Published online 23 April 2009 *Nucleic Acids Research, 2009, Vol. 37, Web Server issue* ***W435–W440***
doi:10.1093/nar/gkp254

ANNIE: integrated *de novo* protein sequence annotation

Hong Sain Ooi[1], Chia Yee Kwo[1], Michael Wildpaner[2], Fernanda L. Sirota[1], Birgit Eisenhaber[3], Sebastian Maurer-Stroh[1], Wing Cheong Wong[1], Alexander Schleiffer[4], Frank Eisenhaber[1,5] and Georg Schneider[1,*]

[1]Bioinformatics Institute, A*Star, Singapore, [2]Google Switzerland GmbH, Zuerich, Switzerland, [3]Experimental Therapeutics Center, A*Star, Singapore, [4]Research Institute of Molecular Pathology, Vienna, Austria and [5]Department of Biological Sciences, National University of Singapore, Singapore

Received January 16, 2009; Revised March 31, 2009; Accepted April 6, 2009

ABSTRACT

Function prediction of proteins with computational sequence analysis requires the use of dozens of prediction tools with a bewildering range of input and output formats. Each of these tools focuses on a narrow aspect and researchers are having difficulty obtaining an integrated picture. ANNIE is the result of years of close interaction between computational biologists and computer scientists and automates an essential part of this sequence analytic process. It brings together over 20 function prediction algorithms that have proven sufficiently reliable and indispensable in daily sequence analytic work and are meant to give scientists a quick overview of possible functional assignments of sequence segments in the query proteins. The results are displayed in an integrated manner using an innovative AJAX-based sequence viewer. ANNIE is available online at: http://annie.bii.a-star.edu.sg. This website is free and open to all users and there is no login requirement.

INTRODUCTION

Advances in sequencing technology have taken the number of available sequences in databases to unprecedented levels (1). Unfortunately, the ability to determine the sequence of a particular gene has not been accompanied by an equally impressive gain in our ability to achieve insights into the biological function (including molecular and celullar) of these sequences. For example, the full genome sequence of the yeast *Saccharomyces cerevisae* became available in 1997 (2); nevertheless more than a decade later, of the 6000+ identified genes there are still over 1000 with uncharacterized function (3). In human, more than half of the genes are functionally characterized incompletely or not at all.

The classic route to functional characterization involving experimental methods from the genetic and biochemical toolbox-like specific knockouts, targeted mutations and a battery of biochemical assays is time consuming (depending on the model organism, it can take years) and costly. Therefore, there is a strong case for using *in silico* methods in a preliminary analysis for functional hypothesis generation to direct experimental planning in the laboratory.

There are literally hundreds of prediction algorithms described in the literature, although only some of those have a sensitivity and selectivity to be applicable for unsupervised function prediction of arbitrary query protein sequences (4). Each method concentrates on some specific structural or functional aspect of a sequence, e.g. the distribution of unstructured regions (5), its amino acid compositional particularities in sequence windows (6) or the existence of globular domains (7,8). The input formats, method of program invocation as well as the result presentation vary widely making it difficult to interconnect results and obtain an integrated picture of a possible functional assignment. Even when concentrating on a smaller set of reliable prediction methods, the results can still easily exceed several Megabytes of textual (ASCII-type) information, integration of which into an overall functional prediction can be a formidable task requiring days of work per sequence. The need for standardizing automated annotations as well as assessing their quality has been recognized by initiatives such as AFP (9).

There have been several attempts to address the interoperability problem (10–12). JAFA (13) is an example of an annotation meta-server that sends a query sequence to several function prediction servers and displays the overlap in Gene Ontology terms (14) as well as providing links to the original results. The ProFunc server (15) combines a range of methods for sequence analysis but requires the

*To whom correspondence should be addressed. Tel: (65) 6478 8264; Fax: (65) 6478 9047; Email: georgs@bii.a-star.edu.sg

