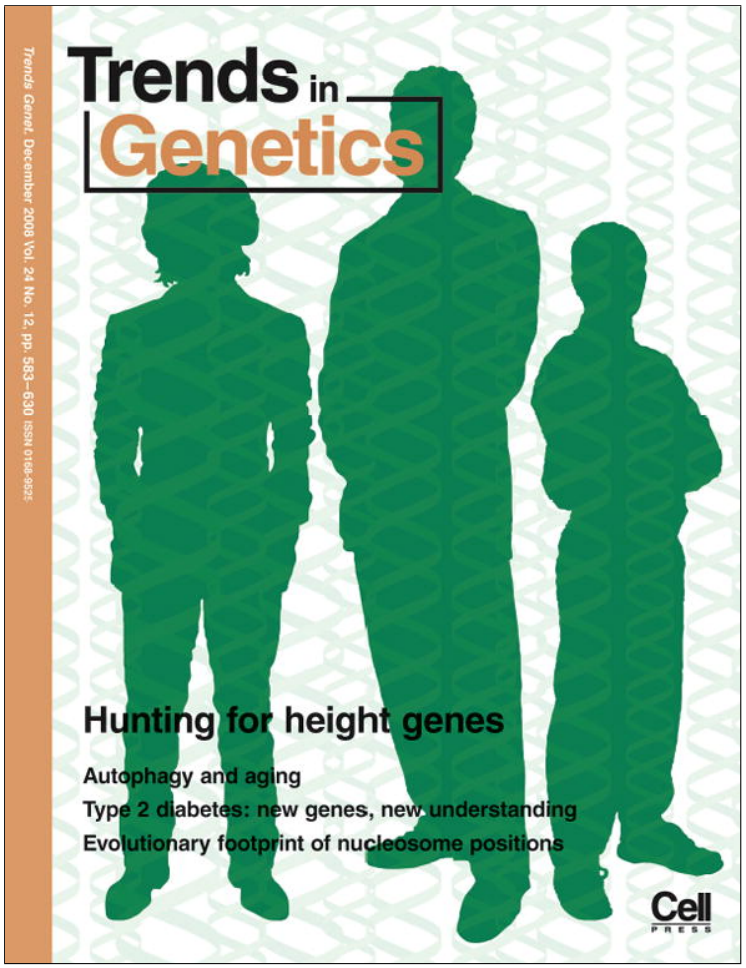

16 Thoma, F. (2005) Repair of UV lesions in nucleosomes–intrinsic properties and remodeling. *DNA Repair (Amst.)* 4, 855–869
17 Ataian, Y. and Krebs, J.E. (2006) Five repair pathways in one context: chromatin modification during DNA repair. *Biochem. Cell Biol.* 84, 490–504
18 Suter, B. and Thoma, F. (2002) DNA-repair by photolyase reveals dynamic properties of nucleosome positioning *in vivo*. *J. Mol. Biol.* 319, 395–406
19 Ura, K. *et al.* (2001) ATP-dependent chromatin remodeling facilitates nucleotide excision repair of UV-induced DNA lesions in synthetic dinucleosomes. *EMBO J.* 20, 2004–2014
20 Beard, B.C. *et al.* (2003) Suppressed catalytic activity of base excision repair enzymes on rotationally positioned uracil in nucleosomes. *Proc. Natl. Acad. Sci. U. S. A.* 100, 7465–7470
21 Hawk, J.D. *et al.* (2005) Variation in efficiency of DNA mismatch repair at different sites in the yeast genome. *Proc. Natl. Acad. Sci. U. S. A.* 102, 8639–8643
22 Siepel, A. *et al.* (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* 15, 1034–1050
23 Cliften, P. *et al.* (2003) Finding functional features in Saccharomyces genomes by phylogenetic footprinting. *Science* 301, 71–76
24 Higasa, K. and Hayashi, K. (2006) Periodicity of SNP distribution around transcription start sites. *BMC Genomics* 7, 66
25 Schones, D.E. *et al.* (2008) Dynamic regulation of nucleosome positioning in the human genome. *Cell* 132, 887–898
26 Ozsolak, F. *et al.* (2007) High-throughput mapping of the chromatin structure of human promoters. *Nat. Biotechnol.* 25, 244–248
27 Warnecke, T. *et al.* (2008) The impact of the nucleosome code on protein-coding sequence evolution in yeast. *PLoS Genet.* (in press)

