



HAL
open science

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

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

► **To cite this version:**

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

HAL Id: hal-02884145

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

Submitted on 29 Jun 2020

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

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

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

26 **Keywords:**

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

28

29 **Abbreviations** ¹

30 **Introduction**

31 Alternative splicing (AS) of pre-mRNA transcripts and alternative transcription initi-
32 ation/termination are essential eukaryotic regulatory processes. They can impact the
33 regulation of gene expression, for instance by introducing changes in the three prime
34 untranslated region⁶¹. Or they can directly modify the content of the coding sequence
35 (CDS)²⁶, leading to different protein isoforms. Virtually all multi-exons genes in verte-
36 brates are subject to AS⁶⁸ and about 25% of the AS events (ASEs) common to human
37 and mouse are conserved in other vertebrates^{3;49;51}. This suggests an important role for
38 AS in expanding the protein repertoire through evolution. AS has also gained interest for
39 medicinal purpose, as the ratio of alternatively spliced isoforms is imbalanced in several
40 cancers^{43;70}.

41 The extent to which the ASEs detected at the gene level actually result in functional
42 protein isoforms in the cell remains largely unknown. Transcriptomics and proteomics
43 studies suggested that most highly expressed human genes have only one single dominant
44 isoform^{22;25}, but the detection rate of these experiments is very difficult to assess³⁷ and
45 likely suffer from strong experimental detection bias⁶⁹. A recent analysis of ribosome
46 profiling data suggested that a major fraction of splice variants is translated, with direct
47 implications on specific cellular functions⁷². Moreover, a large scale assessment of isoforms
48 present in the cell revealed that the majority of isoform pairs share less than 50% of their
49 interactions⁷⁴. From a structural perspective, very few alternatively spliced isoforms
50 have been characterized and are available in the Protein Data Bank (PDB)^{7;27}. It was

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

51 shown that the boundaries of single constitutive exons or of co-occurring exon pairs tend
52 to overlap those of compact structural units, called protein units²⁴. Moreover, tissue-
53 specific alternatively spliced exons are enriched in disordered regions containing binding
54 motifs¹³. It was also suggested that splicing events may induce major fold changes^{9;10},
55 and a few cases of isoforms displaying domain atrophy while retaining some activity have
56 been reported⁵⁴.

57 The elusiveness of the significance of AS for protein function and fold diversification
58 though evolution has stimulated the development of knowledge bases, such as APPRIS⁵⁷
59 and Exon Ontology⁶⁵. They provide functional and sequence-based information at the
60 level of the transcript or the exon. A method reconstructing transcripts' phylogenies was
61 also proposed and proved useful for enhancing transcriptome reconstruction from ESTs
62 and investigating proteins functional features (domains, sites)^{16;17}.

63 In this work, we combine sequence- and structure-based information to shed led on
64 the evolution of AS. We have developed PhyloSofS (Phylogenies of Splicing Isoforms
65 Structures), an automated tool that infers plausible evolutionary scenarios explaining
66 an ensemble of protein coding transcripts observed in a set of species and predicts the
67 tertiary structures of the protein isoforms. We show how PhyloSofS can be used to
68 identify and date ASEs, and also shed light on the molecular mechanisms underlying
69 their functional outcome in the c-Jun N-terminal kinase (JNK) family. This choice was
70 motivated by the fact that JNKs are among the few tens of families for which alternative
71 transcripts performing different functional tasks have been experimentally identified and
72 characterized^{8;36;64}. Moreover, they play essential regulatory roles by targeting specific
73 transcription factors (c-Jun, ATF2...) in response to cellular stimuli. The deregulation of
74 their activity is associated with various diseases (cancer, inflammatory diseases, neuronal
75 disorder...) which makes them important therapeutic targets⁴⁶. About ten JNK splicing
76 isoforms have been documented in the literature³⁹. They were shown to perform different
77 context-specific tasks^{12;30;32;66} and to have different affinities for their substrates^{11;67}. By
78 reconstructing the phylogeny of JNK transcripts across seven species, we identify two
79 ASEs of ancient origin. We further identify key residues that may be responsible for

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

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

95 Computational method

96 PhyloSofS can be applied to single genes or to gene families. Given a gene tree and the
97 observed protein coding transcripts at the leaves (**Fig. 1a**, on the left), it reconstructs
98 a phylogenetic forest embedded in the gene tree (**Fig. 1a**, on the right) representing
99 plausible evolutionary scenarios explaining the transcripts. Each tree in the forest (in
100 orange, green or purple) represents the phylogeny of one transcript. In other words,
101 each root indicates the appearance of a new transcript and its corresponding ASE(s).
102 Transcript losses are possible (triangles in **Fig. 1a**), and the exon usage of a transcript
103 can change along the branches upon the inclusion or exclusion of one or several exons
104 (which we refer to as “mutations”). The underlying evolutionary model is comprised
105 of two levels, following¹⁶. At the level of the gene, exons can be absent, constitutive, or
106 alternative (*i.e.* involved in at least one ASE), whereas at the level of the transcript, exons
107 are either present or absent. The cost associated to the mutation of an exon naturally

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

111 PhyloSofS algorithm seeks to determine the scenario with the smallest number of
112 evolutionary events, following the maximum parsimony principle. It is inspired from that
113 reported in¹⁶. Our main contribution was to develop heuristics in order to treat complex
114 cases in a computationally tractable way. Specifically, we have implemented a multi-
115 start iterative strategy combined with a systematic local exploration around the best
116 current solution to efficiently search the space of phylogenetic forests (see *Methods* and
117 *Supplementary Fig. S1*). Moreover, we have designed a branch-and-bound algorithm
118 adapted to the problem of assigning transcripts between parent and child nodes (see
119 *Methods* and *Supplementary Text S1*). The reconstructed forests are provided with a user-
120 friendly visualization (**Fig. 1b-c**). In addition to phylogenetic reconstruction, PhyloSofS
121 predicts the 3D structures of the protein isoforms. The predictions are performed based
122 on comparative modeling using the HH-suite²⁹. Furthermore, PhyloSofS annotates the
123 generated models with sequence (exon boundaries) and structure (secondary structure,
124 solvent accessibility, model quality) information. For example, it is very easy to visualize
125 the location of each exon on the modeled structure. Here, we present the application
126 of PhyloSofS to the c-Jun N-terminal kinase (JNK) family across 7 species (*H. sapiens*,
127 *M. musculus*, *X. tropicalis*, *T. rubripes*, *D. rerio*, *D. melanogaster* and *C. elegans*). This
128 case represents a high degree of complexity with 60 observed transcripts assembled from
129 a total of 19 different exons. Most of these transcripts comprise more than 10 exons and
130 the number of transcripts per gene per species varies from 1 to 8 (**Fig. 1b-c**).

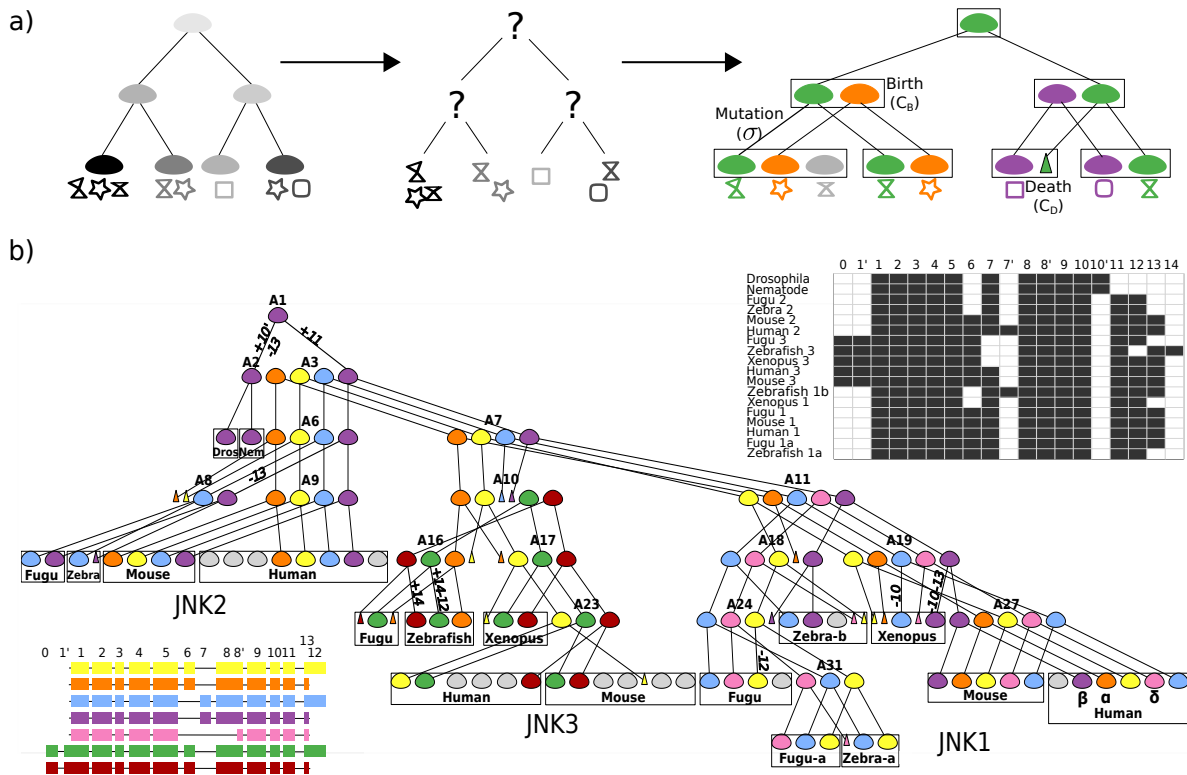


Figure 1: **Transcripts' phylogenies reconstructed by PhyloSofS.** (a) On the left, example of a phylogenetic gene tree where 8 transcripts (represented by geometrical symbols) are observed in 4 current species (leaves of the tree, colored in different grey tones). These data are given as input to PhyloSofS.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

Results

Transcripts' phylogeny for the JNK family.

