Title	In Silico Protein Motif Discovery and Structural Analysis
Authors(s)	Mooney, Catherine; Davey, Norman E.; Martin, Alberto J. M.; Walsh, Ian; Shields, Denis C.; Pollastri, Gianluca
Publication date	2011-06-30
Publication information	Yu, B. and Hinchcliffe, M. (eds.). In Silico Tools for Gene Discovery
Series	Methods in Molecular Biology; 760
Publisher	Springer
Item record/more information	http://hdl.handle.net/10197/13126
Publisher's statement	The final publication is available at www.springerlink.com.
Publisher's version (DOI)	10.1007/978-1-61779-176-5_21

In Silico Protein Motif Discovery and Structural Analysis

Catherine Mooney[*1,2,3],Norman Davey[4], Alberto J.M. Martin[1,5,6], Ian Walsh[1,5,6], Denis C. Shields[1,2,3] and Gianluca Pollastri[1,6]

[1] Complex and Adaptive Systems Laboratory, University College Dublin, Belfield, Dublin 4, Ireland

[2] Conway Institute of Biomolecular and Biomedical Research, University College Dublin, Belfield, Dublin 4, Ireland

[3] School of Medicine and Medical Science, University College Dublin, Belfield, Dublin 4, Ireland

[4] EMBL Structural and Computational Biology Unit, Meyerhofstrasse 1, 69117 Heidelberg, Germany

[5] Biocomputing UP, Department of Biology, University of Padua, Viale G. Colombo 3, I-35131 Padova, Italy

[6] School of Computer Science and Informatics, University College Dublin, Belfield, Dublin 4, Ireland

Email: Catherine Mooney[*]- catherine.mooney@ucd.ie; Norman Davey - norman.davey@embl.de; Alberto J. M. Martin - albertoj@bio.unip.it; Ian Walsh - ian.walsh@bio.unipd.it; Denis C. Shields - denis.shields@ucd.ie; Gianluca Pollastri - gianluca.pollastri@ucd.ie;

[*]Corresponding author

Abstract

A wealth of *in-silico* tools is available for protein motif discovery and structural analysis. The aim of this chapter is to collect some of the most common and useful tools and to guide the biologist in their use. A detailed explanation is provided for the use of Distill, a suite of web-servers for the prediction of protein structural features and the prediction of full-atom 3D models from a protein sequence. Besides this, we also provide pointers to many other tools available for motif discovery and secondary and tertiary structure prediction from a primary amino acid sequence. The prediction of protein intrinsic disorder, and the prediction of functional sites and SLiMs are also briefly discussed. Given that user queries vary greatly in size, scope and character, the trade-offs in speed, accuracy and scale need to be considered when choosing which methods to adopt.

Key words: protein structure prediction; secondary structure; disorder; functional sites; SLiMs.

1. Introduction

Compared with over 10 million known protein sequences (UniProtKB/TrEMBL ***(1)***), as of June 2010 there are only in the region of 60,000 proteins of known structure deposited in the Protein Data Bank (PDB) ***(2)***. As experimental determination of a protein's structure is difficult, expensive and time consuming, the gap between sequence-known and structure-known proteins is continuing to grow rapidly. Currently the only feasible way to bridge this gap is computational modelling. This is especially important for analysis at a genomic or inter-genomic level, where informative structural models need to be generated for thousands of gene products (or portions of them) in a reasonable amount of time.

Computational modelling methods can be divided into two groups: those that use similarity to

proteins of known structure to model all or part of the query protein (comparative or template-based modelling) and *ab initio* or *de novo* prediction methods where no similarity to a protein of known structure can be found. If a close homologue is found (e.g. a protein of known structure with a sequence identity greater than approximately 30% to the query) then a model can be produced with a high degree of confidence in its accuracy ***(3)***. However, many proteins share similar structures even though their sequences may share less than 15% sequence identity ***(4)***. Finding these remote homologues is a much more difficult task. As structural genomic (SG) projects worldwide gather momentum the hope is to populate the protein fold space with a useful 3D model for all protein families using high throughput protein structure determination methods ***(5)***. Providing more accurate templates for more proteins should lead to an increase in protein structure prediction accuracy for many proteins and move them out of the *ab initio*/*de novo* prediction category into the comparative/homology modelling category. As the accuracy of predicted 3D protein models improves they are becoming increasingly more useful in biomolecular and biomedical research. In the absence of an experimental structure there are many applications for which a predicted structure may be of use to biologists. Moult ***(6)*** describes the uses of models at three levels of resolution. At the lowest level of resolution are models which have typically been produced by remote fold recognition relationships and are likely to have many errors, however, they may still be useful for domain boundary, super-family and approximate function identification. Medium resolution models, built using less remote homologues, for instance obtained from a carefully designed PSI-BLAST ***(7)*** search, may be used to identify possible protein-protein interaction sites, the likely role of disease-associated substitutions or the consequences of alternative splicing in protein function. Higher resolution models, where there is a known structure showing at least 30% sequence identity to the query sequence may be useful for molecular replacement in solving a crystal structure, give insight into the impact of mutations in disease, the consequences of missense or