0168-9525/$ – see front matter 
doi:10.1016/j.tig.2008.09.003 Available online 24 October 2008

Genome Analysis

# Inferring molecular function: contributions from functional linkages

**Arturo Medrano-Soto[1,2], Debnath Pal[3] and David Eisenberg[1,2]**

[1] Howard Hughes Medical Institute (HHMI), 675C. E. Young Drive South, Los Angeles, CA 90095, USA
[2] UCLA-DOE Institute for Genomics and Proteomics, 611C. E. Young Drive East, 201 Boyer Hall, Los Angeles, CA 90095, USA
[3] Bioinformatics Center and Supercomputer Education Research Center, C.V. Raman Circle, Indian Institute of Science, Bangalore 560012, Karnataka, India

**In the current era of high-throughput sequencing and structure determination, functional annotation has become a bottleneck in biomedical science. Here, we show that automated inference of molecular function using functional linkages among genes increases the accuracy of functional assignments by ≥8% and enriches functional descriptions in ≥34% of top assignments. Furthermore, biochemical literature supports >80% of automated inferences for previously unannotated proteins. These results emphasize the benefit of incorporating functional linkages in protein annotation.**

## Functional linkages and annotation of protein function

The current flood of complete genome sequences, coupled with the substantial progress of structural genomics, has deluged scientists with myriad protein sequences and structures for which there is often little or no functional information. This flood of data has stimulated the development of a body of computational methods to reveal the likely biological roles of unannotated proteins (for recent reviews, see Refs [1–4]). Functional linkages – genes identified as functionally related by bioinformatic approaches based on genomic context – have mainly been used to gain insights into the cellular processes in which genes participate [5,6]; for instance, in model organisms *Escherichia coli* K12 and *Bacillus subtilis*, ~70% of all pairs of genes within operons share similar biological processes (see Supplementary Material online). However, little attention has been devoted to learning how these relationships might contribute to the specific task of inferring molecular function. A preliminary estimate of the utility of functional linkages is available from the observation that, in *E. coli* K12 and *B. subtilis*, >40% of gene pairs within operons share very similar molecular functions (see Supplementary Material online). This indicates that computational methods aiming to infer or assign a molecular function to a protein can benefit from a better understanding of functional linkages. Here, we use the ProKnow metaserver (http://proknow.mbi.ucla.edu) [7] (Box 1) as a tool to assess the extent to which the quality of assignment of molecular function can be improved by incorporating annotations collected from proteins functionally linked to the query protein. We believe this work is the first attempt to quantify the contribution of information on functional linkages to the inference of molecular function.

## Assessing the contribution of functional linkages to inference of molecular function

We have added a new feature extractor to the ProKnow metaserver, whereby the Gene Ontology (GO; http://www.geneontology.org/) annotations of proteins that are inferred to be functionally linked to the query by methods based on genomic context available in the ProLinks database (http://prolinks.mbi.ucla.edu) [8] are taken into account in the inference process (Box 1). Hereafter, we refer to this feature extractor as the ProLinks module. We evaluated functional assignments using two test sets of proteins extracted from the Protein Data Bank (PDB; http://www.rcsb.org/pdb/) [9] based on protein sequence identity and type of fold. The first set consisted of 599 representative PDB proteins showing <50% sequence

*Corresponding author:* Eisenberg, D. (david@mbi.ucla.edu).

**Box 1. ProKnow 2.0 – a tool for evaluating the contribution of functional linkages to the automated inference of molecular function**

