PconsFam: An Interactive Database of Structure Predictions of Pfam Families

John Lamb [1,†], Aleksandra I. Jarmolinska [2,†], Mirco Michel [1], David Menéndez-Hurtado [1], Joanna I. Sulkowska [2,‡] and Arne Elofsson [1 ‡]

*1 - **Science for Life Laboratory and Department of Biochemistry and biophysics,** Stockholm University, Tomtebodav 23, 171 21 Solna, Sweden*
*2 - **Centre of New Technologies,** University of Warsaw, Banacha 2c, 02097 Warsaw, Poland*

***Correspondence to Arne Elofsson:** arne@bioinfo.se.*
https://doi.org/10.1016/j.jmb.2019.01.047
Edited by Michael Sternberg

Abstract

At present, about half of the protein domain families lack a structural representative. However, in the last decade, predicting contact maps and using these to model the tertiary structure for these protein families have become an alternative approach to gain structural insight. At present, reliable models for several hundreds of protein families have been created using this approach. To increase the use of this approach, we present PconsFam, which is an intuitive and interactive database for predicted contact maps and tertiary structure models of the entire Pfam database. By modeling all possible families, both with and without a representative structure, using the PconsFold2 pipeline, and running quality assessment estimator on the models, we predict an estimation for how confident the contact maps and structures are for each family.

Introduction

In recent years, it has been shown that with the use of evolutionary information and direct coupling analysis [1], it is possible to obtain sufficiently accurate contact prediction of proteins from their sequence and multiple-sequence alignment alone to predict accurate structures of many protein families [2]. At first, the direct coupling analysis methods were limited to very large protein families, but with the use of deep learning methodologies to improve the contact predictions, it is now possible to accurately predict the contacts for families with only a few hundred members [3–5].

Using these predicted contacts, it is then possible to model the structure of a protein using a protein folding program. In PconsFold, both Rosetta [6] and the CNS suite (Crystallography & NMR System) [7] have been used, with CNS being faster by an order of magnitude while only producing models of slightly less quality than Rosetta [8].

In 2017, we developed the PconsFold2 pipeline [9], which uses contact predictions from PconsC3 [4], the CNS-based CONFOLD folding algorithm [10] and most importantly multiple model quality estimations [11,12] to predict the structure of proteins. Here, we present the related web resource PconsFam (http://pconsfam.bioinfo.se), a database with predicted structural information for all Pfam families that could successfully be modeled using the PconsFold2 pipeline. Some families are excluded due to short length (<50) or that no contact maps of sufficient quality could be generated. The dataset is summarized in Table 1.

The PconsFold2 pipeline can predict accurate models [template modeling score (TM score) >0.5] for 51% of the large families (>1000 effective sequences); for smaller families, the fraction of correct models decreases, but they still exist. Therefore, a major challenge for large-scale predictions is to distinguish between correct and incorrect models. Here, we have applied a set of model quality estimation methods [15]. These methods will be discussed further in the Materials and Methods section.

When the PconsFold2 pipeline was applied to 6379 Pfam families of unknown structure, 558 models with a predicted specificity of 90% were created [9]. Out of these 558, 415 had never been reported before.

Table 1. Average and median of different quality metrics for Pfam families with and without known structure

Data set	M_{eff}	PPV	TM	FDR	Pcons	ProQ3
Average values						
All	2286	–	–	0.37	0.22	0.35
Structure	2232	0.45	0.42	0.29	0.28	0.42
No structure	2331	–	–	0.45	0.17	0.30
Median values						
All	200	–	–	0.40	0.19	0.35
Structure	486	0.46	0.41	0.28	0.25	0.45
No structure	111	–	–	0.48	0.16	0.27

Effective number of sequences (M_{eff}) defined by Ekeberg *et al.* [13]; positive predicted value (PPV) defined over 2.5 times sequence length (L) best-ranked contacts; TM score as defined by Zhang *et al.* [14]; false discovery rate (FDR); Pcons score [12], which scores by looking at recurring structures when comparing multiple models; and ProQ3 score as defined by Uziela *et al.* [11].

For each family in Pfam, multiple contact maps are generated, and from these, a set of models are predicted and ranked with quality assessment estimators. The model quality estimation gives an indication of the reliability of the model. The top ranked models for each predicted family can be visualized and are available for download. The full set of contact predictions is available for visualization together with the predicted model in an intuitive and powerful user interface that allows for interaction between the contact maps and the predicted structure.