The observed transcripts were collected from the Ensembl⁷⁵ database (see *Methods*). PhyloSofS algorithm was run for 10^6 iterations on the JNK family gene tree and we retained the most parsimonious evolutionary scenario (cost = 69, see *Methods* for a detailed description of the parameters). The reconstructed forest is comprised of 7 transcript trees (**Fig. 1b**, each tree is colored differently). Each transcript is described as a collection of exons, numbered from 0 to 14 (**Fig. 1b**, top right and bottom left corners, and see *Methods* for more details on the numbering). We could reconstruct a phylogeny for 46 out of the 60 observed transcripts. The 14 “orphan” transcripts (leaves in grey) are not conserved across the studied species, and thus likely result in non-functional protein products. Mutations occurring along the branches of the trees are labelled (**Fig. 1b**, see +/− symbols followed by the number of the included/excluded exon). In total, JNK transcripts' phylogeny comprises 11 mutations.

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

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

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

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

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

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

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

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

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

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

204 About eighty structures of human JNKs are available in the PDB (*Supplementary Table*
205 *S1*). This abundance of structural data can be explained by the fact that JNKs are im-
206 portant therapeutic targets and they were crystallized with different inhibitors. The three
207 paralogues share the same fold, which is highly conserved among protein kinases. The
208 structures are highly redundant, with an average root mean square deviation (RMSD) of
209 $1.96 \pm 0.71 \text{ \AA}$, computed over more than 80% of the protein residues. In order to visualize
210 the correspondence between the gene structure and the protein secondary and tertiary
211 structures, the exons were mapped onto a high-resolution PDB structure (3ELJ¹⁵) of
212 human JNK1 (**Fig. 2**, each exon is colored differently). One can observe that the orga-
213 nization of the protein 3D structure is preserved by the 1D structure of the gene. Most
214 of the secondary structures (10 over 12 α -helices and 7 over 9 β -strands) are completely
215 included in single exons. Moreover, each one of the regions important for the structural
216 stability and/or function of protein kinases (**Fig. 2**, labelled in black) is included in one
217 single exon (see also *Supplementary Table S2*). So are the N-terminal hairpin and the
218 MAPK insert (labelled in grey), two structural motifs specific to the mitogen-activated
219 protein kinase (MAPK) type, to which the JNKs belong. By contrast, binding sites for
220 cofactors and substrates (green circles, see also *Supplementary Table S2*) are comprised
221 of residues belonging to different exons. This is expected as binding sites are comprised
222 of segments that can be very far from each other along the protein sequence. Of note, the
223 block formed by exons 1 to 5, comprising the N-terminal lobe and the A(ctivation)-loop
224 (**Fig. 2**, from blue to white), is constitutively present in all transcripts belonging to
225 the colored trees on **Fig. 1b**. The correspondence was also analyzed for the JNK pro-
226 tein from *Drosophila melanogaster*, whose 3D structure is very similar to that of human
227 JNK1 (*Supplementary Fig. S4*, RMSD of 0.68 \AA). The JNK gene from the *Drosophila*
228 *melanogaster* genome comprises much fewer exons than the human gene. The match be-
229 tween the borders of these exons and the borders of the secondary structures and known
230 important regions is even better in that species. Considering the high degree of conserva-
231 tion of JNK sequences, one may hypothesize that a good match also exists in all studied

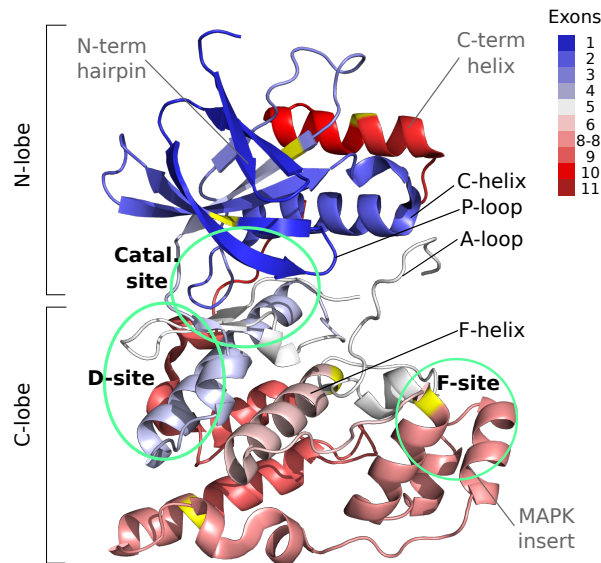


Figure 2: **Exons mapped onto the tertiary structure of human JNK1.** The protein (residues 7 to 364) is represented as a cartoon and the different exons are colored from blue through white to red. The residues in yellow are at the junction of 2 exons. It should be noted that exons 8 and 8' used in PhyloSofS actually correspond to only one genomic exon (see *Methods*). The regions labelled in black are common to kinases and were reported in the literature (see³³) for playing important roles in their structural stability and/or function. The regions labelled in grey are specific to MAP kinases. The green circles indicate the catalytic site and binding sites for JNK cellular partners^{28;44}. The structure was solved by X-ray crystallography at 1.80 Å resolution (PDB code: 3ELJ¹⁵).

232 species. Our observation is in agreement with a previous study establishing a relationship
 233 between exon boundaries and structurally consistent protein regions²⁴.

234 **Properties of the orphan transcripts.**

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

243 content, both types of transcripts contain about 40% of residues predicted in α -helices or
244 β -sheets (**Fig. 3b**). Third, the 3D models generated by PhyloSofS molecular modeling
245 routine for the orphan transcript isoforms are of poorer quality than those for the tran-
246 scripts belonging to a phylogeny (**Fig. 3c-d**). The quality of the models was assessed by
247 computing Procheck⁴² G-factor and Modeller⁴⁷ normalized DOPE score (**Fig. 3c-d**). A
248 model resembling experimental structures deposited in the PDB should have a G-factor
249 greater than -0.5 (the higher the better) and a normalized DOPE score lower than -1 (the
250 lower the better). The distributions obtained for the parented isoforms are clearly shifted
251 toward better values and are more narrow than those for the orphan transcripts. Finally,
252 the proportion of protein residues being exposed to the solvent (relative accessible surface
253 area $rsa > 25\%$) is significantly higher for the orphan isoforms (**Fig. 3e**), as is the pro-
254 portion of hydrophobic residues being exposed to the solvent (**Fig. 3f**). Overall, these
255 observations suggest that simple sequence and structure descriptors enable to distinguish
256 the orphan transcripts from the ones within a phylogeny and that the formers display
257 properties likely reflecting structural instability (large truncations, poorer quality, larger
258 and more hydrophobic surfaces).

259 **Subtle changes in the protein's internal dynamics linked to sub-** 260 **strate differential affinity.**

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

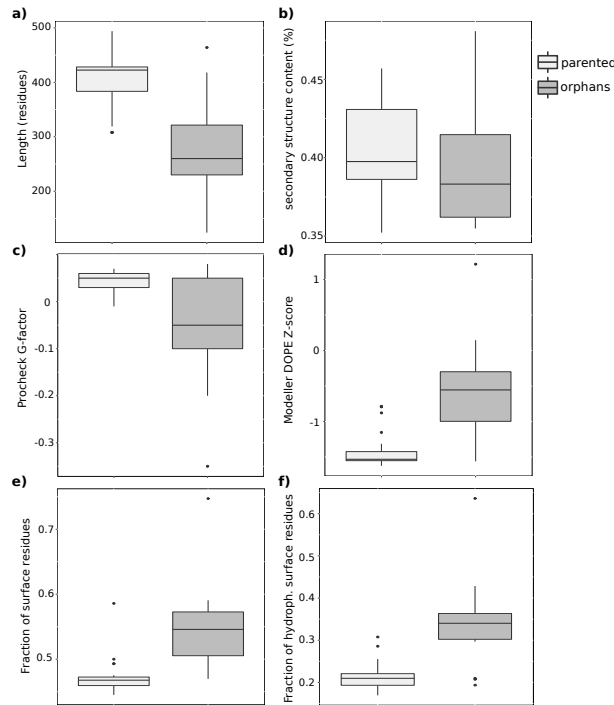


Figure 3: **Structural features of the transcript isoforms.** Distributions are reported for the parented transcripts (in light gray) and the orphan transcripts (in dark grey) in the transcripts' phylogeny (see **Fig. 1b**). **(a)** Length of the transcript (in residues). **(b)** Predicted secondary structure content (in percentages of residues). **(c)** Overall G-factor computed by Procheck⁴². **(d)** Normalized DOPE score computed by Modeller⁴⁷. **(e)** Fraction of protein residues being exposed to the solvent ($rsa > 0.25$). **(f)** Fraction of hydrophobic protein residues being exposed to the solvent ($rsa > 0.25$).