The ProKnow metaserver [7] runs a query protein sequence or 3D structure against a set of programs (or feature extractors). Originally, only PSI–BLAST (http://blast.ncbi.nlm.nih.gov/blast/blast.cgi) [13], RIGOR (http://alpha2.bmc.uu.se/usf) [14], DALI (http://ekhidna.biocenter.helsinki.fi/dali_server) [15], DASEY (http://www.doe-mbi.ucla.edu/Services/FOLD)[16], PROSITE (http://br.expasy.org/prosite/) [17] and DIP (http://dip.mbi.ucla.edu/) [18] were used to obtain clues about the most likely function of a query. Here, we have added a new feature extractor to ProKnow that incorporates information on functional linkages, inferred from genomic context-based methods, available in the PROLINKS [8] database. The outputs of each feature extractor are collected and related to GO [19] annotations. A profile of consensus GO terms is then extracted. Finally, the GO terms obtained are weighted based upon their frequency and the statistical significance reported by the individual feature extractors [7]. The output is a list of GO terms with their definitions, significance and data provenance. ProKnow now also includes an option to plot the ontology tree for the assigned GO terms, as shown in Figure I. The new feature extractor consists of the subset of pairs of proteins in ProLinks meeting four requirements: (i) protein pairs must belong to non-redundant genomes (see Supplementary Material online) to avoid biases introduced by multiple annotations from closely related organisms over-represented in the current set of complete genomes; (ii) protein pairs must be functionally linked by at least one of the methods Rosetta Stone (RS), Gene Cluster (GC), Gene Neighbor (GN) or Phylogenetic Profiles (PP); (iii) linked proteins must show a ProLinks confidence value $\geq$0.5; and (iv) protein pairs must display significant functional similarities (GO distances $\leq$0.05; see Supplementary Material online). In short, with this strategy, we selected the subset of functionally linked proteins with highly related functions using current knowledge. When a query protein has a BLAST hit with a gene in the ProLinks module, the GO annotations of all linked proteins are retrieved and processed in the same way as those from any other feature extractor [7]. In Figure I, information on functional linkages obtained from ProLinks is indicated in green, whereas information contributed by all other feature extractors is indicated in red. Numbers follow the same color convention and they indicate ProKnow confidence scores. The structure shown in the figure was arbitrarily selected as an input example to ProKnow (PDB ID: 1gO3).