nonsense mutations for protein structure and function, identification of orthologous functional relationships and aspects of molecular function which may not be possible from an experimental structure.

Due to space limitations, we cannot cover all web-servers available for protein structural motif discovery and structure prediction, but provide a useful overview of the area. Although we will provide a detailed description of our Distill suite of servers, we will also point the reader to other publicly available, up-to-date, accurate and easy to use *in silico* tools which have the potential to predict structures, structural features or motifs on a genomic scale.

2. Materials

There are many freely available *in silico* tools to aid the active researcher which can not only save time but as many are constantly updated and improved upon, ensure that one's research is in keeping with or at the state-of-the-art (*See* **Note 4**). Below we have listed those we have found to be most useful in our experience.

2.1 Protein Structural Feature Prediction

- Distill (http://distill.ucd.ie/distill/)
- PROTEUS (http://wks16338.biology.ualberta.ca/proteus/)
- Scratch (SSpro and ACCpro) (http://scratch.proteomics.ics.uci.edu/)
- Jpred 3 (http://www.compbio.dundee.ac.uk/www-jpred/)
- PSIPRED (http://bioinf4.cs.ucl.ac.uk:3000/psipred/)
- SABLE (http://sable.cchmc.org/)

2.2 Protein 3D Structure Prediction

- 3Distill (http://distill.ucd.ie/distill/)
- I-TASSER (http://zhanglab.ccmb.med.umich.edu/I-TASSER/)
- HHpred (http://toolkit.tuebingen.mpg.de/hhpred/)
- Robetta (http://robetta.bakerlab.org/)

2.3 Functional Site Prediction for Structured Proteins

- SDPsite (http://bioinf.fbb.msu.ru/SDPsite/index.jsp)
- ConSurf (http://consurf.tau.ac.il)
- Evolutionary Trace (http://mammoth.bcm.tmc.edu/reportmaker)
- SITEHOUND (http://bsbbsinai.org/SHserver/SiteHound/)

2.4 Disorder Prediction

- Spritz (http://distill.ucd.ie/spritz)
- IUPred server (http://iupred.enzim.hu/)

2.5 SLiM Discovery, Rediscovery and Post-Processing

- The ELM Server (http://elm.eu.org/)
- Minimotif Miner (http://mnm.engr.uconn.edu/MNM/)
- SIRW (http://sirw.embl.de/)
- SLiMSearch, SLiMfinder and CompariMotif (http://bioware.ucd.ie/)
- Dilimot (http://dilimot.embl.de/)
- ANCHOR (http://anchor.enzim.hu/)
- Conscore (http://conscore.embl.de)
- PepSite (http://pepsite.embl.de/)

3. Methods

3.1 Prediction of Protein Structural Features

3.1.1 Distill: Protein Structure and Structural Feature Prediction Server

Distill ***(8)*** is a suite of web servers available to the public for protein structure and structural feature prediction. The Distill suite of servers currently contains nine predictors: six predictors of 1D features (i.e. properties which may be represented as a string of the same length as the amino acid sequence - secondary structure ***(9)***, contact

density ***(10)***, local structural motifs ***(11)***, relative solvent accessibility ***(12)***, protein disorder ***(13)*** and protein domain boundary prediction ***(14)***), a coarse contact map and protein topology predictor, a predictor of protein residue contact maps ***(15)*** and the predictor of full-atom 3D models and C*α* traces (3Distill). The servers are based on large-scale ensembles of machine learning systems that include recursive neural networks, support vector machines and monte carlo simulations. They are trained on large, up-to-date, non-redundant subsets of the PDB ***(2)***.