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

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

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

285 To further investigate the potential impact of the inclusion/exclusion of exon 6 or 7 on
286 the dynamical behavior of the protein, we performed all-atom molecular dynamics (MD)
287 simulations of the human isoforms colored in orange and purple on **Figure 1b**. We shall
288 refer to these isoforms as JNK1 α (with exon 6) and JNK1 β (with exon 7), in agreement
289 with the nomenclature found in the literature⁶⁷. JNK1 α and JNK1 β were simulated in
290 explicit solvent for 250 ns (5 replicates of 50 ns, see *Methods*). The backbone atomic
291 fluctuation profiles of the two isoforms are very similar (**Fig. 4a**, orange and purple
292 curves), except for the A-loop which is significantly more flexible in JNK1 α : the region
293 from residue 176 to 188 displays averaged C α fluctuations of 1.55 ± 0.28 Å in JNK1 α
294 and of 0.98 ± 0.16 Å in JNK1 β (**Fig. 4a**). We should stress that this loop displays the
295 highest deviations among the JNK structures available in the PDB and often comprises
296 unresolved residues. The two exons, 6 and 7, have similar backbone flexibility. In the
297 F-helix, the anchor residues for the spines, D207, I214 and M218 adopt stable and very
298 similar conformations (**Fig. 4b**). Moreover, the HRD backbone strain and the associated
299 H-bond pattern are maintained along the simulations of both systems (*Supplementary*
300 *Fig. S8a-b*). Consequently, the observations realized on the static 3D models hold true
301 when simulating their dynamical behavior: the 6/7 variation does not induce any drastic
302 change on the protein's overall shape and behaviour.

303 Nevertheless, we observe differences in the side-chain flexibilities of a few residues
304 lying in the loop following the F-helix between the two isoforms (**Fig. 4b**). On the one
305 hand, in exon 6 (in orange), the polar and positively charged residues H221, K222 and
306 R228 are exposed to the solvent and display large amplitude side-chain motions. These
307 amino acids are 100% conserved in exon 6 across all species (*Supplementary Fig. S5*).
308 On the other hand, in exon 7 (**Fig. 4b**, in purple), G221, G222 and T228 have small side
309 chains with much reduced motions. While G221 is conserved across all species, position
310 222 is variable and position 228 features G, T or S (*Supplementary Fig. S5*). This region
311 of the protein is involved in the binding of substrates (see **Fig. 2**, F-site). Moreover, in

312 both isoforms, we predicted residues 223-230 as directly interacting with cellular partners
313 (see *Methods*). Consequently, one may hypothesize that the differences highlighted here
314 may be crucial for substrate molecular recognition specificity. The positive charges, high
315 fluctuations, high solvent accessibility and high conservation of residues H221, K222 and
316 R228 in JNK1 α support a determinant role for these residues in selectively recognizing
317 specific substrates.

318 **Structural dynamics of a newly identified isoform.**

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

324 JNK1 δ displays a large deletion (of about 80 residues), lacking exons 6, 7 and 8. It
325 does not contain the F-helix, shown to be crucial for kinases structural stability³⁸, nor
326 the MAPK insert, involved in the binding of the phosphatase MKP7⁴⁴ (**Fig. 2**). The 3D
327 model generated by PhyloSofS superimposes well to those of JNK1 α and JNK1 β , with a
328 RMSD lower than 0.5 Å on 245 residues. This is somewhat expected as we use homology
329 modeling. Nevertheless, cases were reported in the literature where homology modeling
330 detected big changes in protein structures induced by exon skipping⁵². In the model of
331 JNK1 δ , the F-helix present in JNK1 α and JNK1 β (residues 207 to 220) is replaced by
332 a loop (residues 282 to 288) corresponding to exon 8' (**Fig. 4c**, indicated by the two
333 stars). The sequence of this loop (exon 8') does not share any significant identity with
334 the F-helix (N-terminal parts of exons 6 and 7), except for the N-terminal residue which
335 is an aspartate, namely D282 (D207 in JNK1 α and JNK1 β). This replacement results in
336 the regulatory spine being intact in JNK1 δ (*Supplementary Fig. S6d*, in red). Moreover,
337 the HRD motif's strained backbone conformation and the associated H-bond pattern,
338 which are stabilized by the aspartate, are maintained (*Supplementary Fig. S7d*). By
339 contrast, the catalytic spine lacks its two anchors (*Supplementary Fig. S6d*, in yellow).

340 JNK1 δ was simulated in explicit solvent for 250 ns (5 replicates of 50 ns). The
341 isoform displays stable secondary structures (*Supplementary Fig. S9*, at the bottom)
342 and atomic fluctuations comparable to those of JNK1 α and JNK1 β (**Fig. 4a**, pink
343 curve to be compared with the purple and orange curves). The C α atomic fluctuations
344 averaged over the loop replacing the F-helix are of 0.88 ± 0.18 Å. This is higher than
345 the values computed for the F-helix in JNK1 α and JNK1 β (0.57 ± 0.10 Å and $0.53 \pm$
346 0.09 Å), but it still indicates a limited flexibility. Moreover, the N-terminal aspartate
347 D282 establishes stable H-bonds with the HRD motif along all but one of the replicates
348 (*Supplementary Fig. S8a*, on the right) and the HRD motif's backbone remains in a
349 strained conformation (*Supplementary Fig. S8b*, on the right), as was observed for JNK1 α
350 and JNK1 β . Consequently, JNK1 δ seems stable in solution, and, as observed on the
351 static 3D model, the absence of the F-helix in this isoform is partially compensated by
352 the presence of D282, which is sufficient to maintain H-bonds with the HRD motif and a
353 resulting backbone strain of the motif, important for kinase structural stability.

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

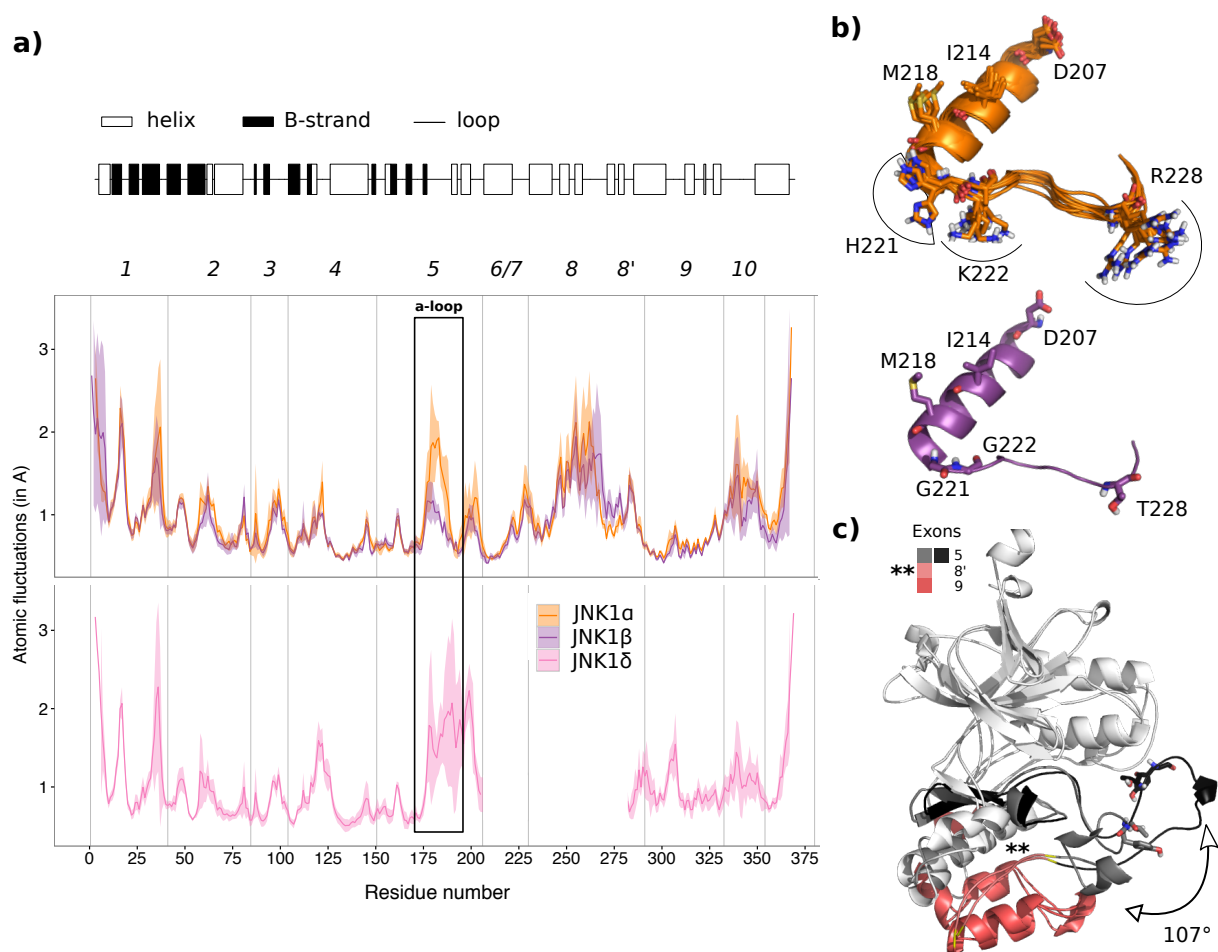


Figure 4: **Dynamical behavior of the human JNK1 isoforms in solution.** (a) The secondary structures for JNK1 α (with exon 6) are depicted on top (the profiles for the 2 other isoforms are very similar, see *Supplementary Fig. S9*). The atomic fluctuations (computed on the $C\alpha$) averaged over 5 50-ns MD replicates are reported for JNK1 α in orange, JNK1 β in purple and JNK1 δ 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 α (in orange) and only 1 for JNK1 β (in purple). (c) Superimposed pair of MD conformations illustrating the amplitude of the A-loop motion in JNK1 δ (see *Materials ad Methods* for details on the calculation of the angle). Exons 5, 8' and 9 are indicated by colors and labels. For clarity, 8' is also indicated by two stars on the structure.

Unresolved residues in the 3D models.

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

At the C-terminus, exons *12* and *13* are completely predicted as intrinsically disordered (*Supplementary Fig. S12a* and *Supplementary Fig. S12b*, blue curve). At the N-terminus, exons *0* and *1'* contain two segments of about 10 residues predicted as disordered protein-binding regions (*Supplementary Fig. S12b*, orange curve), *i.e.* regions unable to form enough favorable intra-chain interactions to fold on their own and likely stabilized upon interaction with a globular protein partner⁵⁰. These exons are present in only two JNK3 transcript isoforms (**Fig. 1b**, colored in dark red and green).