Materials and Methods

PconsFold2

PconsFold2 works in three separate steps (see Fig. 1b). First, multiple-sequence alignments are generated using HHblits and Jackhmmer [16] at different *E*-value cutoffs. Second, PconsC3 is used for contact predictions where each alignment is used to create a distinct contact map. Third, these contact maps, together with predicted secondary structure, are used as input to CONFOLD to generate 50 models for each contact map, resulting in a total of 200 models.

Model quality assessment

Model quality assessment estimators are run on predicted models to give a value of how reliable the models are. This can be seen as a measure of how confident we are that the predicted model is representative of the real structure. The final step in PconsFold2, CONFOLD, ranks the predicted models by its own measure, CNS contact energy (NOE), which is the sum of all violations of the contact restraints from the contact maps used as input. This score is normalized by protein length to be comparable between proteins. Additional quality assessment methods that are used are Pcons [12], which compare recurring structures between multiple models, ProQ3D [11], that uses a deep neural network, and in the cases where we have a known structure the PPV. PPV is defined as true positives divided by the sum of true positives and false positives (see below) of the top 2.5 times sequence length contacts in the predicted contact map when compared to the true contacts as per the known structure. In this context, it is how many of the contacts that we predict are real contacts according to the known structure.

$$\mathrm{PPV} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

A linear combination of these three scores, as defined below, is used to create the fifth model quality assessment score, PcombC. The coefficients were selected by a grid search over a subset of the families with a known structure as defined in Ref. [9].

$$S_{\mathrm{PcombC}} = \frac{0.3}{1.9} \cdot S_{\mathrm{Pcons}} + \frac{0.6}{1.9} \cdot S_{\mathrm{ProQ3D}} + \frac{1.0}{1.9} \cdot \mathrm{PPV}$$

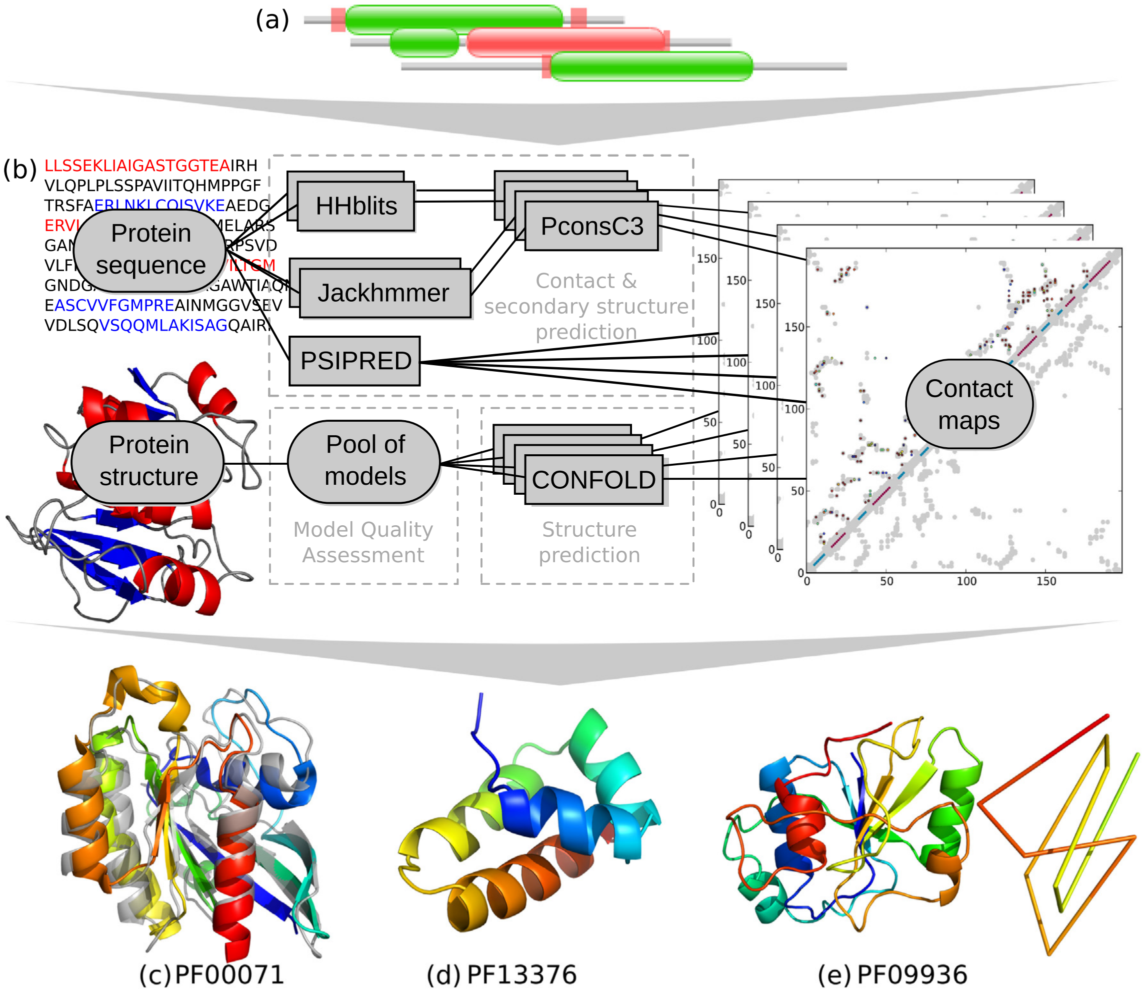
This was used on all families with a known structure to rank the best model. For families without a known structure, the union of FDR, Pcons and ProQ3 was used. Based on the ROC analysis done on a test set of families of known structure, see Ref. [9], a score cutoff of FPR 0.01 and 0.1 was set, and these cutoffs were then used to estimate how many unknown structures that could be predicted accurately (TM score $\geq$0.5). It is therefore enough that one of these methods is above the chosen cutoff to be considered correct.