Structural motifs ***(11)*** are identified by applying multidimensional scaling and clustering to pair-wise angular distances between quadruplets of ϕ - ψ dihedral angle pairs collected from high-resolution protein structures ***(16)***. Structural motif predictions are highly informative and provide a finer-resolution picture of a protein backbone and may be used to improve traditional three class secondary structure, and for the identification of remote homologues ***(17)***. The definition and one-letter code for the fourteen structural motifs are provided on the Distill help page.

Each of the servers take as input a profile obtained from multiple sequence alignments of the protein sequence to its homologues in the UniRef90 database ***(18)*** to leverage evolutionary information. Until recently predictors of 1D structural properties have generally been *ab initio*. However it has been shown that evolutionary information from proteins of known structure can contribute to more accurate 1D prediction ***(12, 17, 19, 20)***. When available, this information, in the form of homologous structures from the PDB, is provided as a further input to all the servers, resulting in greatly improved reliability. For more information on the use of homology during the predictive process see references ***(12)*** and ***(17)***.

In addition, 1D predictions augment the 2D and 3D predictions as follows: secondary structure and solvent accessibility are provided as additional input to the residue contact maps and coarse protein topology predictors; secondary structure, solvent accessibility

and contact density are provided as additional input to the residue contact maps predictor; secondary structure, solvent accessibility, structural motif, contact density, coarse and residue contact maps are provided as additional input to 3Distill (3D). For a more detailed description of the models and training algorithms see ***(8-17)***.

All predictions are freely available through a simple joint web interface and the results are returned by email. In a single submission a user can send protein sequences for over 32,000 residues to all or a selection of the servers. If a template is found in the PDB the sequence identity between the query sequence and the best template is provided (*See* **Note 3**).

3.1.2 Other 1D Structural Feature Prediction Servers

Some other popular secondary structure (SS) and relative solvent accessibility (RSA) prediction servers are **PROTEUS** ***(19)*** and **Scratch** (SSpro and ACCpro) ***(20)*** which include homology to proteins of known structure in the PDB, if available, during the prediction process. **Jpred 3** ***(21)*** will notify the user if there is a homologous sequence available in the PDB prior to prediction, but does not include this information in the prediction process. **PSIPRED** ***(22)*** and **SABLE** ***(23)*** are *ab initio* predictors (*See* **Note 2**). Methods of searching for and incorporating homology information into the prediction process vary between the different servers, see ***(17)*** for further discussion of some of the different methods for homology search and inclusion.

3.2 Three-Dimensional Protein Structure Prediction

3.2.1 3D Prediction by Distill

3Distill ***(8)*** is a server for the prediction of full-atom 3D models of protein structures which accepts queries of up to 250 amino acids in length. 3Distill relies on a fast optimization algorithm guided by a potential based on secondary structure, solvent accessibility, structural motif,

contact density, coarse contact maps and residue contact maps, all predicted by Distill. Note that, when available, homology information is provided to 3Distill which results in substantially improved predictions. 3Distill and the underlying servers have been tuned and generally improved in the lead-up to CASP9. Input into the servers is handled by the same two simple HTML forms for the submission of single and multiple queries as for 1D prediction. 3Distill's outputs come as attachments in PDB format. Five ranked models are returned in PDB file format, each one containing all atoms in the protein except hydrogen. When the query is longer than 250 residues fold predictions by XStout are returned instead of full atom models by 3Distill. An average sized protein takes less than an hour to predict and no user expertise or intervention is required. 3Distill is free for academic use.

3.2.2 Other 3D Structure Prediction Servers

The Critical Assessment of Techniques for Protein Structure Prediction experiment (CASP) evaluates the current state of the art in protein structure prediction ***(24)***. There have been eight experiments to date taking place every two years since 1994. Participants predict the 3D structure, and other structural features, of a set of soon to be known structures, these predictions are then assessed by a panel of experts when the structures are known. Fully automated prediction, by servers, has played an increasingly important role at CASP. Although most protein structure predictions are automated in some way many still require human intervention by experts to get the most accurate results. Fully automated processes have the advantage of being available to the non-expert user and, in general being faster than human approaches, may be used on a genomic scale, something that is more of a requirement these days than just being desirable (*See* **Note 1**). The accuracy of server predictions has significantly increased over the last number of years with servers being ranked in the top five overall in CASP7 and CASP8. Some of the servers that have performed best at

CASP are described below. For a detailed comparison and in-depth discussion of all methods that participated in CASP8 see the special edition of the journal "Proteins: Structure, Function, and Bioinformatics" ***(24)*** and look out for the results of CASP9 which took place in 2010.