Alternative phylogenies and robustness to parameter changes

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

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

399 To further assess the robustness of our results, we ran PhyloSofS algorithm with differ-
400 ent parameters and analyzed the output phylogenies. The three main parameters of the
401 algorithm are the costs C_B , C_D and σ , associated to the creation (birth), the loss (death)
402 and the mutation of a transcript, respectively. The forests described above were obtained
403 by setting the death cost to zero ($C_D = 0$) and the ratio between the birth and mutation
404 costs below 2 ($C_B/\sigma = 3/2 = 1.5$). The choice of not penalizing death was motivated
405 by the fact that the transcriptome data and annotations we are working with may be
406 incomplete. Indeed, the different genomes available in Ensembl are not annotated with
407 the same accuracy. The choice of tolerating few mutations within each tree was moti-
408 vated by the fact that several pairs of transcripts from the same species differ by only two
409 exons (**Fig. 1b**, compare exon compositions in the bottom left corner). Varying slightly
410 the birth-to-mutation ratio while maintaining the death cost to zero did not impact the
411 results (**Fig. 5** and *Supplementary Table S3*, compare combinations 5.0.3 and 5.0.4 with
412 3.0.2). A bigger parameter change, resulting in a birth-to-mutation ratio larger than 2,
413 had a minor impact on the results (**Fig. 5**, combination 5.0.2). The numbers of orphans
414 and deaths were slightly modified, but the main information contained in the phylogenies
415 remained the same (*Supplementary Table S3*). Penalizing death reduced the number of
416 deaths, as expected, and increased the number of orphans (**Fig. 5**, combinations 5.3.2,
417 3.3.2 and 2.2.2), while the resulting scenarios were less parsimonious (*Supplementary Ta-*
418 *ble S3*). For example, the JNK1 δ transcript was created twice, at the internal nodes
419 A24 and A27 (*Supplementary Table S3*). The two mutually exclusive events involving

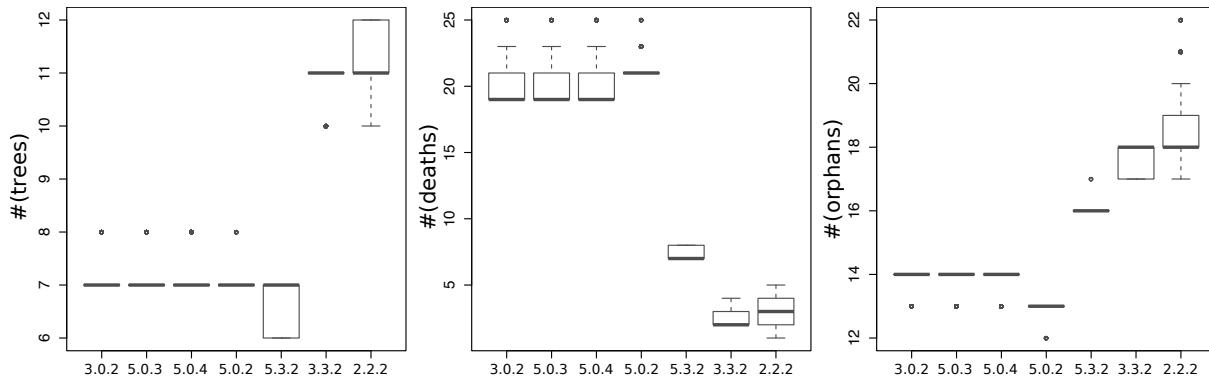


Figure 5: **Statistical analysis of the phylogenies obtained with different parameters.** The numbers of trees, deaths and orphans are given in y-axis for seven combinations of parameters. The parameter values are indicated on the x-axis, in the following order: $C_B.C_D.\sigma$. The distributions are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

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

436 Complexity and comparison with other methods

437 The size of the search space for the transcripts' phylogeny reconstruction grows exponen-
 438 tially with the number of observed transcripts (leaves). To explore that space, the heuris-
 439 tic algorithm implemented in PhyloSofS relies on a multi-start iterative procedure and
 440 on the computation of a lower bound to early filter out unlikely scenarios (see *Methods*).
 441 As the algorithm finds better and better solutions, the filtering procedure becomes more
 442 and more efficient. Transcripts assignment (**Fig. 1a**, in the middle) is performed only
 443 when the computed lower bound passes the filter. At each internal node, k (among n_l)
 444 transcripts from the left child must be paired with k (among n_r) transcripts from the right
 445 child. We solve this problem with a branch-and-bound algorithm whose complexity is of
 446 $\mathcal{O}(n^3)$ for $n_l = n_r = n$ (see details in *Supplementary Text S1*). Hence, it requires about
 447 $(s-1)n^3$ operations to pair up transcripts from the leaves to the root, with s the number of
 448 current species ($s-1$ being the number of ancestral species). Another heuristic algorithm,
 449 making use of a neighbor-joining operation, has been proposed in the literature¹⁷. This al-
 450 gorithm requires to look at up to $\binom{(s-1)n}{2} + \binom{(s-1)(n-1)}{2} + \binom{(s-1)(n-1)(n-2)}{2} + \dots + 1 \simeq \mathcal{O}(s^3n^3)$
 451 possible pairs of transcripts at the level just below the root, and recursively applies a
 452 neighbor-joining procedure to each one of these pairs down to the leaves. One advantage
 453 of this algorithm is that it constrains the space of considered phylogenies and thus may
 454 in practice be more efficient than PhyloSofS. However, it does not consider transcript
 455 mutations and hence cannot reconstruct ancestral transcripts (see Fig. 1A in¹⁷). More-
 456 over, the information about transcript ancestry is virtually lost as trees can be merged
 457 by the neighbor-joining operation (see Fig. 1B in¹⁷). As a consequence, this algorithm
 458 is not suitable for inferring evolutionary scenarios explaining observed transcripts nor for
 459 dating AS events.

460 Discussion

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

463 been subject to much debate^{48;56}. The present work contributes to elaborating strategies
464 to answer it, by crossing sequence analysis and phylogenetic inference with molecular
465 modeling. We report the first joint analysis of the evolution of alternative splicing across
466 several species and of its structural impact on the produced isoforms. The analysis was
467 performed on the JNK family, which represents a high interest for medicinal research and
468 for which a number of human isoforms have been described and biochemically character-
469 ized.

470 Firstly, our results allowed dating an ASE consisting of two mutually exclusive ho-
471 mologous exons (6 and 7) in the ancestor common to mammals, amphibians and fishes.
472 We find that the most ancient of these two exons is exon 7. By characterizing in de-
473 tails the structural dynamics of two human isoforms, JNK1 α and JNK1 β , bearing one or
474 the other exon, we could emphasize subtle changes associated to this ASE and identify
475 residues that may be responsible for the selectivity of the JNK isoforms toward their
476 substrates. Secondly, we highlighted an isoform that was not previously described in the
477 literature, namely JNK1 δ . Despite displaying a large deletion (about 80 residues), it is
478 conserved across several species and short MD simulations suggest that it is stable in
479 solution. According to the APPRIS database v20⁵⁷, there are 4 peptides matching this
480 isoform in publicly available proteomics data. By comparison, the other human JNK1
481 isoforms with a phylogeny have between 5 and 7 matching peptides, while the orphan
482 transcripts identified by our analysis have between zero and 2 matching peptides, sug-
483 gesting that JNK1 δ is indeed translated and stable in solution. Hence, considering that
484 the catalytic site is intact in JNK1 δ , we propose that this isoform might be catalytically
485 competent and that the large amplitude motion of the A-loop observed in the simulations
486 might facilitate the activation of the protein by exposing a couple of tyrosine and threo-
487 nine residues that are targeted by MAPK kinases. The validation of this hypothesis would
488 require further calculations and experiments that fall beyond the scope of this study. Al-
489 ready, this interesting result suggests that our approach could be used to identify and
490 characterize new isoforms, that may play a role in the cell and thus serve as therapeutic
491 targets. Thirdly, we found characteristics specific to the JNK3 isoforms, namely the ab-

492 sence of exon 7 and the presence of two exons (*0* and *1*) containing regions predicted
493 to be disordered and involved in interactions. These observations suggest specific compe-
494 tences or functions for this gene. Studies investigating the gain/loss of alternative splice
495 forms associated to gene duplication at large scale^{1;58} have highlighted a wide diversity
496 of cases and have suggested that it depends on the specific cellular context of each gene.
497 Although we did not have a sufficient sample resolution to confirm it with RNA-Seq data,
498 JNK3 is reported to be specifically expressed in the heart brain and testes⁶⁷

499 Our approach enables to go beyond a description of transcript variability across species
500 and/or across genes. Indeed, by reconstructing phylogenies, we do not only cluster tran-
501 scripts but we also add a temporal dimension to the analysis. Previous methods reported
502 in the literature were only applied on simple cases¹⁶ and/or largely simplified the evolu-
503 tion model to increase computational efficiency¹⁷, such that they could not be used for
504 ancestral transcripts' reconstruction. PhyloSofS algorithm makes the reconstruction of
505 transcripts' phylogenies feasible for any gene family. For the JNK family, the execution
506 of 1 million iterations took about two weeks on a single CPU. This case represents a high
507 level of complexity as most of the transcripts contain more than 10 exons (the average
508 number of exons per gene being estimated at 8.8 in the human genome⁶⁰) and up to 8
509 transcripts are observed within each species (it is estimated that about 4 distinct-coding
510 transcripts per gene are expressed in human⁵⁷). To reduce the computing time, the user
511 can easily parallelize the multi-start iterative search on multiple cores and s/he has the
512 possibility to give as input a previously computed value for the lower bound (to increase
513 the efficiency of the cut). We should stress that the problem of pairing transcripts across
514 homologous and paralogous genes between different species, addressed here, is much more
515 complex than that of inferring the transcripts' phylogeny of each gene separately. Indeed,
516 in the former case, the problem size is bigger, one needs to reconcile the gene tree with
517 the species tree, and the sequences are more divergent.