Table 1. Sequence analytic algorithms

Algorithm	Description	Parameters
CAST (37)	Algorithm for low-complexity region (LCR) detection and selective masking	Threshold: 40
IUPred (5)	Prediction method for recognizing ordered and intrinsically unstructured/disordered regions in proteins	Prediction type: long disorder
SAPS (6)	Statistical analysis of protein sequences with respect to amino acid composition and simple sequence motifs	n/a
SEG (38)	Prediction of low complexity regions	Three parameter sets: Window-size 12, Locut 2.2, Hicut 2.5 Window-size 25, Locut 3.0, Hicut 3.3 Window-size 45, Locut 3.4, Hicut 3.75
Big-∏ (27–29)	Prediction of protein GPI lipid anchor cleavage sites	Taxon-specific learning set
NMT (30,31)	Prediction of N-terminal N-myristoylation of proteins	Taxon-specific parameter set
PrePS – FT (32)	Farnesylation prediction	n/a
PrePS – GGT1 (32)	Geranylgeranylation prediction	n/a
PrePS – GGT2 (32)	Rab geranylgeranylation Prediction	n/a
PeroPS/PTS1 (33,34)	Prediction of peroxisomal targeting signal 1	Taxon-specific prediction function
DAS-TMfilter (39)	Prediction of transmembrane regions	Quality cutoff: 0.72
HMMTOP (40)	Transmembrane topology prediction using Hidden Markov models	n/a
PHOBIUS (41)	Combined transmembrane topology and signal peptide predictor	n/a
TMHMM (42)	Transmembrane helix predictor	n/a
IMP-COIL (43)	Prediction of coiled-coil regions, modified implementation of the algorithm Lupas *et al.* by F. Eisenhaber	n/a
PROSITE (44)	Pattern search in the PROSITE database	n/a
PROSITE-Profile (44)	Profile search in the PROSITE database	n/a
HMMER (7)	Profile Hidden Markov Models	SMART (8) with *e*-value cutoff of 0.001
IMPALA (45)	Tool to compare a query sequence against a library of position-specific scoring matrices	Wolf-library (*e*-value cutoff: 0.001) (46), Aravind-library (*e*-value cutoff: 1e-5) (47)
RPS-BLAST against CDD (48)	Reverse-position-specific BLAST against the Conserved Domain Database (CDD)	*e*-value cutoff: 0.001

3D structure of the query to be known in advance. There are also a number of databases that provide sequence annotations from various sources like UniProtKB (16) or Ensembl (17) as well as some services that predict a limited set of features for a given input sequence such as SMART (8), InterProScan (18,19) or TarO (20). It should be noted that, frequently, database annotations contain errors and, especially, function descriptions propagated by sequence similarity criteria might be dubious. Therefore, tools for *de novo* sequence annotation are important for reducing the dependence on potentially misleading or incomplete database comments (21,22).

ANNIE is unique in that it has been developed by a collaboration of sequence analysis as well as computer science experts. It provides over 20 of the most useful algorithms (Table 1) covering the first two steps of segment-based sequence analysis (23) that have proven indispensable in daily sequence analytic work for functional discovery (24–26). Of particular value is the inclusion of predictors for a number of post-translational modifications (27–32) as well as targeting signals (33,34) developed in-house.

The results of all algorithms are displayed in an integrated manner using a newly developed interactive sequence viewer as well as a number of views highlighting the distribution of features across sets of sequences. ANNIE enables scientists to gain a quick overview of possible functional assignments in protein sequence sets.

METHODS

Algorithms

Segment-based sequence analysis (23) starts with the assumption that proteins are chains of functional units which can be analyzed independently. The overall function arises from the synthesis of the functions predicted for each individual module.

The procedure first uses algorithms for the detection of nonglobular regions, which are segments with a compositional bias or repetitive patterns that often represent linker regions, fibrillar segments, flexible binding sites or points of post-translational modifications (35). The subsequent step is to run algorithms for the identification of known globular domains. These domains are conserved within groups of homologous proteins and are often associated with enzymatic or ligand-binding function. In the last step, it is assumed that the remaining parts of the sequence represent yet uncharacterized globular domains that need to be characterized within the homologous family concept. Iterative heuristic have to be applied to uncover weak links in sequence space and collect a family of protein sequence segments that contain yet unknown globular domains (36).

ANNIE provides a selection of algorithms covering the first two steps of this approach. Table 1 lists the algorithms which have been integrated together with a short description, references and the preselected runtime parameters. These parameters have been chosen so as to provide a reasonable compromise between the need to give a comprehensive and sensitive overview of sometimes weak signals and the ability of scientists not trained in sequence analysis to discard false positives. It should be noted that further relaxed parameterization might produce more prediction results; yet, their interpretation might require expert knowledge and experience. ANNIE is based on our extensive in-house sequence analytic pipeline ANNOTATOR, which is used to analyze proteomes and detect distant evolutionary relationships using computationally intensive iterative heuristics (36). The engine behind ANNIE has been in use for several years and has annotated millions of sequences. The online help pages contain a detailed description of each individual algorithm.

