Journal of Structural Biology 194 (2016) 112–123

Contents lists available at ScienceDirect

Journal of Structural Biology

journal homepage: www.elsevier.com/locate/yjsbi

Accurate prediction of helix interactions and residue contacts in membrane proteins

Peter Hönigschmid [a], Dmitrij Frishman [a,b,c,*]

[a] Department of Bioinformatics, Technische Universität München, Wissenschaftszentrum Weihenstephan, Maximus-von-Imhof Forum 3, 85354 Freising, Germany
[b] Helmholtz Zentrum Munich, German Research Center for Environmental Health (GmbH), Institute of Bioinformatics and Systems Biology, 85764 Neuherberg, Germany
[c] Laboratory of Bioinformatics, RASA Research Center, St Petersburg State Polytechnical University, St Petersburg 195251, Russia

ARTICLE INFO

Article history:
Received 12 August 2015
Received in revised form 1 February 2016
Accepted 2 February 2016
Available online 3 February 2016

Keywords:
Protein structure prediction
Sequence analysis
Molecular interactions
Machine learning

ABSTRACT

Accurate prediction of intra-molecular interactions from amino acid sequence is an important prerequisite for obtaining high-quality protein models. Over the recent years, remarkable progress in this area has been achieved through the application of novel co-variation algorithms, which eliminate transitive evolutionary connections between residues. In this work we present a new contact prediction method for α-helical transmembrane proteins, MemConP, in which evolutionary couplings are combined with a machine learning approach. MemConP achieves a substantially improved accuracy (precision: 56.0%, recall: 17.5%, MCC: 0.288) compared to the use of either machine learning or co-evolution methods alone. The method also achieves 91.4% precision, 42.1% recall and a MCC of 0.490 in predicting helix–helix interactions based on predicted contacts. The approach was trained and rigorously benchmarked by cross-validation and independent testing on up-to-date non-redundant datasets of 90 and 30 experimental three dimensional structures, respectively. MemConP is a standalone tool that can be downloaded together with the associated training data from http://webclu.bio.wzw.tum.de/MemConP.

1. Background

Protein sequence-structure gap (Rost and Sander, 1996), already quite dramatic for globular proteins, is even more pronounced for membrane proteins, with merely two thousand atomic structures available (Kozma et al., 2013; Tusnady et al., 2005a) for over one million amino acid sequences containing at least one predicted transmembrane (TM) region (The UniProt, 2014). The bulk of this huge discrepancy stems from the challenge to crystallize membrane proteins, as they are likely to lose their original structure when removed from their natural lipid environment due to their strongly hydrophobic surfaces, flexibility, and lack of stability (Carpenter et al., 2008). The low number of known 3D structures also limits our ability to increase the structural coverage of membrane proteins by template-based structure prediction methods. On the other hand, sequence-based methods to predict the topology of TM proteins, while highly accurate and useful, are unable to shed light on their spatial architecture.

Perhaps the only sequence-based approach able to provide information about the spatial arrangement of polypeptide chains and, in particular, useful constraints for 3D structure modeling, involves predicting contacts between amino acid residues. Prediction methods of the first generation exploited the idea of compensatory residue substitutions as an indication of a residue contact and utilized statistical methods of varying degree of sophistication to identify correlated mutations between pairs of positions in a multiple alignment (reviewed in (Fuchs et al., 2007)). More recent methods additionally applied machine learning algorithms to extract information about potential contacts form multidimensional data, such as evolutionary profiles, physico-chemical properties of amino acids, and other sequence specific features (Punta and Rost, 2005). However, all these methods, without exception, were designed to predict residue contacts in soluble proteins.

For a very long time sparseness of structural data precluded the application of contact prediction techniques to TM proteins. Not surprisingly, methods trained on globular proteins produce extremely poor results when applied to membrane protein sequences due to their very specific biophysical properties, most notably the fact that their exterior is much more hydrophobic than the interior due to the interaction with the lipid environment. In 2009 we developed the first contact predictor (TMHcon)

[*] Corresponding author at: Department of Bioinformatics, Technische Universität München, Wissenschaftszentrum Weihenstephan, Maximus-von-Imhof Forum 3, 85354 Freising, Germany.

E-mail address: d.frishman@wzw.tum.de (D. Frishman).

http://dx.doi.org/10.1016/j.jsb.2016.02.005
1047-8477/© 2016 Elsevier Inc. All rights reserved.

specifically geared towards α-helical membrane proteins, which employed a neural network trained on sequence features and correlation measures, which dramatically outperformed earlier methods used for globular proteins in terms of precision and recall (Fuchs et al., 2009).

Since the release of TMHcon the number of experimentally determined three-dimensional structures of TM proteins that can be used for training prediction algorithms increased significantly, from a mere 160 high resolution structures (non-redundant at 40% sequence identity) in 2009 to over 330 today. Concomitantly, the recent availability of more sensitive database search methods, such as HHblits (Remmert et al., 2012), allows to create better evolutionary sequence profiles by detecting more homologous sequences to be included in the multiple sequence alignment. Finally, and most importantly, there has been a quantum leap in our ability to detect compensatory mutations, which are indicative of structural contacts. While earlier methods assessed residue co-variation between each pair of positions in a multiple sequence alignment individually using simple correlation measures, such as mutual information, recent methods rely on global statistical models. These models attempt to infer causative correlations from the entire alignment and are thus able to distinguish between direct structural contacts and transitive connections between residues. The two pioneer approaches based on these novel ideas are mean-field direct coupling analysis (mfDCA), implemented as EVFold (Marks et al., 2011), and the estimation of a sparse inverse covariance matrix, as used in PSICOV (Jones et al., 2012). For both methods an accelerated implementation called Freecontact (Kajan et al., 2014) is available. Recently improved methods to predict residue contacts in soluble proteins have been released, which either employ enhanced algorithms (CCMpred, (Seemayer et al., 2014)), or combine several co-evolution methods (PconsC2 (Skwark et al., 2014), MetaPSICOV (Jones et al., 2015)).

