomics  
442 data. By comparison, the human JNK1 isoforms identified as orphans by our phylogenetic analysis are  
443 also annotated as *minor* in the APPRIS database and have between zero and 2 matching peptides.  
444 The other human JNK1 isoforms, which possess a phylogeny and are described in the literature [21],  
445 are annotated as *alternative* or *principal* and have between 5 and 7 matching peptides. Our analysis  
446 showed that JNK1 $\delta$  remains stable in solution and that its catalytic site is intact. We propose that  
447 JNK1 $\delta$  might be catalytically competent and that the large amplitude motion of the A-loop observed  
448 in the simulations might facilitate the activation of the protein by exposing a couple of tyrosine and  
449 threonine residues that are targeted by MAPK kinases. The validation of this hypothesis would  
450 require further calculations and experiments that fall beyond the scope of this study. Already, this  
451 interesting result suggests that our approach could be used to identify and characterize new isoforms,  
452 that may play a role in the cell and thus serve as therapeutic targets.

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

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

468 Under-annotation of transcripts is a potential source of error coming from the input data. It can  
469 impact the phylogenetic reconstruction by missing distant evolutionary relationships. To deal with  
470 this issue, we set the cost associated to transcript death to zero. This enables to construct trees  
471 that can relate transcripts possibly very far from each other in the phylogeny (*i.e.* expressed in  
472 very distant species, because some species in between are under-annotated). This parameter may  
473 be tuned by the user depending on the quality and reliability of the input data. A second source of  
474 error comes from annotated transcripts supposedly non-functional. We expect that these transcripts  
475 are likely not conserved across species and thus will be attributed the status of orphans in the phy-  
476 logenetic reconstruction. Moreover, we have emphasized an independent source of evidence coming  
477 from their structural characterization which can help us flag them. The reliability of the transcript  
478 expression data clearly constitutes a present limitation of the method. However, as experimental  
479 evidence accumulate and precise quantitative data become available, computational methods such  
480 as PhyloSofS will become instrumental in assessing the contribution of AS in protein evolution. The  
481 present work opens the way to such assessment at large-scale.

482 To efficiently search the space of possible phylogenies, the algorithm implemented in PhyloSofS  
483 relies on a multi-start iterative procedure and on the computation of a lower bound that enables  
484 to early eliminate unsuitable candidate solutions (see *Methods*). For the JNK family, the execution  
485 of 1 million iterations took about two weeks on a single CPU. This case represents a high level of  
486 complexity as most of the transcripts contain more than 10 exons (the average number of exons per



487 gene being estimated at 8.8 in the human genome [44]) and up to 8 transcripts are observed within  
488 each species (it is estimated that about 4 distinct-coding transcripts per gene are expressed in human  
489 [40]). To reduce the computing time, the user can easily parallelize the multi-start iterative search  
490 on multiple cores and he/she has the possibility to give as input a previously computed value for the  
491 lower bound (to increase the efficiency of the cut). This implementation makes feasible, for the first  
492 time, the reconstruction of transcripts' phylogenies for any gene family.

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

## 505 **Methods**

### 506 **PhyloSofS workflow**

507 PhyloSofS takes as input a binary tree (called a gene tree) describing the phylogeny of the gene(s) of  
508 interest for a set of species (**Fig. 1a**, on the left), and the ensemble of transcripts observed in these

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

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

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

521 PhyloSofS comes with helper functions for the visualization of the output transcripts' phylogeny(ies)  
522 and of the isoforms' molecular models. The program is implemented in Python.

### 523 **Step a. Transcripts' phylogenies reconstruction**

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

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

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

Table II: **Exon states and associated costs  $\sigma$ .**

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	$\sigma$	$\sigma$	excluded from $t_i^s$ but included in some transcript(s) of $s$
(1,1)	$\sigma$	$\sigma$	0	0	included in $t_i^s$ but excluded from some transcript(s) of $s$
(1,2)	0	$\sigma$	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.

535 **Input data.** The input consists in a gene tree with the observed transcripts at the leaves (**Fig.**  
 536 **5a**). The gene is represented by an ensemble  $E$  of  $n_e$  exons. The identification and alignment of the  
 537  $n_e$  homologous exons between the different transcripts must be performed prior to the application of  
 538 the method (see below for details on data preprocessing for the JNK family). The  $n_s$  transcripts of  
 539 species  $s$  are described by a binary table  $T^s$  of  $n_e \times n_s$  elements, where  $T_{i,j}^s = 1$  if exon  $i$  is included  
 540 in transcript  $j$  (colored squares on **Fig. 5a**), 0 if it is excluded (white squares).

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  $\varphi_i$  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.

541 **Exon states at the gene level.** For a given species  $s$ , a vector  $g^s$  of length  $n_e$  encodes the state  
542 of each exon by the values  $\{0, 1, 2\}$  for absent, alternative and constitutive, respectively (**Fig. 5b**,  
543 white, black/white and black squares). At the leaves (current species), the components of  $g^s$  are  
544 calculated as:

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

545 The  $g^s$  vectors for internal nodes (ancestral species) are determined by using Sankoff's algorithm  
546 [45]. Dollo's parsimony principle is also respected, such that an exon cannot be created twice [46].  
547 If different exon states have equal cost, we follow the priority rule  $2 > 0 > 1$ .

548 **Forest structure.** Each internal node of the gene tree, representing an ancestral species, is ex-  
549 panded in several subnodes, representing the transcripts of the gene in this ancestral species (**Fig.**

550 **5c).** There exist three types of subnodes: binary (two transcript children), left (one transcript child  
551 in the node's left child) and right (one transcript child in the node's right child). Left and right  
552 subnodes imply that a transcript death occurred along the branch. A forest structure  $S$  is fixed by  
553 setting  $n_b$ ,  $n_l$  and  $n_r$  the respective numbers of binary, left and right subnodes for every internal  
554 node of the gene tree. The cost associated to structure  $S$  is calculated as  $C_S = C_{birth}(S) + C_{death}(S)$ ,  
555 where  $C_{birth}(S)$  and  $C_{death}(S)$  are the total costs of creation and loss of transcripts, expressed as:

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

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

556 **Transcripts' phylogeny.** A transcripts' phylogeny determines the pairings of transcripts at each  
557 level of the forest structure (**Fig. 5d**). The cost of the phylogeny  $\varphi$  complying with the structure  $S$   
558 is calculated as:

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

559 where  $\Gamma(A)$  is computed for each tree  $A$  of  $\varphi$  by evaluating the changes of exon states along the  
560 branches of  $\varphi$ :

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

561 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  
562  $l$  and  $\Gamma(t_i^k \rightarrow t_j^l) = \sum_{e \in E} \sigma((T_{e,i}^k; g_e^k), (T_{e,j}^l; g_e^l))$ , with  $g_e^y \in \{0, 1, 2\}$  the state of exon  $e$  at the level of  
563 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  $y$ .