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

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

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

1 Materials and Methods

1.1 PhyloSofS workflow

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

- It reconstructs a forest of phylogenetic trees describing plausible evolutionary scenarios that can explain the observed transcripts (**Fig. 1a**, on the right). The forest is embedded in the input gene tree and is reconstructed by using the maximum parsimony principle. The root of a tree corresponds to the creation of a new transcript, each leaf stands for an observed transcript and a dead end (indicated by a triangle on **Fig. 1a**, on the right) indicates a transcript loss. Transcripts can mutate along the branches of the trees.
- It predicts the three-dimensional structures of the protein isoforms corresponding to the observed transcripts by using homology modeling. The 3D models are then annotated with quality measures and with exon labels.

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

1.1.1 Step a. Transcripts' phylogenies reconstruction

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

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

577 We model transcripts evolution as a two-level process, at the gene and transcript
 578 levels, as described by Christinat and Moret¹⁶. At the level of the gene, each exon can be
 579 either absent, alternative or constitutive. This status is inferred from the occurrence of
 580 the exon in the transcripts. Hence, for a given species s , a vector g^s of length n_e encodes
 581 the state of each exon by the values $\{0, 1, 2\}$ for absent, alternative and constitutive,
 582 respectively (*Supplementary Fig. S1b*, white, black/white and black squares). At the
 583 leaves (current species), the components of g^s are calculated as:

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

584 As in¹⁶, the g^s vectors for internal nodes (ancestral species) are determined by using
 585 Sankoff's algorithm⁶². Dollo's parsimony principle is also respected, such that an exon
 586 cannot be created twice². If different exon states have equal cost, we follow the priority
 587 rule $1 > 0 > 2$.

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

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

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

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

Each cell is associated to the evolution of the state of the exon e from a parent transcript to a child transcript. The first and second terms in parenthesis correspond to the status of the exon at the transcript and gene levels, respectively. Only evolutionary changes taking place at the transcript level, without being directly induced by a gene-level change, are penalized (σ). Zeros highlighted in bold indicate transcript-level changes being a direct consequence of a gene-level change. This table of costs was taken from¹⁶.

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

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

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

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

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

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

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

612 along the branches of φ :

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

613 where t_i^k is the parent transcript, i^{th} subnode of node k , t_j^l is the child transcript, j^{th}
 614 subnode of node 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
 615 state of exon e at the level of the gene at node y and $T_{e,x}^y \in \{0, 1\}$ the state of exon e at
 616 the level of the x^{th} transcript of node y . The evolution costs σ are given in Table II.

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

620 **Initialization:**

621 $C_{min} \leftarrow \infty$

622 Choose the forest structure S_0 that maximizes the n_b values

623 **Iteration:**

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

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

626 Find the most parsimonious phylogeny φ_i given structure S_i

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

628 $C_{min} \leftarrow C_{\varphi_i}$

629 **end if**

630 **end if**

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

632 **end for**

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

637 φ_i . Forest structures that are too costly are simply discarded, without calculating the
638 corresponding phylogenies. As the algorithm finds better and better solutions, the cut
639 becomes more and more efficient. The phylogeny φ_i is reconstructed by using dynamic
640 programming. Sankoff's algorithm is applied bottom up to compute the minimum pairing
641 costs between transcripts (*Supplementary Fig. S1d*, each transcript is represented by a
642 matrix of costs). At each internal node, the pairings are determined by using a specific
643 version of the branch-and-bound algorithm⁴¹ (see *Supplementary Text S1*). If the re-
644 constructed phylogeny is more parsimonious than those previously visited ($C_{\varphi_i} < C_{min}$),
645 then the minimum cost C_{min} is updated. There may be more than one phylogeny with
646 minimum cost that comply with a given structure S_i . The next forest structure S_j will be
647 randomly chosen among the immediate neighbors of S_i (*Supplementary Fig. S1d*). Two
648 structures are immediate neighbors if each one of them can be obtained by an elemen-
649 tary operation applied to only one node of the other one (*Supplementary Fig. S13*). If
650 the phylogeny φ_j is such that $C_{\varphi_j} < C_{min}$, then the next forest structure will be chosen
651 among the neighbors of S_j , which serves as a new "base" for the search. Otherwise, the
652 algorithm continues to sample the neighborhood of S_i . This step-by-step search is applied
653 until no better solution can be found. At this point, a new random jump is performed.
654 The total number of iterations t_{max} is given as input by the user (1 by default).

655 We should stress that PhyloSofS algorithm is designed to deal with much more complex
656 cases than those reported in¹⁶ in a computationally tractable way. Hence, it differs from
657 the algorithm reported in¹⁶ in several respects. First, our multi-start iterative strategy
658 relies on random jumps in the forest structure space combined with systematic local
659 exploration around the best current solution, while Christinat and Moret¹⁶ proposed an
660 exhaustive generation and evaluation of forest structures. Secondly, we have designed a
661 branch-and-bound algorithm specifically adapted to the problem of determining the best
662 phylogeny complying with a given forest structure (see *Supplementary Text S1*). Both
663 aspects contribute to PhyloSofS efficiency in reconstructing transcripts' phylogenies.

664 PhyloSofS generates PDF files displaying the computed transcripts' phylogenies using
665 a Python driver to the Graphviz²³ DOT format.

666 1.1.2 Step b. Isoforms structures prediction

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

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

677 Step 1 makes extensive use of the HH-suite²⁹ and can be decomposed in: (a) search
678 for homologous sequences and building of a multiple sequence alignment (MSA), by using
679 HHblits⁵⁵, (b) addition of secondary structure predictions, obtained by PSIPRED³⁴, to
680 the MSA, (c) generation of a profile hidden markov model (HMM) from the MSA, (d)
681 search of a database of profile HMMs for homologous proteins, using HHsearch⁶³. Step 3
682 is performed by Modeller⁴⁷ with default options. Step 5 consists in: (a) inserting the num-
683 bers of the exons in the β -factor column of the PDB file of the 3D model, (b) computing
684 the proportion of residues predicted in well-defined secondary structures by PSIPRED³⁴,
685 (c) assessing the quality of the model with Procheck⁴² and with the normalized DOPE
686 score from Modeller, (d) determining the by-residue solvent accessible surface areas with
687 Naccess³¹ and computing the proportions of surface residues and of hydrophobic surface
688 residues.

1.2 Retrieval and pre-processing of JNK annotated transcriptome data

The peptide sequences of all splice variants from the JNK family observed in human, mouse, *Xenopus tropicalis*, zebrafish, fugu, *Drosophila melanogaster* and nematode were retrieved from Ensembl⁷⁵ release 84 (March 2016) along with the phylogenetic gene tree. Only the transcripts containing an open reading frame and not annotated as undergoing nonsense mediated decay or lacking 3' or 5' truncation were retained. The isoforms sharing the same amino acid sequence were merged. The homologous exons between the different genes in the different species were identified by aligning the sequences with MAFFT³⁵, and projecting the alignment on the human annotation. They do not necessarily represent exons definition based on the genomic sequence and this can be explained by two reasons. First, the gene structure may be different from one species to another. For instance, the third and fourth exons of human JNK1 genes are completely covered by a single exon in the *Drosophila melanogaster* JNK gene (*Supplementary Fig. S4*). In cases like this, we keep the highest level of resolution and define two exons (*e.g.* numbered 3 and 4). Secondly, it may happen that a transcript contains only a part of an exon in a given species translated in another frame. In that case, we define two exons sharing the same number but distinguished by the prime symbol (*e.g.* exons 8 and 8'). In total, 64 transcripts comprised of 38 exons were given as input to PhyloSofS.

1.3 PhyloSofS' parameter setting

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

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

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

724 **1.4 Analysis of JNK tertiary structures.**

725 The list of experimental structures deposited in the PDB for the human JNKs was re-
726 trieved from UniProt⁵. The structures were aligned with PyMOL¹⁹ and the RMSD
727 between each pair was computed. Residues comprising the catalytic site were defined
728 from the complex between human JNK3 and adenosine mono-phosphate (PDB code:
729 4KKE, resolution: 2.2 Å), as those located less than 6 Å away from the ligand. Residues
730 comprising the D-site and the F-site were defined from the complexes between human
731 JNK1 and the scaffolding protein JIP-1 (PDB code: 1UKH, resolution: 2.35 Å²⁸) and the
732 catalytic domain of MKP7 (PDB code: 4YR8, resolution: 2.4 Å⁴⁴), respectively. They
733 were detected as displaying a change in relative solvent accessibility $>1 \text{ \AA}^2$ upon binding.

734 The I-TASSER webserver^{59;73;76} was used to try and model the regions for which no
735 structural templates could be found. DISOPRED⁷¹ and IUPred²¹ were used to predict
736 intrinsic disorder. JET2⁴⁰ was used to predict binding sites at the surface of the isoforms.

737 **1.5 Molecular dynamics simulations of human isoforms.**

738 The 3D coordinates of the human JNK1 isoforms JNK1 α (369 res., containing exon 6),
739 JNK1 β (369 res., containing exon 7) and JNK1 δ (304 res., containing neither exon 6 nor
740 exon 7) were predicted by PhyloSofS pipeline. The 3 systems were prepared with the
741 LEAP module of AMBER 12¹⁴, using the ff12SB forcefield parameter set: (*i*) hydrogen
742 atoms were added, (*ii*) the protein was hydrated with a cuboid box of explicit TIP3P
743 water molecules with a buffering distance up to 10Å, (*iii*) Na⁺ and Cl⁻ counter-ions were

744 added to neutralize the protein.

