FEBS Letters 589 (2015) 3516–3526

FEBS Letters

journal homepage: www.FEBSLetters.org

Review

Computational prediction of protein interfaces: A review of data driven methods

Li C. Xue [a,*], Drena Dobbs [b,c], Alexandre M.J.J. Bonvin [a], Vasant Honavar [d,e,f,g,h,i]

[a] Faculty of Science – Chemistry, Bijvoet Center for Biomolecular Research, Utrecht Univ., Utrecht 3584 CH, The Netherlands
[b] Department of Genetics, Development & Cell Biology, Iowa State Univ., Ames, IA 50011, USA
[c] Bioinformatics & Computational Biology Program, Iowa State Univ., Ames, IA 50011, USA
[d] College of Information Sciences & Technology, Pennsylvania State Univ., University Park, PA 16802, USA
[e] Genomics & Bioinformatics Program, Pennsylvania State Univ., University Park, PA 16802, USA
[f] Neuroscience Program, Pennsylvania State Univ., University Park, PA 16802, USA
[g] The Huck Institutes of the Life Sciences, Pennsylvania State Univ., University Park, PA 16802, USA
[h] Center for Big Data Analytics & Discovery Informatics, Pennsylvania State Univ., University Park, PA 16802, USA
[i] Institute for Cyberscience, Pennsylvania State Univ., University Park, PA 16802, USA

ARTICLE INFO

Article history:
Received 6 July 2015
Revised 1 October 2015
Accepted 2 October 2015
Available online 13 October 2015

Edited by A. Valencia

Keywords:
Protein–protein interaction
Machine learning
Docking
Partner-specific interface prediction
Cross validation on protein level
Cross validation on instance level
Evaluation caveats

ABSTRACT

Reliably pinpointing which specific amino acid residues form the interface(s) between a protein and its binding partner(s) is critical for understanding the structural and physicochemical determinants of protein recognition and binding affinity, and has wide applications in modeling and validating protein interactions predicted by high-throughput methods, in engineering proteins, and in prioritizing drug targets. Here, we review the basic concepts, principles and recent advances in computational approaches to the analysis and prediction of protein–protein interfaces. We point out caveats for objectively evaluating interface predictors, and discuss various applications of data-driven interface predictors for improving energy model-driven protein–protein docking. Finally, we stress the importance of exploiting binding partner information in reliably predicting interfaces and highlight recent advances in this emerging direction.

© 2015 The Authors. Published by Elsevier B.V. on behalf of the Federation of European Biochemical Societies. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

1. Introduction

Proteins are the principal catalytic agents, structural elements, signal transmitters, transporters and molecular machines in cells [52]. But individual proteins do not function alone; they must interact with other molecules to carry out their cellular roles. Alterations in protein–protein interfaces often lead to disease, and hence protein interfaces have become one of the most popular new targets for rational drug design [35,60]. In addition to practical applications in drug design, reliable identification of protein–protein interfaces is important for basic research on the mechanisms of macromolecular recognition.

Many biochemical and/or biophysical experimental methods have been used to identify and characterize protein–protein interfaces at the level of individual atoms or residues. Widely used techniques include: X-ray crystallography [66] and nuclear magnetic resonance (NMR) spectroscopy [22], both of which are capable of determining interfaces at the atomic level; alanine scanning mutagenesis, which can determine interfaces at the residue level; various mass spectrometry-based approaches, such as chemical cross-linking and hydrogen/deuterium (H/D) exchange, which typically report the location of interfaces at lower resolution, but are capable of identifying individual interfacial residues [27,38]; and various NMR-based approaches [70], such as chemical shift perturbations, cross-saturation, and H/D exchange, which determine interfaces at the residue or atomic level (for an recent summary, see [63]).

These experiments are extremely valuable and have contributed greatly to our knowledge of protein recognition mechanisms. However, technical challenges, such as difficulties in expressing and purifying aggregation-prone protein samples, obtaining high quality crystals, as well as the protein size constraints (for NMR), make such experiments both labor-intensive and time-consuming. Because high throughput experimental characterization of protein interfaces is not yet possible, reliable computational approaches to identify interfacial residues are especially valuable.

* Corresponding author.
E-mail address: L.Xue@uu.nl (L.C. Xue).

http://dx.doi.org/10.1016/j.febslet.2015.10.003
0014-5793/© 2015 The Authors. Published by Elsevier B.V. on behalf of the Federation of European Biochemical Societies.
This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Based on the extent to which a method relies on experimental data, protein–protein interface prediction methods can be classified into two broad strategies: (1) data-driven or knowledge-based methods, which heavily depend on the availability of experimental data to make predictions, either by using homologous data as templates or by extracting interaction patterns from data into statistical models; (2) protein–protein docking (see a review by [69]), that typically use physics-based and/or geometric models to search for putative conformations with low interaction energy and high surface complementarity. The data-driven interface prediction methods include: (1) homology-based methods, which assume that interfaces are conserved among homologs and exploit experimentally determined interfaces of homologs as templates to infer those of query proteins [34,67,72]; (2) machine learning based methods, which use a dataset of experimentally determined interfaces to train interface predictors and use the trained models to predict interfacial residues of query proteins (see reviews by [9,20,79]; and (3) co-evolution based statistical models, which operate under the assumption that interacting residues at the interface are likely to co-evolve and use a large multiple sequence alignment (MSA) to identify such residues [24,28,46] (also see [47] for a general review of co-evolution based methods for intra-protein contact predictions and their applications to protein structure prediction).