564 The evolution costs  $\sigma$  are given in Table II.

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

568 **Initialization:**

569  $C_{min} \leftarrow \infty$

570 Choose the forest structure  $S_0$  that maximizes the  $n_b$  values

571 **Iteration:**

572 **for**  $i = 0$  to  $t_{max} - 1$  **do**

573 **if**  $C_{S_i} < C_{min}$  **then**

574 Find the most parsimonious phylogeny  $\varphi_i$  given structure  $S_i$

575 **if**  $C_{\varphi_i} < C_{min}$  **then**

576  $C_{min} \leftarrow C_{\varphi_i}$

577 **end if**

578 **end if**

579 Choose forest structure  $S_{i+1}$  by setting  $n_b$ ,  $n_l$  and  $n_r$  at every internal node

580 **end for**

581 To efficiently search the space of all possible forest structures (**Fig. 5c**), PhyloSofS relies on a  
582 multi-start iterative procedure. Random jumps in the search space are performed until a suitable  
583 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  
584 lower bound for the cost  $C_{\varphi_i}$  of the phylogeny  $\varphi_i$ . Forest structures that are too costly are simply  
585 discarded, without calculating the corresponding phylogenies. As the algorithm finds better and  
586 better solutions, the cut becomes more and more efficient. The phylogeny  $\varphi_i$  is reconstructed by  
587 using dynamic programming. Sankoff's algorithm is applied bottom up to compute the minimum  
588 pairing costs between transcripts (**Fig. 5d**, each transcript is represented by a matrix of costs). At  
589 each internal node, the pairings are determined by using a specific version of the branch-and-bound  
590 algorithm [47] (see *Supplementary Text S1*). If the reconstructed phylogeny is more parsimonious  
591 than those previously visited ( $C_{\varphi_i} < C_{min}$ ), then the minimum cost  $C_{min}$  is updated. There may

592 be more than one phylogeny with minimum cost that comply with a given structure  $S_i$ . The next  
593 forest structure  $S_j$  will be randomly chosen among the immediate neighbors of  $S_i$  (**Fig. 5d**). Two  
594 structures are immediate neighbors if each one of them can be obtained by an elementary operation  
595 applied to only one node of the other one (**S12 Fig**). If the phylogeny  $\varphi_j$  is such that  $C_{\varphi_j} < C_{min}$ ,  
596 then the next forest structure will be chosen among the neighbors of  $S_j$ , which serves as a new  
597 "base" for the search. Otherwise, the algorithm continues to sample the neighborhood of  $S_i$ . This  
598 step-by-step search is applied until no better solution can be found. At this point, a new random  
599 jump is performed. The total number of iterations  $t_{max}$  is given as input by the user (1 by default).

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

## 602 **Step b. Isoforms structures prediction**

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

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

612 **Search for templates.** Step 1 makes extensive use of the HH-suite [49] and can be decomposed  
613 in: (a) search for homologous sequences and building of a multiple sequence alignment (MSA), by  
614 using HHblits [50], (b) addition of secondary structure predictions, obtained by PSIPRED [51], to  
615 the MSA, (c) generation of a profile hidden markov model (HMM) from the MSA, (d) search of a  
616 database of profile HMMs for homologous proteins, using HHsearch [52].

617 **3D model building.** Step 3 is performed by Modeller [31] with default options.

618 **Annotation of the models.** Step 5 consists in: (a) inserting the numbers of the exons in the  
619  $\beta$ -factor column of the PDB file of the 3D model, (b) computing the proportion of residues predicted  
620 in well-defined secondary structures by PSIPRED [51], (c) assessing the quality of the model with  
621 Procheck [30] and with the normalized DOPE score from Modeller, (d) determining the by-residue  
622 solvent accessible surface areas with Naccess [53] and computing the proportions of surface residues  
623 and of hydrophobic surface residues.

## 624 **Application of PhyloSofS to the JNK family**

625 **Retrieval and pre-processing of transcriptome data.** The peptide sequences of all splice  
626 variants from the JNK family observed in human, mouse, xenope, zebrafish, fugu, drosophila and  
627 nematode were retrieved from Ensembl [23] release 84 (March 2016) along with the phylogenetic  
628 gene tree. Only the transcripts containing an open reading frame and not annotated as undergoing  
629 nonsense mediated decay or lacking 3' or 5' truncation were retained. The homologous exons between  
630 the different genes in the different species were identified by aligning the sequences with MAFFT  
631 [54], and projecting the alignment on the human annotation. The isoforms resulting in the same  
632 amino acid sequence were merged. In total, 64 transcripts comprised of 38 exons were given as input  
633 to PhyloSofS.



634 **Exon numbering.** The set of homologous exons used in PhyloSofS were defined so as to account for  
635 all the variations occurring between the observed transcripts in any species. They do not necessarily  
636 represent exons definition based on the genomic sequence, for two reasons. First, the structure of  
637 the genes may be different from one species to another. For instance, the third and fourth exons of  
638 human JNK1 genes are completely covered by only one exon in the drosophila JNK gene (**S2 Fig**).  
639 In that case, we keep the highest level of resolution and define two exons (*3* and *4*). Second, it may  
640 happen that a transcript contains only a part of an exon in a given species translated in another  
641 frame. In that case, we define two exons sharing the same number but distinguished by the prime  
642 symbol, *e.g.* exons *8* and *8'*.

643 **Reconstruction of the transcripts' phylogeny.** To set the parameters, two criteria were taken  
644 into consideration. First, the different genomes available in Ensembl are not annotated with the  
645 same accuracy and the transcriptome data and annotations may be incomplete. This may challenge  
646 the reconstruction of transcripts' phylogenies across species. To cope with this issue, we chose not  
647 to penalize transcript death ( $C_D=0$ ). Second, the JNK genes are highly conserved across the seven  
648 studied species (**Table I**), indicating that this family has not diverged much through evolution.  
649 Consequently, we set the transcript mutation and birth costs to  $\sigma = 2$  and  $C_B = 3$  ( $C_B < \sigma \times 2$ ).  
650 This implies that few mutations will be tolerated along a phylogeny. Prior to the phylogenetic  
651 reconstruction, PhyloSofS removed 19 exons that appeared in only one transcript (default option),  
652 reducing the number of transcripts to 60. This pruning enables to limit the noise contained in the  
653 input data and to more efficiently reconstruct phylogenies. PhyloSofS algorithm was then run for  
654  $10^6$  iterations.

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

657 parameter) for every isoform.

## 658 **Analysis of JNK tertiary structures.**

659 The list of experimental structures deposited in the PDB for the human JNKs was retrieved  
660 from UniProt [39]. The structures were aligned with PyMOL [55] and the RMSD between each  
661 pair was computed. Residues comprising the catalytic site were defined from the complex between  
662 human JNK3 and adenosine mono-phosphate (PDB code: 4KKE, resolution: 2.2 Å), as those located  
663 less than 6 Å away from the ligand. Residues comprising the D-site and the F-site were defined  
664 from the complexes between human JNK1 and the scaffolding protein JIP-1 (PDB code: 1UKH,  
665 resolution: 2.35 Å [26]) and the catalytic domain of MKP7 (PDB code: 4YR8, resolution: 2.4 Å  
666 [27]), respectively. They were detected as displaying a change in relative solvent accessibility  $>1 \text{ \AA}^2$   
667 upon binding.