I-TASSER ***(25)*** was ranked first in the server category of the CASP8 experiment. It is free for academic use, no expert knowledge is required and prediction from a protein sequence takes in the region of 24-48 hours for full 3D structure and function prediction. The I-TASSER pipeline includes four general steps: template identification; structure reassembly; atomic model construction; and final model selection. In cases where no appropriate template is identified the whole structure is predicted by *ab initio*. The success of I-TASSER is primarily due to the use of information from multiple templates.

HHpred ***(26)*** is primarily an interactive function and structure prediction server. For example, the user can search various databases, manually select templates or correct errors in the proposed target-template alignment. The prediction pipeline is as follows: build a multiple sequence alignment for the target sequence; search for homologous templates; re-rank the potential templates with a neural network; generate sets of multiple alignments with successively lower sequence diversities for the target sequence and the templates; rank target-template alignments of various alignment diversities with neural network; choose template(s); and run MODELLER ***(27)***. Some user expertise in the area of alignment/template selection is useful as users have the option to intervene at this step before the 3D model is built. Predictions are fast, talking less than an hour for a protein of average size.

David Baker's **Robetta** ***(28)*** is one of the best known, consistently most accurate and most popular of all protein structure prediction servers. The server parses protein chains into putative domains and predicts these domains either *ab initio* or by homology modelling. However, the popularity of the server and computational requirements result in long waiting times before the prediction process even starts, and public users are restricted to submitting one protein

sequence at a time.

3.3 Functional Site Prediction for Structured Proteins

Predicting functionally important amino acids or active sites of proteins is a good starting point for structure-based function prediction. Most predictors use sequence conservation as an indication of functional importance with some newer predictors incorporating structural information. **SDPsite** ***(29)*** predicts functional sites using conserved positions and specificity-determining positions (SDP residues which are conserved within sub-groups of a protein family but differ between groups). The server takes as input a multiple sequence alignment and a phylogenetic tree of the proteins in the alignment.

The **ConSurf** Server ***(30)*** takes as input a protein sequence, multiple sequence alignment or PDB file. The PDB file can be uploaded, in which case the functional site of a predicted protein model can be predicted, or if the structure is known the PDB ID can be entered. If the input is a protein sequence or multiple sequence alignment the output includes a sequence/multiple sequence alignment coloured according to the conservation scores and a phylogenetic tree. If a PDB structure is provided the output is a PDB file with the predicted functionally important residues highlighted. ConSurf is free for academic use, easy to use, fast and requires no expert knowledge.

Evolutionary Trace ***(31)*** captures the extent of evolutionary pressure at a given position in a protein sequence and ranks the amino acids by their relative evolutionary importance. There are two tools available: the ET Viewer which takes a PDB ID as input and displays a colour map of the structure showing the ranked residues; and the ET Report Maker which takes either a PDB ID or UniProt accession number as input and returns a detailed report which includes information about protein sequence, structure, suggested mutations and substitutions for selective functional site knock out. The Evolutionary Trace Server is free for academic use.

SITEHOUND ***(32)*** takes as input a protein structure in PDB format and identifies regions

corresponding to putative ligand binding sites. These sites are characterised by favourable noncovalent interactions with a chemical probe. The selection of different chemical probes results in the identification of different types of binding site. Currently, carbon and phosphate probes are available to identify binding sites for drug like molecules and phosphorylated ligands respectively. The output is a list of residues which correspond to the putative binding sites.

3.4 Disorder Prediction

Many proteins or protein regions fail to fold into fixed tertiary structures. Over the last ten years these Intrinsically Unstructured (IU)/Disordered proteins have been shown to be important functionally leading to an alternative view of protein function to the traditional sequence-structure-function paradigm ***(33)***. **Spritz** ***(13)*** is a web server for the prediction of intrinsically disordered regions in protein sequences. Spritz is available as part of the Distill suite of servers described above and predicts ordered/disordered residues using two specialised binary classifiers both implemented with probabilistic soft-margin support vector machines or C-SVM. The SVM-LD (LD: long disorder) classifier is trained on a subset of non redundant sequences known to contain only long disordered protein fragments (>=30*AA*). The SVM-SD (SD: short disorder) classifier is trained instead on a subset of non redundant sequences with only short disordered fragments.

The **IUPred** server ***(34)*** predicts disorder based on the difference between estimates of the pairwise energy content for globular proteins which have the potential to form a large number of favourable interactions compared with disordered proteins which do not form sufficiently favourable interactions to adopt a stable structure due to their amino acid composition. For a comprehensive list of other disorder predictors see: http://www.disprot.org/predictors.php.

