Available online at www.sciencedirect.com
ScienceDirect
Current Opinion in Structural Biology

Challenges for the prediction of macromolecular interactions

Mark N Wass, Alessia David and Michael JE Sternberg

Macromolecular interactions are central to most cellular processes. Experimental methods generate diverse data on these interactions ranging from high throughput protein–protein interactions (PPIs) to the crystallised structures of complexes. Despite this, only a fraction of interactions have been identified and therefore predictive methods are essential to fill in the numerous gaps. Many predictive methods use information from related proteins. Accordingly, we review the conservation of interface and ligand binding sites within protein families and their association with conserved residues and Specificity Determining Positions. We then review recent developments in predictive methods for the identification of PPIs, protein interface sites and small molecule ligand binding sites. The challenges that are still faced by the community in these areas are discussed.

Address
Structural Bioinformatics Group, Division of Molecular Biosciences, Imperial College London, South Kensington, London SW7 2AZ, UK

Corresponding author: Sternberg, Michael JE (m.sternberg@imperial.ac.uk)

Current Opinion in Structural Biology 2011, **21**:382–390

This review comes from a themed issue on
Sequences and topology
Edited by Julian Gough and Keith Dunker

Available online 14 April 2011

0959-440X/$ – see front matter

DOI 10.1016/j.sbi.2011.03.013

Introduction

The cell is a crowded environment [1] in which proteins, DNA and small molecules interact in specific ways to perform their biological functions. For proteins, many of their functions are performed through interactions with other proteins and ligands, which may be their substrates or regulators of their activity. Thus central to understanding the interactions occurring in the cell is the knowledge of which proteins and ligands interact together and the locations of their interfaces.

This review first considers the conservation of protein–protein interactions (PPIs) between globular proteins and also protein–ligand binding sites across structural space. We then consider three challenges associated with the prediction of macromolecular interactions (Figure 1) and recent progress that has been made in each of these areas: first prediction of PPIs, second prediction of protein interfaces for globular proteins of known 3D structure and third prediction of ligand binding sites for small molecules such as metal ions, ATP and FAD excluding larger molecules such as other proteins, RNA and DNA.

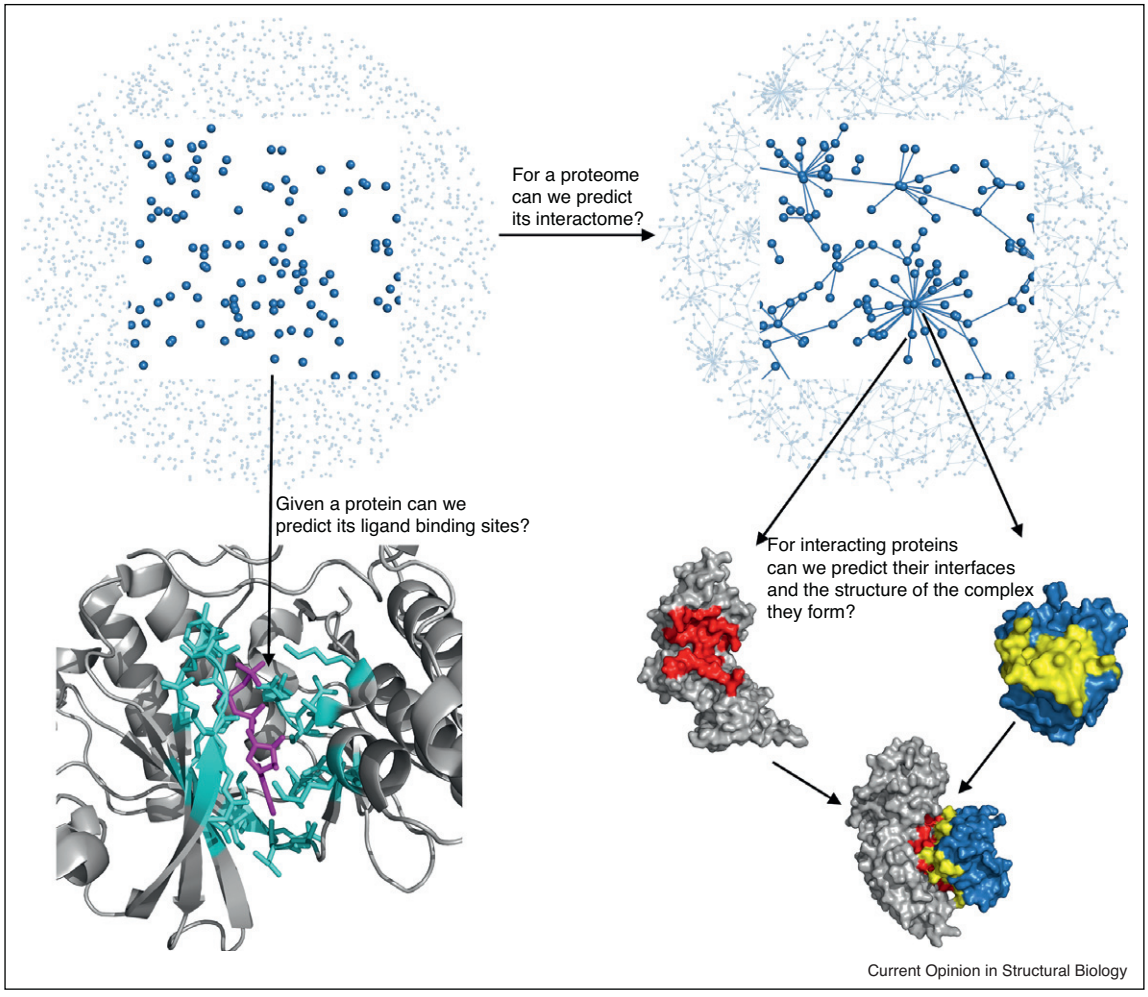
While experimental methods continue to generate data on interactions, there is still an essential need for predictive methods. It has recently been estimated that in humans there may be as many as 600 000 PPI [2] and so to date we have identified only a fraction of human PPIs. This is supported by the work of Ranea *et al.*, (see prediction of PPI section) who suggest that new experimental methods may be required to characterise many interactions [3•]. Experimental data for protein complexes, which structurally characterise their interfaces, are even sparser. Therefore docking methods, which predict the structure of the complex formed by interacting proteins, and other programs that predict the location of interfaces on proteins are an essential area of development. Here we focus on new methods for the identification of interface sites, including methods that are associated with docking protocols.