745 The systems were minimized, thermalized and equilibrated using the SANDER mod-
746 ule of AMBER 12. The following minimization procedure was applied: (i) 10,000 steps
747 of minimization of the water molecules keeping protein atoms fixed, (ii) 10,000 steps of
748 minimization keeping only protein backbone fixed to allow protein side chains to relax,
749 (iii) 10,000 steps of minimization without any constraint on the system. Heating of the
750 system to the target temperature of 310 K was performed at constant volume using the
751 Berendsen thermostat⁶ and while restraining the solute C_α atoms with a force constant of
752 $10 \text{ kcal/mol/\AA}^2$. Thereafter, the system was equilibrated for 100 ps at constant volume
753 (NVT) and for further 100 ps using a Langevin piston (NPT)⁴⁵ to maintain the pressure.
754 Finally the restraints were removed and the system was equilibrated for a final 100 ps
755 run.

756 Each system was simulated during 250 ns (5 replicates of 50 ns, starting from different
757 initial velocities) in the NPT ensemble using the PMEMD module of AMBER 12. The
758 temperature was kept at 310 K and pressure at 1 bar using the Langevin piston coupling
759 algorithm. The SHAKE algorithm was used to freeze bonds involving hydrogen atoms,
760 allowing for an integration time step of 2.0 fs. The Particle Mesh Ewald (PME) method¹⁸
761 was employed to treat long-range electrostatics. The coordinates of the system were
762 written every ps.

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

773 residue was assigned to a loop. The amplitude of the motion of the A-loop compared
774 to the rest of the protein was estimated by computing the angle between the geometric
775 center of residues 189-192, residue 205 and either residue 211 in the isoforms JNK1 α and
776 JNK1 β or residue 209 in the isoform JNK1 δ . Only C- α atoms were considered.

777 1.6 RNA-Seq Data integration

778 To obtain additional support for transcript isoform expression, we queried the Bgee
779 database (v.14)⁴ for a list of all RNA-Seq experiments related to the selected species.
780 Using SRA tools, we downloaded raw sequences from *H. sapiens* (224 samples), *M. mus-*
781 *culus* (155 samples), *Xenopus tropicalis* (69 samples) and *D. rerio* (67 samples) and then
782 aligned the reads using STAR v.2.5.3a²⁰ with default parameters. *T. rubripes* is not an-
783 notated in Bgee and was not integrated in this part of the analysis. Reads overlapping
784 exon-exon boundaries (*e.g.* splice-junction reads) next to alternative splicing events pro-
785 vide direct evidence for the expression of specific transcripts isoforms. Combined with
786 sample annotation, they could also inform on tissue specific isoform expression. We thus
787 considered all reads included within one of the JNK genes and monitored the alignment
788 of splice junctions between different exons as support for the transcripts isoforms: exons
789 5-6 and 6-8 for JN1 α , exons 5-7 and 7-8 for JNK1 β , and exons 5-8' for JNK1 δ .

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

794 References

795 [1] Abascal, F., Tress, M. L., and Valencia, A. (2015). The evolutionary fate of alterna-
796 tively spliced homologous exons after gene duplication. *Genome Biol Evol*, 7(6):1392–
797 1403.

- 798 [2] Alekseyenko, A. V., Lee, C. J., and Suchard, M. A. (2008). Wagner and Dollo: a
799 stochastic duet by composing two parsimonious solos. *Syst. Biol.*, 57(5):772–784.
- 800 [3] Barbosa-Morais, N. L., Irimia, M., Pan, Q., Xiong, H. Y., Gueroussov, S., Lee, L. J.,
801 Slobodeniuc, V., Kutter, C., Watt, S., Colak, R., Kim, T., Misquitta-Ali, C. M., Wil-
802 son, M. D., Kim, P. M., Odom, D. T., Frey, B. J., and Blencowe, B. J. (2012). The evolu-
803 tionary landscape of alternative splicing in vertebrate species. *Science*, 338(6114):1587–
804 1593.
- 805 [4] Bastian, F., Parmentier, G., Roux, J., Moretti, S., Laudet, V., and Robinson-Rechavi,
806 M. (2008). chapter Bgee: Integrating and Comparing Heterogeneous Transcriptome
807 Data Among Species, pages 124–131.
- 808 [5] Bateman, A., Martin, M. J., O’Donovan, C., Magrane, M., Apweiler, R., Alpi, E.,
809 Antunes, R., Arganiska, J., Bely, B., Bingley, M., Bonilla, C., Britto, R., Bursteinas,
810 B., Chavali, G., Cibrian-Uhalte, E., Silva, A. D., De Giorgi, M., Dogan, T., Fazzini,
811 F., Gane, P., Castro, L. G., Garmiri, P., Hatton-Ellis, E., Hieta, R., Huntley, R.,
812 Legge, D., Liu, W., Luo, J., MacDougall, A., Mutowo, P., Nightingale, A., Orchard, S.,
813 Pichler, K., Poggioli, D., Pundir, S., Pureza, L., Qi, G., Rosanoff, S., Saidi, R., Sawford,
814 T., Shypitsyna, A., Turner, E., Volynkin, V., Wardell, T., Watkins, X., Zellner, H.,
815 Cowley, A., Figueira, L., Li, W., McWilliam, H., Lopez, R., Xenarios, I., Bougueleret,
816 L., Bridge, A., Poux, S., Redaschi, N., Aimo, L., Argoud-Puy, G., Auchincloss, A.,
817 Axelsen, K., Bansal, P., Baratin, D., Blatter, M. C., Boeckmann, B., Bolleman, J.,
818 Boutet, E., Breuza, L., Casal-Casas, C., de Castro, E., Coudert, E., CuChe, B., Doche,
819 M., Dornevil, D., Duvaud, S., Estreicher, A., Famiglietti, L., Feuermann, M., Gasteiger,
820 E., Gehant, S., Gerritsen, V., Gos, A., Gruaz-Gumowski, N., Hinz, U., Hulo, C., Jungo,
821 F., Keller, G., Lara, V., Lemercier, P., Lieberherr, D., Lombardot, T., Martin, X.,
822 Masson, P., Morgat, A., Neto, T., Nospikel, N., Paesano, S., Pedruzzi, I., Pilbout, S.,
823 Pozzato, M., Pruess, M., Rivoire, C., Roechert, B., Schneider, M., Sigrist, C., Sonesson,
824 K., Staehli, S., Stutz, A., Sundaram, S., Tognolli, M., Verbregue, L., Veuthey, A. L.,
825 Wu, C. H., Arighi, C. N., Arminski, L., Chen, C., Chen, Y., Garavelli, J. S., Huang,

- 826 H., Laiho, K., McGarvey, P., Natale, D. A., Suzek, B. E., Vinayaka, C., Wang, Q.,
827 Wang, Y., Yeh, L. S., Yerramalla, M. S., and Zhang, J. (2015). UniProt: a hub for
828 protein information. *Nucleic Acids Res.*, 43(Database issue):D204–212.
- 829 [6] Berendsen, H., Postma, J., van Gunsteren, W., DiNola, A., and Haak, J. (1984).
830 Molecular dynamics with coupling to an external bath. *The Journal of chemical physics*,
831 81(8):3684–3690.
- 832 [7] Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H.,
833 Shindyalov, I. N., and Bourne, P. E. (2000). The Protein Data Bank. *Nucleic Acids*
834 *Res.*, 28(1):235–242.
- 835 [8] Bhuiyan, S. A., Ly, S., Phan, M., Huntington, B., Hogan, E., Liu, C. C., Liu, J., and
836 Pavlidis, P. (2018). Systematic evaluation of isoform function in literature reports of
837 alternative splicing. *BMC Genomics*, 19(1):637.
- 838 [9] Birzele, F., Csaba, G., and Zimmer, R. (2008a). Alternative splicing and protein
839 structure evolution. *Nucleic Acids Res.*, 36(2):550–558.
- 840 [10] Birzele, F., Kuffner, R., Meier, F., Oefinger, F., Potthast, C., and Zimmer, R.
841 (2008b). ProSAS: a database for analyzing alternative splicing in the context of protein
842 structures. *Nucleic Acids Res.*, 36(Database issue):D63–68.
- 843 [11] Bogoyevitch, M. A. and Kobe, B. (2006). Uses for JNK: the many and varied
844 substrates of the c-Jun N-terminal kinases. *Microbiol. Mol. Biol. Rev.*, 70(4):1061–
845 1095.
- 846 [12] Brecht, S., Kirchhof, R., Chromik, A., Willesen, M., Nicolaus, T., Raivich, G.,
847 Wessig, J., Waetzig, V., Goetz, M., Claussen, M., Pearse, D., Kuan, C. Y., Vaudano,
848 E., Behrens, A., Wagner, E., Flavell, R. A., Davis, R. J., and Herdegen, T. (2005).
849 Specific pathophysiological functions of JNK isoforms in the brain. *Eur. J. Neurosci.*,
850 21(2):363–377.
- 851 [13] Buljan, M., Chalancon, G., Eustermann, S., Wagner, G. P., Fuxreiter, M., Bateman,