3.5 Short Linear Motifs (SLiMs)

Short linear motifs (SLiMs) are abundant protein microdomains that play a central role in cell

regulation. SLiMs, also referred to as linear motifs, minimotifs or Eukaryotic Linear Motifs (ELMs, in eukaryotes) typically act as protein ligands and mediate many biological processes including cell signalling, post-translational modification (PTM) and trafficking target proteins to specific subcellular localisations (numerous excellent reviews of motif biology are available ***(35-37)***). Several organizations, such as the Eukaryotic Linear Motif resource (ELM) ***(38, 39)*** and Minimotif Miner (MnM) ***(40, 41)*** are actively curating the available SLiM literature and currently 200 classes of motifs are known, yet without a doubt many more remain to be discovered. SLiMs are defined by a conduciveness to convergently evolve, their preferential occurrence in disordered regions and their short length. Each of these attributes contributes to the difficulty of motif discovery, both experimentally and computationally, however despite the challenges several useful motif discovery tools are available.

3.5.1 Motif Rediscovery

The ELM Server ***(38, 39)*** searches the ELM database for regular expression matches and discover putatively functional novel instances of known SLiMs. Returned motifs are filtered to exclude motifs occurring in globular regions of proteins using information from Pfam ***(42, 43)***, SMART ***(44)*** and the PDB when available. **Minimotif Miner** ***(40, 41)*** searches an input protein for matches to the MnM dataset, scoring motifs based on surface accessibility, conservation and fold enrichment (based on the ratio of observed motifs to expected motifs). **SIRW** is a web-server that calculates motif enrichment, using the Fisher's exact test, in a set of proteins with a particular keyword or Gene Ontology (GO) ***(45)*** terms. Similarly, **SLiMSearch** uses the masking and statistical methods of the SLiMfinder tool ***(46)*** to search for motifs in an input dataset.

3.5.2 *De-Novo* Motif Discovery

Dilimot ***(47)*** and **SLiMfinder** ***(46)***, using motif over-representation, attempt *de novo* computational discovery of SLiMs in datasets of proteins. Dilimot masks globular regions and

enriches for convergently evolved motifs by removing all but one representative homologous region. Returned motifs are scored using a binomial scoring scheme. Finally, conservation of the motif in several species is incorporated into a final combined score. SLiMfinder excludes under-conserved residues, non-disordered regions predicted using IUPred ***(34)*** and UniProt ***(1)*** annotated features such as domains. Motifs are scored using an extension of binomial statistics allowing the consideration of homologous motif instances and correction for multiple testing. **ANCHOR** ***(48)*** attempts the difficult task of *de novo* motif discovery from primary sequence by predicting disordered binding regions. These are regions that undergo a disorder-to-order transition on binding to a structured partner. ANCHOR uses the same pairwise energy estimation approach as IUPred to identify protein segments that reside in disordered regions but are unable to form enough favourable intra-chain interactions to fold on their own and therefore are likely to require an interaction with a globular protein partner to gain stabilizing energy.

3.5.3 Post-Processing

After discovery of a novel motif there are multiple steps that that can help increase confidence of functionality. **CompariMotif** ***(49)*** searches for matches to known functional motifs. Novel motifs are compared against motif databases using shared Information Content, allowing the best matches to be easily identified in large comparisons. Currently, the ELM ***(38, 39)*** and MnM ***(40, 41)*** databases, as well as several other specialized datasets are available to search. Conservation is one of the strongest classifiers of novel motif functionality and several tools are available to score the conservation of motif occurrences. For example, **Conscore** ***(50)*** uses an information content based scoring scheme which incorporates phylogeny information to weight sequences and Dinkel *et al* ***(51)*** introduced an Average Conservation Score. **PepSite** ***(52)*** can be used to scan known interactors for binding sites for a discovered motif. Using spatial position-specific scoring matrices (PSSMs) created from known 3D structures of motif/protein complexes Pepsite scores the surface of target and suggests potential binding site and rough orientation of

the motif.

3.5.3 Biological Uses

Several examples of experimentally validated motifs discovered by *in silico* methods are available. Neduva *et al* ***(47)*** applied Dilimot to discover and verify a protein phosphatase 1 binding motif (DxxDxxxD) and a motif that binds Translin (VxxxRxYS). Keyword enrichment has been used to discover novel KEN box ***(53)***, KEPE ***(54)*** and EH1 motifs ***(55)***. Two 14-3-3 motifs in EFF-1 were discovered using MnM and subsequently experimentally validated ***(40, 41)***. For more on SLiM discovery see the review by Davey *et al* ***(56)***.