User-interface

There are two input methods allowing the user to either paste sequences in FASTA-format (a single sequence can also be pasted without a description line) or upload them from a corresponding FASTA-formatted file. There is currently a limit of 10 sequences per annotation run which might be increased in the future depending on actual usage patterns and the availability of compute server resources.

It is highly recommended to include taxonomic information in the classical NCBI square bracket notation at the end of the description line (e.g. [*Homo sapiens*]) in order for ANNIE to automatically choose the correct parameterization for predictors of post-translational modifications and targeting signals. Additionally, this will enable the user to view the taxonomic distribution of the uploaded sequence set.

The annotation process is started by pressing the corresponding 'Annotate' button. Requests are queued and, upon availability of resources, sent to a cluster of dedicated CPUs for execution of algorithms and parsing of output. The user will be directed to a page containing the current as well as past results. If an (optional) email address is provided, a message containing a link will be sent once all algorithms have completed. This gives the user access to past annotations for at least 72 h, after which they will be deleted.

There are a number of views that allow the user to look at different aspects of the annotation. Upon submission of an annotation request the user will normally click on the corresponding result folder and be presented with a view displaying the uploaded sequences with links to individual results. If a certain algorithm is still queued or running a special symbol will be displayed and the page reloads periodically until all algorithms have terminated (under average load this should take no more than 1 min).

Result view

Following the links for individual algorithms will display the corresponding result together with links to external resources where applicable (e.g. domain descriptions for HMMER). Each result also provides access for validation purposes to the 'raw' unparsed data generated by the executable.

Interactive sequence view

Clicking on the protein sequence symbol starts the interactive sequence view (Figure 1). The results of individual algorithms are displayed as rectangles projected onto the sequence ruler. Hovering over regions will display information specific to the result (e.g. *e*-values of globular domain model hits). Right-clicking on a region will allow examination of the particular feature in greater detail with algorithm-specific information as well as a compositional analysis of the sequence stretch.

Figure 1 displays the interactive sequence view of Dysferlin (49,50), a protein involved in a number of hereditary myopathies (it is provided as a sample sequence on the main page). The characteristic C2-domains (51) have been detected by a number of distinct tools (HMMER against Smart, IMPALA against Wolf-Library, PROSITE-Profile search, RPS-Blast against CDD) giving enhanced confidence to that particular finding. The detection of a C-terminal membrane-embedded region by three different methods also lends plausibility to the claim that Dysferlin is a transmembrane protein. It should be noted that there is a seventh C2-domain not shown in this view between residues 1338 and 1437 (the *e*-value = 0.025 is above the default threshold of 0.001),

Due to the AJAX-based technology of the viewer, zooming and panning is almost instantaneous, allowing fast and concise drill-down to a particular region. Additional feature-specific information can be obtained by right-clicking on a region. This will lead to a detailed compositional analysis of the sequence stretch and, were applicable, include alignment data as well as links to external resources.

Set view

Uploading several sequences at once opens up the possibility to analyze the frequency of certain features within that set of sequences. ANNIE provides a special view called 'Histogram' (Figure 2). This view displays features found with diverse algorithms sorted by the number of occurrences. Clicking on the name of the feature will link to all the sequences in which it has been detected.

A third view called 'Taxonomy' (Figure 3) shows the taxonomic distribution of sequences within the set.

CONCLUSIONS AND OUTLOOK

We have presented ANNIE, a comprehensive *de novo* protein annotation system that integrates a large number of indispensable algorithms used in everyday sequence analytic work. The results of individual algorithms can be accessed separately or displayed together in an interactive