Topology

The knotted topology was established using an implementation of the Alexander polynomial described in the KnotProt database [17]. To form a closed chain we used random closure method: two random points on a large sphere ae randomly chosen and connected by line segments to the endpoints of a chain, and to each other by the an (auxiliary) arc. The closed backbone (chain) is then reduced using a KMT algorithm, and the Alexander polynomial is computed. The structure is called knotted if the same type of knot appears in more than 40% of cases.

User interface

An overview of the interface can be seen in Fig. 2. For each family, contacts predicted by the workflow



(c) PF00071 (d) PF13376 (e) PF09936

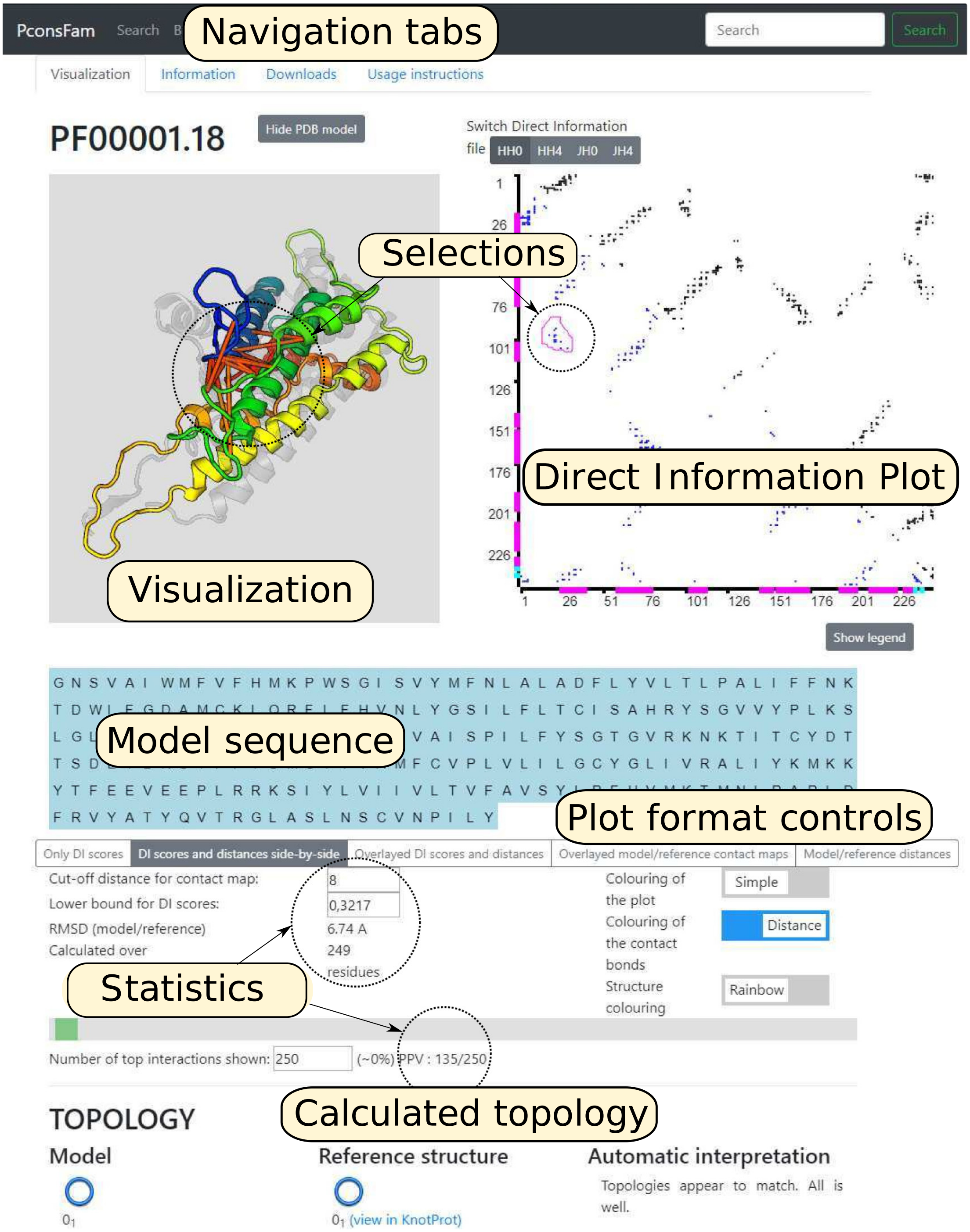
Fig. 1. PconsFold2 pipeline. All Pfam families (a) are used as input for the PconsFold2 pipeline (b), this results in models for both families with known structure (c) and without known structure (d). In addition, knotted topology (e) is also predicted.

are plotted as a heatmap of values between different residues. Automatically assigned secondary structure elements are marked along the axes. Heatmap values can be either presented as binary, or through a rainbow color map. Additional plot modes include a side-by-side comparison with model contact map (with a user specified C-β distance between amino acids) and an overlay of predicted and observed contacts. Predicted contacts can be visualized on the structure, as colored bonds (with different styles available). The user can highlight some regions on the plot by selecting them either on the structure or on the sequence. When a reference structure for a given family is available, it