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

## 671 **Molecular dynamics simulations of human isoforms.**

672 **Set up of the systems.** The 3D coordinates of the human JNK1 isoforms JNK1 $\alpha$  (369 res.,  
673 containing exon 6), JNK1 $\beta$  (369 res., containing exon 7) and JNK1 $\delta$  (304 res., containing neither  
674 exon 6 nor exon 7) were predicted by PhyloSofS pipeline. The 3 systems were prepared with the  
675 LEAP module of AMBER 12 [62], using the ff12SB forcefield parameter set: (i) hydrogen atoms  
676 were added, (ii) the protein was hydrated with a cuboid box of explicit TIP3P water molecules with  
677 a buffering distance up to 10Å, (iii) Na<sup>+</sup> and Cl<sup>-</sup> counter-ions were added to neutralize the protein.

678 **Minimization, heating and equilibration.** The systems were minimized, thermalized and equi-  
679 librated using the SANDER module of AMBER 12. The following minimization procedure was  
680 applied: (i) 10,000 steps of minimization of the water molecules keeping protein atoms fixed, (ii)  
681 10,000 steps of minimization keeping only protein backbone fixed to allow protein side chains to re-  
682 lax, (iii) 10,000 steps of minimization without any constraint on the system. Heating of the system to  
683 the target temperature of 310 K was performed at constant volume using the Berendsen thermostat  
684 [63] and while restraining the solute  $C_{\alpha}$  atoms with a force constant of  $10 \text{ kcal/mol}/\text{\AA}^2$ . Thereafter,  
685 the system was equilibrated for 100 *ps* at constant volume (NVT) and for further 100 *ps* using a  
686 Langevin piston (NPT) [64] to maintain the pressure. Finally the restraints were removed and the  
687 system was equilibrated for a final 100 *ps* run.

688 **Production of the trajectories.** Each system was simulated during 250 ns (5 replicates of 50 ns,  
689 starting from different initial velocities) in the NPT ensemble using the PMEMD module of AMBER  
690 12. The temperature was kept at 310 K and pressure at 1 bar using the Langevin piston coupling  
691 algorithm. The SHAKE algorithm was used to freeze bonds involving hydrogen atoms, allowing for  
692 an integration time step of 2.0 fs. The Particle Mesh Ewald (PME) method [65] was employed to  
693 treat long-range electrostatics. The coordinates of the system were written every *ps*.

694 **Analysis of the trajectories** Standard analyses of the MD trajectories were performed with the  
695 *ptraj* module of AMBER 12. The calculation of the root mean square deviation (RMSD) over all  
696 atoms indicated that it took between 5 and 20 ns for the systems to relax. Consequently, the last 30  
697 ns of each replicate were retained for further analysis, totaling 150 000 snapshots for each system. The  
698 fluctuations of the C- $\alpha$  atoms were recorded along each replicate. For each residue or each system,  
699 we report the value averaged over the 5 replicates and the standard deviation (see **Fig. 4a**). The  
700 secondary structures were assigned by DSSP algorithm over the whole conformational ensembles.

701 For each residue, the most frequent secondary structure type was retained (see **Fig. 4a** and **S6**  
702 **Fig**). If no secondary structure was present in more than 50% of the MD conformations, then the  
703 residue was assigned to a loop. The amplitude of the motion of the A-loop compared to the rest of  
704 the protein was estimated by computing the angle between the geometric center of residues 189-192,  
705 residue 205 and either residue 211 in the isoforms JNK1 $\alpha$  and JNK1 $\beta$  or residue 209 in the isoform  
706 JNK1 $\delta$ . Only C- $\alpha$  atoms were considered.

707 **Acknowledgments** We thank Y. Christinat for providing information on the algorithm he devel-  
708 oped for the reconstruction of transcript phylogenies.

## 709 **References**

- 710 1. Wang ET, Sandberg R, Luo S, Khrebtkova I, Zhang L, Mayr C, et al. Alternative isoform  
711 regulation in human tissue transcriptomes. *Nature*. 2008 Nov;456(7221):470–476.
- 712 2. Ward AJ, Cooper TA. The pathobiology of splicing. *J Pathol*. 2010 Jan;220(2):152–163.
- 713 3. Lim KH, Ferraris L, Filloux ME, Raphael BJ, Fairbrother WG. Using positional distribution  
714 to identify splicing elements and predict pre-mRNA processing defects in human genes. *Proc*  
715 *Natl Acad Sci USA*. 2011 Jul;108(27):11093–11098.
- 716 4. Mudge JM, Frankish A, Fernandez-Banet J, Alioto T, Derrien T, Howald C, et al. The origins,  
717 evolution, and functional potential of alternative splicing in vertebrates. *Mol Biol Evol*. 2011  
718 Oct;28(10):2949–2959.
- 719 5. Barbosa-Morais NL, Irimia M, Pan Q, Xiong HY, Gueroussov S, Lee LJ, et al. The evolutionary  
720 landscape of alternative splicing in vertebrate species. *Science*. 2012 Dec;338(6114):1587–1593.
- 721 6. Merkin J, Russell C, Chen P, Burge CB. Evolutionary dynamics of gene and isoform regulation  
722 in Mammalian tissues. *Science*. 2012 Dec;338(6114):1593–1599.
- 723 7. Gonzalez-Porta M, Frankish A, Rung J, Harrow J, Brazma A. Transcriptome analysis of  
724 human tissues and cell lines reveals one dominant transcript per gene. *Genome Biol*. 2013  
725 Jul;14(7):R70.
- 726 8. Ezkurdia I, Rodriguez JM, Carrillo-de Santa Pau E, Vazquez J, Valencia A, Tress ML. Most  
727 highly expressed protein-coding genes have a single dominant isoform. *J Proteome Res*. 2015  
728 Apr;14(4):1880–1887.
- 729 9. Kim MS, Pinto SM, Getnet D, Nirujogi RS, Manda SS, Chaerkady R, et al. A draft map of  
730 the human proteome. *Nature*. 2014 May;509(7502):575–581.

- 731 10. Hao Y, Colak R, Teyra J, Corbi-Verge C, Ignatchenko A, Hahne H, et al. Semi-supervised  
732 Learning Predicts Approximately One Third of the Alternative Splicing Isoforms as Functional  
733 Proteins. *Cell Rep.* 2015 Jul;12(2):183–189.
- 734 11. Weatheritt RJ, Sterne-Weiler T, Blencowe BJ. The ribosome-engaged landscape of alternative  
735 splicing. *Nat Struct Mol Biol.* 2016 Dec;23(12):1117–1123.
- 736 12. Birzele F, Kuffner R, Meier F, Oefinger F, Potthast C, Zimmer R. ProSAS: a database for  
737 analyzing alternative splicing in the context of protein structures. *Nucleic Acids Res.* 2008  
738 Jan;36(Database issue):D63–68.
- 739 13. Birzele F, Csaba G, Zimmer R. Alternative splicing and protein structure evolution. *Nucleic  
740 Acids Res.* 2008 Feb;36(2):550–558.
- 741 14. Christinat Y, Moret BM. Inferring transcript phylogenies. *BMC Bioinformatics.* 2012 Jun;13  
742 Suppl 9:S1.
- 743 15. Manning AM, Davis RJ. Targeting JNK for therapeutic benefit: from junk to gold? *Nat Rev  
744 Drug Discov.* 2003 Jul;2(7):554–565.
- 745 16. Kyriakis JM, Avruch J. Mammalian MAPK signal transduction pathways activated by stress  
746 and inflammation: a 10-year update. *Physiol Rev.* 2012 Apr;92(2):689–737.
- 747 17. Hirosumi J, Tuncman G, Chang L, Gorgun CZ, Uysal KT, Maeda K, et al. A central role for  
748 JNK in obesity and insulin resistance. *Nature.* 2002 Nov;420(6913):333–336.
- 749 18. Hunot S, Vila M, Teismann P, Davis RJ, Hirsch EC, Przedborski S, et al. JNK-mediated in-  
750 duction of cyclooxygenase 2 is required for neurodegeneration in a mouse model of Parkinson's  
751 disease. *Proc Natl Acad Sci USA.* 2004 Jan;101(2):665–670.