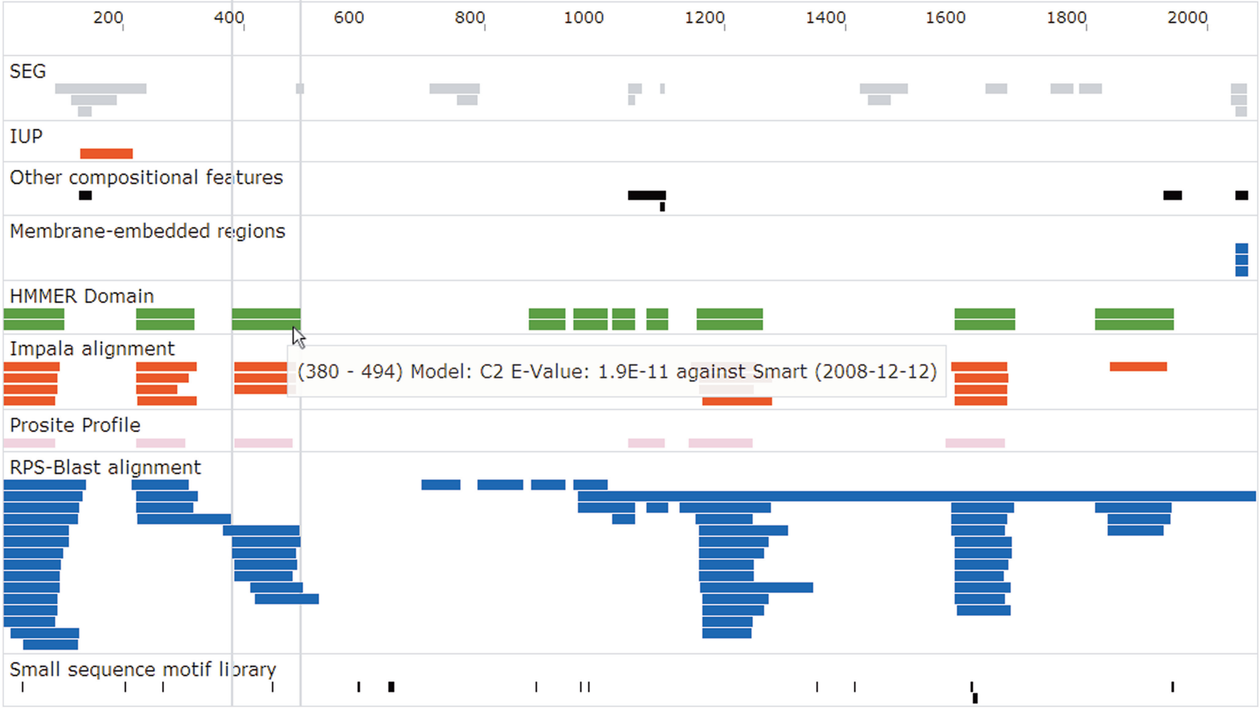
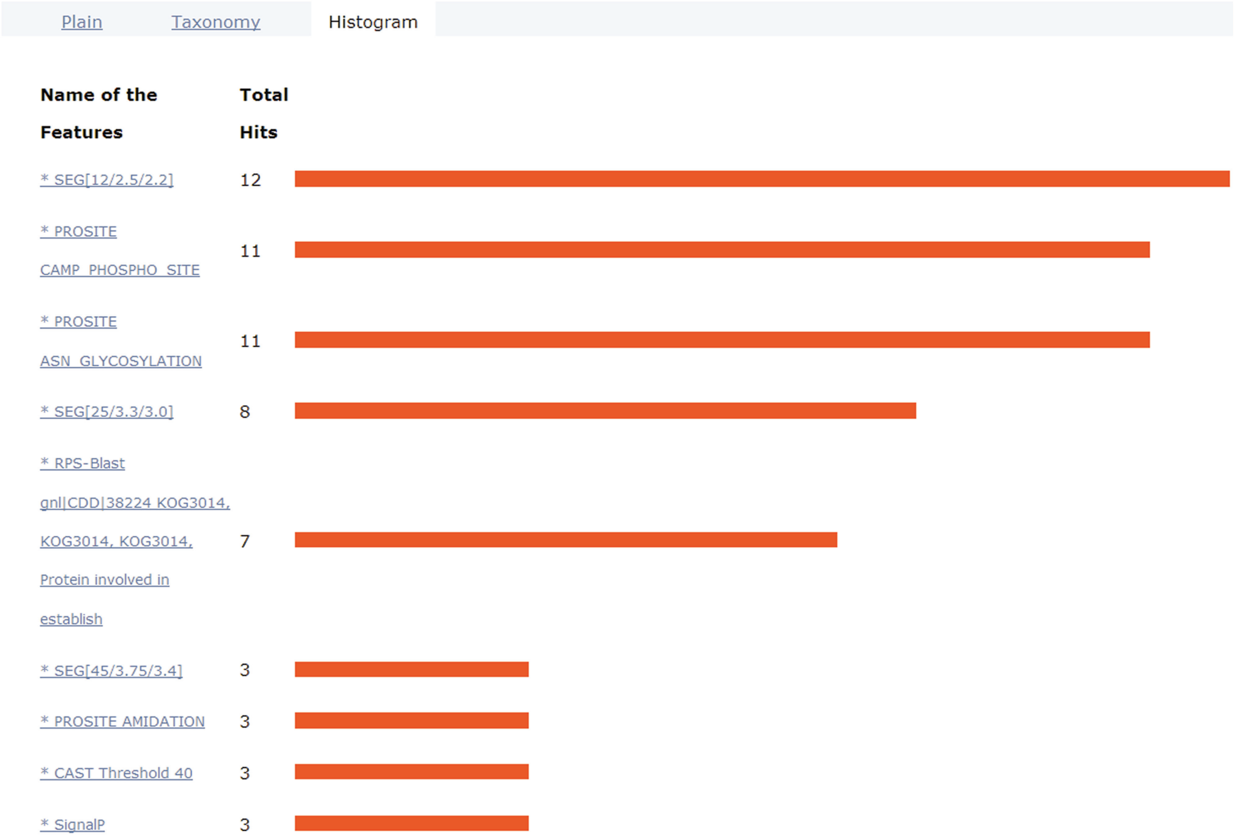
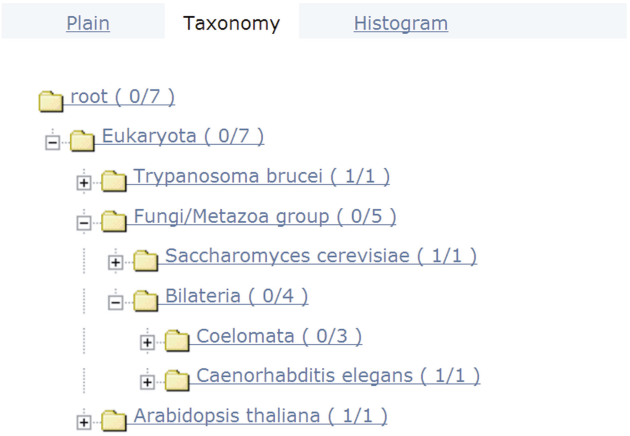