is visualized in gray and superposed on the model structure. RMSD between matching parts of the structures is calculated and displayed below. When a reference protein is present, two additional plot modes are available—overlay of contact maps of model and native structure and a side-by-side comparison of their distance maps. For plot modes that include predicted contacts the PPV is calculated and indicates how well the currently used predictions are represented by the structure. Other tabs on the family details page include "Information" (about the family), and "Downloads." More detailed description of the capabilities of the user interface is available under the "Usage Instructions" tab.

Results

Overview

PconsFam contains predictions for 13,617 Pfam families, of which 6492 have a member of known structure. Using the independent model quality



PF00001.18

G N S V A I W M F V F H M K P W S G I S V Y M F N L A L A D F L Y V L T L P A L I F F N K
T D W L E G D A M C K I Q R E I E H V N L Y G S I L F L T C I S A H R Y S G V V Y P L K S
L G L V A I S P I L F Y S G T G V R K N K T I T C Y D T
T S D M F C V P L V L I L G C Y G L I V R A L I Y K M K K
Y T F E E V E E P L R R K S I Y L V I I V L T V F A V S Y
F R V Y A T Y Q V T R G L A S L N S C V N P I L Y

Cut-off distance for contact map: 8
Lower bound for DI scores: 0,3217
RMSD (model/reference) 6.74 A
Calculated over 249 residues

Colouring of the plot: Simple
Colouring of the contact bonds: Distance
Structure colouring: Rainbow

Number of top interactions shown: 250 (~0%) PPV : 135/250

TOPOLOGY

Model: 0_1

Reference structure: 0_1 (view in KnotProt)

Automatic interpretation: Topologies appear to match. All is well.