4. Notes

1. The number of protein sequences is growing at an ever increasing pace and many *in silico* methods are available for the efficient annotation of these sequences. Given that user queries vary greatly in size, scope and character, when choosing which methods to adopt the speed, accuracy and scale of the method need to be considered. As a first approximation, especially in the case of structure prediction, the greater the accuracy, the slower the processing time will be. The larger the scale of a query (e.g. when genomic-scale predictions are necessary), the harder it will be to obtain the most accurate answers available, unless one has access to a large amount of computational resources and has the time to download and set up one of the methods that are available for local installation.
2. When deciding which prediction method to use, the main consideration to make is whether there is a homologue for the query in the PDB. If so, normally, methods incorporating homology are significantly more accurate. Another consideration is the scale of predictions to be performed. All servers handle predictions on a small scale (tens of queries), some (e.g. Distill) facilitate predictions on a larger scale (hundreds of queries). If genomic or especially multi-genomic scale predictions are needed it may be necessary to resort to one of the methods that

can be downloaded and run locally. In all cases, when possible, consensus predictions are desirable, that is, polling multiple methods for the same query and comparing the results. Where methods agree, generally predictions are to be considered more reliable.

3. Greater confidence can be placed in the accuracy of structure predictions if there is high sequence similarity between the query sequence and a protein of known structure which can act as a template. However, it is worth remembering that even with little or no sequence similarity, proteins may share the same structure and therefore a low sequence identity template does not imply that the prediction is inaccurate. In Distill we find templates via the SAMD program ***(17)***, which may yield informative templates even for very low sequence identity.

4. During our research experiences we have often encountered resistance from some experimental researchers in exploiting all the computational tools that are available to simplify their jobs. Even in the absence of resistance, countless times we have observed ageing, outdated tools being adopted when far better ones were freely available and ready to use. Within the limits of this chapter, we hope we have made a small step towards solving this problem and bringing the power of novel predicting methods to its full fruition.

Acknowledgements

CM is supported by Science Foundation Ireland (SFI) grant 08/IN.1/B1844. ND is supported by an EMBL Interdisciplinary Postdoc (EIPOD) fellowship. CM, GP, IW and AJMM were partly supported by SFI grant 05/RFP/CMS0029, grant RP/2005/219 from the Health Research Board of Ireland, a UCD President's Award 2004 and UCD Seed Funding 2009 award SF371.

References

1. The UniProt Consortium (2008) The Universal Protein Resource (UniProt). Nucleic Acids Res **36**, D190-D195.

2. Berman H, Westbrook J, Feng Z, Gilliland G, Bhat T, Weissig H, Shindyalov I, and Bourne P

(2000) The Protein Data Bank. Nucleic Acids Res **28**, 235-242.

3. Aloy P, Pichaud M, and Russell R (2005) Protein complexes: structure prediction challenges for the 21st century. Curr Opin Struct Biol **15**, 15-22.

4. Chothia C and Lesk A (1986) The relation between the divergence of sequence and structure in proteins. EMBO J **5**, 823-826.

5. Chandonia J and Brenner S (2006) The impact of structural genomics: expectations and outcomes. Science **311**, 347.

6. Moult J (2008) Comparative modeling in structural genomics. Structure **16**, 14-16.

7. Altschul S, Madden T, Schaffer A, Zhang J, Zhang Z, Miller W, and Lipman D (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res **25**, 3389.

8. Baù D, Martin A, Mooney, C Vullo A, Walsh I, and Pollastri G (2006) Distill: a suite of web servers for the prediction of one-, two- and three-dimensional structural features of proteins. BMC Bioinformatics **7**, 402.

9. Pollastri G and McLysaght A (2005) Porter: a new, accurate server for protein secondary structure prediction. Bioinformatics **21**, 1719-20.

10. Vullo A, Walsh I, and Pollastri G (2006) A two-stage approach for improved prediction of residue contact maps. BMC Bioinformatics **7**, 180.

11. Mooney C, Vullo A, and Pollastri G (2006) Protein structural motif prediction in multidimensional phi-psi space leads to improved secondary structure prediction. J Comput Biol **13**, 1489-1502.