The different classes of interface prediction methods have different respective strengths and weaknesses, and can be combined in ways that exploit this. Data-driven methods are capable of integrating heterogeneous experimental data and are usually quite computationally efficient. But because most data-driven methods are based on statistical rules extracted from training datasets, they typically predict interfaces at the residue level and can suffer from high false positive rates. *Ab initio* docking programs can predict 3D structures of protein–protein complexes at the atomic level, but usually are computationally demanding and don't consider relevant non-physicochemical information, such as residue conservation and correlated mutations, which can be extracted from the existing wealth of sequence data.

We note that the different strategies are not necessarily mutually exclusive. For example, machine learning algorithms are also widely used in homology based methods to integrate templates of varying quality. Also, statistical potentials derived from experimental interface data are often used in scoring functions of docking programs. Further, data-driven docking approaches such as HADDOCK [16] have been developed to make use of interface predictions, or any available experimental information on the target system to guide the docking process [62]. Increasingly, the state-of-the-art approaches leverage heterogeneous data sources and integrate multiple analysis and modeling strategies.

This review focuses on data-driven methods. Over the past two decades, the protein interface prediction field has advanced considerably and several reviews have been published along the way [9,20,79]. The most recent review by [19] summarized and classified the majority of existing methods on a broad scope, covering not only general protein–protein interface predictions, but also specific areas such as paratope prediction, epitope prediction, and antibody-specific epitope prediction. Our aim here is to provide an entry point for researchers and practitioners who are new to this field. Hence, we focus on introducing basic concepts, practical technical details (e.g., statistical comparison of multiple methods, handling unbalanced dataset, and useful resources) and the rationale behind representative methods. We stress the added value of considering binding partner information in interface analyses and prediction, and highlight a recent significant advance – partner-specific prediction methods – and their application to improve and guide computational docking. Most importantly, while none of the previous reviews has emphasized objective evaluations, we point out an important caveat, i.e., cross-validation over proteins *vs.* over sliding windows (or surface patches). This caveat is a serious one and reoccurs even in the recent literature. Using a concrete example, we illustrate how the evaluation over sliding windows gives artificially high performance. We conclude with a discussion of key challenges and promising future directions in the field.

2. Data-driven approaches for protein interface prediction

In the past two decades, a broad range of computational methods for protein–protein interface prediction have been proposed in the literature. Some representative methods are summarized in Table 1 (also see reviews by [9,20,79]. These methods can be grouped into two major categories: homology-based approaches and template-free machine learning-based approaches.

2.1. Homology-based methods

Homology-based approaches infer biological properties of a query protein from its homologs based on the assumption that homologs share significant similarity in sequence, structure and functional sites. Whenever close homologs are available, homology-based (also called template-based) methods usually provide the most reliable results compared with other methods, and have been successfully applied in many areas, such as protein structure prediction [48], the prediction of protein interaction partners [76], and function annotation [45].

The potential value of using homologs to infer interfacial residues was unclear for several years because several published studies disagreed as to whether or not interfacial residues are conserved among homologs [6,23,61]. The relatively small (and different) datasets used in these studies contributed to this discrepancy. More important, however, is the finding that in contrast to proteins in *stable* complexes, which tend to have a single dominant interface, proteins in *transient* complexes tend to use different interfaces for binding different partners. By taking into account specific binding partner information, our group demonstrated that the locations of interfaces in transient complexes are highly conserved, even though the sequences (i.e., the identities of the amino acids) in these interfaces are not usually conserved [72]. Based on this *partner-specific* interface conservation, we designed one of the first partner-specific interface predictors, PS-HomPPI [72]. Given a query protein and its specific binding partner, PS-HomPPI searches the PDB (Protein Data Bank, www.rcsb.org) [4] for homologous interacting proteins and uses these selected homologs as templates for mapping experimentally determined interfacial residues onto the query protein sequences. For each predicted interfacial residue pair, PS-HomPPI also reports the average, minimum and maximum CA–CA (alpha carbon – alpha carbon) distances calculated from the templates. Two important steps guarantee the reliability of PS-HomPPI: (i) PS-HomPPI automatically classifies the templates into one of three categories, Safe Zone, Twilight Zone and Dark Zone, based on the similarity of the templates to the query protein, and uses templates from the best available zone; (ii) PS-HomPPI uses multiple templates to reduce the negative impact of occasionally choosing an incorrect (non-homologous) template.

Other published homology-based methods are non-partner-specific (NPS) methods, i.e., they do not consider the specific binding partner information when making predictions. Representative methods include NPS-HomPPI [72], PredUS [78], PriSE [34] and IBIS [67]. NPS homology-based methods search the PDB database for homologs of a query protein and map the *union* of the interfaces in homologs with *all* possible binding partners of the query protein. One exception is PriSE [34], a *local* structural homology-based

Download English Version:

https://daneshyari.com/en/article/10869910

Download Persian Version:

https://daneshyari.com/article/10869910

Daneshyari.com