200
400
600
800
1000
1200
1400
1600
1800
2000
SEG
IUP
Other compositional features
Membrane-embedded regions
HMMER Domain
Impala alignment
(380 - 494) Model: C2 E-Value: 1.9E-11 against Smart (2008-12-12)
Prosite Profile
RPS-Blast alignment
Small sequence motif library

Figure 1. Interactive sequence view. This figure shows an exemplary interactive sequence view using the sequence of Dysferlin. The sequence features found by the various programs are organized in panes that coalesce findings with similar functional significance. The different color coding is just for the purpose of easing navigation.



Name of the Features	Total Hits
* SEG[12/2.5/2.2]	12
* PROSITE CAMP_PHOSPHO_SITE	11
* PROSITE ASN_GLYCOSYLATION	11
* SEG[25/3.3/3.0]	8
* RPS-Blast gnl|CDD|38224 KOG3014, KOG3014, KOG3014, Protein involved in establish	7
* SEG[45/3.75/3.4]	3
* PROSITE AMIDATION	3
* CAST Threshold 40	3
* SignalP	3

Figure 2. Histogram view. This view shows the occurrence of sequence features in the sequence set under investigation. The features are sorted by their number of incidences in the set. Clicking on the link provided with the feature name will generate the sublist of sequences with this feature. In this example of Eco1-type proteins, the top four entries in the histogram are related to low-complexity regions as well as short motifs from PROSITE that are less reliable predictions. The fifth entry indicates the occurrence of the KOG3014 domain model that is characteristic for the Eco1-class of proteins necessary for the establishment of sister chromatid cohesion in mitosis.



- root (0/7)
  - Eukaryota (0/7)
    - Trypanosoma brucei (1/1)
    - Fungi/Metazoa group (0/5)
      - Saccharomyces cerevisiae (1/1)
      - Bilateria (0/4)
        - Coelomata (0/3)
        - Caenorhabditis elegans (1/1)
    - Arabidopsis thaliana (1/1)

Figure 3. Taxonomy view. The taxonomic distribution of the sequence set is displayed. The numbers in brackets refer to the number of sequences below a branch in the taxonomic tree and those assigned to a particular taxon. For the given Eco1 example set, this view shows that it contains one plant sequence (*Arabidopsis thaliana*) together with a trypanosome, one fungal sequence and four from *Bilateria*.