- 752 19. Brecht S, Kirchhof R, Chromik A, Willesen M, Nicolaus T, Raivich G, et al. Specific patho-  
753 physiological functions of JNK isoforms in the brain. *Eur J Neurosci.* 2005 Jan;21(2):363–377.
- 754 20. Tuncman G, Hirosumi J, Solinas G, Chang L, Karin M, Hotamisligil GS. Functional in vivo  
755 interactions between JNK1 and JNK2 isoforms in obesity and insulin resistance. *Proc Natl  
756 Acad Sci USA.* 2006 Jul;103(28):10741–10746.
- 757 21. Waetzig V, Herdegen T. Context-specific inhibition of JNKs: overcoming the dilemma of  
758 protection and damage. *Trends Pharmacol Sci.* 2005 Sep;26(9):455–461.
- 759 22. Bogoyevitch MA, Kobe B. Uses for JNK: the many and varied substrates of the c-Jun N-  
760 terminal kinases. *Microbiol Mol Biol Rev.* 2006 Dec;70(4):1061–1095.
- 761 23. Yates A, Akanni W, Amode MR, Barrell D, Billis K, Carvalho-Silva D, et al. Ensembl 2016.  
762 *Nucleic Acids Res.* 2016 Jan;44(D1):D710–716.
- 763 24. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, et al. The Protein Data  
764 Bank. *Nucleic Acids Res.* 2000 Jan;28(1):235–242.
- 765 25. Huse M, Kuriyan J. The conformational plasticity of protein kinases. *Cell.* 2002  
766 May;109(3):275–282.
- 767 26. Heo YS, Kim SK, Seo CI, Kim YK, Sung BJ, Lee HS, et al. Structural basis for the se-  
768 lective inhibition of JNK1 by the scaffolding protein JIP1 and SP600125. *EMBO J.* 2004  
769 Jun;23(11):2185–2195.
- 770 27. Liu X, Zhang CS, Lu C, Lin SC, Wu JW, Wang ZX. A conserved motif in JNK/p38-specific  
771 MAPK phosphatases as a determinant for JNK1 recognition and inactivation. *Nat Commun.*  
772 2016;7:10879.

- 773 28. Chamberlain SD, Redman AM, Wilson JW, Deanda F, Shotwell JB, Gerding R, et al. Op-  
774 timization of 4,6-bis-anilino-1H-pyrrolo[2,3-d]pyrimidine IGF-1R tyrosine kinase inhibitors to-  
775 wards JNK selectivity. *Bioorg Med Chem Lett.* 2009 Jan;19(2):360–364.
- 776 29. Gelly JC, Lin HY, de Brevern AG, Chuang TJ, Chen FC. Selective constraint on human  
777 pre-mRNA splicing by protein structural properties. *Genome Biol Evol.* 2012;4(9):966–975.
- 778 30. Laskowski RA, MacArthur MW, Moss DS, Thornton JM. PROCHECK: a program to check  
779 the stereochemical quality of protein structures. *Journal of Applied Crystallography.* 1993  
780 Apr;26(2):283–291. Available from: <http://dx.doi.org/10.1107/s0021889892009944>.
- 781 31. Marti-Renom MA, Stuart AC, Fiser A, Sanchez R, Melo F, Sali A. Comparative protein  
782 structure modeling of genes and genomes. *Annu Rev Biophys Biomol Struct.* 2000;29:291–325.
- 783 32. Kornev AP, Haste NM, Taylor SS, Eyck LF. Surface comparison of active and inactive pro-  
784 tein kinases identifies a conserved activation mechanism. *Proc Natl Acad Sci USA.* 2006  
785 Nov;103(47):17783–17788.
- 786 33. Oruganty K, Talathi NS, Wood ZA, Kannan N. Identification of a hidden strain switch provides  
787 clues to an ancient structural mechanism in protein kinases. *Proc Natl Acad Sci USA.* 2013  
788 Jan;110(3):924–929.
- 789 34. Nicolas A, Raguenes-Nicol C, Ben Yaou R, Ameziane-Le Hir S, Cheron A, Vie V, et al. Becker  
790 muscular dystrophy severity is linked to the structure of dystrophin. *Hum Mol Genet.* 2015  
791 Mar;24(5):1267–1279.
- 792 35. Meszaros B, Simon I, Dosztanyi Z. Prediction of protein binding regions in disordered proteins.  
793 *PLoS Comput Biol.* 2009 May;5(5):e1000376.



- 794 36. Reyes A, Anders S, Weatheritt RJ, Gibson TJ, Steinmetz LM, Huber W. Drift and conservation  
795 of differential exon usage across tissues in primate species. *Proc Natl Acad Sci USA*. 2013  
796 Sep;110(38):15377–15382.
- 797 37. Melamud E, Moulton J. Stochastic noise in splicing machinery. *Nucleic Acids Res*. 2009  
798 Aug;37(14):4873–4886.
- 799 38. Abascal F, Ezkurdia I, Rodriguez-Rivas J, Rodriguez JM, del Pozo A, Vazquez J, et al. Al-  
800 ternatively Spliced Homologous Exons Have Ancient Origins and Are Highly Expressed at the  
801 Protein Level. *PLoS Comput Biol*. 2015 Jun;11(6):e1004325.
- 802 39. Bateman A, Martin MJ, O'Donovan C, Magrane M, Apweiler R, Alpi E, et al. UniProt: a hub  
803 for protein information. *Nucleic Acids Res*. 2015 Jan;43(Database issue):D204–212.
- 804 40. Rodriguez JM, Maietta P, Ezkurdia I, Pietrelli A, Wesselink JJ, Lopez G, et al. APPRIS: an-  
805 notation of principal and alternative splice isoforms. *Nucleic Acids Res*. 2013 Jan;41(Database  
806 issue):D110–117.
- 807 41. Abascal F, Tress ML, Valencia A. The evolutionary fate of alternatively spliced homologous  
808 exons after gene duplication. *Genome Biol Evol*. 2015 Apr;7(6):1392–1403.
- 809 42. Roux J, Robinson-Rechavi M. Age-dependent gain of alternative splice forms and biased dupli-  
810 cation explain the relation between splicing and duplication. *Genome Res*. 2011 Mar;21(3):357–  
811 363.
- 812 43. Ellis JD, Barrios-Rodiles M, Colak R, Irimia M, Kim T, Calarco JA, et al. Tissue-specific alter-  
813 native splicing remodels protein-protein interaction networks. *Mol Cell*. 2012 Jun;46(6):884–  
814 892.