- 852 A., and Babu, M. M. (2012). Tissue-specific splicing of disordered segments that embed
853 binding motifs rewires protein interaction networks. *Mol. Cell*, 46(6):871–883.
- 854 [14] Case, D., Darden, T., Cheatham III, T., Simmerling, C., Wang, J., Duke, R., Luo,
855 R., Walker, R., Zhang, W., Merz, K., et al. (2012). Amber 12. *University of California,*
856 *San Francisco*, 1(2):3.
- 857 [15] Chamberlain, S. D., Redman, A. M., Wilson, J. W., Deanda, F., Shotwell, J. B.,
858 Gerding, R., Lei, H., Yang, B., Stevens, K. L., Hassell, A. M., Shewchuk, L. M.,
859 Leesnitzer, M. A., Smith, J. L., Sabbatini, P., Atkins, C., Groy, A., Rowand, J. L.,
860 Kumar, R., Mook, R. A., Moorthy, G., and Patnaik, S. (2009). Optimization of 4,6-
861 bis-anilino-1H-pyrrolo[2,3-d]pyrimidine IGF-1R tyrosine kinase inhibitors towards JNK
862 selectivity. *Bioorg. Med. Chem. Lett.*, 19(2):360–364.
- 863 [16] Christinat, Y. and Moret, B. M. (2012). Inferring transcript phylogenies. *BMC*
864 *Bioinformatics*, 13 Suppl 9:S1.
- 865 [17] Christinat, Y. and Moret, B. M. (2013). A transcript perspective on evolution.
866 *IEEE/ACM Trans Comput Biol Bioinform*, 10(6):1403–1411.
- 867 [18] Darden, T., York, D., and Pedersen, L. (1993). Particle mesh ewald: An $n\log(n)$
868 method for ewald sums in large systems. *The Journal of Chemical Physics*, 98:10089–
869 10092.
- 870 [19] DeLano, W. (2002). The PyMOL Molecular Graphics System.
871 <http://www.pymol.org>.
- 872 [20] Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut,
873 P., Chaisson, M., and Gingeras, T. R. (2013). Star: ultrafast universal rna-seq aligner.
874 *Bioinformatics*, 29(1):15–21.
- 875 [21] Dosztanyi, Z., Csizmok, V., Tompa, P., and Simon, I. (2005). IUPred: web server
876 for the prediction of intrinsically unstructured regions of proteins based on estimated
877 energy content. *Bioinformatics*, 21(16):3433–3434.

- 878 [22] Ezkurdia, I., Rodriguez, J. M., Carrillo-de Santa Pau, E., Vazquez, J., Valencia, A.,
879 and Tress, M. L. (2015). Most highly expressed protein-coding genes have a single
880 dominant isoform. *J. Proteome Res.*, 14(4):1880–1887.
- 881 [23] Gansner, E. R. and North, S. C. (2000). An open graph visualization system and its
882 applications to software engineering. *SOFTWARE - PRACTICE AND EXPERIENCE*,
883 30(11):1203–1233.
- 884 [24] Gelly, J. C., Lin, H. Y., de Brevern, A. G., Chuang, T. J., and Chen, F. C. (2012).
885 Selective constraint on human pre-mRNA splicing by protein structural properties.
886 *Genome Biol Evol*, 4(9):966–975.
- 887 [25] Gonzalez-Porta, M., Frankish, A., Rung, J., Harrow, J., and Brazma, A. (2013).
888 Transcriptome analysis of human tissues and cell lines reveals one dominant transcript
889 per gene. *Genome Biol.*, 14(7):R70.
- 890 [26] Graveley, B. R. (2001). Alternative splicing: increasing diversity in the proteomic
891 world. *Trends Genet*, 17(2):100–107.
- 892 [27] Hegyi, H., Kalmar, L., Horvath, T., and Tompa, P. (2011). Verification of alternative
893 splicing variants based on domain integrity, truncation length and intrinsic protein
894 disorder. *Nucleic Acids Res.*, 39(4):1208–1219.
- 895 [28] Heo, Y. S., Kim, S. K., Seo, C. I., Kim, Y. K., Sung, B. J., Lee, H. S., Lee, J. I.,
896 Park, S. Y., Kim, J. H., Hwang, K. Y., Hyun, Y. L., Jeon, Y. H., Ro, S., Cho, J. M.,
897 Lee, T. G., and Yang, C. H. (2004). Structural basis for the selective inhibition of
898 JNK1 by the scaffolding protein JIP1 and SP600125. *EMBO J.*, 23(11):2185–2195.
- 899 [29] Hildebrand, A., Remmert, M., Biegert, A., and Soding, J. (2009). Fast and accurate
900 automatic structure prediction with HHpred. *Proteins*, 77 Suppl 9:128–132.
- 901 [30] Hirosumi, J., Tuncman, G., Chang, L., Gorgun, C. Z., Uysal, K. T., Maeda, K.,
902 Karin, M., and Hotamisligil, G. S. (2002). A central role for JNK in obesity and
903 insulin resistance. *Nature*, 420(6913):333–336.

- 904 [31] Hubbard, S. and Thornton, J. (1992-6). [http://www.bioinf.manchester.ac.uk/
905 naccess/](http://www.bioinf.manchester.ac.uk/naccess/).
- 906 [32] Hunot, S., Vila, M., Teismann, P., Davis, R. J., Hirsch, E. C., Przedborski, S., Rakic,
907 P., and Flavell, R. A. (2004). JNK-mediated induction of cyclooxygenase 2 is required
908 for neurodegeneration in a mouse model of Parkinson's disease. *Proc. Natl. Acad. Sci.
909 U.S.A.*, 101(2):665–670.
- 910 [33] Huse, M. and Kuriyan, J. (2002). The conformational plasticity of protein kinases.
911 *Cell*, 109(3):275–282.
- 912 [34] Jones, D. T. (1999). Protein secondary structure prediction based on position-specific
913 scoring matrices. *J. Mol. Biol.*, 292(2):195–202.
- 914 [35] Katoh, K. and Standley, D. M. (2013). MAFFT multiple sequence alignment software
915 version 7: improvements in performance and usability. *Mol. Biol. Evol.*, 30(4):772–780.
- 916 [36] Kelemen, O., Convertini, P., Zhang, Z., Wen, Y., Shen, M., Falaleeva, M., and
917 Stamm, S. (2013). Function of alternative splicing. *Gene*, 514(1):1–30.
- 918 [37] Kim, M. S., Pinto, S. M., Getnet, D., Nirujogi, R. S., Manda, S. S., Chaerkady, R.,
919 Madugundu, A. K., Kelkar, D. S., Isserlin, R., Jain, S., Thomas, J. K., Muthusamy, B.,
920 Leal-Rojas, P., Kumar, P., Sahasrabudde, N. A., Balakrishnan, L., Advani, J., George,
921 B., Renuse, S., Selvan, L. D., Patil, A. H., Nanjappa, V., Radhakrishnan, A., Prasad,
922 S., Subbannayya, T., Raju, R., Kumar, M., Sreenivasamurthy, S. K., Marimuthu, A.,
923 Sathe, G. J., Chavan, S., Datta, K. K., Subbannayya, Y., Sahu, A., Yelamanchi, S. D.,
924 Jayaram, S., Rajagopalan, P., Sharma, J., Murthy, K. R., Syed, N., Goel, R., Khan,
925 A. A., Ahmad, S., Dey, G., Mudgal, K., Chatterjee, A., Huang, T. C., Zhong, J., Wu,
926 X., Shaw, P. G., Freed, D., Zahari, M. S., Mukherjee, K. K., Shankar, S., Mahadevan,
927 A., Lam, H., Mitchell, C. J., Shankar, S. K., Satishchandra, P., Schroeder, J. T.,
928 Sirdeshmukh, R., Maitra, A., Leach, S. D., Drake, C. G., Halushka, M. K., Prasad,
929 T. S., Hruban, R. H., Kerr, C. L., Bader, G. D., Iacobuzio-Donahue, C. A., Gowda, H.,

- 930 and Pandey, A. (2014). A draft map of the human proteome. *Nature*, 509(7502):575–
931 581.
- 932 [38] Kornev, A. P., Haste, N. M., Taylor, S. S., and Eyck, L. F. (2006). Surface comparison
933 of active and inactive protein kinases identifies a conserved activation mechanism. *Proc.*
934 *Natl. Acad. Sci. U.S.A.*, 103(47):17783–17788.
- 935 [39] Kyriakis, J. M. and Avruch, J. (2012). Mammalian MAPK signal transduction path-
936 ways activated by stress and inflammation: a 10-year update. *Physiol. Rev.*, 92(2):689–
937 737.
- 938 [40] Laine, E. and Carbone, A. (2015). Local Geometry and Evolutionary Conserva-
939 tion of Protein Surfaces Reveal the Multiple Recognition Patches in Protein-Protein
940 Interactions. *PLoS Comput. Biol.*, 11(12):e1004580.
- 941 [41] Land, A. H. and Doig, A. G. (1960). An automatic method of solving discrete
942 programming problems. *Econometrica*, 28:497–520.
- 943 [42] Laskowski, R. A., MacArthur, M. W., Moss, D. S., and Thornton, J. M. (1993).
944 PROCHECK: a program to check the stereochemical quality of protein structures.
945 *Journal of Applied Crystallography*, 26(2):283–291.
- 946 [43] Lim, K. H., Ferraris, L., Filloux, M. E., Raphael, B. J., and Fairbrother,
947 W. G. (2011). Using positional distribution to identify splicing elements and pre-
948 dict pre-mRNA processing defects in human genes. *Proc. Natl. Acad. Sci. U.S.A.*,
949 108(27):11093–11098.
- 950 [44] Liu, X., Zhang, C. S., Lu, C., Lin, S. C., Wu, J. W., and Wang, Z. X. (2016). A
951 conserved motif in JNK/p38-specific MAPK phosphatases as a determinant for JNK1
952 recognition and inactivation. *Nat Commun*, 7:10879.
- 953 [45] Loncharich, R., Brooks, B., and Pastor, R. (1992). Langevin dynamics of peptides:
954 The frictional dependence of isomerization rates of n-acetylalanyl-N'-methylamide.
955 *Biopolymers*, 32(5):523–535.