AJAX-based sequence viewer. There are additional views for assessing the frequency of certain features across a set of sequences as well as revealing its taxonomic distribution.

New algorithms appearing in the literature are constantly being evaluated as to their potential contribution for function discovery and are eventually integrated. Future work will also see the inclusion of algorithms from the third step of segment-based sequence analysis if the necessary computational resources can be obtained.

ACKNOWLEDGEMENTS

We would like to thank all colleagues that have contributed at various stages to the development of the software for protein sequence studies in the Eisenhaber group especially Georg Kraml, Miklos Kozlovszky, Florian Leitner, Georg Neuberger, Maria Novatchkova, Gernot Stocker, James Tay, Sun Tian and Wong Chee Hong.

FUNDING

Bioinformatics Insitute, A*Star, Singapore; Boehringer Ingelheim (2001-2007); and the Austrian Gen-AU BIN program (2004-2007) when the Eisenhaber group was located at the Research Institute of Molecular Pathology in Vienna (Austria). Funding for open access charge: Bioinformatics Institute, A*Star, Singapore.

Conflict of interest statement. None declared.

REFERENCES

1. Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J. and Wheeler,D.L. (2008) GenBank. *Nucleic Acids Res.*, **36**, D25–D30.
2. Cherry,J.M., Ball,C., Weng,S., Juvik,G., Schmidt,R., Adler,C., Dunn,B., Dwight,S., Riles,L., Mortimer,R.K. *et al.* (1997) Genetic and physical maps of Saccharomyces cerevisiae. *Nature*, **387**, 67–73.
3. Peña-Castillo,L. and Hughes,T.R. (2007) Why are there still over 1000 uncharacterized yeast genes? *Genetics*, **176**, 7–14.
4. Ponting,C.P. (2001) Issues in predicting protein function from sequence. *Brief Bioinform.*, **2**, 19–29.
5. Dosztányi,Z., Csizmók,V., Tompa,P. and Simon,I. (2005) The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins. *J. Mol. Biol.*, **347**, 827–839.
6. Brendel,V., Bucher,P., Nourbakhsh,I.R., Blaisdell,B.E. and Karlin,S. (1992) Methods and algorithms for statistical analysis of protein sequences. *Proc. Natl Acad. Sci. USA*, **89**, 2002–2006.
7. Eddy,S.R. (1998) Profile hidden Markov models. *Bioinformatics*, **14**, 755–763.
8. Letunic,I., Doerks,T. and Bork,P. (2009) SMART 6: recent updates and new developments. *Nucleic Acids Res.*, **37**, D229–D232.
9. Rodrigues,A.P.C., Grant,B.J., Godzik,A. and Friedberg,I. (2007) The 2006 automated function prediction meeting. *BMC Bioinformatics*, **8 (Suppl. 4)**, S1–S4.
10. Letondal,C. (2001) A web interface generator for molecular biology programs in Unix. *Bioinformatics*, **17**, 73–82.
11. Hull,D., Wolstencroft,K., Stevens,R., Goble,C., Pocock,M.R., Li,P. and Oinn,T. (2006) Taverna: a tool for building and running workflows of services. *Nucleic Acids Res.*, **34**, W729–W732.
12. Wilkinson,M.D. and Links,M. (2002) BioMOBY: an open source biological web services proposal. *Brief Bioinform*, **3**, 331–341.
13. Friedberg,I., Harder,T. and Godzik,A. (2006) JAFA: a protein function annotation meta-server. *Nucleic Acids Res.*, **34**, W379–W381.
14. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.
15. Laskowski,R.A., Watson,J.D. and Thornton,J.M. (2005) ProFunc: a server for predicting protein function from 3D structure. *Nucleic Acids Res.*, **33**, W89–W93.
16. Boutet,E., Lieberherr,D., Tognolli,M., Schneider,M. and Bairoch,A. (2007) UniProtKB/Swiss-Prot. *Methods Mol. Biol.*, **406**, 89–112.
17. Hubbard,T.J.P., Aken,B.L., Ayling,S., Ballester,B., Beal,K., Bragin,E., Brent,S., Chen,Y., Clapham,P., Clarke,L. *et al.* (2009) Ensembl 2009. *Nucleic Acids Res.*, **37**, D690–D697.
18. Mulder,N. and Apweiler,R. (2007) InterPro and InterProScan: tools for protein sequence classification and comparison. *Methods Mol Biol.*, **396**, 59–70.
19. Hunter,S., Apweiler,R., Attwood,T.K., Bairoch,A., Bateman,A., Binns,D., Bork,P., Das,U., Daugherty,L., Duquenne,L. *et al.* (2009) InterPro: the integrative protein signature database. *Nucleic Acids Res.*, **37**, D211–D215.
20. Overton,I.M., van Niekerk,C.A.J., Carter,L.G., Dawson,A., Martin,D.M.A., Cameron,S., McMahon,S.A., White,M.F., Hunter,W.N., Naismith,J.H. *et al.* (2008) TarO: a target optimisation system for structural biology. *Nucleic Acids Res.*, **36**, W190–W196.
21. Gilks,W.R., Audit,B., De Angelis,D., Tsoka,S. and Ouzounis,C.A. (2002) Modeling the percolation of annotation errors in a database of protein sequences. *Bioinformatics*, **18**, 1641–1649.
22. Gilks,W.R., Audit,B., de Angelis,D., Tsoka,S. and Ouzounis,C.A. (2005) Percolation of annotation errors through hierarchically structured protein sequence databases. *Math. Biosci.*, **193**, 223–234.
23. Eisenhaber,F. (2006) Prediction of protein function. *Discovering Biomolecular Mechanisms with Computational Biology*. Springer US, pp. 39–54.
24. Rea,S., Eisenhaber,F., O'Carroll,D., Strahl,B.D., Sun,Z.W., Schmid,M., Opravil,S., Mechtler,K., Ponting,C.P., Allis,C.D. *et al.* (2000) Regulation of chromatin structure by site-specific histone H3 methyltransferases. *Nature*, **406**, 593–599.
25. Ivanov,D., Schleiffer,A., Eisenhaber,F., Mechtler,K., Haering,C.H. and Nasmyth,K. (2002) Eco1 is a novel acetyltransferase that can acetylate proteins involved in cohesion. *Curr. Biol.*, **12**, 323–328.
26. Schleiffer,A., Kaitna,S., Maurer-Stroh,S., Glotzer,M., Nasmyth,K. and Eisenhaber,F. (2003) Kleisins: a superfamily of bacterial and eukaryotic SMC protein partners. *Mol. Cell*, **11**, 571–575.