Fig. 2. User interface detailing results for the PF00001 family. The default Visualization's tab contains structure visualization of the model(s) (superposed with reference structure if available), Direct Information (DI) plot (which can also display contact maps), and the sequence and topology of the models. Range and format of displayed contacts can be changed, and contacts between residues can be visualized as bonds on the structure. RMSD between model and reference, and a PPV score indicating overlap between residues pairs and structural contacts in the model are also shown. Other tabs contain additional information about the family and download links for calculated data.

assessment estimators, Pcons and ProQ3D, it is clear that there is a correlation between model quality and effective sequences (see Figs. S1 and S2). The more effective sequences a family contains, the better the predicted model for that family. In Figs. S3, S4, and S5, examples of proteins are shown.

Comparison with other resources

A similar resource is the GREMLIN database [18], which builds contact prediction based on PFAM 27.0 [19].

In contrast, PconsFam is currently based on Pfam 29.0 and with its modular nature; each of the three steps (generate alignments, generate contact maps, and generate models) can be changed independently to faster/more accurate tools.

GREMLIN shows predicted contact maps with an option to overlay with the pdb structure if one exists. We extend on this by using a tool that can visualize the predicted contacts on the models. In our database, both contact maps and predicted structure can be investigated in detail and downloaded. We have also used a deep learning methodology for contact predictions. In general, this provides a better coverage of small protein families.

Over 7% of proteins deposited in Protein Data Bank (PDB) possess a non-trivial topology [20]. They are knotted [17], slipknotted [21], and linked [22] or contain lassos [23], although for a long time it was believed that proteins should not be entangled [24]. Today, a non-trivial topology such as knotting, which is hard to detect by visual inspection, is easily detected using algorithms [25–28]. The complexity of encountered knots varies from the simplest trefoil knot (3_1) [24] to Stevedore knot (6_1) [29]. Apart from the topology, knotted proteins differ also by the depth of the knot (the lengths of the tails outside the knotted core) and thus are divided into shallow (tails possess less than 10 amino acids) and deeply knotted. However, it is still not clear how proteins can self-tie and how knotted a protein backbone can be [30,31]. Thus, the PconsFam reports also the information about the topology of modeled structures. This information can be used as an additional descriptor to validate structures [32]. Figure 1e shows a correctly predicted structure from PF09936 family with the so-called trefoil knot (a three-crossings knot, shown in a simplified form next to the structure).

Discussion

PconsFam is a novel tool and an informative interface that enables the examination of multiple models and contacts maps of Pfam families of both known and unknown structure. To allow this comparison, it uses the same pipeline to predict both contact maps and models for both groups. ProQ3D [11] and Pcons [12] are used as quality estimation methods for all families to evaluate FDR. In addition, PPV and the linear combination PcombC are used to rank models for families of known structure. PconsFam is a complement to existing resources and aims to provide an easy accessibility to contact maps and predicted models of the Pfam database.

Modularity of pipeline

As PconsFold2 is highly modular, each of the three steps in the pipeline can be changed independently. Any alignment tool can be used to generate the alignments, and any contact prediction tool can be used. This opens up the possibility to run different tools for different data sets where there are known tools that work better for specific data.

Future Directions

Currently, the input to CONFOLD is both predicted contact maps from PconsC3 and predicted secondary structure from PSIPRED. We are currently developing a new predictor, PconsC4 [33], which uses a deep learning approach to predict both contact maps and secondary structure without the dependencies in PconsC3 and PSIPRED. In addition to performing better than PconsC3, it is also faster by orders of magnitudes [33]. This would both speed up the pipeline and increase the quality of contact maps.

Conclusions

Here, we present an intuitive and interactive web interface for contact maps and models for Pfam families. We have used the modular PconsFold2 pipeline to predict both multiple contact maps and multiple models for all eligible families. All these are presented visually and interactively, together with quality assessment scores to highlight the confidence in both contact maps and models.

Acknowledgment

This work was supported by grants from the Swedish Research Council (VR-NT 2016-03798 to A.E.), National Science Centre (No. 2012/07/E/NZ1/01900 to J.I.S. and No. 2018/29/N/NZ2/02897 to A.I.J.).