- 956 [46] Manning, A. M. and Davis, R. J. (2003). Targeting JNK for therapeutic benefit:
957 from junk to gold? *Nat Rev Drug Discov*, 2(7):554–565.
- 958 [47] Marti-Renom, M. A., Stuart, A. C., Fiser, A., Sanchez, R., Melo, F., and Sali, A.
959 (2000). Comparative protein structure modeling of genes and genomes. *Annu Rev*
960 *Biophys Biomol Struct*, 29:291–325.
- 961 [48] Melamud, E. and Moulton, J. (2009). Stochastic noise in splicing machinery. *Nucleic*
962 *Acids Res.*, 37(14):4873–4886.
- 963 [49] Merkin, J., Russell, C., Chen, P., and Burge, C. B. (2012). Evolutionary dynamics
964 of gene and isoform regulation in Mammalian tissues. *Science*, 338(6114):1593–1599.
- 965 [50] Meszaros, B., Simon, I., and Dosztanyi, Z. (2009). Prediction of protein binding
966 regions in disordered proteins. *PLoS Comput. Biol.*, 5(5):e1000376.
- 967 [51] Mudge, J. M., Frankish, A., Fernandez-Banet, J., Alioto, T., Derrien, T., Howald,
968 C., Reymond, A., Guigo, R., Hubbard, T., and Harrow, J. (2011). The origins, evolu-
969 tion, and functional potential of alternative splicing in vertebrates. *Mol. Biol. Evol.*,
970 28(10):2949–2959.
- 971 [52] Nicolas, A., Raguenees-Nicol, C., Ben Yaou, R., Ameziane-Le Hir, S., Cheron, A.,
972 Vie, V., Claustres, M., Leturcq, F., Delalande, O., Hubert, J. F., Tuffery-Giraud, S.,
973 Giudice, E., and Le Rumeur, E. (2015). Becker muscular dystrophy severity is linked
974 to the structure of dystrophin. *Hum. Mol. Genet.*, 24(5):1267–1279.
- 975 [53] Oruganty, K., Talathi, N. S., Wood, Z. A., and Kannan, N. (2013). Identification
976 of a hidden strain switch provides clues to an ancient structural mechanism in protein
977 kinases. *Proc. Natl. Acad. Sci. U.S.A.*, 110(3):924–929.
- 978 [54] Prakash, A. and Bateman, A. (2015). Domain atrophy creates rare cases of functional
979 partial protein domains. *Genome Biol.*, 16:88.
- 980 [55] Remmert, M., Biegert, A., Hauser, A., and Soding, J. (2011). HHblits: lightning-

- 981 fast iterative protein sequence searching by HMM-HMM alignment. *Nat. Methods*,
982 9(2):173–175.
- 983 [56] Reyes, A., Anders, S., Weatheritt, R. J., Gibson, T. J., Steinmetz, L. M., and Huber,
984 W. (2013). Drift and conservation of differential exon usage across tissues in primate
985 species. *Proc. Natl. Acad. Sci. U.S.A.*, 110(38):15377–15382.
- 986 [57] Rodriguez, J. M., Maietta, P., Ezkurdia, I., Pietrelli, A., Wesselink, J. J., Lopez, G.,
987 Valencia, A., and Tress, M. L. (2013). APPRIS: annotation of principal and alternative
988 splice isoforms. *Nucleic Acids Res.*, 41(Database issue):D110–117.
- 989 [58] Roux, J. and Robinson-Rechavi, M. (2011). Age-dependent gain of alternative splice
990 forms and biased duplication explain the relation between splicing and duplication.
991 *Genome Res.*, 21(3):357–363.
- 992 [59] Roy, A., Kucukural, A., and Zhang, Y. (2010). I-TASSER: a unified platform for
993 automated protein structure and function prediction. *Nat Protoc*, 5(4):725–738.
- 994 [60] Sakharkar, M. K., Chow, V. T., and Kanguane, P. (2004). Distributions of exons
995 and introns in the human genome. *In Silico Biol. (Gedruckt)*, 4(4):387–393.
- 996 [61] Sandberg, R., Neilson, J. R., Sarma, A., Sharp, P. A., and Burge, C. B. (2008). Pro-
997 liferating cells express mrnas with shortened 3' untranslated regions and fewer microRNA
998 target sites. *Science*, 320(5883):1643–1647.
- 999 [62] Sankoff, D. (1975). Minimal mutation trees of sequences. *SIAM Journal on Applied*
1000 *Mathematics*, 28(1):35–42.
- 1001 [63] Soding, J. (2005). Protein homology detection by HMM-HMM comparison. *Bioin-*
1002 *formatics*, 21(7):951–960.
- 1003 [64] Stamm, S., Ben-Ari, S., Rafalska, I., Tang, Y., Zhang, Z., Toiber, D., Thanaraj,
1004 T. A., and Soreq, H. (2005). Function of alternative splicing. *Gene*, 344:1–20.
- 1005 [65] Tranchevent, L. C., Aube, F., Dulaurier, L., Benoit-Pilven, C., Rey, A., Poret, A.,
1006 Chautard, E., Mortada, H., Desmet, F. O., Chakrama, F. Z., Moreno-Garcia, M. A.,

- 1007 Goillot, E., Janczarski, S., Mortreux, F., Bourgeois, C. F., and Auboeuf, D. (2017).
1008 Identification of protein features encoded by alternative exons using Exon Ontology.
1009 *Genome Res.*, 27(6):1087–1097.
- 1010 [66] Tuncman, G., Hirosumi, J., Solinas, G., Chang, L., Karin, M., and Hotamisligil,
1011 G. S. (2006). Functional in vivo interactions between JNK1 and JNK2 isoforms in
1012 obesity and insulin resistance. *Proc. Natl. Acad. Sci. U.S.A.*, 103(28):10741–10746.
- 1013 [67] Waetzig, V. and Herdegen, T. (2005). Context-specific inhibition of JNKs: overcom-
1014 ing the dilemma of protection and damage. *Trends Pharmacol. Sci.*, 26(9):455–461.
- 1015 [68] Wang, E. T., Sandberg, R., Luo, S., Khrebtkova, I., Zhang, L., Mayr, C.,
1016 Kingsmore, S. F., Schroth, G. P., and Burge, C. B. (2008). Alternative isoform regu-
1017 lation in human tissue transcriptomes. *Nature*, 456(7221):470–476.
- 1018 [69] Wang, X., Codreanu, S. G., Wen, B., Li, K., Chambers, M. C., Liebler, D. C., and
1019 Zhang, B. (2018). Detection of proteome diversity resulted from alternative splicing is
1020 limited by trypsin cleavage specificity. *Mol Cell Proteomics*, 17(3):422–430.
- 1021 [70] Ward, A. J. and Cooper, T. A. (2010). The pathobiology of splicing. *J. Pathol.*,
1022 220(2):152–163.
- 1023 [71] Ward, J. J., McGuffin, L. J., Bryson, K., Buxton, B. F., and Jones, D. T. (2004). The
1024 DISOPRED server for the prediction of protein disorder. *Bioinformatics*, 20(13):2138–
1025 2139.
- 1026 [72] Weatheritt, R. J., Sterne-Weiler, T., and Blencowe, B. J. (2016). The ribosome-
1027 engaged landscape of alternative splicing. *Nat. Struct. Mol. Biol.*, 23(12):1117–1123.
- 1028 [73] Yang, J., Yan, R., Roy, A., Xu, D., Poisson, J., and Zhang, Y. (2015). The I-TASSER
1029 Suite: protein structure and function prediction. *Nat. Methods*, 12(1):7–8.
- 1030 [74] Yang, X., Coulombe-Huntington, J., Kang, S., Sheynkman, G. M., Hao, T., Richard-
1031 son, A., Sun, S., Yang, F., Shen, Y. A., Murray, R. R., Spirohn, K., Begg, B. E.,
1032 Duran-Frigola, M., MacWilliams, A., Pevzner, S. J., Zhong, Q., Trigg, S. A., Tam,

- 1033 S., Ghamsari, L., Sahni, N., Yi, S., Rodriguez, M. D., Balcha, D., Tan, G., Costanzo,
1034 M., Andrews, B., Boone, C., Zhou, X. J., Salehi-Ashtiani, K., Charloteaux, B., Chen,
1035 A. A., Calderwood, M. A., Aloy, P., Roth, F. P., Hill, D. E., Iakoucheva, L. M., Xia,
1036 Y., and Vidal, M. (2016). Widespread expansion of protein interaction capabilities by
1037 alternative splicing. *Cell*, 164(4):805–817.
- 1038 [75] Yates, A., Akanni, W., Amode, M. R., Barrell, D., Billis, K., Carvalho-Silva, D.,
1039 Cummins, C., Clapham, P., Fitzgerald, S., Gil, L., Giron, C. G., Gordon, L., Hourlier,
1040 T., Hunt, S. E., Janacek, S. H., Johnson, N., Juettemann, T., Keenan, S., Lavidas, I.,
1041 Martin, F. J., Maurel, T., McLaren, W., Murphy, D. N., Nag, R., Nuhn, M., Parker,
1042 A., Patricio, M., Pignatelli, M., Rahtz, M., Riat, H. S., Sheppard, D., Taylor, K.,
1043 Thormann, A., Vullo, A., Wilder, S. P., Zadissa, A., Birney, E., Harrow, J., Muffato,
1044 M., Perry, E., Ruffier, M., Spudich, G., Trevanion, S. J., Cunningham, F., Aken, B. L.,
1045 Zerbino, D. R., and Flicek, P. (2016). Ensembl 2016. *Nucleic Acids Res.*, 44(D1):D710–
1046 716.
- 1047 [76] Zhang, Y. (2008). I-TASSER server for protein 3D structure prediction. *BMC*
1048 *Bioinformatics*, 9:40.