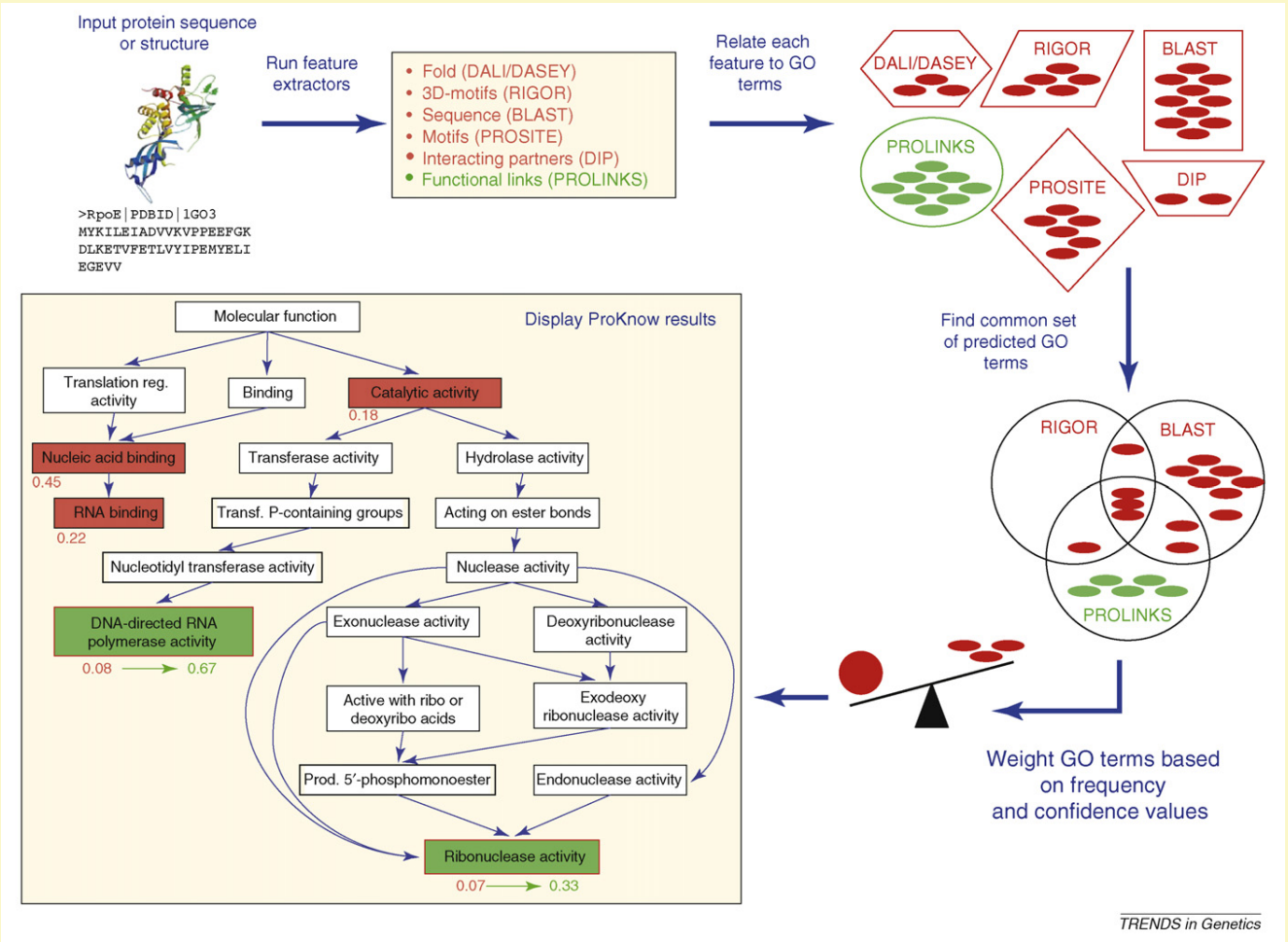


**Figure I.** ProKnow 2.0 flow chart.

identity; the second set consisted of 1740 representative PDB proteins with unique folds as defined in the domain dictionary created with the distance-matrix alignment algorithm (DALI; see Supplementary Material online). To assess the contribution of functional linkages, we compared the inference performance when ProKnow was run in two modes: (i) the ProLinks module is activated (ON); and (ii) the ProLinks module is deactivated (OFF). Then, for each mode, the coverage (number of inferred GO terms matching current GO annotations in PDB divided by the total number of GO annotations currently existing in PDB) and accuracy (the number of inferred GO terms matching

**Table 1. Coverage and accuracy of automated assignments of molecular function when functional linkages are considered**

| Type of PDB structure | Test sets | | | |
|---|---|---|---|---|
| | PDB 50 | | PDB fold | |
| | ProLinks ON | ProLinks OFF | ProLinks ON | ProLinks OFF |
| Structures with assigned functions[a] | 372 | | 469 | |
| Structures with GO annotations in PDB[b] | 361 (97%) | | 449 (96%) | |
| Structures without GO annotations in PDB | 11 (3%) | | 20 (4%) | |
| Structures with both assignments and GO annotations in PDB[c] | 353 (98%) | 357 (99%) | 422 (94%) | 442 (98%) |
| Structures with no matches between assignments and GO annotations in PDB | 8 (2%) | 4 (1%) | 27 (6%) | 7 (1%) |
| Coverage | 81% | 97% | 78% | 96% |
| Accuracy | 97%[d] | 89%[d] | 94%[e] | 85%[e] |

[a]The number of PDB structures in each dataset having molecular function GO assignments with contributions from functional linkages (ProLinks module in ProKnow).
[b]The number of proteins with at least one molecular function GO term annotated in PDB.
[c]These cases comprise the great majority and indicate how well the ProKnow metaserver can recover annotations without using information from other closely related proteins.
[d,e]Note how the inclusion of functional linkages ('ProLinks ON') increases the accuracy of function assignment by 8–9%.

current GO annotations in PDB divided by the total number of inferred GOs) were determined. Coverage indicates the proportion of current GO annotations in PDB that is recovered by ProKnow inferences. Accuracy indicates the proportion of ProKnow inferences matching current annotations in PDB. As previously reported [7], we identified matches between inferred and previously annotated GO terms by pairwise comparisons of their full ontology hierarchies and by observing the depth of the matching nodes. To prevent the introduction of undesired biases by self-entries and highly similar proteins in UniProt (http://www.uniprot.org/) [10] to the query, we discarded from our analysis all protein sequences showing ≥40% sequence identity with the query protein (see Supplementary Material online for an explanation of why a desirable sequence identity threshold of ≤30% could not be used in our evaluations).