Conflict of Interest Statement: None declared.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jmb.2019.01.047.

Received 13 December 2018;
Received in revised form 30 January 2019;
Accepted 31 January 2019
Available online 21 February 2019

Keywords:
contact maps;
structure prediction;
folding pipeline;
Pfam

†Contributed equally.

‡Contributing authors.

Abbreviations used:
CNS suite, Crystallography & NMR System; TM score, template modeling score; FDR, false discovery rate; PPV, positive predicted value.

References

[1] M. Weigt, R. White, H. Szurmant, J. Hoch, T. Hwa, Identification of direct residue contacts in protein–protein interaction by message passing, Proc. Natl. Acad. Sci. U. S. A. 106 (1) (2009) 67–72.

[2] J.I. Sułkowska, F. Morcos, M. Weigt, T. Hwa, J.N. Onuchic, Genomics-aided structure prediction, Proc. Natl. Acad. Sci. 109 (26) (2012) 10340–10345, https://doi.org/10.1073/pnas.1207864109 (arXiv:http://www.pnas.org/content/109/26/10340.full.pdf, URL http://www.pnas.org/content/109/26/10340.abstract).

[3] M. Skwark, D. Raimondi, M. Michel, A. Elofsson, Improved contact predictions using the recognition of protein like contact patterns, PLoS Comput. Biol. 10 (11) (2014), e1003889. https://doi.org/10.1371/journal.pcbi.1003889.

[4] M. Michel, M. Skwark, D. Menendez Hurtado, M. Ekeberg, A. Elofsson, Predicting accurate contacts in thousands of pfam domain families using pconsc3, Bioinformatics 33 (18) (2017) 2859–2866, https://doi.org/10.1093/bioinformatics/btx332.

[5] S. Wang, S. Sun, Z. Li, R. Zhang, J. Xu, Accurate de novo prediction of protein contact map by ultra-deep learning model, PLoS Comput. Biol. 13 (1) (2017), e1005324. https://doi.org/10.1371/journal.pcbi.1005324.

[6] K. Simons, R. Bonneau, I. Ruczinski, D. Baker, Ab initio protein structure predictions of CASP III targets using ROSETTA, Proteins Struct. Funct. Genet. (Suppl. 3) (1999) 171–176.

[7] F. Morcos, A. Pagnani, B. Lunt, A. Bertolino, D. Marks, C. Sander, R. Zecchina, J. Onuchic, T. Hwa, M. Weigt, Direct-coupling analysis of residue coevolution captures native contacts across many protein families, Proc. Natl. Acad. Sci. U. S. A. 108 (49) (2011) 1293–1301, https://doi.org/10.1073/pnas.1111471108.

[8] M. Michel, S. Hayat, M. Skwark, C. Sander, D. Marks, A. Elofsson, Pconsfold: improved contact predictions improve protein models, Bioinformatics 30 (17) (2014) i482–i488, https://doi.org/10.1093/bioinformatics/btu458.

[9] M. Michel, D. Menendez Hurtado, K. Uziela, A. Elofsson, Large-scale structure prediction by improved contact predictions and model quality assessment, Bioinformatics 33 (14) (2017) i23–i29, https://doi.org/10.1093/bioinformatics/btx239.

[10] B. Adhikari, D. Bhattacharya, R. Cao, J. Cheng, CONFOLD: residue–residue contact-guided ab initio protein folding, Proteins 83 (8) (2015) 1436–1449, https://doi.org/10.1002/prot.24829.

[11] K. Uziela, D. Menendez Hurtado, N. Shu, B. Wallner, A. Elofsson, Proq3d: improved model quality assessments using deep learning, Bioinformatics (2017). https://doi.org/10.1093/bioinformatics/btw819.

[12] J. Lundström, L. Rychlewski, J. Bujnicki, A. Elofsson, Pcons: a neural network based consensus predictor that improves fold recognition, Protein Sci. 10 (11) (2001) 2354–2365.

[13] M. Ekeberg, C. Lovkvist, Y. Lan, M. Weigt, E. Aurell, Improved contact prediction in proteins: using pseudolikelihoods to infer Potts models, Phys. Rev. E Stat. Nonlinear Soft Matter Phys. 87 (1-1) (2013), 012707.