Here we introduce a novel computational method, MemConP (**Mem**brane **Con**tact **P**rediction), which is specifically geared towards predicting residue contacts and helix interactions in TM proteins. The tool takes advantage of the recent surge in the number of 3D structures, more sensitive sequence analysis techniques, and vastly improved approaches to residue co-variation. It employs the random forest classification algorithm, which utilizes a large number of decision trees, each trained on a randomly chosen subset of training data and features. The resulting ensemble of classifiers determines the outcome by a majority voting. The random forest approach is used to combine several sequence-derived (evolutionary profiles, amino acid properties) and structure-derived (predicted TM topology) features with the mfDCA approach offered by Freecontact. We also introduce a new highly non-redundant dataset for training machine learning methods on TM proteins, as well as a new independent test dataset, which can serve for performance comparison with future methods. We compare the performance of MemConP with several recent predictive techniques, which employ residue co-evolution.

2. Methods

2.1. Definition of transmembrane segments, residue contacts, and helix interactions

For comparison of our method with other techniques we used the definition of TM regions obtained from the PDBTM database (Kozma et al., 2013). PDBTM definitions were also utilized for benchmarking of contact predictions. For benchmarking the quality of helix interaction predictions we rely both on PDBTM as well as on TM topology predictions produced by PolyPhobius (Kall et al., 2005).

To make our method comparable to the already existing and future ones (including our own previous work (Fuchs et al., 2009), we used the definition of residue contacts based on the Euclidean distance between any two atoms of less than 5.5 Å. A pair of helices was defined to be interacting if there was at least one residue contact between them. Another common contact definition is the distance between the C_β atoms of two residues of less than 8 Å. Performance measures for this alternative definition are reported in Supplementary materials.

2.2. Datasets

We used four datasets to train and benchmark the predictor: *OldTrain*, *OldTest*, *NewTrain* and *NewTest*. The first two datasets, *OldTrain* and *OldTest*, were used by all recent TM helix contact prediction methods and thus served as comparison datasets. *OldTrain* (introduced by (Fuchs et al., 2009)) originally consisted of 62 redundancy reduced X-ray structures of TM proteins extracted from PDBTM, TOPDB (Tusnady et al., 2008), and OPM (Lomize et al., 2006), with a resolution better than 3.5 Å and possessing at least three TM segments. We omitted the entry *2a79* from this dataset, as the corresponding topology data was deleted from PDBTM. *OldTest* was introduced by (Wang et al., 2011) and contains 21 TM proteins, of which none has a sequence identity above 40% to any other protein in this dataset, nor to those in *OldTrain*.

To create the *NewTrain* and *NewTest* datasets, used to train and test our final predictor, atomic coordinates and the annotation of membrane-spanning regions were extracted from the PDBTM database. PDBTM contains 3D structure information of experimentally solved TM protein structures, including atomic coordinates and the annotation of TM regions generated by TMDET (Tusnady et al., 2005b). We used the "Redundant Alpha" dataset of June 2015 from PDBTM containing 7374 protein chains as the initial dataset of transmembrane proteins. In order to produce a training dataset which is not biased towards an overrepresented family of proteins, and a test dataset which is totally independent from the training data, the initial dataset had to be redundancy reduced. Unfortunately, all existing approaches are aimed towards redundancy reduction of globular proteins. These methods take into account global or local sequence similarity using a substitution matrix which is designed for globular proteins and not optimized for highly hydrophobic TM segments. We therefore applied a very rigorous procedure to reduce redundancy both within and between our training and test datasets, incorporating structural similarity and PFAM family/clan membership (Finn et al., 2014) in addition to sequence similarity. Specifically, we calculated the length-independent measure of structural similarity, the so called TM-score, using the TMalign method (Zhang, 2005). The PFAM family/clan membership was added as an additional criterion to eliminate similarity between multi-domain proteins as well as to address those cases where even structural similarity comparison fails. Two proteins were declared similar if they (i) either shared a sequence identity of more than 35%, (ii) or displayed a TM-score below 0.5, which, according to the authors, implies that they share the same fold, (iii) or belonged to the same PFAM family or clan. To minimize the bias towards a specific type of TM proteins we grouped all proteins in this initial dataset according to the number of TM segments they possess. Subsequently protein chains were drawn from each of these groups, one at a time, and added to the *NewTest* dataset. At the same time, the sequences in the same group, which were similar to the drawn protein, were removed from the initial dataset. Upon achieving a certain pre-defined size of the *NewTest* dataset, N_{test}, the procedure was continued and the drawn proteins added to the *NewTrain* dataset until the initial dataset was depleted, automatically yielding a certain size of the NewTrain dataset, N_{train}. By applying the described procedure we