The results are summarized in Table 1 and indicate a lower boundary of 8% for the contribution in accuracy of functional linkages to the inference of molecular function. That is, without using functional linkages, 85% of all ProKnow molecular function assignments actually match current annotations, whereas if information on functional linkages is included, these percentages increase to 94% (see last two rows in Table 1). When functional linkages are included in the analysis, we assign fewer GO terms, resulting in a concomitant decrease in coverage. This is because the ProKnow metaserver reports only the set of common GO terms contributed by the different feature extractors [7] and, thus, the addition of the ProLinks module often results in the recovery of fewer GO annotations. However, the increase in accuracy indicates that even though we assign fewer GO terms when we incorporate information on functional linkages, at least 94% of the time they correspond to previously annotated GO terms. For more details, see Supplementary Material online.

Furthermore, for ~50% of the cases, including functional linkages in the inferences yields a different top scoring GO term than when functional linkages are ignored. In these cases, ≥80% of the highest scoring assignments (or ≥34% of total inferences) are more specific and informative after taking functional linkages into consideration. That is, the best assignment corresponds to a more detailed annotation with deeper level of description in the GO hierarchy (see Supplementary Material online).

Table 1 also highlights two types of inferences that deserve special attention. The third row shows that there are few proteins (≤4%) with inferred functions, but with no annotations in PDB; and the fifth row indicates a small number of false positives (≤6%). After careful case-by-case examination of the biochemical literature, we found that for >80% of such protein structures, there is evidence directly supporting the inferred activity or, if the evidence is not direct, it nevertheless is related to the assigned function. For example, ProKnow inferred that the cell division protein FtsA (PDB: 1e4f) from *Thermotoga maritima* binds ATP (GO: 0005524). Although this protein currently carries no molecular function annotations in PDB, it has already been shown to bind ATP [11]. As another example, the protein MoeA (PDB: 1fc5) participates in molybdenum biosynthesis and we inferred that it binds guanosine-5′-triphosphate (GTP) (GO: 0005525). Although we found no direct evidence of GTP binding, MoeA actually binds a GTP derivative and contains inferred active sites very similar to those found in related proteins that do bind GTP [12] (see Supplementary Material online for more examples and a comprehensive list of cases).

## Functional linkages substantially contribute to inference of molecular function

Our results illustrate both that the accuracy of inference of molecular function is improved (by ≥8%) when information on functionally linked proteins is taken into consideration and that a substantial number of the best assignments (≥34%) contributed by functional linkages involve more specific and informative annotations compared with assignments that do not consider functional linkages in the analysis (≤9%). These findings emphasize the benefit of formally incorporating functional linkages as an intrinsic part of the inference process. There is, however, the caveat that the percentage of query proteins with contributions from functional linkages is not high (27–62% depending on the test set), although this is partly a consequence of the removal of most proteins related to the query (showing ≥40% sequence identity) from UniProt during our evaluations. The greater the number of proteins homologous to the query, the better the chances of extracting information on functional linkages from ProLinks and, thus, the more likely it is for functional linkages to contribute to the inferences.

That we found evidence in the biochemical literature supporting most (>80%) of our inferences for previously unannotated proteins and false-positive candidates (see

Supplementary Material online) attests to the power of methodologies for function inference. As a matter of fact, some of our inferred functions for these cases have already been incorporated in the most recent release of PDB. Because the reliability of comparative genomics approaches to infer functional linkages increases as more genomes are sequenced, we anticipate that the contribution of functional linkages to inference of protein molecular function will continue to grow in the future and that automated function assignment will show a steady increase in power.

**Acknowledgements**
The authors thank C. Miller and R. Llewellyn for valuable discussion, D. Cascio, T. Holton and A. Lisker for technical support and the NIH, DOE and HHMI for financial support.