Detailed knowledge of ligand binding sites is limited to proteins with known structures in which the ligand is also present. Many methods have used residue conservation and surface clefts for prediction of binding sites. Here we review recent methods that exploit the data available from ligand-bound structures present in the Protein Data Bank (PDB).

Conservation of interfaces and ligand binding sites

Central in structural bioinformatics is the quantification of the level of sequence identity required for the conservation of certain features such as structure [4], function [5,6] or interfaces. Aloy and Russell observed that interfaces are generally structurally conserved down to approximately 30–0% sequence identity [7]. They also estimated that there would be up to 10 000 different interaction types [8•]. Analysis of complexes in the PDB over the past 20 years [9] shows that the number of structurally distinct interfaces is increasing more rapidly than the number of protein families in SCOP (Structural Classification of Proteins [10]) and the rate of growth is inline with a total number of interactions in agreement with the 10 000 proposed by Aloy and Russell. Honig and co-workers have recently extended this topic by assessing the conservation of interface residues across structural space [11]. They show that for groups of related proteins, which interact with different proteins and with different geometries, there are conserved

Figure 1



Challenges associated with the prediction of macromolecular interactions.

regions on the protein surface that form the interfaces for the interactions. They propose that it is the location of surface residues responsible for forming interactions that is conserved rather than the geometries of interactions.

Later we report on algorithms to predict the location in a protein of ligand binding regions. Many approaches are based on inheritance from homologous structures with known ligand binding sites. However, the principles underpinning this approach have not been established. Accordingly, here we report the key findings of our analysis (David A, Wass MN, Sternberg MJ, unpublished) of the conservation of ligand binding sites within protein families. We analysed the conservation of ligand binding sites within SCOP superfamilies. For each superfamily, domains with biologically relevant ligands were aligned and the number of distinct binding sites was identified. We found that ligand binding sites are highly conserved within protein superfamilies, and that the number of binding sites within the superfamilies is generally small (Figure 2). For superfamilies with a single binding site, the site is conserved in most members of the superfamily (greater than 90% conservation for over 80% of superfamilies). For superfamilies with more than one binding site, there are varying degrees of conservation. However, at least one binding site is highly conserved in the majority of superfamilies (Figure 2). For example, for 51 of the 64 superfamilies with two binding sites, one of the binding sites is conserved in more than 60% of members of the superfamily, while the second site has more variable levels of conservation within superfamilies (Figure 2b). Thus this study helps to explain the success of methods to predict ligand binding residues by inheritance from a homologue.

Rausell *et al.* [12••] have analysed the inter-relationships of residues required for ligand binding and for PPIs within homologous families of proteins. They considered Specificity Determining Positions (SDPs) which are positions in a multiple sequence alignment for a protein family that are invariant within subfamilies but vary between different