- 815 44. Sakharkar MK, Chow VT, Kanguene P. Distributions of exons and introns in the human  
816 genome. *In Silico Biol (Gedruckt)*. 2004;4(4):387–393.
- 817 45. Sankoff D. Minimal Mutation Trees of Sequences. *SIAM Journal on Applied Mathematics*.  
818 1975;28(1):35–42. Available from: <http://dx.doi.org/10.1137/0128004>.
- 819 46. Alekseyenko AV, Lee CJ, Suchard MA. Wagner and Dollo: a stochastic duet by composing  
820 two parsimonious solos. *Syst Biol*. 2008 Oct;57(5):772–784.
- 821 47. Land AH, Doig AG. An Automatic Method of Solving Discrete Programming Problems.  
822 *Econometrica*. 1960;28:497–520.
- 823 48. Gansner ER, North SC. An open graph visualization system and its applications to software  
824 engineering. *SOFTWARE - PRACTICE AND EXPERIENCE*. 2000;30(11):1203–1233.
- 825 49. Hildebrand A, Remmert M, Biegert A, Soding J. Fast and accurate automatic structure  
826 prediction with HHpred. *Proteins*. 2009;77 Suppl 9:128–132.
- 827 50. Remmert M, Biegert A, Hauser A, Soding J. HHblits: lightning-fast iterative protein sequence  
828 searching by HMM-HMM alignment. *Nat Methods*. 2011 Dec;9(2):173–175.
- 829 51. Jones DT. Protein secondary structure prediction based on position-specific scoring matrices.  
830 *J Mol Biol*. 1999 Sep;292(2):195–202.
- 831 52. Soding J. Protein homology detection by HMM-HMM comparison. *Bioinformatics*. 2005  
832 Apr;21(7):951–960.
- 833 53. Hubbard S, Thornton J; 1992-6. <http://www.bioinf.manchester.ac.uk/naccess/>.
- 834 54. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improve-  
835 ments in performance and usability. *Mol Biol Evol*. 2013 Apr;30(4):772–780.

- 836 55. DeLano WL. The PyMOL Molecular Graphics System; 2002. [Http://www.pymol.org](http://www.pymol.org).
- 837 56. Yang J, Yan R, Roy A, Xu D, Poisson J, Zhang Y. The I-TASSER Suite: protein structure  
838 and function prediction. *Nat Methods*. 2015 Jan;12(1):7–8.
- 839 57. Roy A, Kucukural A, Zhang Y. I-TASSER: a unified platform for automated protein structure  
840 and function prediction. *Nat Protoc*. 2010 Apr;5(4):725–738.
- 841 58. Zhang Y. I-TASSER server for protein 3D structure prediction. *BMC Bioinformatics*. 2008  
842 Jan;9:40.
- 843 59. Ward JJ, McGuffin LJ, Bryson K, Buxton BF, Jones DT. The DISOPRED server for the  
844 prediction of protein disorder. *Bioinformatics*. 2004 Sep;20(13):2138–2139.
- 845 60. Dosztanyi Z, Csizmok V, Tompa P, Simon I. IUPred: web server for the prediction of intrin-  
846 sically unstructured regions of proteins based on estimated energy content. *Bioinformatics*.  
847 2005 Aug;21(16):3433–3434.
- 848 61. Laine E, Carbone A. Local Geometry and Evolutionary Conservation of Protein Surfaces  
849 Reveal the Multiple Recognition Patches in Protein-Protein Interactions. *PLoS Comput Biol*.  
850 2015 Dec;11(12):e1004580.
- 851 62. Case D, Darden T, Cheatham III T, Simmerling C, Wang J, Duke R, et al. AMBER 12.  
852 University of California, San Francisco. 2012;1(2):3.
- 853 63. Berendsen HJC, Postma JPM, van Gunsteren WF, DiNola A, Haak JR. Molecular dynamics  
854 with coupling to an external bath. *The Journal of chemical physics*. 1984;81(8):3684–3690.
- 855 64. Loncharich RJ, Brooks BR, Pastor RW. Langevin dynamics of peptides: The frictional depen-  
856 dence of isomerization rates of N-acetylalanyl-N'-methylamide. *Biopolymers*. 1992;32(5):523–  
857 535.

- 858 65. Darden T, York D, Pedersen L. Particle mesh Ewald: An  $N \log(N)$  method for Ewald sums in  
859 large systems. *The Journal of Chemical Physics*. 1993;98:10089–10092.
- 860 66. Chimnaronk S, Sitthiroongruang J, Srisucharitpanit K, Srisaisup M, Ketterman AJ, Boon-  
861 serm P. The crystal structure of JNK from *Drosophila melanogaster* reveals an evolutionarily  
862 conserved topology with that of mammalian JNK proteins. *BMC Struct Biol*. 2015 Sep;15:17.
- 863 67. Knighton DR, Zheng JH, Ten Eyck LF, Ashford VA, Xuong NH, Taylor SS, et al. Crystal  
864 structure of the catalytic subunit of cyclic adenosine monophosphate-dependent protein kinase.  
865 *Science*. 1991 Jul;253(5018):407–414.
- 866 68. Brown NR, Noble ME, Endicott JA, Johnson LN. The structural basis for specificity of sub-  
867 strate and recruitment peptides for cyclin-dependent kinases. *Nat Cell Biol*. 1999 Nov;1(7):438–  
868 443.

## 869 **Supporting Information Captions**

### 870 **S1 Text**

### 871 **S1 Table**

### 872 **3D structures of human JNKs deposited in the Protein Data Bank.**

### 873 **S2 Table**

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

### 880 **S1 Fig**

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

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

## 895 **S2 Fig**

### 896 **Comparison of exon mapping onto JNK tertiary structure between human and drosophila.**

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

## 901 **S3 Fig**

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

## 908 **S4 Fig**

909 **Catalytic and regulatory spines.** (**a**) The PKA kinase (PDB code: 2CPK [67]) is taken as a  
910 reference to illustrate the spines identified in all kinases in [32]. (**b-d**) 3D models for the human  
911 isoforms JNK1 $\alpha$  (**b**), JNK1 $\beta$  (**c**) and JNK1 $\delta$  (**d**). The residues comprising the spines are displayed as  
912 sticks and transparent surfaces and are labelled. The catalytic spine is in yellow. It is anchored by  
913 two hydrophobic residues located in the F-helix (in all panels, except for **d**, JNK1 $\delta$ ). The regulatory  
914 spine is in red. It is anchored by the N-terminal aspartate of the F-helix, namely D220 in PKA (**a**)

915 and D207 in JNK (b-c) that forms an H-bond with Y164 (a) or H149 (b-c). In JNK1 $\delta$ , the aspartate  
916 is replaced by D282 and the H-bond with H149 is maintained (d).