**Supplementary data**
Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.tig.2008.10.001.

**References**
1 Friedberg, I. (2006) Automated protein function prediction–the genomic challenge. *Brief. Bioinform.* 7, 225–242
2 Ofran, Y. *et al.* (2005) Beyond annotation transfer by homology: novel protein-function prediction methods to assist drug discovery. *Drug Discov. Today* 10, 1475–1482
3 Rost, B. *et al.* (2003) Automatic prediction of protein function. *Cell. Mol. Life Sci.* 60, 2637–2650
4 Whisstock, J.C. and Lesk, A.M. (2003) Prediction of protein function from protein sequence and structure. *Q. Rev. Biophys.* 36, 307–340
5 Eisenberg, D. *et al.* (2000) Protein function in the post-genomic era. *Nature* 405, 823–826
6 Enault, F. *et al.* (2005) Phydbac "Gene Function Predictor": a gene annotation tool based on genomic context analysis. *BMC Bioinformatics* 6, 247
7 Pal, D. and Eisenberg, D. (2005) Inference of protein function from protein structure. *Structure* 13, 121–130
8 Bowers, P.M. *et al.* (2004) Prolinks: a database of protein functional linkages derived from coevolution. *Genome Biol.* 5, R35
9 Berman, H. *et al.* (2007) The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data. *Nucleic Acids Res.* 35, D301–D303
10 Wu, C.H. *et al.* (2006) The Universal Protein Resource (UniProt): an expanding universe of protein information. *Nucleic Acids Res.* 34, D187–D191
11 van den Ent, F. and Lowe, J. (2000) Crystal structure of the cell division protein FtsA from *Thermotoga maritima*. *EMBO J.* 19, 5300–5307
12 Schrag, J.D. *et al.* (2001) The crystal structure of *Escherichia coli* MoeA, a protein from the molybdopterin synthesis pathway. *J. Mol. Biol.* 310, 419–431
13 Altschul, S.F. *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25, 3389–3402
14 Kleywegt, G.J. (1999) Recognition of spatial motifs in protein structures. *J. Mol. Biol.* 285, 1887–1897
15 Holm, L. and Sander, C. (1998) Touring protein fold space with Dali/FSSP. *Nucleic Acids Res.* 26, 316–319
16 Mallick, P. *et al.* (2002) The directional atomic solvation energy: an atom-based potential for the assignment of protein sequences to known folds. *Proc. Natl. Acad. Sci. U. S. A.* 99, 16041–16046
17 Hulo, N. *et al.* (2006) The PROSITE database. *Nucleic Acids Res.* 34, D227–D230
18 Salwinski, L. *et al.* (2004) The Database of Interacting Proteins: 2004 update. *Nucleic Acids Res.* 32, D449–D451
19 Ashburner, M. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* 25, 25–29

0168-9525/$ – see front matter 
doi:10.1016/j.tig.2008.10.001 Available online 24 October 2008

Genome Analysis

# Co-evolution of the branch site and SR proteins in eukaryotes

**Mireya Plass[1], Eneritz Agirre[1], Diana Reyes[2], Francisco Camara[2] and Eduardo Eyras[1,3]**

[1] Computational Genomics Group, Universitat Pompeu Fabra, PRBB, Dr. Aiguader 88 E08003, Barcelona, Spain
[2] Centre for Genomic Regulation, PRBB, Dr. Aiguader 88 E08003, Barcelona, Spain
[3] Catalan Institution for Research and Advanced Studies (ICREA), Passeig de Lluis Companys 23 E08010, Barcelona, Spain

**Serine–arginine-rich (SR) proteins are essential for splicing in metazoans but are absent in yeast. By contrast, many fungi have SR protein homologs with variable arginine-rich regions analogous to the arginine–serine-rich (RS) domain in metazoans. The density of RS repeats in these regions correlates with the conservation of the branch site signal, providing evidence for an ancestral origin of SR proteins and indicating that the SR proteins and the branch site co-evolved.**

## Variation of the splicing signals across eukaryotes

Splicing is a key step in eukaryotic gene expression that requires the precise definition of the exon–intron boundaries by the splicing signals. Yeasts have a strong consensus across six nucleotides at the 5′ splice site (5′ss) and across seven nucleotides at the branch site (BS) [1,2]. By contrast, metazoans have a much weaker consensus signal at both sites (Figure 1a and Supplementary Material online). Understanding the direction of the evolution between weak and strong consensus might enable us to obtain insight into the origin of alternative splicing.

Recent analyses show that weak consensus signals are widespread across eukaryotic groups, indicating that the common ancestor had splicing signals similar to those of metazoans [3–5]. For instance, the 5′ss and the BS are highly conserved in the *Saccharomycetaceae* but not in other fungi such as *Rhizopus oryzae* and *Batrachochytrium dendrobatidis* (Figure 1a). Similarly, the strength of the polypyrimidine tract (PPT), which is important for the

*Corresponding author:* Eyras, E. (eduardo.eyras@upf.edu).