27. Eisenhaber,B., Bork,P. and Eisenhaber,F. (1999) Prediction of potential GPI-modification sites in proprotein sequences. *J. Mol. Biol.*, **292**, 741–758.
28. Eisenhaber,B., Wildpaner,M., Schultz,C.J., Borner,G.H.H., Dupree,P. and Eisenhaber,F. (2003) Glycosylphosphatidylinositol lipid anchoring of plant proteins. Sensitive prediction from sequence- and genome-wide studies for Arabidopsis and rice. *Plant Physiol.*, **133**, 1691–1701.
29. Eisenhaber,B., Schneider,G., Wildpaner,M. and Eisenhaber,F. (2004) A sensitive predictor for potential GPI lipid modification sites in fungal protein sequences and its application to genome-wide studies for Aspergillus nidulans, Candida albicans, Neurospora crassa, Saccharomyces cerevisiae and Schizosaccharomyces pombe. *J. Mol. Biol.*, **337**, 243–253.
30. Maurer-Stroh,S., Eisenhaber,B. and Eisenhaber,F. (2002) N-terminal N-myristoylation of proteins: prediction of substrate proteins from amino acid sequence. *J. Mol. Biol.*, **317**, 541–557.
31. Maurer-Stroh,S., Eisenhaber,B. and Eisenhaber,F. (2002) N-terminal N-myristoylation of proteins: refinement of the sequence motif and its taxon-specific differences. *J. Mol. Biol.*, **317**, 523–540.
32. Maurer-Stroh,S. and Eisenhaber,F. (2005) Refinement and prediction of protein prenylation motifs. *Genome Biol.*, **6**, R55.
33. Neuberger,G., Maurer-Stroh,S., Eisenhaber,B., Hartig,A. and Eisenhaber,F. (2003) Motif refinement of the peroxisomal targeting signal 1 and evaluation of taxon-specific differences. *J. Mol. Biol.*, **328**, 567–579.
34. Neuberger,G., Maurer-Stroh,S., Eisenhaber,B., Hartig,A. and Eisenhaber,F. (2003) Prediction of peroxisomal targeting signal 1 containing proteins from amino acid sequence. *J. Mol. Biol.*, **328**, 581–592.
35. Eisenhaber,B. and Eisenhaber,F. (2007) Posttranslational modifications and subcellular localization signals: indicators of sequence regions without inherent 3D structure? *Curr. Protein Pept. Sci.*, **8**, 197–203.
36. Schneider,G., Neuberger,G., Wildpaner,M., Tian,S., Berezovsky,I. and Eisenhaber,F. (2006) Application of a sensitive collection heuristic for very large protein families: evolutionary relationship between adipose triglyceride lipase (ATGL) and classic mammalian lipases. *BMC Bioinformatics*, **7**, 164.
37. Promponas,V.J., Enright,A.J., Tsoka,S., Kreil,D.P., Leroy,C., Hamodrakas,S., Sander,C. and Ouzounis,C.A. (2000) CAST: an iterative algorithm for the complexity analysis of sequence tracts. Complexity analysis of sequence tracts. *Bioinformatics*, **16**, 915–922.
38. Wootton,J.C. (1994) Non-globular domains in protein sequences: automated segmentation using complexity measures. *Comput. Chem.*, **18**, 269–285.
39. Cserzo,M., Eisenhaber,F., Eisenhaber,B. and Simon,I. (2004) TM or not TM: transmembrane protein prediction with low false positive rate using DAS-TMfilter. *Bioinformatics*, **20**, 136–137.
40. Tusnády,G.E. and Simon,I. (2001) The HMMTOP transmembrane topology prediction server. *Bioinformatics*, **17**, 849–850.
41. Käll,L., Krogh,A. and Sonnhammer,E.L.L. (2004) A combined transmembrane topology and signal peptide prediction method. *J. Mol. Biol.*, **338**, 1027–1036.
42. Krogh,A., Larsson,B., von Heijne,G. and Sonnhammer,E.L. (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.*, **305**, 567–580.
43. Lupas,A., Van Dyke,M. and Stock,J. (1991) Predicting coiled coils from protein sequences. *Science*, **252**, 1162–1164.
44. Hulo,N., Bairoch,A., Bulliard,V., Cerutti,L., Cuche,B.A., de Castro,E., Lachaize,C., Langendijk-Genevaux,P.S. and Sigrist,C.J.A. (2008) The 20 years of PROSITE. *Nucleic Acids Res.*, **36**, D245–D249.
45. Schäffer,A.A., Wolf,Y.I., Ponting,C.P., Koonin,E.V., Aravind,L. and Altschul,S.F. (1999) IMPALA: matching a protein sequence against a collection of PSI-BLAST-constructed position-specific score matrices. *Bioinformatics*, **15**, 1000–1011.
46. Wolf,Y.I., Brenner,S.E., Bash,P.A. and Koonin,E.V. (1999) Distribution of protein folds in the three superkingdoms of life. *Genome Res.*, **9**, 17–26.
47. Chervitz,S.A., Aravind,L., Sherlock,G., Ball,C.A., Koonin,E.V., Dwight,S.S., Harris,M.A., Dolinski,K., Mohr,S., Smith,T. *et al.* (1998) Comparison of the complete protein sets of worm and yeast: orthology and divergence. *Science*, **282**, 2022–2028.
48. Marchler-Bauer,A., Panchenko,A.R., Shoemaker,B.A., Thiessen,P.A., Geer,L.Y. and Bryant,S.H. (2002) CDD: a database of conserved domain alignments with links to domain three-dimensional structure. *Nucleic Acids Res.*, **30**, 281–283.
49. Matsuda,C., Aoki,M., Hayashi,Y.K., Ho,M.F., Arahata,K. and Brown,R.H. (1999) Dysferlin is a surface membrane-associated protein that is absent in Miyoshi myopathy. *Neurology*, **53**, 1119.
50. Han,R. and Campbell,K.P. (2007) Dysferlin and muscle membrane repair. *Curr. Opin. Cell Biol.*, **19**, 409–416.
51. Nalefski,E.A. and Falke,J.J. (1996) The C2 domain calcium-binding motif: structural and functional diversity. *Protein Sci.*, **5**, 2375–2390.