12. Pollastri G, Martin A, Mooney C, and Vullo A (2007) Accurate prediction of protein secondary structure and solvent accessibility by consensus combiners of sequence and structure information. BMC Bioinformatics **8**, 201.

13. Vullo A, Bortolami O, Pollastri G, and Tosatto S (2006) Spritz: a server for the prediction of intrinsically disordered regions in protein sequences using kernel machines. Nucleic Acids Res

34, W164.

14. Walsh I, Martin A, Mooney C, Rubagotti E, Vullo A, and Pollastri G (2009) Ab initio and homology based prediction of protein domains by recursive neural networks. BMC Bioinformatics **10**, 195.

15. Walsh I, Baù D, Martin A, Mooney C, Vullo A, and Pollastri G (2009) Ab initio and template-based prediction of multi-class distance maps by two-dimensional recursive neural networks. BMC Struct Biol **9**, 5.

16. Sims G, Choi I, and Kim S (2005) Protein conformational space in higher order ψ-ϕ maps. Proc Natl Acad Sci USA **18**, 618-621.

17. Mooney C and Pollastri G (2009) Beyond the Twilight Zone: automated prediction of structural properties of proteins by recursive neural networks and remote homology information. Proteins **77**, 181-190.

18. Suzek B, Huang H, McGarvey P, Mazumder R, and Wu C (2007) UniRef: comprehensive and non-redundant UniProt reference clusters. Bioinformatics **23**, 1282.

19. Montgomerie S, Sundararaj S, Gallin W, and Wishart D (2006) Improving the accuracy of protein secondary structure prediction using structural alignment. BMC Bioinformatics 7, 301. 15

20. Cheng J, Randall A, Sweredoski M, and Baldi P (2005) SCRATCH: a protein structure and structural feature prediction server. Nucleic Acids Res **33**, W72.

21. Cole C, Barber J, and Barton G (2008) The Jpred 3 secondary structure prediction server. Nucleic Acids Res **36**, W197-W201.

22. Jones D (1999) Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol **292**, 195-202.

23. Adamczak R, Porollo A, and Meller J (2005) Combining prediction of secondary structure and solvent accessibility in proteins. Proteins **59**, 467-475.

24. Moult J, Fidelis K, Kryshtafovych A, Rost B, and Tramontano A (2009) Critical assessment of methods of protein structure prediction - Round VIII. Proteins **77**, 1-4.

25. Zhang Y (2009) I-TASSER: Fully automated protein structure prediction in CASP8. Proteins **77**, 100.

26. Hildebrand A, Remmert M, Biegert A, and Söding J (2009) Fast and accurate automatic structure prediction with HHpred. Proteins **77**, 128-132.

27. Eswar N, Webb B, Marti-Renom M, Madhusudhan M, Eramian D, Shen M, Pieper U, and Sali A (2007) Comparative protein structure modeling using Modeller. Curr Protoc Protein Sci **50**:2.9.1-2.9.31.

28. Raman S, Vernon R, Thompson J, Tyka M, Sadreyev R, Pei J, Kim D, Kellogg E, DiMaio F, Lange O, et al. (2009) Structure prediction for CASP8 with all-atom refinement using Rosetta. Proteins **77**, 89-99.

29. Kalinina O, Gelfand M, and Russell R (2009) Combining specificity determining and conserved residues improves functional site prediction. BMC Bioinformatics **10**, 174.

30. Landau M, Mayrose I, Rosenberg Y, Glaser F, Martz E, Pupko T, and Ben-Tal N (2005) ConSurf 2005: the projection of evolutionary conservation scores of residues on protein structures. Nucleic Acids Res **33**, W299.

31. Morgan D, Kristensen D, Mittelman D, and Lichtarge O (2006) ET viewer: an application for predicting and visualizing functional sites in protein structures. Bioinformatics **22**, 2049.

32. Hernandez M, Ghersi D, and Sanchez R (2009) SITEHOUND-web: a server for ligand binding site identification in protein structures. Nucleic Acids Res **37**, W413-W416.

33. Dyson H and Wright P (2005) Intrinsically unstructured proteins and their functions. Nat Rev Mol Cell Bio **6**, 197-208.

34. Dosztanyi Z, Csizmok V, Tompa P, and Simon I (2005) IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. Bioinformatics **21**, 3433.

35. Diella F, Haslam N, Chica C, Budd A, Michael S, Brown N, Travé G, and Gibson T (2008) Understanding eukaryotic linear motifs and their role in cell signaling and regulation. Front Biosci

13, 6580-6603.

36. Neduva V and Russell R (2006) Peptides mediating interaction networks: new leads at last. Curr Opin Biotech **17**, 465-471.

37. Neduva V and Russell R (2005) Linear motifs: evolutionary interaction switches. FEBS Lett **579**, 3342-3345.

38. Puntervoll P, Linding R, Gemund C, Chabanis-Davidson S, Mattingsdal M, Cameron S, Martin D, Ausiello G, Brannetti B, Costantini A, et al. (2003) ELM server: a new resource for investigating short functional sites in modular eukaryotic proteins. Nucleic Acids Res 31, 3625.

39. Gould C, Diella F, Via A, Puntervoll P, Gemund C, Chabanis-Davidson S, Michael S, Sayadi A, Bryne J, Chica C, et al. (2010) ELM: the status of the 2010 eukaryotic linear motif resource. Nucleic Acids Res **38**, D167.

40. Balla S, Thapar V, Verma S, Luong T, Faghri T, Huang C, Rajasekaran S, del Campo J, Shinn J, Mohler W, et al. (2006) Minimotif Miner: a tool for investigating protein function. Nature Methods **3**, 175-177.

41. Rajasekaran S, Balla S, Gradie P, Gryk M, Kadaveru K, Kundeti V, Maciejewski M, Mi T, Rubino N, Vyas J, et al. (2009) Minimotif miner 2nd release: a database and web system for motif search. Nucleic Acids Res **37**, D185.

42. Bateman A, Birney E, Cerruti L, Durbin R, Etwiller L, Eddy S, Griffiths-Jones S, Howe K, Marshall M, and Sonnhammer E (2002) The Pfam protein families database. Nucleic Acids Res **30**, 276.

43. Finn R, Mistry J, Tate J, Coggill P, Heger A, Pollington J, Gavin O, Gunasekaran P, Ceric G, Forslund K, et al. (2009) The Pfam protein families database. Nucleic Acids Res **36**, 281-288.

44. Letunic I, Doerks T, and Bork P (2008) SMART 6: recent updates and new developments. Nucleic Acids Res **1**, 4.

45. Ashburner M, Ball C, Blake J, Botstein D, Butler H, Cherry J, Davis A, Dolinski K, Dwight S, Eppig J, et al. (2000) Gene Ontology: tool for the unification of biology. Nature genetics **25**, 25-

29.

46. Edwards R, Davey N, and Shields D (2007) SLiMFinder: a probabilistic method for identifying over-represented, convergently evolved, short linear motifs in proteins. PloS One **2**, e967.

47. Neduva V, Linding R, Su-Angrand I, Stark A, De Masi F, Gibson T, Lewis J, Serrano L, and Russell R (2005) Systematic discovery of new recognition peptides mediating protein interaction networks. PLoS Biology **3**, 2090.

48. Mészáros B, Simon I, and Dosztányi Z (2009) Prediction of protein binding regions in disordered proteins. PLoS Comput Biol **5**, 5.

49. Edwards R, Davey N, and Shields D (2008) CompariMotif: quick and easy comparisons of sequence motifs. Bioinformatics **24**, 1307.

50. Chica C, Labarga A, Gould C, López R, and Gibson T (2008) A tree-based conservation scoring method for short linear motifs in multiple alignments of protein sequences. BMC Bioinformatics **9**, 229.

51. Dinkel H and Sticht H (2007) A computational strategy for the prediction of functional linear peptide motifs in proteins. Bioinformatics **23**, 3297. 17

52. Petsalaki E, Stark A, García-Urdiales E, and Russell R (2009) Accurate prediction of peptide binding sites on protein surfaces. PLoS Comput Biol **5**, e1000335.

53. Michael S, Trave G, Ramu C, Chica C, and Gibson T (2008) Discovery of candidate KEN-box motifs using cell cycle keyword enrichment combined with native disorder prediction and motif conservation. Bioinformatics **24**, 453.

54. Diella F, Chabanis S, Luck K, Chica C, Ramu C, Nerlov C, and Gibson T (2009) KEPE-a motif frequently superimposed on sumoylation sites in metazoan chromatin proteins and transcription factors. Bioinformatics **25**, 1.

55. Copley R (2005) The EH 1 motif in metazoan transcription factors. BMC Genomics 6, 169.

56. Davey N, Edwards R, and Shields D (2010) Computational identification and analysis of

protein short linear motifs. Front Biosci **15**, 801-825.