[14] Y. Zhang, M. Devries, J. Skolnick, Structure modeling of all identified g protein-coupled receptors in the human genome, PLoS Comput. Biol. 2 (2) (2006), e13. https://doi.org/10.1371/journal.pcbi.0020013.

[15] A. Kryshtafovych, B. Monastyrskyy, K. Fidelis, T. Schwede, A. Tramontano, Assessment of model accuracy estimations in CASP12, Proteins 86 (Suppl. 1) (2018) 345–360, https://doi.org/10.1002/prot.25371.

[16] S. Eddy, Accelerated profile HMM searches, PLoS Comput. Biol. 7 (10) (2011), e1002195.

[17] M. Jamroz, W. Niemyska, E.J. Rawdon, A. Stasiak, K.C. Millett, P. Sułkowski, J.I. Sulkowska, Knotprot: a database of proteins with knots and slipknots, Nucleic Acids Res. 43 (D1) (2014) D306–D314, https://doi.org/10.1093/nar/gku1059.

[18] H. Kamisetty, S. Ovchinnikov, D. Baker, Assessing the utility of coevolution-based residue–residue contact predictions in a sequence- and structure-rich era, Proc. Natl. Acad. Sci. 110 (39) (2013) 15674–15679 (arXiv:http://www.pnas.org/content/110/39/15674.full.pdf).

[19] E. Sonnhammer, S. Eddy, R. Durbin, Pfam: a comprehensive database of protein domain families based on seed alignments, Proteins Struct. Funct. Genet. 28 (1997) 405–420.

[20] J.I. Sulkowska, P. Sułkowski, Entangled proteins: knots, slipknots, links, and lassos, The Role of Topology in Materials, Springer 2018, pp. 201–226.

[21] N.P. King, E.O. Yeates, T.O. Yeates, Identification of rare slipknots in proteins and their implications for stability and folding, J. Mol. Biol. 373 (1) (2007) 153–166.

[22] P. Dabrowski-Tumanski, A.I. Jarmolinska, W. Niemyska, E.J. Rawdon, K.C. Millett, J.I. Sulkowska, Linkprot: a database collecting information about biological links, Nucleic Acids Res. (2016) D243–D249.

[23] P. Dabrowski-Tumanski, W. Niemyska, P. Pasznik, J.I. Sulkowska, Lassoprot: server to analyze biopolymers with lassos, Nucleic Acids Res. 44 (W1) (2016) W383–W389.

[24] M.L. Mansfield, Are there knots in proteins? Nat. Struct. Mol. Biol. 1 (4) (1994) 213.

[25] W.R. Taylor, A deeply knotted protein structure and how it might fold, Nature 406 (6798) (2000) 916.

[26] P. Virnau, L.A. Mirny, M. Kardar, Intricate knots in proteins: function and evolution, PLoS Comput. Biol. 2 (9) (2006) e122.

[27] R.C. Lua, A.Y. Grosberg, Statistics of knots, geometry of conformations, and evolution of proteins, PLoS Comput. Biol. 2 (5) (2006) e45.

[28] K.C. Millett, E.J. Rawdon, A. Stasiak, J.I. Sułkowska, Identifying knots in proteins, Biochem. Soc. Trans. 41 (2) (2013) 533–537 (arXiv:http://www.biochemsoctrans.org/content/41/2/533.full.pdf).

[29] D. Bölinger, J.I. Sułkowska, H.-P. Hsu, L.A. Mirny, M. Kardar, J.N. Onuchic, P. Virnau, A Stevedore's protein knot, PLoS Comput. Biol. 6 (4) (2010), e1000731.

[30] J.I. Sułkowska, P. Sułkowski, J. Onuchic, Dodging the crisis of folding proteins with knots, Proc. Natl. Acad. Sci. 106 (9) (2009) 3119–3124 (pnas–0811147106).

[31] P. Dabrowski-Tumanski, J.I. Sulkowska, To tie or not to tie? That is the question, Polymers 9 (9) (2017) 454.

[32] F. Khatib, M.T. Weirauch, C.A. Rohl, Rapid knot detection and application to protein structure prediction, Bioinformatics 22 (14) (2006) e252–e259.

[33] M. Michel, D. Hurtado Menendez, A. Elofsson, Pconsc4: fast, free, easy, and accurate contact predictions, Bioinformatics (2018).