## 917 **S5 Fig**

918 **Hydrogen-bond pattern associated to the HRD motif backbone strain.** (a) The CDK-  
919 substrate complex (PDB code: 1QMZ [68]) is taken as a reference to illustrate the hidden strain  
920 identified in kinases structures in [33]. (b-d) 3D models for the human isoforms JNK1 $\alpha$  (b), JNK1 $\beta$   
921 (c) and JNK1 $\delta$  (d). Hydrogen-bonds are formed between 3 components: the aspartate in the F-helix,  
922 the HRD motif in the catalytic loop and the DFG motif in the activation loop. The 3 component  
923 are conserved across the whole protein kinase family.

## 924 **S6 Fig**

925 **Secondary structures for the human JNK1 isoforms.** The secondary structures were recorded  
926 along MD simulations of human JNK1 $\alpha$  (on top), JNK1 $\beta$  (in the middle) and JNK1 $\delta$  (at the bottom).  
927 For each residue, the most frequent secondary structure type is depicted. If a residue does not  
928 adopt a well-defined secondary structure, either  $\alpha$ -helix or  $\beta$ -sheet, for more than 50% of the MD  
929 conformations, then it is assigned to a loop.

## 930 **S7 Fig**

931 **H-bond pattern and backbone strain of the HRD motif.** (a) Persistence of the H-bonds  
932 formed between H149 and R150 from the HRD motif in the catalytic loop and D207 from the F-helix  
933 and D169 from the DFG motif in the catalytic site, recorded along the 5 replicates of MD simulations  
934 for JNK1 $\alpha$ , JNK1 $\beta$  and JNK1 $\delta$ . (b) Ramachandran plot showing the torsion angle values ( $\phi$ ,  $\psi$ )  
935 of R150 from the HRD motif in the MD conformations of the 3 isoforms.

936 **S8 Fig**

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

940 **S9 Fig**

941 **Transcripts' phylogeny reconstructed by PhyloSofS for the JNK family.** The forest is  
942 comprised of 7 phylogenetic trees, 21 deaths and 14 orphan leaves. The cost of the phylogeny is 69  
943 (with  $C_B = 3$ ,  $C_D = 0$  and  $\sigma = 2$ ). The legend is the same as in **Fig. 1b**. Compared to **Figure**  
944 **1b**, there is one additional subnode (in orange) in the internal nodes A18 and A24 and one JNK1  
945 transcript in *fugu* is the leaf of the orange tree instead of the yellow one.

946 **S10 Fig**

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

952 **S11 Fig**

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



957 **S12 Fig**

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

965

## Figures

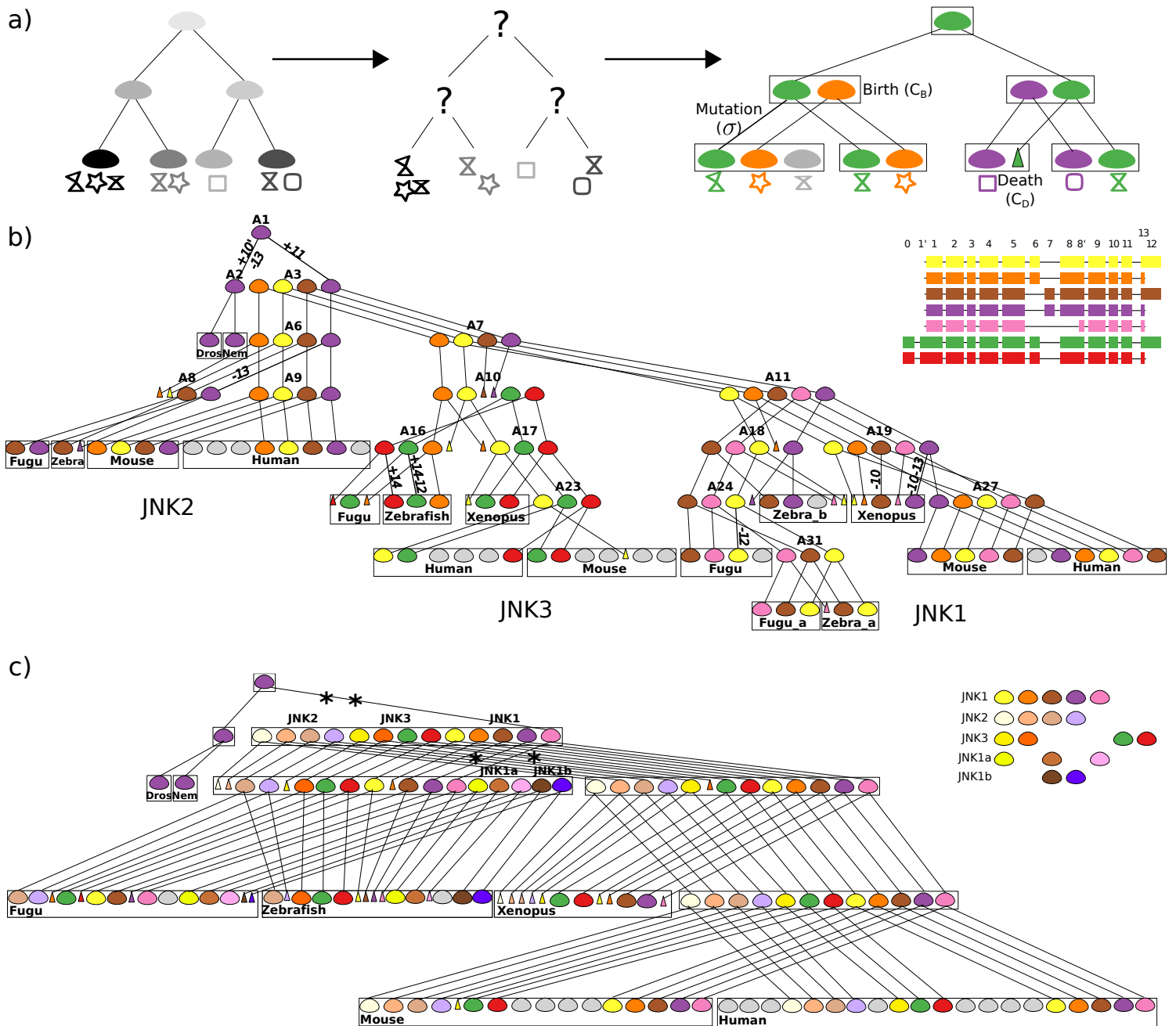


Figure 1:

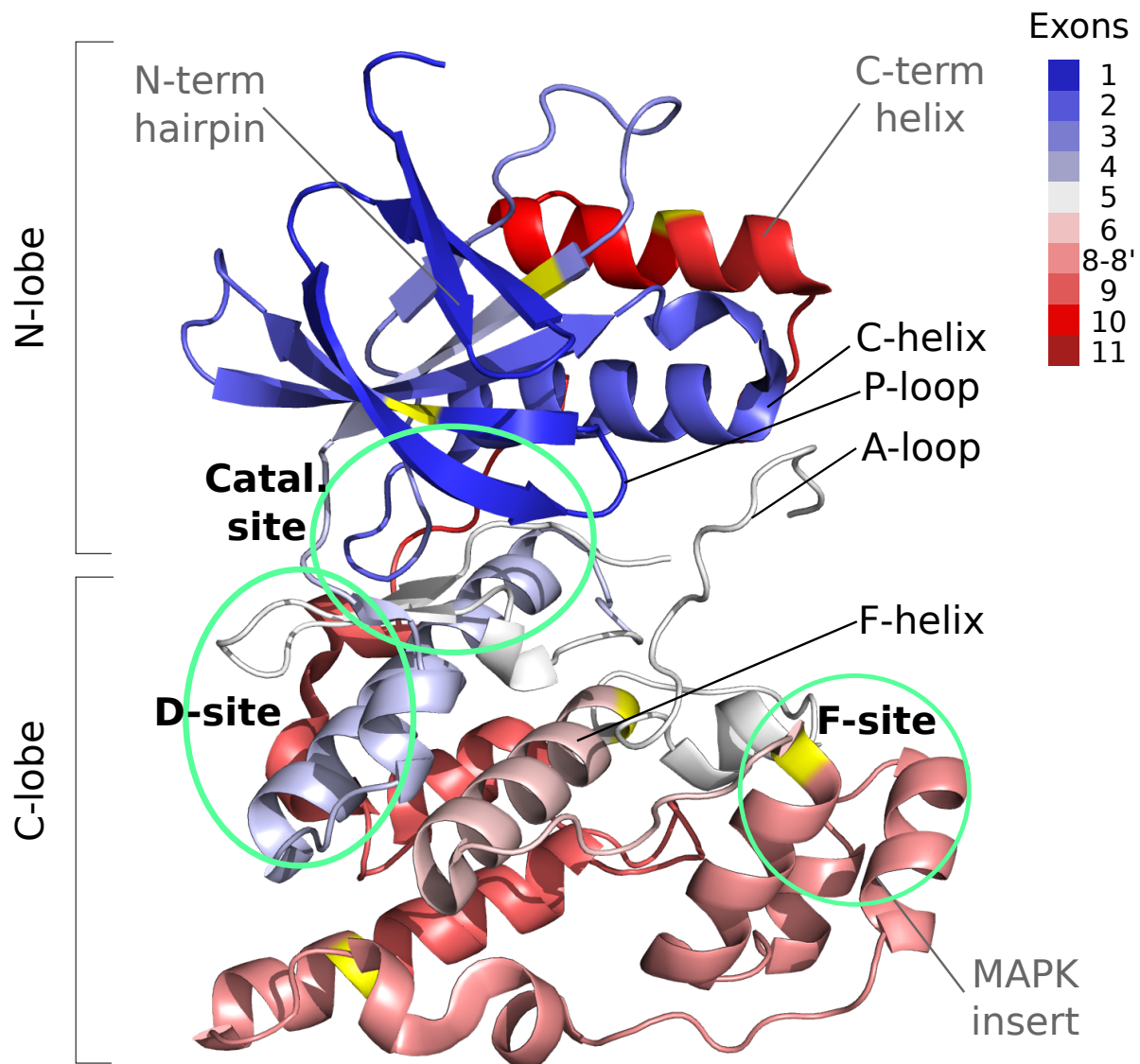


Figure 2:

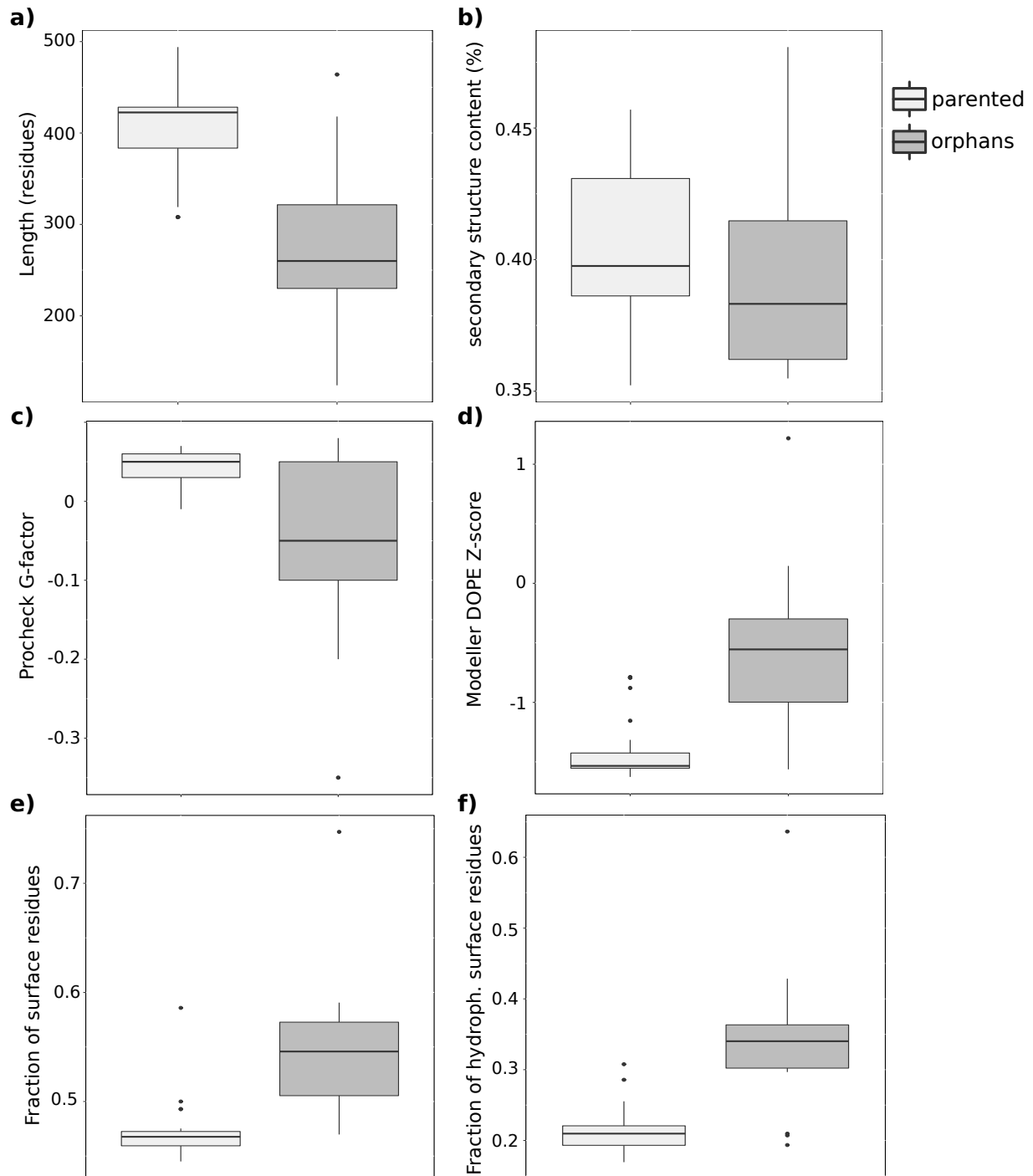


Figure 3:

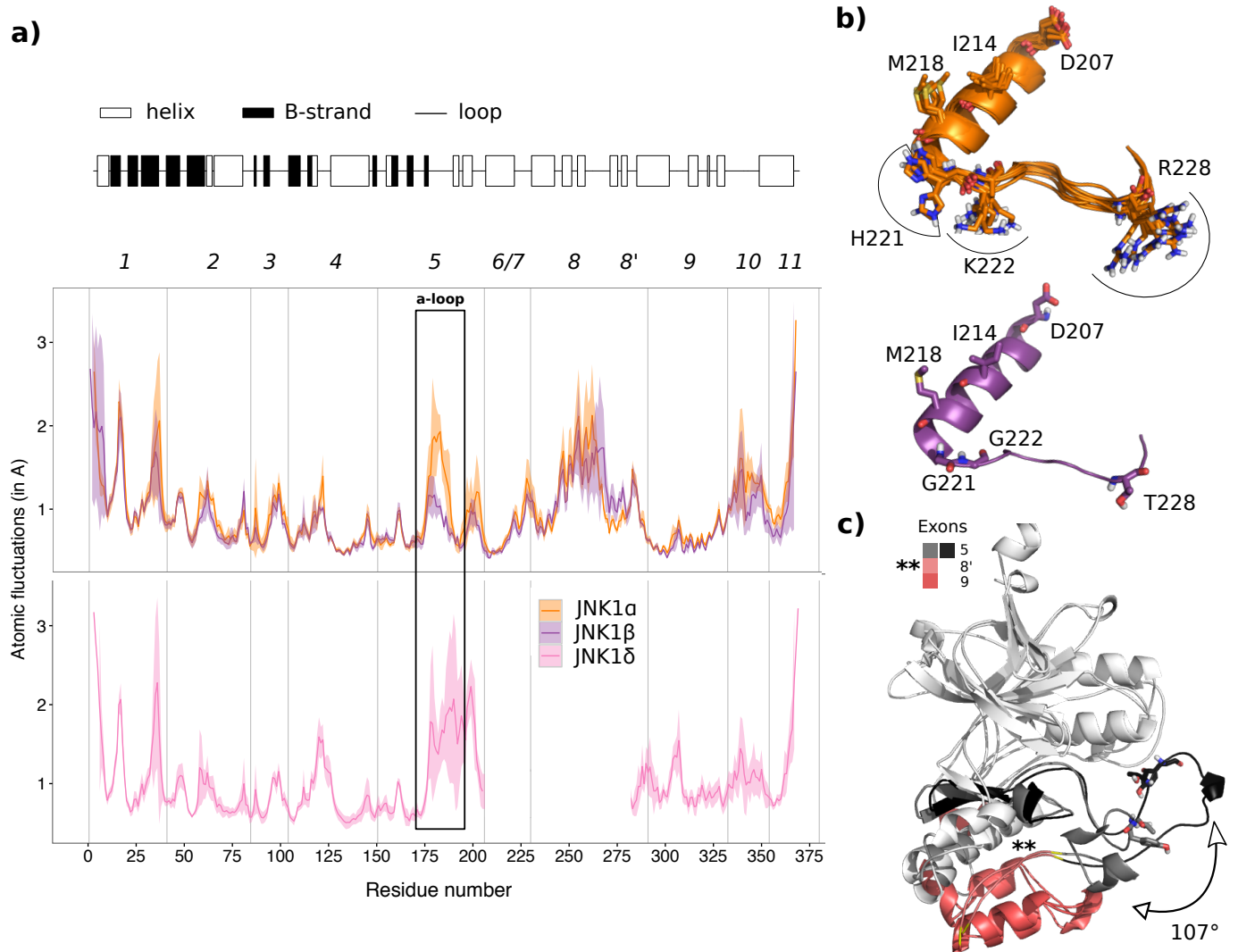


Figure 4:

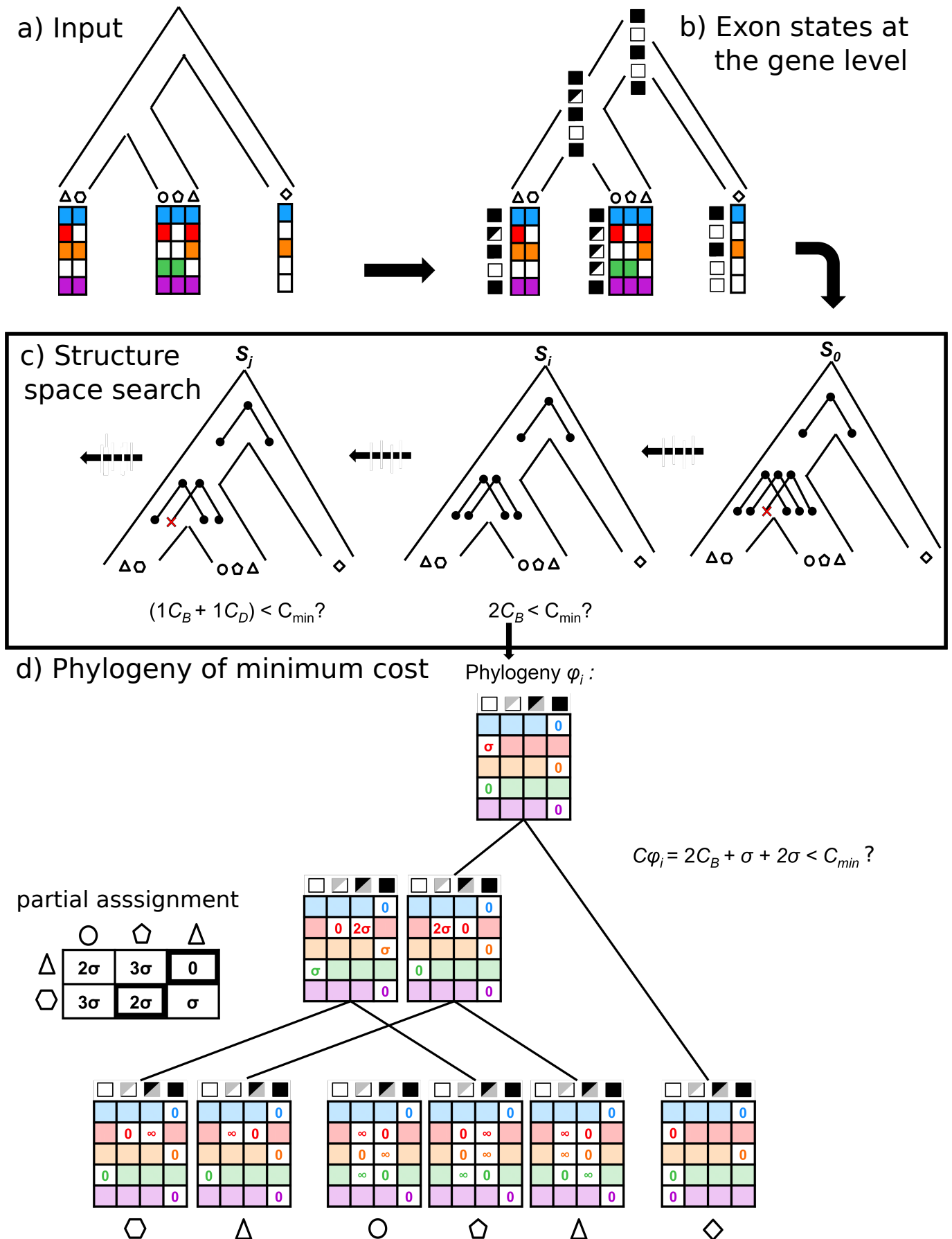


Figure 5: