Modelling cofactors in comparative protein structure models by evolutionary inference

Inauguraldissertation

zur

Erlangung der Würde eines Doktors der Philosophie

vorgelegt der

Philosophisch-Naturwissenschaftlichen Fakultät

der Universität Basel

von

Tiziano Gallo Cassarino

aus Italien

Basel, 2014

Genehmigt von der Philosophisch-Naturwissenschaftlichen Fakultät

auf Antrag von

Prof. Dr. Torsten Schwede

Prof. Dr. Olivier Michielin

Basel, den 20.05.2014

Prof. Dr. Jörg Schibler

Dekan

Table of contents

Summary

Proteins perform their role through the interactions they establish with other proteins and with small molecules, like ions or organic cofactors. The identification of these partners and of the mechanisms involved in their functional interactions can provide helpful insights into the molecular details of the protein annotation and for the development of new drugs. As many proteins lack of experimental structures and of annotated ligands, computational methods are required in order to predict these details and to guide the direction of experimental investigation.

In this context, our main aim is to enhance protein functional annotation and to improve comparative models by inferring their potential binding cofactors. Moreover, we want to evaluate the current state-of-the-art methods for binding site prediction in order to understand their advantages and limitations for future developments. Additionally, we aimed to improve the assessment of binding site prediction methods by creating an automated system of continuous model evaluation. Finally, we created a new binding site descriptor for the de novo ligand and binding site prediction in protein models.

The content of this thesis is organized as follows. Chapter 1 introduces protein structure, binding sites and experimental techniques for structure determination; moreover, we illustrate the current approaches to model protein structures and to predict their ligand binding sites. In chapter 2, we describe the assessment of the ligand binding site predictions within the 9th edition of the Critical Assessment of protein Structure Prediction (CASP) experiment, while in chapter 3 we discuss the latest developments in the 10th round. Within chapter 4 we illustrate the evolution of this assessment into the Continuous Automated Model EvaluatiOn (CAMEO) Ligand Binding category and we describe the homology predictor, which is used as reference for the comparison of the other methods registered to CAMEO. Chapter 5 presents the new SWISS-MODEL server, which employs a base ligand modelling pipeline to place potential small molecules partners, inferred from the target's template, into the built models. Motivated by the performances of the previous method and by the results seen in the last CASP editions, in chapter 6 we present a new method to model ligands, especially ions and organic cofactors, into comparative models; this approach is based on the analysis of the similarities between a target and its homologous proteins. In chapter 7, we describe a novel descriptor for ligand binding sites, based on moment invariants and developed for the de novo prediction of ligands. Finally, in chapter 8 we draw the general conclusions of the work presented in this thesis.

1. Introduction

Protein structure

Many of the biological functions performed by living organisms are mediated by proteins, which can catalyse reactions (for instance, the production of metabolites), have a structural or mechanical role (like in the muscle fibre), propagate signals (e.g. the kinases), act as sensors of metabolites (as for neurotransmitter receptors), transport or store small molecules (e.g. oxygen in the haemoglobin). Proteins can be generally classified as "membrane proteins", when they act as receptors or as channels to allow the passage of charged molecules through a membrane, as "fibrous proteins", when they have a structural role, or as "globular proteins" in all the remaining cases.

Primary and secondary structure

A protein is a linear polymer composed of a chain of amino acids, called "residues", translated from a mRNA molecule, so that each protein has a well defined amino acidic sequence, indicated as the "primary structure". Each protein's residue is made of a central C-alpha carbon covalently bound to an amminic group, an acidic group - which together form the protein backbone - and a variable side-chain.

The first two groups are condensed together by a peptide bond, which has partial double-bond behaviour due to the resonance between a neutral and a charged conformation. This characteristic does not allow the rotation of the bond itself, so that the residues' C-alphas are almost coplanar [1]. Additionally, due to sterical constraint between the CO and NH groups, the peptide bond reduces the degree of freedom of the backbone, which can rotate only around the two dihedral angles phi and psi, defined between N-C-alpha and C-C-alpha respectively. The value of these angles can be distributed only within a finite set of combinations, traditionally described by the so called "Ramachandran plot" and recently refined by Ting and colleagues [2].

The variable side-chain is used to identify the amino acids in 20 "standard" types and to classify them in different chemical categories based on several properties, like for example charge or size. However, they can be broadly categorized in hydrophobic (non-polar) and hydrophilic (polar). Hydrophobic residues do not interact favourably with water molecules, so they are more often found in the core of a water-soluble protein; for the opposite reason, hydrophilic amino

acids can be exposed to the solvent - where an interaction with small molecules can occur - or can be located in the protein core, where they contribute to the structural stability of the protein by forming salt bridges with other residues [3]. Moreover, some residues have particular features. For example, cysteine can bind another residue of the same type to form disulfide bridges; glycine confers more flexibility to the surrounding structure as has only a hydrogen atom as side-chain; finally, proline has a cyclic structure that increases the conformational rigidity of the backbone.

The variability in residues chemical properties and in their position along the protein sequence determines the structure and the biological function of the protein itself. Each residue can interact with the other amino acids by different non-covalent bonds, which might be hydrogen bonds, ionic bonds or Van der Waals interactions; all of these are weaker than a typical covalent bond, but they can act together to create a strong bonding network. The hydrogen bond in particular is involved in the stability of the two simplest and common structural patterns that can be found in proteins, that is, the alpha-helix and the beta-strand [4].

The first is a right-handed helical conformation characterized by a hydrogen bond present every four residues between the CO and the NH groups of the backbone, creating a complete turn every 3.6 amino acids. Left-handed helices exist in nature, but they are less energetically favourable because of the steric clashes between the backbone and the side-chains. Usually alpha-helices can range from four to forty residues in length and are more frequent in proteins that cross a lipid membrane.

The beta-strand, instead, is a fully extended backbone region characterized by several hydrogen bonds between the CO and the NH groups of residues located further apart in the protein sequence than in the alpha-helix. Two or more strands can organize themselves in a beta-sheet, with a twisted and pleated shape, where the side-chains are oriented to both sides of the sheet. In the parallel beta-sheets, the strands point to the same direction; in anti-parallel beta-sheets, strands point to opposite directions; finally, in mixed beta-sheets, both strand directions are present. Alpha-helices and beta-strands are connected by loops, which are structural motifs that do not create a regular pattern and in which the involved residues are positioned in close proximity.

These three structural units (helices, strands and loops) constitute the "secondary structure" of a protein and the combinations of these elements are known as "protein folds". From the functional point of view, groups of secondary structures can give rise to three-dimensional

elements, named "domains", which are able to fold in a stable manner independently from the rest of the protein. Proteins might contain several domains, whose length usually ranges from 40 to 350 residues, and the same type of domain - which defines a particular function - may appear in a variety of different proteins [3].

Tertiary and quaternary structure

The next level of complexity is defined by the "tertiary structure", which refers to the overall three-dimensional structure of a protein chain. The tertiary structure is the result of a thermodynamical process, called "protein folding", which is guided by the cooperative interaction of the residues. The forces driving this process are mainly hydrogen bonds [5] and hydrophobic effects, in which the non-polar side-chains tend to pack within the protein core in order to avoid any exposition to the surrounding water [6].

Some proteins are able to function as single chains and, therefore, they are indicated as “monomers”; however, many others need to assemble in complexes called “oligomers”, which are stabilized by non-covalent bonds interacting at the chains interfaces. When these assemblies are composed of many copies of the same chain, they are called "homo-oligomers"; otherwise, assemblies consisting of at least two different chains are indicated as "hetero-oligomers". This level of structural organization is referred to as the "quaternary structure" of a protein, while the single chains are called "subunits". A summary of all protein structural levels is shown in Figure 1.1.

Protein oligomers perform, or regulate, their function by changing the conformation of individual chains or their relative orientation to each other. One example of this behaviour is haemoglobin, a hetero-oligomer in which the allosteric regulation of its function is achieved by the relative orientation of the subunits [7].

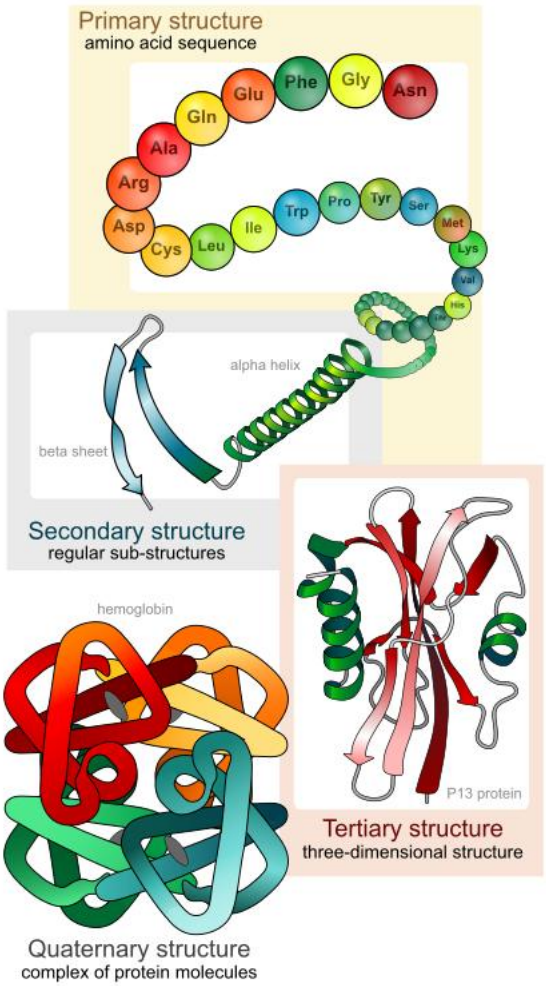

Figure 1.1 Illustration of the hierarchical structural levels of a protein. From Wikipedia, retrieved 2014 March 31, from http://en.wikipedia.org/wiki/Protein_structure

Experimental techniques to determine protein structure

The most important experimental methods to decipher the above mentioned structural levels of a protein are X-ray crystallography [8], Nuclear Magnetic Resonance spectroscopy (NMR) [9] and Electron Microscopy (EM) [10].

In the X-ray crystallography, a purified protein crystal is irradiated with X-ray beams in order to reconstruct the precise atom positions. The directions and intensities of the X-rays which are diffracted by the electrons in the crystal are measured from the so called "diffraction pattern", which can be converted, through a Fourier transform, to an electron density map. By combining the knowledge about the target amino acidic sequence with proteins' geometrical constraints, it is possible to reconstruct atom positions and to build a model of the protein. Although producing high-quality crystals is a time-consuming process and membrane proteins in particular do not

crystallize, the X-ray crystallography is still considered the gold standard method mainly because of its high accuracy and for the fact that the protein function is preserved in the resulting crystal [8].

In NMR spectroscopy, the magnetic properties of the atom nuclei are used to determine the structure and the dynamics of a target molecule. The proteins, usually suspended in a buffer solution, are placed within a strong magnetic field and irradiated with varying radio wave pulses. The measured variable is the resonance of the nuclei possessing a spin, i.e. those which produce a magnetic moment, like hydrogen (1H), carbon (13C) and nitrogen (15N). Depending on the atom type, on the surrounding atoms and on their distances, the resonance frequency of an atom can change and this information is used to infer the structure of the target molecule [9].

The electron microscopy method uses a beam of electrons to illuminate a sample and to produce a magnified image, which has a much higher resolution than an image produced with conventional light. This technique is based on the high scattering power of electrons; for this reason, the sample must be a very thin crystal. Moreover, the possibility to focus the electrons by an electric or magnetic field allows retaining the crystallographic phase information in the resulting image. However, biological material is sensitive to radiation and, for this reason, the electron dose must be limited, at a cost of a small signal to noise ratio. The most used approach to create an image is the single particle averaging, in which several 2D images of the molecule densities are collected and averaged; then, by applying a back projection in real space, the three dimensional density of the sample is assembled. A second approach, in which the diffraction pattern of a two-dimensional crystal is produced, is more commonly applied to determine the structure of membrane proteins. The main limitation of electron microscopy consists in the need of a relatively large array of ordered macromolecules to achieve a resolution around 3.5 Angstroms [10].

The structural data of biological macromolecules obtained through any of the three techniques described above is deposited by experimentalists in the database "Protein Data Bank" (PDB) [11]. At the moment of its inauguration, in 1971, the PDB contained only 7 structures, but since 1980 the number of entries started to increase substantially. The reason of this growth mainly resided in the improvement of the crystallographic techniques and in the emergence of new methods to determine the structure of a protein, as for instance NMR. Recently, structural genomic initiatives like the Protein Structure Initiative are increasing even more the number of deposited structures, which has reached almost 100'000 entries. Apart from atomic coordinates, other types of information are deposited in the PDB, including experimental details, raw density

maps and quaternary structures, to name a few. To increase the robustness of the service to the public, three mirror sites are available: RCSB [12], PDBe [13] and PDBj [14]. Finally, since several structures were deposited long ago and were refined with different types of algorithms, an updated and optimized version of the PDB entries is now available through PDBredo [15].

Ligand binding sites

Apart from the role played by the structural conformation, a critical element that defines the function of a protein is the “binding site”, that is, the portion of the protein surface through which it interacts with either other proteins or small non-protein molecules, for example ions, organic ligands or nucleic acids. These interactions can be stable, i.e. they are required to stabilize the structure and to perform the function (for instance, in the case of quaternary assemblies), or transient, as for example, when the protein binds to the substrate during an enzymatic reaction or to a signalling molecule (as in Figure 1.2).

Knowing the ligands bound by a protein and the residues involved in these interactions can provide a significant help in the identification of the protein function and in the understanding of its mechanism of action at the atomic level. Moreover, the information regarding the ligand preferences of a protein can constitute a valuable insight for protein mutational experiments, structure-based drug design and virtual screening.

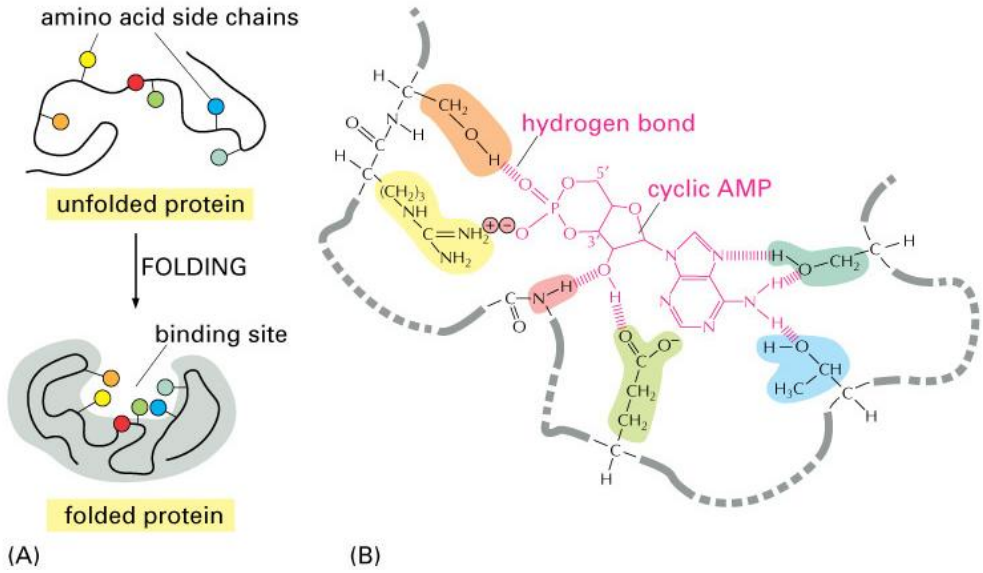

Figure 1.2 (**A**) The protein folding shapes and brings together a group of residues that constitute the protein binding site. (**B**) An example of the molecular interactions, hydrogen bonds and ionic interactions, between a small molecule (in pink cyclic AMP) and a protein binding site. From [16].

Protein structure prediction

Sequence-structure gap

The dependency of the protein's three-dimensional structure on its sequence was revealed for the first time by the work of Anfinsen [17], in which he performed denaturation experiments on the ribonuclease enzyme showing the relationship between the conformation of a protein and its biological function. The major finding of Anfinsen consisted in the fact that the enzyme in a denaturated (or unfolded) state spontaneously regained its native activity under particular buffer conditions. More in particular, a pivotal role in the transition to the functional conformation of the enzyme was played by a decrease in free energy of the system. Afterwards, similar experiments showed that, while many proteins can fold in their native state under proper conditions, other proteins need the help of assistant proteins, called "chaperons", to reach the correct conformation and to avoid uncontrolled aggregation within the cell.

The relationship between the sequence and the structure was further investigated by the work of Chothia and Lesk [18], who showed a non-linear correlation between the divergence of the protein sequence and the structure core in a set of evolutionary related proteins solved by X-ray crystallography (Figure 1.3). This observation implicated that the success of protein structure prediction involving evolutionary related sequences depends on the extent of the sequence identity between the target protein and its homologs. However, because of convergent evolution, even a distantly related protein with overall low sequence identity to the sequence of interest can turn out to be a useful template for modelling the active site [18].

Despite the rapid increase in the number of experimentally determined structures, the number of sequences identified by Next-Generation Sequencing (NGS) techniques grows even faster. Consequently, the difference between protein structures and sequences, also called "sequence-structure gap", is constantly widening. Structural genomic initiatives, as for instance the Protein Structure Initiative (PSI) [19], are trying to reduce the uncovered protein space by determining the structures of proteins with less than 30% sequence identity to existing structures. In the attempt to fill the sequence-structure gap, several computational methods were developed for building models of proteins with still unknown structure; these can be classified in "de novo" methods, in which a candidate structure is selected from a set of pre-generated models, and "template-based" approaches, which adopt the sequence-to-structure relationship to find the best structure for a given protein sequence.

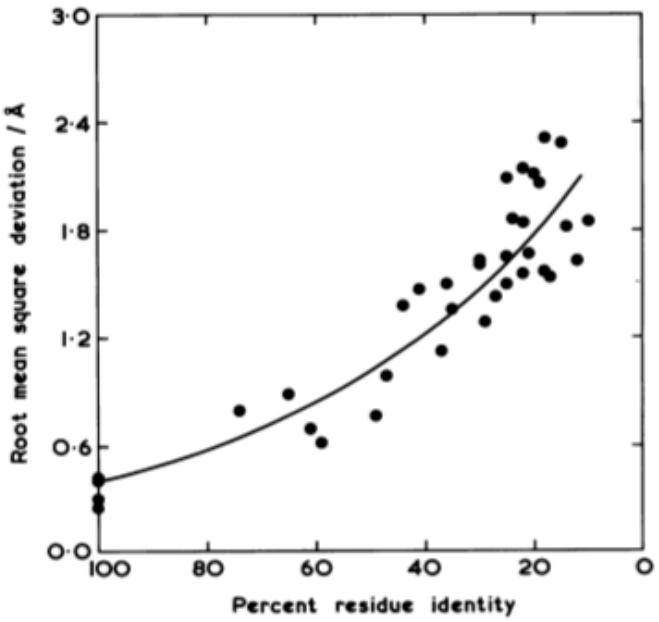

Figure 1.3 The non-linear correlation between the residue identity percentage and the root mean squared deviation in the common cores of the backbone atoms from 32 pairs of homologous proteins.

Template-based structure prediction

The template-based structure prediction methods take advantage of existing structures to generate a model of a protein, also called "target", at atomic resolution, weighting more evolutionary information than physics-based energy functions. The first step of this procedure is the identification and alignment of sequences taken from the structures found to be evolutionary related to the target protein. In the next step, the target sequence is modelled on the selected structure, called "template", and the model is subsequently refined. Finally, the quality of the built model is evaluated in order to assess whether the model is reliable or not. In case of a negative answer, the above procedure must be repeated to find a suitable model [20].

The initial step of the prediction is the most crucial one, since it significantly affects the quality of the model; previously, it was performed by local alignment tools, for example BLAST [21], which can be used to generate accurate alignments when the sequence identity between the target and the templates is above 50%. Below this threshold, more sensitive and sophisticated methods based on sequence-profile [22], sequence-HMM [23] or HMM-HMM [24] alignments showed to be more successful. Protein threading methods can be applied in case only remote homologs are found and, in particular, for homologs with sequence identity in the range called “twilight zone” [25]. An example of a tool implementing this approach is RaptorX, which assigns

more weight to the sequence features when a high sequence identity is measured, while it gives higher priority to the structural properties in case of remote structures [26].

After a template and its alignment are selected, these are used to generate the three-dimensional coordinates of the target protein. The main approaches employed during this stage can be divided in "fragment assembly" and "satisfaction of spatial restraints". According to the former method, the conserved structural elements are initially copied from the template and, in a later stage, the variable regions are remodelled; instead, in the satisfaction of spatial restraints method (an example of which is the software MODELLER [27]), the probability density function derived from geometrical criteria are used as spatial constraints to drive the global energy minimization of the model's atom coordinates. Overall, the higher is the sequence identity between the target and the template, the more successful become template-based approaches.

De novo structure prediction

The de novo structure prediction infers the structure of a protein either on the basis of the principles that guide molecular interactions, or by doing a statistical analysis of the native structure conformational features. In the former case, the method of prediction is named “physics-based”, while in the latter the method is called “knowledge-based”. In general, the de novo approach samples the structural conformational space by using a scoring function based on one of the two above mentioned methods and generates a set of candidate structures, called “decoys”, which are then filtered to select the native-like conformations. Even though the de novo approach does not achieve fold level quality in many cases [28], a successful example of this procedure is represented by ROSETTA [29]. Finally, although template-based methods are preferred when a suitable template is found, de novo methods can be useful for modelling targets with none or low template coverage, as well as for model refining.

Ligand binding site prediction

Several approaches of binding site prediction have been proposed in the last decade; these can be subdivided on the basis of the main information employed, which can be: target sequence conservation [30-35], protein surface geometry [36-42] or functional annotation from evolutionary related proteins [43-49]. Depending on the available data, different methods can be applied. In case the structure of the target protein is unknown or cannot be modelled, only the sequence conservation-based approach can be used; otherwise, the clefts on the protein surface can be

investigated to identify potential ligand binding sites; finally, the functional annotation-based methods can only be used in case homologous proteins are found.

In the approach last mentioned above, the fundamental steps of ligand binding site prediction consist in: (i) finding the target's homologs, (ii) identify their functional sites, (iii) determine the corresponding residues in the target and (iv) transfer to these residues the annotations found for the homologs. While some methods rely on the alignment between the target and the homologous sequences (for example [43]), others superpose the homologous structures to the target model (as in [44, 46, 48, 49]) to identify the functional residues in the target and to transfer the available annotation.

To assess the performances of these different methods, each two years the Critical Assessment of protein Structure Prediction (CASP) Function prediction (FN), evaluates the accuracy of the participant methods. Recently, algorithms based on the homology transfer approach have shown excellent results in the ligand binding site prediction [50, 51]. To tackle the challenges involved in the precise evaluation of binding site predictions emerged during the last CASP editions, an automated server, the Continuous Automated Model EvaluatiOn (CAMEO) Ligand binding site for the ligand binding site prediction assessment was developed (http://cameo3d.org/lb/).

Objectives

The main focus of this thesis is to improve the information contained in the models built by the SWISS-MODEL server, by introducing a new ligand modelling pipeline. Secondly, we examined and assessed the current methods available for predicting binding sites and for modelling ligands into protein models.

In the next chapters, we first show our assessment of the current state-of-the-art methods for the CASP9 and CASP10 editions. Then, we describe the method developed to assess the predictions of the servers registered to CAMEO. Afterwards, we illustrate the approach used for ligand modelling and implemented in SWISS-MODEL. Finally, we describe a method to represent binding site geometries, called “moment invariants”, which we studied to develop a future de novo ligand binding site predictor.

References

1. Corey, R.B. and L. Pauling, *Fundamental dimensions of polypeptide chains.* Proc R Soc Lond B Biol Sci, 1953. 141(902): p. 10-20.
2. Ting, D., et al., *Neighbor-dependent Ramachandran probability distributions of amino acids developed from a hierarchical Dirichlet process model.* PLoS Comput Biol, 2010. 6(4): p. e1000763.
3. Alberts, B., J.H. Wilson, and T. Hunt, *Molecular biology of the cell*. 5th ed. 2008, New York: Garland Science. xxxiii, 1601, 90 p.
4. Berg, J.M., J.L. Tymoczko, and L. Stryer, *Biochemistry*. 5th ed. 2002, New York: W.H. Freeman.
5. Rose, G.D., et al., *A backbone-based theory of protein folding.* Proc Natl Acad Sci U S A, 2006. 103(45): p. 16623-33.
6. Pace, C.N., et al., *Forces contributing to the conformational stability of proteins.* FASEB J, 1996. 10(1): p. 75-83.
7. Hilser, V.J., J.O. Wrabl, and H.N. Motlagh, *Structural and energetic basis of allostery.* Annu Rev Biophys, 2012. 41: p. 585-609.
8. Rhodes, G., *Crystallography made crystal clear : a guide for users of macromolecular models*. 3rd ed. Complementary science series. 2006, Amsterdam ; Boston: Elsevier/Academic Press. xxv, 306 p.
9. Keeler, J., *Understanding NMR spectroscopy*. 2nd ed. 2010, Chichester, U.K.: John Wiley and Sons. xiii, 511 p.
10. Glaeser, R.M., *Electron crystallography of biological macromolecules*. 2007, Oxford ; New York: Oxford University Press. xv, 476 p.
11. Bernstein, F.C., et al., *The Protein Data Bank: a computer-based archival file for macromolecular structures.* J Mol Biol, 1977. 112(3): p. 535-42.
12. Berman, H.M., et al., *The Protein Data Bank.* Nucleic Acids Res, 2000. 28(1): p. 235-42.
13. Gutmanas, A., et al., *PDBe: Protein Data Bank in Europe.* Nucleic Acids Res, 2014. 42(Database issue): p. D285-91.
14. Kinjo, A.R., et al., *Protein Data Bank Japan (PDBj): maintaining a structural data archive and resource description framework format.* Nucleic Acids Res, 2012. 40(Database issue): p. D453-60.
15. Joosten, R.P., et al., *PDB_REDO: constructive validation, more than just looking for errors.* Acta Crystallogr D Biol Crystallogr, 2012. 68(Pt 4): p. 484-96.
16. Alberts, B., *Essential cell biology*. 2nd ed. 2004, New York, NY: Garland Science Pub. xxi, 740, 102 p.
17. Anfinsen, C.B., *Principles that govern the folding of protein chains.* Science, 1973. 181(4096): p. 223-30.
18. Chothia, C. and A.M. Lesk, *The relation between the divergence of sequence and structure in proteins.* EMBO J, 1986. 5(4): p. 823-6.
19. Dessailly, B.H., et al., *PSI-2: structural genomics to cover protein domain family space.* Structure, 2009. 17(6): p. 869-81.
20. Schwede, T. and M.C. Peitch, *Computational Structural Biology: Methods and Applications*. 1st ed. 2008: World Scientific Publishing Company.
21. Altschul, S.F., et al., *Basic local alignment search tool.* J Mol Biol, 1990. 215(3): p. 403-10.
22. Marti-Renom, M.A., M.S. Madhusudhan, and A. Sali, *Alignment of protein sequences by their profiles.* Protein Sci, 2004. 13(4): p. 1071-87.
23. Karplus, K., C. Barrett, and R. Hughey, *Hidden Markov models for detecting remote protein homologies.* Bioinformatics, 1998. 14(10): p. 846-56.

24. Remmert, M., et al., *HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment.* Nat Methods, 2012. 9(2): p. 173-5.
25. Rost, B., *Twilight zone of protein sequence alignments.* Protein Eng, 1999. 12(2): p. 85-94.
26. Peng, J. and J. Xu, *RaptorX: exploiting structure information for protein alignment by statistical inference.* Proteins, 2011. 79 Suppl 10: p. 161-71.
27. Sali, A. and T.L. Blundell, *Comparative protein modelling by satisfaction of spatial restraints.* J Mol Biol, 1993. 234(3): p. 779-815.
28. Kinch, L., et al., *CASP9 assessment of free modeling target predictions.* Proteins, 2011. 79 Suppl 10: p. 59-73.
29. Leaver-Fay, A., et al., *ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules.* Methods Enzymol, 2011. 487: p. 545-74.
30. Capra, J.A. and M. Singh, *Predicting functionally important residues from sequence conservation.* Bioinformatics, 2007. 23(15): p. 1875-82.
31. Fischer, J.D., C.E. Mayer, and J. Soding, *Prediction of protein functional residues from sequence by probability density estimation.* Bioinformatics, 2008. 24(5): p. 613-20.
32. Pupko, T., et al., *Rate4Site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues.* Bioinformatics, 2002. 18 Suppl 1: p. S71-7.
33. Casari, G., C. Sander, and A. Valencia, *A method to predict functional residues in proteins.* Nat Struct Biol, 1995. 2(2): p. 171-8.
34. del Sol, A., F. Pazos, and A. Valencia, *Automatic methods for predicting functionally important residues.* J Mol Biol, 2003. 326(4): p. 1289-302.
35. Pazos, F., A. Rausell, and A. Valencia, *Phylogeny-independent detection of functional residues.* Bioinformatics, 2006. 22(12): p. 1440-8.
36. Laurie, A.T. and R.M. Jackson, *Q-SiteFinder: an energy-based method for the prediction of protein-ligand binding sites.* Bioinformatics, 2005. 21(9): p. 1908-16.
37. Binkowski, T.A. and A. Joachimiak, *Protein functional surfaces: global shape matching and local spatial alignments of ligand binding sites.* BMC Struct Biol, 2008. 8: p. 45.
38. Ghersi, D. and R. Sanchez, *EasyMIFS and SiteHound: a toolkit for the identification of ligand-binding sites in protein structures.* Bioinformatics, 2009. 25(23): p. 3185-6.
39. Huang, B. and M. Schroeder, *LIGSITEcsc: predicting ligand binding sites using the Connolly surface and degree of conservation.* BMC Struct Biol, 2006. 6: p. 19.
40. Glaser, F., et al., *A method for localizing ligand binding pockets in protein structures.* Proteins, 2006. 62(2): p. 479-88.
41. Capra, J.A., et al., *Predicting protein ligand binding sites by combining evolutionary sequence conservation and 3D structure.* PLoS Comput Biol, 2009. 5(12): p. e1000585.
42. Berezin, C., et al., *ConSeq: the identification of functionally and structurally important residues in protein sequences.* Bioinformatics, 2004. 20(8): p. 1322-4.
43. Lopez, G., et al., *firestar--advances in the prediction of functionally important residues.* Nucleic Acids Res, 2011. 39(Web Server issue): p. W235-41.
44. Brylinski, M. and J. Skolnick, *A threading-based method (FINDSITE) for ligand-binding site prediction and functional annotation.* Proc Natl Acad Sci U S A, 2008. 105(1): p. 129-34.
45. Wass, M.N., L.A. Kelley, and M.J. Sternberg, *3DLigandSite: predicting ligand-binding sites using similar structures.* Nucleic Acids Res, 2010. 38(Web Server issue): p. W469-73.
46. Roche, D.B., S.J. Tetchner, and L.J. McGuffin, *FunFOLD: an improved automated method for the prediction of ligand binding residues using 3D models of proteins.* BMC Bioinformatics, 2011. 12: p. 160.
47. Roy, A. and Y. Zhang, *Recognizing protein-ligand binding sites by global structural alignment and local geometry refinement.* Structure, 2012. 20(6): p. 987-97.

48. Yang, J., A. Roy, and Y. Zhang, *Protein-ligand binding site recognition using complementary binding-specific substructure comparison and sequence profile alignment.* Bioinformatics, 2013. 29(20): p. 2588-95.
49. Oh, M., K. Joo, and J. Lee, *Protein-binding site prediction based on three-dimensional protein modeling.* Proteins, 2009. 77 Suppl 9: p. 152-6.
50. Schmidt, T., et al., *Assessment of ligand-binding residue predictions in CASP9.* Proteins, 2011. 79 Suppl 10: p. 126-36.
51. Gallo Cassarino, T., L. Bordoli, and T. Schwede, *Assessment of ligand binding site predictions in CASP10.* Proteins, 2014. 82 Suppl 2: p. 154-63.

2. Assessment of ligand-binding residue predictions in CASP9

This chapter has been published with the title:

"*Assessment of ligand binding residue predictions in CASP9*", Tobias Schmidt[1,2], Jürgen Haas[1,2], Tiziano Gallo Cassarino[1,2], and Torsten Schwede[1,2]. Proteins, 2011. 79 Suppl 10: p. 126-36.

[1]. Biozentrum, University of Basel, Switzerland

[2]. SIB Swiss Institute of Bioinformatics, Basel, Switzerland

Contribution: I analysed the targets and assessed the biologically relevance of their ligands.

Abbreviations

MCC: Matthews' Correlation Coefficient

TBM: Template-Based Modelling

FM: Free Modelling

Abstract

Interactions between proteins and their ligands play central roles in many physiological processes. The structural details for most of these interactions, however, have not yet been characterized experientially. Therefore, various computational tools have been developed to predict the location of binding sites and the amino acid residues interacting with ligands. In this manuscript, we assess the performance of 33 methods participating in the ligand binding site prediction category in CASP9. The overall accuracy of ligand binding site predictions in CASP9 appears rather high (average MCC of 0.62 for the ten top performing groups), and compared to previous experiments more groups performed equally well. However, this should be seen in context of a strong bias in the test data towards easy template based models. Overall, the top performing methods have converged to a similar approach using ligand binding site inference

from related homologous structures, which limits their applicability for difficult “de novo” prediction targets. Here, we present the results of the CASP9 assessment of the ligand binding site category, discuss examples for successful and challenging prediction targets in CASP9, and finally suggest changes in the format of the experiment to overcome the current limitations of the assessment.

Introduction

To perform their functions, proteins interact with a plethora of small molecules within the cell. Most of these interactions are unspecific and transient in nature (e.g. interactions with water and ions), some are persistent and may play a structural or functional role (e.g. certain metal ions), and others might be transient but nevertheless highly specific, often resulting in essential changes of the protein or the ligand (e.g. enzyme-substrate complexes or receptor-ligand complexes). Hence, the identification of a protein’s functionally important residues, such as ligand binding sites or catalytic active residues, is a crucial step towards the goal of understanding the protein’s molecular function and its biological role in the cell. Although protein ligand interactions are crucial for the function of a protein, in many cases they are unknown. While the kind of ligands interacting with a protein is often known from biochemical analyses, elucidating the structural details of these interactions requires elaborate and time-consuming studies by X-ray crystallography or NMR. Therefore, computational tools have been developed aiming at predicting the precise location of binding sites, and specifically which amino acid residues in a protein are directly interacting with ligands. Various approaches for the prediction of ligand binding sites have been proposed,[1] both from structure and from sequence, based on sequence conservation [2-7], geometric criteria of the protein surface [8-12] or homology transfer from known structures.[13-17]

The function prediction category (FN) was introduced in the 6th Critical Assessment of Protein Structure Prediction (CASP), where predictions for Gene Ontology molecular function terms, Enzyme Commission numbers, and ligand binding site residues were evaluated. [18, 19] Since very little new functional information becomes available during and after the experiment, the first two categories were difficult to assess. Therefore, since CASP8 the prediction task has been to identify functionally important residues such as ligand binding residues or catalytic residues. [20] Here, we present the assessment of 33 groups participating in the recent CASP9 experiment. In the ligand binding site prediction category (FN), the sequence of a protein with unknown

structure was provided to predictors. The task was to predict the residues directly involved in ligand binding in the experimental control structure. This approach differs significantly from typical ligand binding studies (like docking or virtual screening), where the chemical identity of the ligand is given, and the correct geometric orientation of the molecule in the receptor protein is to be determined. [11, 21-24] In CASP however, the chemical identity of the ligand is unknown at the time of prediction, and only the interacting residues are predicted.

In summary, all top performing groups have applied a similar approach, using ligand information derived from homologous structures in the PDB.[25] In comparison to CASP8 [20], we could not observe a significant progress by the top groups, but rather a larger number of groups performing at the same level. We believe that this observation is caused on one side by the bias in the data set to “easy” template based predictions with only a very small number of difficult “de novo” targets in recent rounds of CASP. This gives strong advantage to methods using PDB information directly, but discourages the development of methods addressing the more challenging “de novo” cases. Another limiting factor is the binary format of the prediction task, which does not allow specifying probabilities for specific residues or differentiating between types of ligands.

Materials and Methods

Prediction targets

All CASP9 target structures were analyzed for non-solvent non-peptidic ligand groups in the deposited protein structures. Based on literature information, UniProt [26] annotations, structures of closely related homologues (Table SI, Supplementary Information), and conservation of functionally important residues, we aimed at identifying ligands with biological / functional relevance for the specific protein. All targets, including those containing ligands classified as “non-biologically relevant”, were further analyzed to indentify cases where a ligand clearly mimicked the interactions of known biologically relevant ligands for this target.

Binding site definition

For each prediction target, binding site residues were defined as those residues in direct contact with the ligand in the target structure, i.e. all protein residues with at least one heavy atom within a certain distance from any heavy atom of the ligand. The distance cutoff was defined by the

CASP organizers as the sum of the Van der Waals radii of the involved atoms plus a tolerance of 0.5 Å. In addition, different tolerance values ranging from 0 to 2.0 Å were evaluated.

In cases where multiple chains with bound ligands were present in the target structure (e.g. homo-oligomeric assemblies), the definition of the binding site residues for individual chains were combined into a single binding site definition. For targets where ligands were observed to bind in the interface between multiple chains, the oligomeric structure as defined by the authors and PISA [27] (5 cases) or only PISA (1 case) was used for the binding site definition. Analysis of structures and ligand binding sites were performed using OpenStructure (version 1.1). [28]

For targets in which only part of the relevant ligand was present, the binding site definition was extended to include the entire biologically relevant ligand. In these cases, two separate evaluations of the prediction performance were conducted. The first, denoted as 'extended binding site', all atoms of the partial and the extended ligand were used to define the binding site in the same way as described above. The second, denoted as 'partial binding site', only atoms of the partial ligand were used to define the binding site, whereas all residues exclusively in contact with the extended part of the ligand were treated as neutral and excluded from the evaluation.

Binding site prediction evaluation

As in the previous assessment,[20] binding site prediction performance was measured using the Matthews Correlation Coefficient[29] (MCC) which accounts both for over and under predictions. For each target, residue predictions were classified as true positives (TP: correctly predicted binding site residues), true negatives (TN: correctly predicted non-binding site residues), false negatives (FN: incorrectly under predicted binding site residues), false positives (FP: incorrectly over predicted non-binding site residues) based on the binding site definition described before. The MCC was computed using Eq. 1:

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP) \cdot (TP + FN) \cdot (TN + FP) \cdot (TN + FN)}}$$

The MCC ranges from +1 (perfect prediction) over 0 (random prediction) to -1 (inverse prediction). Empty submissions which did not include any binding site predictions and missing predictions were assigned a MCC score of zero.

To reduce the effects of target difficulty on the ranking, MCC scores were standardized by computing Z scores among all predictions P for a given target T using Eq. 2:

$$Z_{P,T} = \frac{MCC_{P,T} - \overline{\overline{MCC_T}}}{\sigma_T}$$

In this equation, $MCC_{P,T}$ is the raw MCC score for target T given by predictor P, $\overline{MCC_T}$ is the mean MCC score for target T, σ_T is the standard deviation of MCC scores for target T. The overall performance for each predictor was computed as the mean of Z scores over all targets, which was subsequently used for obtaining a final ranking.

In addition to the MCC score, we computed the recently published binding site distance test (BDT) [30]. BDT takes the actual three dimensional locations of the predicted residues into account and scores residues differently, according to the distance between the predicted and the observed binding site. Predictions close to the binding site score higher than more distant predictions. The BDT score ranges from 0, for a random prediction to 1, for a perfect prediction.

Robustness and significance

Statistical significance of the ranking and robustness with regard to composition of the target data set was assessed using two different methods. First, two-tailed Student's paired t-tests as well as Wilcoxon signed rank tests [31] between all predictor groups were performed based on MCC scores for each target. Both T-tests and Wilcoxon signed rank tests were performed using R (version 2.11.1). [32] Second, bootstrapping was performed, where scores were computed on a randomly selected subset of ¾ of all targets (i.e. 23 of 30 targets). 75 rounds of bootstrapping were executed for different target subsets, and for each bootstrapping experiment, mean, minimum and maximum Z scores per group were calculated as previously described. Additionally, the rank for each prediction group was calculated and mean, minimum and maximum ranks over all bootstrapping experiments were computed.

To assess the performance of groups on different types of ligands, we have analyzed the prediction performance separately on targets including only metal ions (10 targets) and on targets including only non-metal ligands (17 targets). Mixed targets including both metal and non-metal ligands (3 targets) were not considered in this sub-analysis.

Results and Discussions

Overall performance

In the CASP9 protein binding sites prediction category (FN), the predictors were given a protein sequence with unknown structure and asked to identify the residues involved in ligand binding. According to the CASP format, the predictions were binary and thus, classified each residue as either binding-site or non-binding-site residue. As defined by the organizers, only protein-small molecule interactions were considered in this category. The assessment of this category consisted of the following three steps: (1) identification of biologically relevant ligands in the target structures, (2) definition of binding site residues, (3) assessment of the prediction performance.

One dominant factor in assessing the correctness of ligand binding site prediction is the availability of experimental data, and the evaluation of the biological relevance of the specific ligand binding. Whether a certain ligand is observed in an experimental structure is first and foremost determined by the specific purification procedure, by the experimentalist's choice of using this ligand for a co-crystallization experiment, and the specific experimental conditions (ligand concentration, pH and buffer conditions, ionic strength, precipitant etc.). If a ligand is not observed in a specific experimental structure, it could still bind under different conditions, i.e. it cannot be considered as a "true negative" data point for the assessment. On the other hand, if a certain ligand is observed in a target structure, we can classify the residues within this structure into "binding" and "non binding" with regard to this specific ligand. Note that a target protein might be able to bind different ligands under different experimental conditions, and only a subset of them might be present in the target structure at hand. For example, the structure of an enzyme might be crystallized in complex with the cofactor, but without substrate or product molecules.

Although the identification of ligands in CASP9 was based only on experimentally observed ligands, it was still not straightforward to categorize their biological relevance. Although in 73% of the target structures in CASP9 various ligands were present, most of them were not considered biologically relevant but rather as originating e.g. from solvent, crystallization precipitant, or buffers. For the assessment, however, we included only ligands which we considered to be biologically relevant. The decision on biological relevance was done by manual curation, primarily based on the type and location of the ligand, literature information, and UniProt[26] annotations. In addition, information from structurally closely related homologues and conservation of functionally important residues was used to guide the selection process. Using this approach, 16 target structures with biologically relevant ligands were selected out of the 109 targets available in CASP9 for the assessment.

In addition, we have analyzed all remaining heteroatomic groups, if they occupied binding sites which mimicked the interactions of a known biologically relevant ligand for this protein. In these cases, we defined an “extended binding site” consisting of all residues in contact with the known biologically relevant ligand. We were careful to include only targets where the assignment was unambiguous, in order to avoid the inclusion of false binding site definitions. Using this approach, the number of target structures in the FN category was extended by 14, yielding a total of 30 targets in this category (Table I).

Table I Summary of CASP9 targets with bound ligands.

Target	PDB	Partial Ligand	Extended Ligand	Chemical Class	Interface	CASP Category
T0515	3MT1	SO4	PLP, LYS	Non-metal	A-B	TBM
T0516	3NO6	IMD	PF1	Non-metal		TBM
T0518	3NMB	NA		Metal		TBM
T0521	3MSE	CA, CA		Metal		TBM
T0524	3MWX	GOL	GAL	Non-metal		TBM
T0526	3NRE	PEG	GLA	Non-metal		TBM
T0529	3MWT	MN		Metal		TBM
T0539	2L0B	ZN, ZN		Metal		TBM
T0547	3NZP	PLP	PLP, LYS	Non-metal	A-B	TBM
T0548	3NNQ	ZN		Metal		TBM
T0565	3NPF	CSA	DGL, ALA	Non-metal		TBM
T0570	3NO3	MG, GOL		Metal, Non-metal		TBM
T0582	3O14	ZN		Metal		TBM
T0584	3NF2	SO4	DST, IPR	Non-metal		TBM
T0585	3NE8	ZN		Metal		TBM
T0591	3NRA	LLP		Non-metal	A-B	TBM
T0597	3NIE	ANP		Non-metal		TBM
T0599	3OS6	SO4	ISC	Non-metal		TBM
T0604	3NLC	FAD		Non-metal		TBM / FM
T0607	3PFE	ZN	ZN, BES	Metal, Non-metal		TBM
T0609	3OS7	TLA	GAL	Non-metal		TBM
T0613	3OBI	EDO	GAR, NHS	Non-metal		TBM
T0615	3NQW	MN, SO4	MN, GPX	Metal, Non-metal		TBM
T0622	3NKL	SO4	NAD	Non-metal		TBM
T0625	3ORU	ZN		Metal		TBM
T0629	2XGF	FE, FE, FE, FE, FE, FE, FE		Metal	A-B-C	FM
T0632	3NWZ	COA		Non-metal	A-B-C	TBM
T0635	3N1U	CA		Metal		TBM

T0636	3P1T	TLA	HSA, PLP	Non-metal	A-B	TBM
T0641	3NYI	STE		Non-metal		TBM

Within the selected targets, ten were found in complex with metal ions (Ca, Fe, Mg, Mn, Na, Zn), and further 17 targets in complex with non-metal ligands (Table I). The latter included amino acids and derivatives, nucleotides, sugars, fatty acids and others. Additionally, in three cases non-metal ligands were coordinated to metal ions (Mg, Mn, Zn). In most of the targets, the ligand binding site was located within a monomer, while for six targets the ligand was bound in the interface between multiple chains: T0515, T0547, T0591, T0636 (dimeric structures), T0629 (trimeric structure) and T0632 (tetrameric structure). The ligands were bound between all chains of the oligomeric structure, except for T0632 where the ligand is bound to only three of the four chains. Following the identification of biologically relevant ligands, the binding site residues for those targets were defined as those residues directly in contact with the ligand. Atoms were considered to be in contact if they were within a distance of the sum of their van der Waals radii plus a tolerance distance.

The list of binding site residues used in the assessment for each target is provided in Table SI (Supplementary Material). The tolerance distance was defined as 0.5 Å by the CASP organizers. We tested the influence of different values for the tolerance distance of the binding site definition and their influence on the assessment of prediction performance. No significant differences in the overall prediction performances were observed for different tolerance distances (Fig. S1, Supporting Information).

The majority of FN targets in CASP9 were classified as template based modeling targets (TBM), and only two targets were free modeling (FM) targets: (1) target T0629, where the ligand binding domain had no template structure (Fig. 8C), (2) target T0604, where the ligand was bound between two domains where one was a template based modeling (constituting 90% of the binding site residues) and one a free modeling domain (constituting 10% of the binding site residues). This strong bias in the data set has direct consequences for the assessment, as it is to be expected that template-based prediction methods will perform much better than “de novo” methods in this context.

In total, 33 groups made predictions in the CASP9 FN category. A summary of the predictions is given in Figure 1. Among the participating groups, 18 were registered as “human predictors” and 15 as “servers” (Table II). Most groups predicted at least 25 of the assessed 30 targets, i.e. 12

groups (6 humans, 6 servers) predicted between 25 and 29 of the assessed targets and 15 groups (6 humans, 9 servers) predicted all 30 targets; 6 human groups returned predictions for only 6 or less targets. Binding site prediction performance was measured using Z-scores of Matthews correlation coefficients (see Methods).[1]

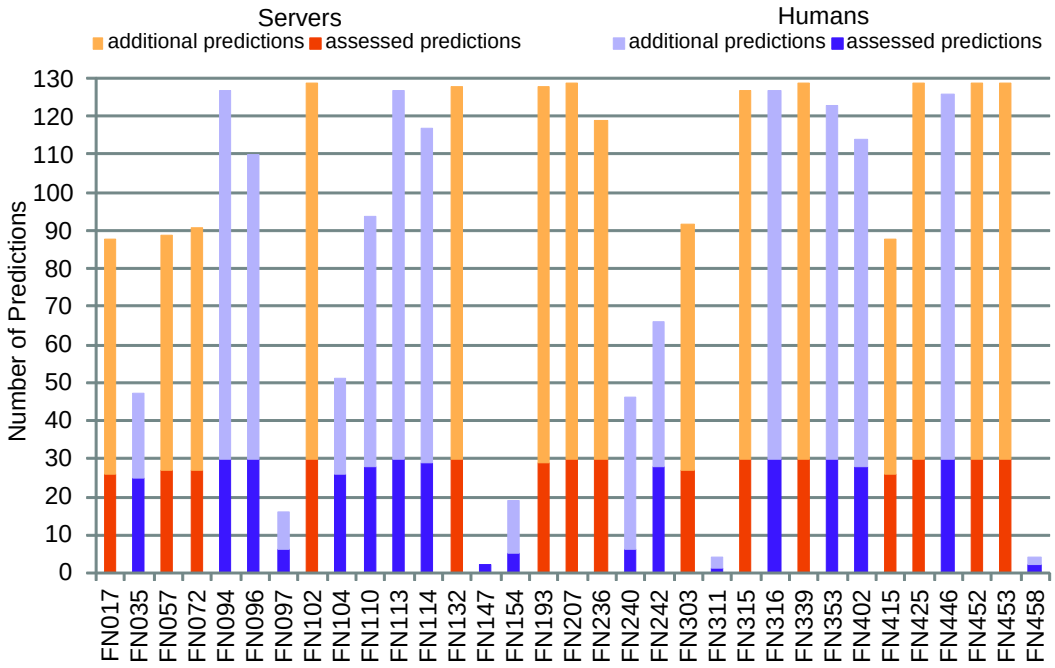

Figure 1. Overview of predictions per group. Predictions for targets which were assessed in the FN category (i.e. targets with a relevant binding site) are displayed in dark colours, additional predictions which were not assessed (i.e. targets without an experimentally confirmed binding site) are displayed in light colours. Human groups are shown in purple, servers in orange.

Table II Groups participating in the FN category in CASP9.

ID	Rank	Name	Type	Group
FN017	22	3DLIGANDSITE1	S	Michael Sternberg
FN035	5	CNIO-FIRESTAR	H	Gonzalo Lopez
FN057	21	3DLIGANDSITE3	S	Michael Sternberg
FN072	23	3DLIGANDSITE4	S	Michael Sternberg
FN094	8	MCGUFFIN	H	Liam McGuffin
FN096	1	ZHANG	H	Yang Zhang
FN097	30	KOCHANCZYK	H	Marek Kochanczyk
FN102	15	BILAB-ENABLE	S	Shugo Nakamura
FN104	7	JONES-UCL	H	David Jones

[1] As described in Materials and Methods, the authors decided that assigning a MCC score of zero to empty submissions which did not include any binding site predictions and to missing predictions would most appropriately reflect a “real life” prediction situation in the assessment. Please note that this policy has consequences for the final ranking as it penalizes methods which are not able to make predictions for some targets, and encourages the risky development of novel methods as there is no implicit penalty for making predictions for challenging targets.

FN110	6	STERNBERG	H	Michael Sternberg
FN113	9	FAMSSEC	H	Katsuichiro Komatsu
FN114	10	LEE	H	Jooyoung Lee
FN132	27	MN-FOLD	S	Chris Kauffman
FN147	28	GENESILICO	H	Janusz Bujnicki
FN154	33	JAMMING	H	Gabriel del Rio
FN193	24	MASON	S	Huzefa Rangwala
FN207	26	ATOME2_CBS	S	Jean-Luc Pons
FN236	12	GWS	S	Jooyoung Lee
FN240	32	TMD3D	H	Hiroshi Tanaka
FN242	4	SEOK	H	Chaok Seok
FN303	20	FINDSITE-DBDT	S	Jeffrey Skolnick
FN311	31	ALADEGAP	H	Kei Yura
FN315	3	FIRESTAR	S	Gonzalo Lopez
FN316	18	LOVELL_GROUP	H	Simon Lovell
FN339	2	I-TASSER_FUNCTION	S	Yang Zhang
FN353	17	SAMUDRALA	H	Ram Samudrala
FN402	13	TASSER	H	Jeffrey Skolnick
FN415	25	3DLIGANDSITE2	S	Michael Sternberg
FN425	19	INTFOLD-FN	S	Liam McGuffin
FN446	16	KIHARALAB	H	Daisuke Kihara
FN452	11	SEOK-SERVER	S	Chaok Seok
FN453	14	HHPREDA	S	Johannes Soeding
FN458	29	BILAB-SOLO	H	Mizuki Morita

The comparison between all groups is shown in Figure 2 where the error bars indicate minimum and maximum Z scores obtained by bootstrapping on a randomly selected subset of three-fourth of the targets. The error bars indicate a fluctuation in the average Z score for each group. However, in case of a correlated movement in the score, this would not influence the groups ranking. Therefore, the rank for each prediction group was computed in each bootstrapping experiment and the average, minimum and maximum rank over all bootstrapping experiments is shown in Figure 3.

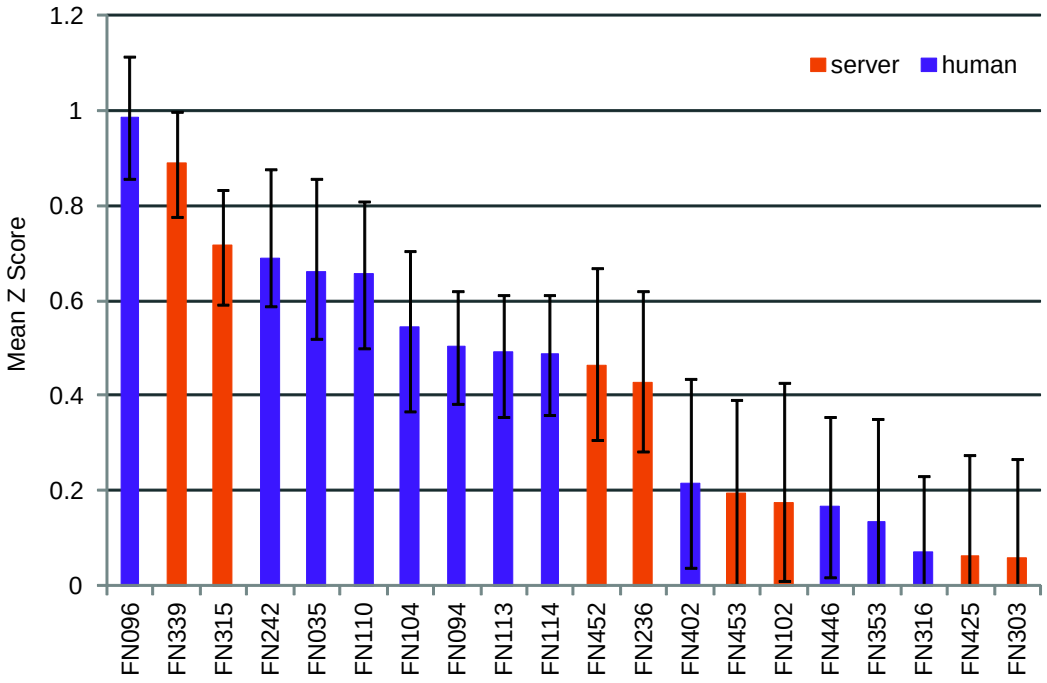

Figure 2. Mean Z scores over all targets for the top 20 predictor groups. Error bars show minimum and maximum average Z scores obtained from bootstrapping experiment. Human predictor groups are shown in purple, servers in orange.

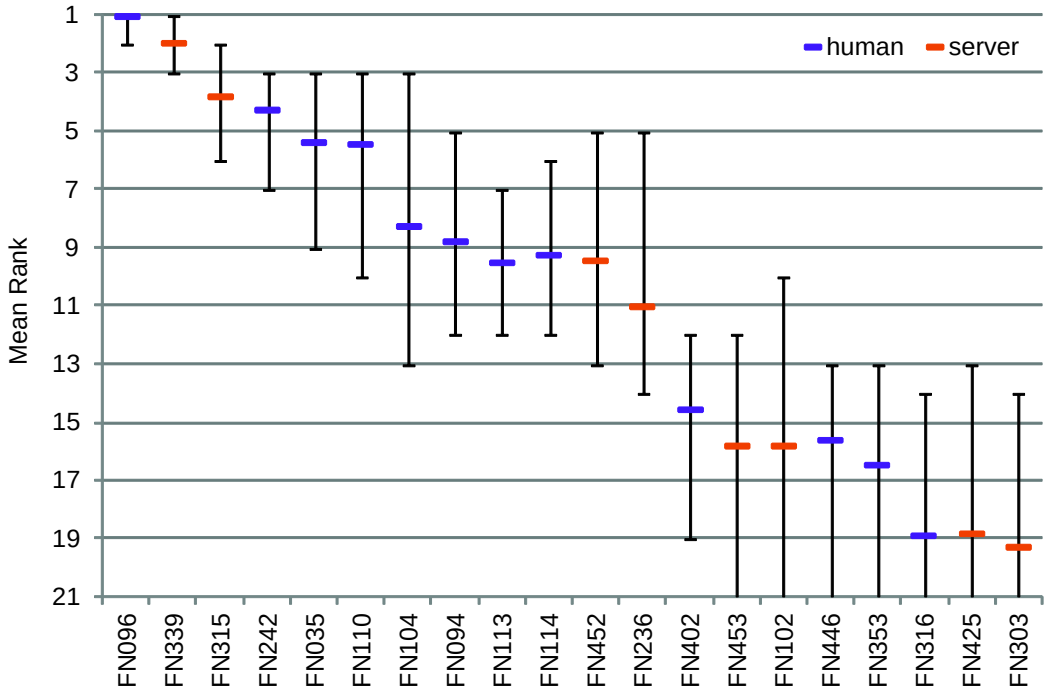

Figure 3. Mean rank based on bootstrapping experiment for the top 20 predictor groups**.** Error bars show minimum and maximum rank obtained from bootstrapping experiment. Human predictors are shown in purple, servers in orange.

The top 12 predictors clearly distinguished themselves from the following 21 groups and show a significantly better performance. Two predictors from the Zhang group (FN096, Zhang and FN339, I-TASSER_FUNCTION) show a better performance in terms of MCC compared to the following 10 groups, whereas the performance among those is comparable. Since many predictors seemed to perform similarly, statistical tests were used to assess the significance of the differences between these groups. Paired t-tests on all targets between all pairs of predictors were performed. The results are shown in Table III, with cells shaded according to computed P values. According to the t-test, the differences between the top ranked group (FN096, Zhang) and groups FN339 (I-TASSER_FUNCTION), FN242 (Seok) and FN035 (CNIO-Firestar) are not statically significant, while the differences between FN096 and the remaining predictors are significant. In addition, the non-parametric Wilcoxon signed rank test was performed, which yielded comparable results to the t-tests (Table SII, Supplementary Information).

Recently, McGuffin and coworkers published an alternative binding site distance test (BDT) [30]. Opposed to MCC, BDT takes the actual three dimensional positions of the predicted residues into account and scores residues differently, according to the distance between the predicted and the observed binding site. Hence, BDT limits the boundary effects originating from ambiguous definition of binding sites. When applying the BDT score on the predictions (Fig. S2, Supporting Information), for the top ranked groups no significant deviations to the MCC based prediction assessment were observed. [2]

Table III. P-values computed by paired t-Test of all against all predictors. Significant differences between two groups are indicated by cells with white background. For clarity, only the 12 top performing predictors are shown, sorted by their overall performance.

	FN096	FN339	FN315	FN242	FN035	FN110	FN104	FN094	FN113	FN114	FN452	FN236
FN096	-	0.24	0.01	0.08	0.06	0.01	0.01	0.00	0.00	0.00	0.00	0.00
FN339	0.24	-	0.27	0.20	0.28	0.20	0.05	0.04	0.02	0.05	0.02	0.02
FN315	0.01	0.27	-	0.81	0.56	0.63	0.17	0.20	0.03	0.14	0.12	0.07
FN242	0.08	0.20	0.81	-	0.85	0.90	0.31	0.28	0.27	0.19	0.10	0.09
FN035	0.06	0.28	0.56	0.85	-	0.88	0.44	0.52	0.38	0.45	0.45	0.31
FN110	0.01	0.20	0.63	0.90	0.88	-	0.33	0.28	0.27	0.30	0.33	0.18
FN104	0.01	0.05	0.17	0.31	0.44	0.33	-	0.88	0.88	0.89	0.93	0.93
FN094	0.00	0.04	0.20	0.28	0.52	0.28	0.88	-	0.99	0.98	0.94	0.79
FN113	0.00	0.02	0.03	0.27	0.38	0.27	0.88	0.99	-	0.99	0.95	0.76
FN114	0.00	0.05	0.14	0.19	0.45	0.30	0.89	0.98	0.99	-	0.96	0.56
FN452	0.00	0.02	0.12	0.10	0.45	0.33	0.93	0.94	0.95	0.96	-	0.83
FN236	0.00	0.02	0.07	0.09	0.31	0.18	0.93	0.79	0.76	0.56	0.83	-

As described earlier, for 14 targets, the partial binding sites were individually extended around the observed ligand to reflect a binding site accommodating the most probable biologically

[2] The largest change in ranking by 3 positions would be for group FN110.

relevant ligand. To investigate the influence of this extension, the assessment was performed both on all residues of the extended binding site and separately on all the residues of the partial binding site while treating the residues exclusively in the extended binding site as “neutral” for the analysis. For the top ranked groups no significant differences in the overall prediction performances were observed between partial and extended binding site definitions (Fig.3 Supporting Information)[3].

Assessment by type of binding sites

In addition to the overall performance, subsets of the targets were evaluated individually, according to the ligand chemotype. The distinct chemical properties of metal ions and organic ligands give raise to diverse binding sites. Thus, it could be expected that various prediction methods perform differently. To address this question, we have analyzed the prediction performance separately on all targets including only metal ligands (10 targets) and on targets including only non-metal ligands (17 targets). The mean Z-score per group separated into metal and non-metal targets are shown in Figure 4. Within the top 10 groups most of them show a better performance for non-metal targets, with the exception of FN242 (Seok) and FN114 (Lee). Especially group FN114 shows a better performance on metal ligands, compared to an average performance on the full set of targets.

Among the CASP9 FN targets, in six cases the ligand binds in the interface between multiple chains of an oligomeric protein complex. Although, the number of interface targets is very limited, we were interested in the question if the prediction of ligand binding sites of interface targets is more difficult than non-interface targets. We compared the average prediction performance, both according to mean MCC values, as well as the number of very good predictions (MCC > 0.85), for interface vs. non-interface targets. No significant difference was observed, thus on average, in those target categories it seems equally difficult to predict the binding site residues. However, it should be considered that four of the six targets are “trivial” oligomers, where a simple blast query returns a homologues template-ligand complex with the correct oligomeric state.

[3] The largest difference was observed for group FN113 which would change rank by 3 positions.

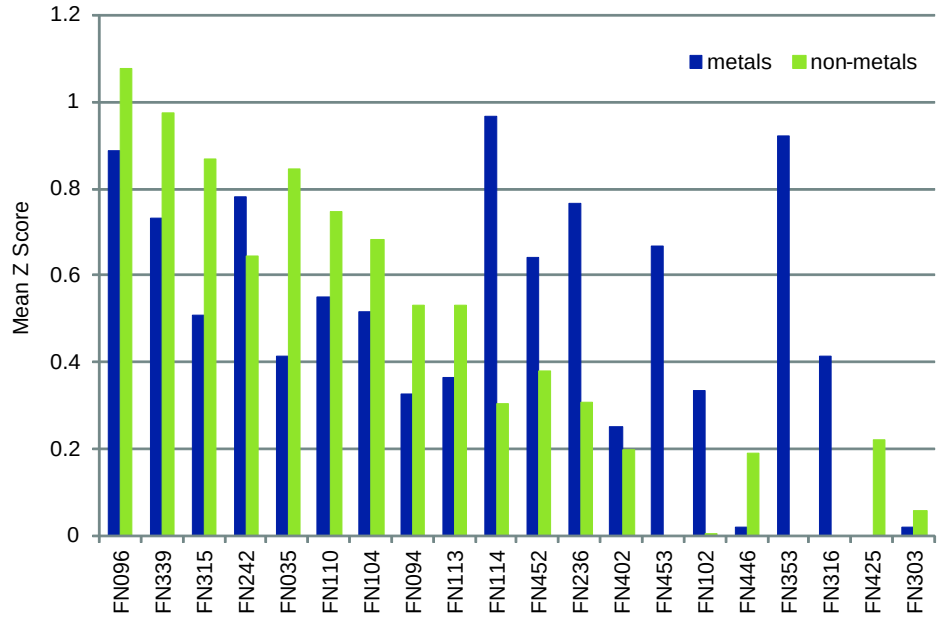

Figure 4. Mean Z scores of the top 20 groups, separated by the ligand chemotype. Metals are shown in blue, non-metals are shown in green.

Human versus server prediction

Looking at the top 10 groups, 8 of them were registered as “humans”, and only 2 as “servers”. Overall, there is a striking difference between the average performance of human groups and server groups with a mean Z score of 0.47 and 0.15, respectively. Although predictor groups registered as “human” performed considerably better than “servers”, the role of human beings in the prediction process was difficult to evaluate.

Several aspects seemed to contribute to this observation: Human predictors had access to multiple servers for structure modeling and various server binding site predictions, while server predictors have to rely on their own method only. While human predictors can make use of additional annotation from biological knowledgebases and scientific literature, servers have to rely on structured machine-readable information. A major bottleneck in this context seems the lack of consistent annotation of ligands found in PDB entries with respect to their biological relevance. It appears that human predictors benefit from the longer prediction time mainly by their ability to distinguish relevant from irrelevant ligand predictions.

Prediction methods have converged to similar approach

When comparing the methods of the top performing groups, it seems they have converged to similar approaches, which are based on homology transfer from related structures in the PDB.

By identifying homologous protein structures with bound ligands, putative binding site residues in the target model are classified by spatial proximity after alignment or superposition. The methods differ in their specific implementations with regards to the underlying structure databases (PDB vs. curated binding site libraries), target representation (alignment to structure vs. full atomic models), superposition to related structures to identify putative binding sites, and the use of residue conservation information in the prediction process.

The major draw-back of these homology-based inference methods is that they rely on the availability of related protein structures with bound ligands and are thus unable to make predictions for novel proteins without prior ligand information.

Although many groups have used similar approaches to make their predictions, we observed a surprising heterogeneity of performance within targets. As shown in Figure 5 (and Figure S4), the 12 top performing groups show overall a similar spectrum of results, with a few nearly perfectly predicted targets and some poorly predicted targets.

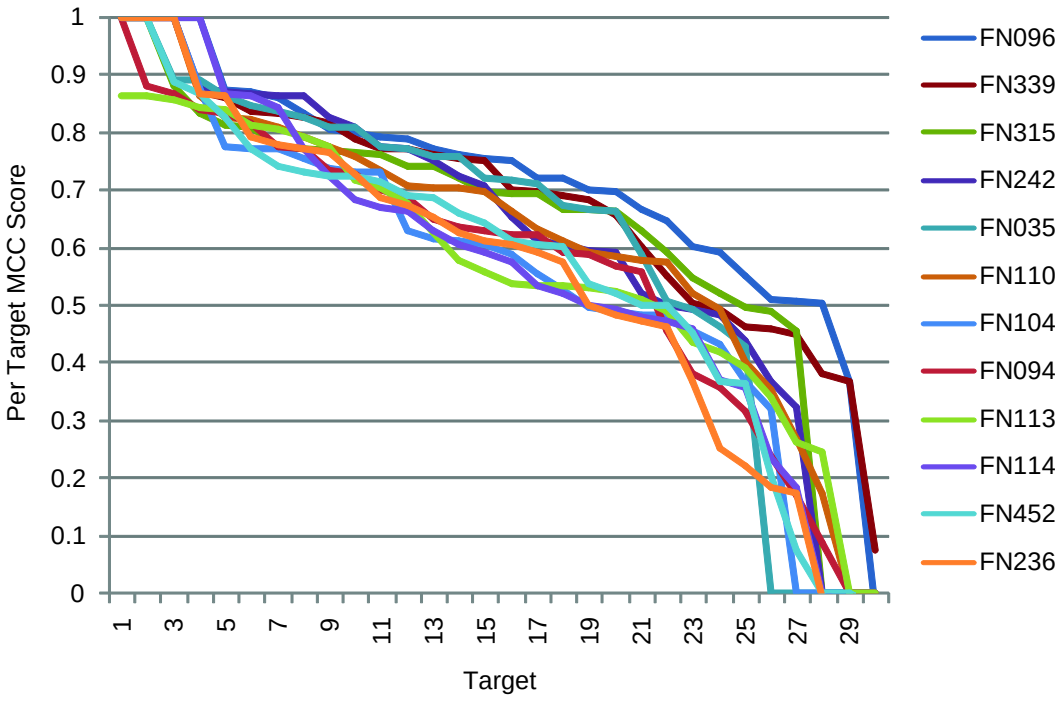

Figure 5. MCC scores for the 12 top performing groups for all targets. Targets were sorted by their respective MCC score, individually for each group.

Interestingly, when analyzing the results for individual targets, at least one good prediction was achieved across all groups (MCC value of at least 0.56; on average 0.84; see Fig. 6), and even predictors with a poor overall performance, can yield the best individual prediction for certain

targets, as shown in Figure 7. Thus, either the performance of the different methods is highly target specific, or there is a considerable random component in the prediction process in combination with a strong influence by the small and biased target data set.

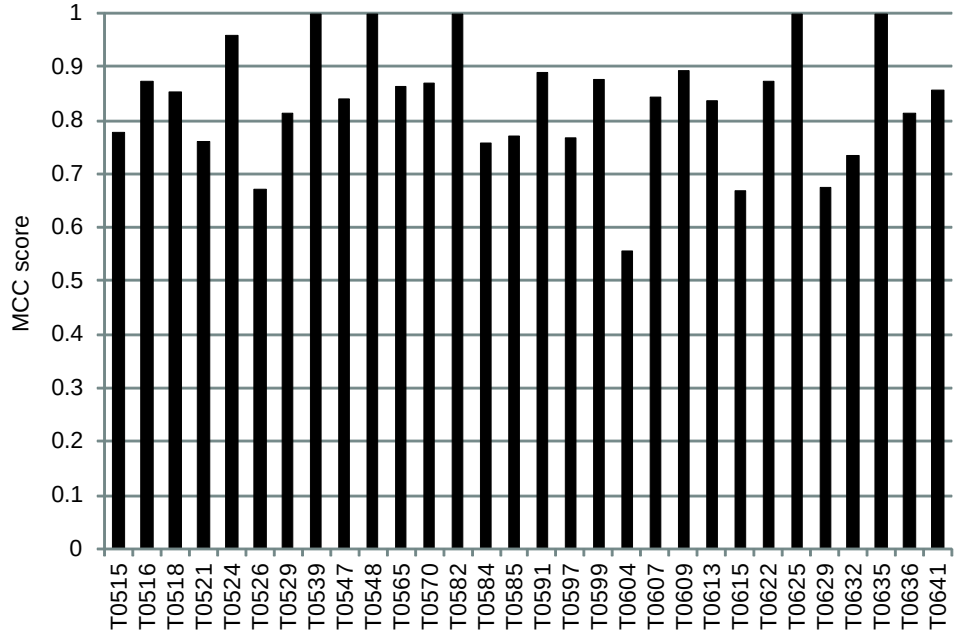

Figure 6. Overall target difficulty. MCC value of the best overall prediction for each target.

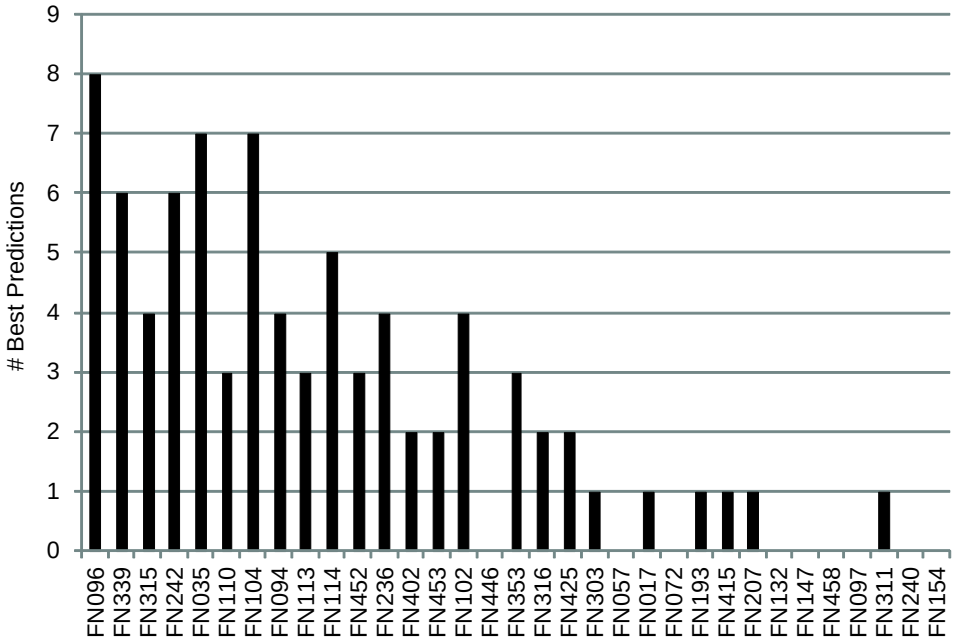

Figure 7. Number of targets where a particular group returned the best prediction. Groups are sorted by their overall performance. For one target, multiple groups can perform equally.

Prediction examples

Obviously, target T0604 was the most difficult target in the FN category in CASP9, with a maximum MCC score of 0.56 for the best prediction, and an average score of 0.29. The protein is a putative FAD dependent oxidoreductase with a bound FAD molecule (PDB: 3nlc). The protein is monomeric and forms a large binding pocket for the ligand. The structure is shown in Figure 8(A) together with the binding site predictions of group FN035 (CNIO-FIRESTAR) as one of the best predictions for this target. The top performing methods were able to accurately predict the lower part of the binding site around the adenine moiety, whereas all of them failed for the upper part of the binding site around the flavin moiety. This stems from the fact that this target structure has only remote homologues, which differ significantly in the flavin binding site region. This example clearly demonstrates the limitations of prediction methods that are based on homology transfer.

Target T0629 is the only target in the current ligand binding target set which was classified as free modeling target and thus has no template structure. The protein (PDB: 2xgf) is the bacteriophage T4 long tail fiber receptor-binding tip. It contains a long fiber like structure which is formed by three chains and binds seven iron atoms. Each iron atom is complexed with six histidine residues. Each protein chain contributes two histidines to each binding site, where the two histidines are in a His-X-His motive, with X being any of Ser, Thr or Gly. The target structure is shown in Figure 8(C) together with the binding site predictions of group FN114 (LEE), the best predictor for this target among the top 10. Common to all predictions for this target is that they correctly predicted a subset of the seven binding sites – most likely due to local similarity to another metal binding protein with a His-X-His motif, but no predictor identified all sites correctly.

The structure of target T0632 (PDB:3nwz) is a homo-tetramer which binds coenzyme-A. This ligand is interacting with three of the four chains of the protein, which seems to present a challenge for binding site residue prediction observed by a low average MCC of 0.22. An excellent prediction was obtained by group FN096 (Zhang) with an MCC of 0.72, which is depicted in Figure 8(B) along with the target structure. Many residues were well predicted despite originating from different chains. In this prediction, the largest errors originate from missing some binding site residues due to an elongated terminus compared to structurally closely related templates.

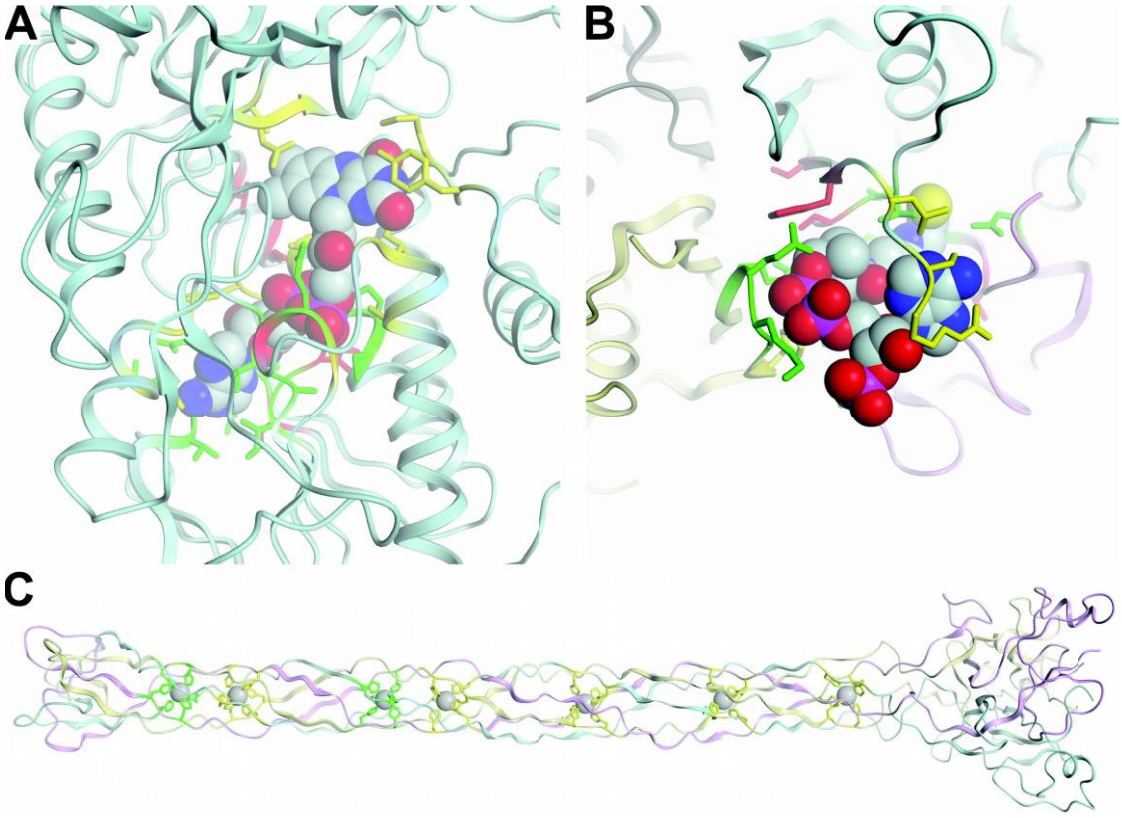

Figure 8. Examples of binding site predictions. All ligands are shown in spheres render mode. The protein backbone is shown in cartoon mode with each chain colored separately. All side chains of observed and predicted binding site residues are shown in licorice sticks. Correctly predicted residues (true positives) are colored in green, incorrectly under predicted binding site residues (false negatives) in yellow and incorrectly over predicted non-binding site residues (false positives) in red. (A) Target T0604 with predictions of group FN035. (B) Predictions of group FN096 for target T0632. (C) Group FN114's predictions for target T0629.

Conclusion

The task of predicting binding sites from a protein's sequence is of high relevance for life science research, ranging from functional characterization of novel proteins to applications in drug design, and consequently the ligand binding site prediction category in CASP has received increasing attention over the past years. In CASP9 it attracted a total of 33 predictors - ten more groups than in CASP8. In contrast to the previous CASPs, where only three predictors yielded reliable predictions,[20] in this assessment nearly half of the prediction groups yielded reliable predictions for the majority of targets. Two groups (FN096, Zhang; FN339, I-TASSER_FUNCTION) performed better than the rest (when accounting for missing target predictions in the assessment), while the following ten prediction groups performed comparably well. This is not very surprising with respect to the observation that in this round all top

performing groups based their methods on approaches, which are similar to the best performing strategy in previous CASP experiments (i.e. Sternberg[33] and LEE[15]).

Limitation of the current format and recommendations for future experiments

The very low number of target structures with relevant ligands is a major limitation to the assessment as it does not allow to draw significant conclusions on the specific strengths and weakness of different prediction methods, e.g. with regard to target difficulty or type of the ligands. Only 30 of the total 109 CASP9 targets (28%) were considered to have a biologically relevant ligand bound in the target structures and were thus assessed in the FN category. It is likely that some of the remaining target proteins would bind interesting ligands under different experimental conditions, but such conclusions cannot be made with the available data. In the previous CASP8 experiment, the total number of targets in this category was 27, illustrating that this is a recurring problem - and not specific to this round of CASP. Another rather drastic limitation of the FN category is the binary prediction format which classifies residues as either ligand binding/non-binding based on a hard distance cutoff. Consequently, all ligands are currently treated uniformly, independent of their chemical type, and all potential binding sites are treated uniformly, independent of their affinity (or binding probability) for different ligands. Moreover, most targets in the FN category were straightforward TBM targets with numerous, closely related template structures, and only one of the 30 targets was categorized as free modeling (FM). However, exactly this class of target structures is of highest interest for computational ligand binding site prediction, where no obvious information about the location of their binding sites is available. We would like to suggest the following modifications to the assessment of ligand binding site predictions to enable the community to benefit even further from future rounds of this experiment:

- In order to accumulate a sufficiently large number of prediction targets, the assessment of this category should be done continuously based on a weekly PDB pre-release. This would allow assessing the performance in different ranges of target difficulty, similar to other CASP categories, and facilitate analyzing the strengths and weakness of different approaches. During the CASP meeting in Asilomar, we have suggested that the CAMEO project (Continuous Automated Model EvaluatiOn) of the Protein Model Portal [34] could contribute to this effort.

- Binding sites differ chemically and structurally from each other e.g. a metal ion binding site has different characteristics compared to e.g. a sugar binding site. We therefore

suggest that the assessment of binding site residue predictions should be made according to chemotype categories of the ligand expected to be bound. We would like to propose the following categories: “metal ions” (e.g. Na, Ca, Zn, Fe, Mn, Mg, etc.), “inorganic anions” (e.g. SO4, PO4), “DNA/RNA” for poly-ribonucleic acid binding sites, and “organic ligands” for cofactors, substrates and receptor agonists/antagonists (e.g. NAD, FAD, ATP, SAM, CoA, PLP, etc.). More fine grained assessment categories might be necessary if more specific prediction methods emerge in the future.

- The binary prediction of binding site residues should be replaced by a continuous probability measure, thus reflecting the likelihood for a residue to be involved in binding a ligand of a certain type. For example a certain residue might be predicted as having a high probability to bind a metal ion, but a low probability to bind an organic ligand. The assessment of continuous prediction variable (e.g. using ROC type analysis) would better reflect the spectrum of “high affinity” and “low affinity” sites of different types.

- The experimentalist solving a protein structure typically will have more insights and experimental evidence for the biological role and relevance of ligands observed in a protein structure than the information which is publicly available to assessors during the CASP experiment. It would therefore be beneficial to capture the information about the biological role of “HETATM” records during PDB deposition.

Predicting binding sites from a protein’s sequence has the potential for yielding high impact on life science research – if the predictions are specific and accurate enough to help addressing relevant biological questions. We hope that with the suggested modifications, the assessment of ligand binding site predictions will be more suited to evaluate the current state of the art of prediction methods, identify possible bottlenecks, and further stimulate the development of new methods.

Acknowledgements

The authors would like to thank the experimental groups for providing the target structures for the CASP9 experiments, and all predictors for their participation. We are especially grateful to Mike Sternberg and Johannes Söding for fruitful discussions on ligand binding site prediction and assessment. This work was partially supported by the SIB Swiss Institute of Bioinformatics and by grant U01 GM093324-01 from the National Institute of General Medical Sciences.

Supplementary material

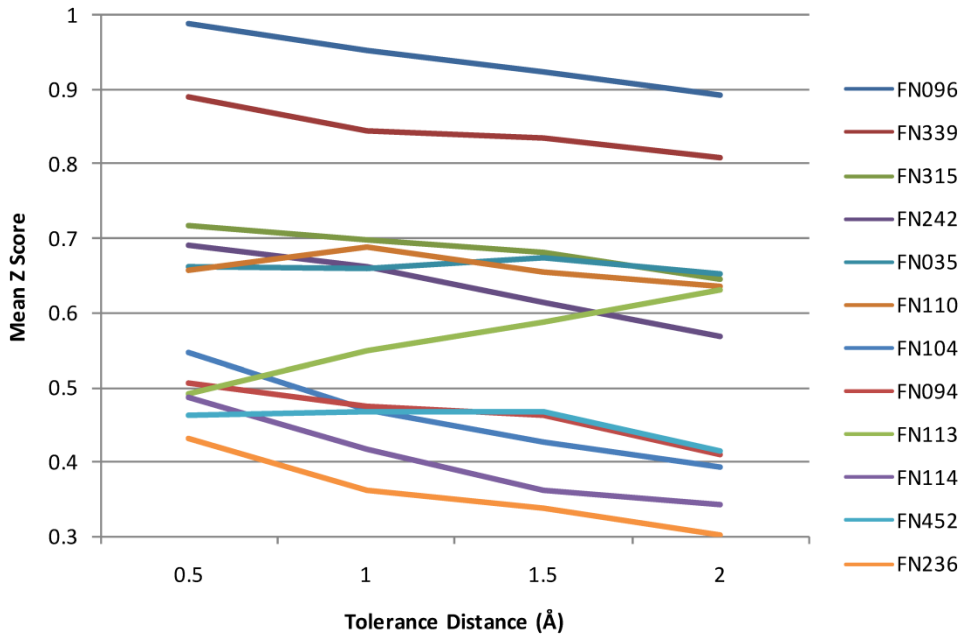

Figure S1. Influence of different binding site definitions on the prediction performance of the top 12 predictors. Mean Z score are shown for different tolerance distance used for binding site definition. All residues with a least one atom within the sum of the van der Waals radius plus the tolerance distance were considered to belong to the binding site.

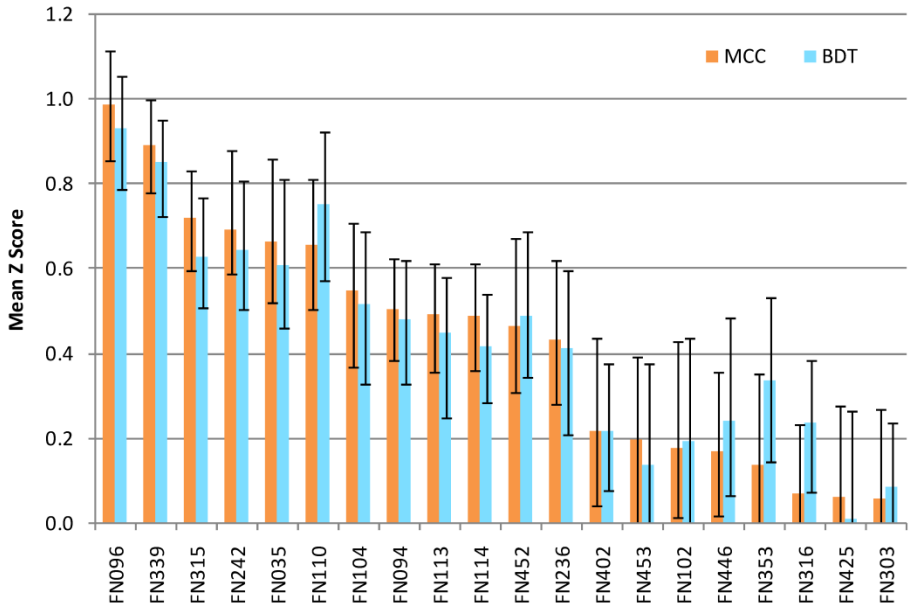

Figure S2. Comparison between the overall prediction performance evaluated using the Mathews Correlation Coefficient (MCC, in orange) and the Binding site Distance Test (BDT, in cyan). Overall prediction performance is shown in mean Z Scores over all targets.

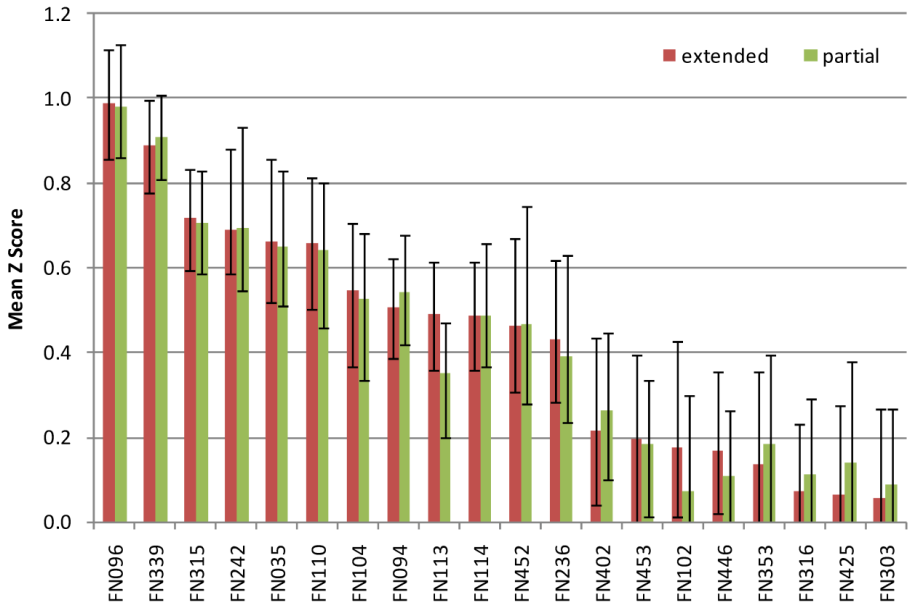

Figure S3. Comparison of the overall prediction performance observed with the extended (red) and the partial (green) binding site definitions. Overall prediction performance is shown in mean Z Scores over all targets.

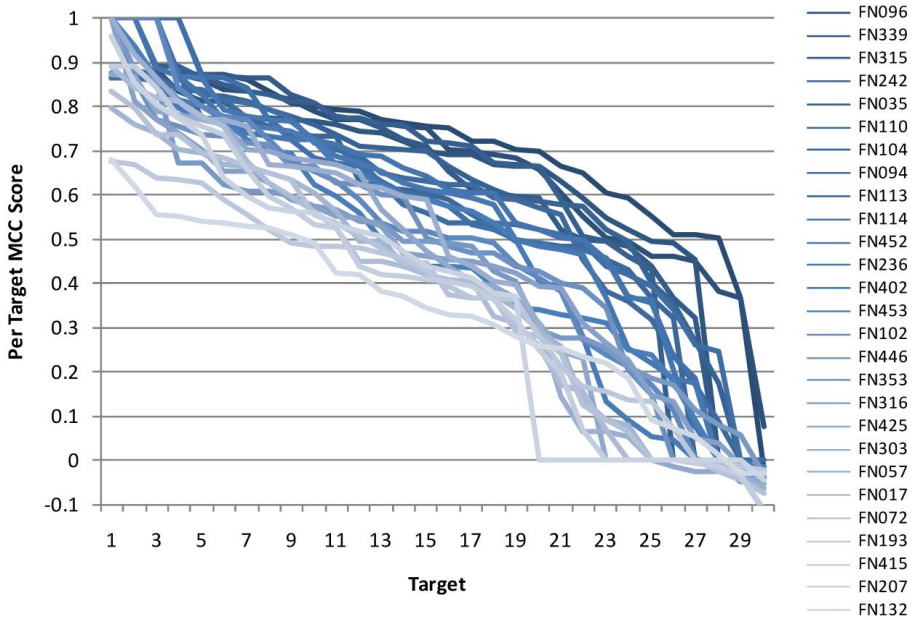

Figure S4. MCC scores for all groups with at least 10 predictions for all targets. Targets were sorted by their respective MCC score, individually for each group.

Table SI Definition of binding site residues used in the assessment. [a] PDB id for the deposited target structure which was used to define the partial binding site. [b] PDB id of the protein structure related to the target structure, used in addition to define the extended binding site.

Target	Target PDB id [a]	Related PDB id [b]	Partial Binding Site	Extended Binding Site	Neutral Residues
T0515	3MT1	1TWI	198, 199, 232, 233, 324	35, 37, 162, 198, 199, 230, 232, 233, 269, 273, 324, 328	35, 37, 162, 230, 269, 273, 328
T0516	3NO6	2QCX	44, 47, 48, 51, 113, 137, 141, 208, 211	44, 47, 48, 51, 113, 137, 138, 141, 167, 171, 174, 208, 211	138, 167, 171, 174
T0518	3NMB		132, 133, 134, 162, 164, 271, 273	132, 133, 134, 162, 164, 271, 273	
T0521	3MSE		48, 50, 52, 54, 59, 117, 121, 123, 128	48, 50, 52, 54, 59, 117, 121, 123, 128	
T0524	3MWX	1SO0	73, 74, 177, 203, 230, 269, 271, 283	62, 63, 73, 74, 100, 101, 177, 203, 230, 269, 271, 283, 285	62, 63, 100, 101, 285
T0526	3NRE	1NS0	56, 83, 148, 173, 200, 241, 253	43, 56, 77, 83, 148, 173, 200, 241, 253	43, 77
T0529	3MWT		389, 390, 391, 533	389, 390, 391, 533	
T0539	2L0B		33, 36, 51, 53, 56, 59, 70, 73	33, 36, 51, 53, 56, 59, 70, 73	
T0547	3NZP	1TWI	84, 86, 87, 132, 231, 233, 236, 273, 274, 320, 321, 322, 323, 483, 519	84, 86, 87, 132, 231, 233, 236, 273, 274, 320, 321, 322, 323, 452, 483, 484, 519	452, 484
T0548	3NNQ		58, 62, 95, 98	58, 62, 95, 98	
T0565	3NPF	3H41	191, 193, 202, 203, 262, 263	54, 80, 191, 193, 194, 202, 203, 204, 221, 222, 262, 263	54, 80, 194, 204, 221, 222
T0570	3NO3		30, 59, 61, 123, 156, 158, 178, 222	30, 59, 61, 123, 156, 158, 178, 222	
T0582	3O14		58, 60, 64, 94	58, 60, 64, 94	
T0584	3NF2	1RQI	55, 58, 87, 104	55, 58, 87, 90, 91, 94, 103, 104, 155, 183, 184, 221, 248	90, 91, 94, 103, 155, 183, 184, 221, 248
T0585	3NE8		13, 28, 82, 84, 115	13, 28, 82, 84, 115	
T0591	3NRA		109, 110, 111, 135, 185, 189, 217, 219, 251, 252, 260, 283	109, 110, 111, 135, 185, 189, 217, 219, 251, 252, 260, 283	
T0597	3NIE		36, 37, 38, 39, 40, 41, 42, 44, 57, 59, 73, 112, 114, 117, 120, 158, 160, 163, 174	36, 37, 38, 39, 40, 41, 42, 44, 57, 59, 73, 112, 114, 117, 120, 158, 160, 163, 174	
T0599	3OS6	3HWO	213, 214, 215, 364, 377, 381	212, 213, 214, 215, 276, 304, 328, 347, 348, 362, 364, 377, 381	212, 276, 304, 328, 347, 348, 362

T0604	3NLC		113, 114, 116, 117, 118, 137, 138, 166, 167, 170, 171, 172, 174, 175, 177, 178, 180, 241, 242, 243, 272, 273, 274, 277, 280, 352, 364, 365, 515, 516, 523, 524, 527	113, 114, 116, 117, 118, 137, 138, 166, 167, 170, 171, 172, 174, 175, 177, 178, 180, 241, 242, 243, 272, 273, 274, 277, 280, 352, 364, 365, 515, 516, 523, 524, 527	
T0607	3PFE	2ZOF	96, 129, 163, 190, 442	96, 129, 162, 163, 165, 190, 191, 205, 340, 410, 412, 413, 414, 442	162, 165, 191, 205, 340, 410, 412, 413, 414
T0609	3OS7	1Z45	67, 69, 108, 184, 288	67, 69, 80, 81, 108, 184, 245, 286, 288, 300	80, 81, 245, 286, 300
T0613	3OBI	1C2T	177, 178, 225, 229, 230	173, 174, 175, 176, 177, 178, 183, 192, 193, 194, 223, 225, 226, 229, 230	173, 174, 175, 176, 183, 192, 193, 194, 223, 226
T0615	3NQW	1VJ7	33, 36, 62, 63, 98, 102, 123, 127, 139, 147	26, 33, 36, 62, 63, 98, 102, 123, 127, 128, 130, 131, 139, 143, 147	26, 128, 130, 131, 143
T0622	3NKL	2VT2	10, 11, 12, 72	7, 9, 10, 11, 12, 33, 34, 35, 38, 69, 70, 71, 72, 77, 81	7, 9, 33, 34, 35, 38, 69, 70, 71, 77, 81
T0625	3ORU		126, 143, 207	126, 143, 207	
T0629	2XGF		73, 75, 105, 107, 119, 121, 156, 158, 170, 172, 179, 181, 188, 190	73, 75, 105, 107, 119, 121, 156, 158, 170, 172, 179, 181, 188, 190	
T0632	3NWZ		76, 83, 109, 110, 117, 118, 119, 120, 134, 136, 137, 138, 139, 164, 166, 167	76, 83, 109, 110, 117, 118, 119, 120, 134, 136, 137, 138, 139, 164, 166, 167	
T0635	3N1U		25, 27, 118	25, 27, 118	
T0636	3P1T	1GEX	22, 47, 145, 172, 197, 301, 306, 313	22, 47, 75, 76, 145, 169, 171, 172, 194, 196, 197, 205, 206, 225, 301, 306, 313	75, 76, 169, 171, 194, 196, 205, 206, 225
T0641	3NYI		30, 65, 66, 67, 96, 97, 127, 128, 129, 164, 166, 179, 204, 241, 272, 275, 279, 286	30, 65, 66, 67, 96, 97, 127, 128, 129, 164, 166, 179, 204, 241, 272, 275, 279, 286	

Table SII P-values computed by Wilcoxon Signed-Rank Test of all against all predictors. Significant differences between two groups are indicated by cells with white background. For clarity, only the 12 top performing predictors are shown, sorted by their overall performance.

	FN096	FN339	FN315	FN242	FN035	FN110	FN104	FN094	FN113	FN114	FN452	FN236
FN096	-	0.11	0.01	0.16	0.14	0.00	0.01	0.00	0.00	0.01	0.00	0.00
FN339	0.11	-	0.26	0.27	0.49	0.12	0.06	0.04	0.02	0.06	0.01	0.02
FN315	0.01	0.26	-	0.76	0.23	0.79	0.39	0.10	0.05	0.18	0.16	0.09
FN242	0.16	0.27	0.76	-	0.91	0.46	0.23	0.12	0.21	0.21	0.05	0.08
FN035	0.14	0.49	0.23	0.91	-	0.92	0.39	0.10	0.10	0.18	0.18	0.17
FN110	0.00	0.12	0.79	0.46	0.92	-	0.52	0.24	0.13	0.50	0.41	0.29
FN104	0.01	0.06	0.39	0.23	0.39	0.52	-	0.78	0.71	0.37	0.73	0.50
FN094	0.00	0.04	0.10	0.12	0.10	0.24	0.78	-	0.68	0.94	0.67	0.72
FN113	0.00	0.02	0.05	0.21	0.10	0.13	0.71	0.68	-	0.87	0.95	0.64
FN114	0.01	0.06	0.18	0.21	0.18	0.50	0.37	0.94	0.87	-	0.83	0.49
FN452	0.00	0.01	0.16	0.05	0.18	0.41	0.73	0.67	0.95	0.83	-	0.87
FN236	0.00	0.02	0.09	0.08	0.17	0.29	0.50	0.72	0.64	0.49	0.87	-

References

1. Gherardini, P.F. and M. Helmer-Citterich, *Structure-based function prediction: approaches and applications.* Brief Funct Genomic Proteomic, 2008. 7(4): p. 291-302.
2. Berezin, C., et al., *ConSeq: the identification of functionally and structurally important residues in protein sequences.* Bioinformatics, 2004. 20(8): p. 1322-4.
3. Casari, G., C. Sander, and A. Valencia, *A method to predict functional residues in proteins.* Nat Struct Biol, 1995. 2(2): p. 171-8.
4. del Sol, A., F. Pazos, and A. Valencia, *Automatic methods for predicting functionally important residues.* J Mol Biol, 2003. 326(4): p. 1289-302.
5. Fischer, J.D., C.E. Mayer, and J. Soding, *Prediction of protein functional residues from sequence by probability density estimation.* Bioinformatics, 2008. 24(5): p. 613-20.
6. Innis, C.A., *siteFiNDER|3D: a web-based tool for predicting the location of functional sites in proteins.* Nucleic Acids Res, 2007. 35(Web Server issue): p. W489-94.
7. Pazos, F., A. Rausell, and A. Valencia, *Phylogeny-independent detection of functional residues.* Bioinformatics, 2006. 22(12): p. 1440-8.
8. Glaser, F., et al., *A method for localizing ligand binding pockets in protein structures.* Proteins, 2006. 62(2): p. 479-88.
9. Hendlich, M., F. Rippmann, and G. Barnickel, *LIGSITE: automatic and efficient detection of potential small molecule-binding sites in proteins.* J Mol Graph Model, 1997. 15(6): p. 359-63, 389.
10. Hernandez, M., D. Ghersi, and R. Sanchez, *SITEHOUND-web: a server for ligand binding site identification in protein structures.* Nucleic Acids Res, 2009. 37(Web Server issue): p. W413-6.
11. Huang, N., B.K. Shoichet, and J.J. Irwin, *Benchmarking sets for molecular docking.* J Med Chem, 2006. 49(23): p. 6789-801.
12. Laskowski, R.A., *SURFNET: a program for visualizing molecular surfaces, cavities, and intermolecular interactions.* J Mol Graph, 1995. 13(5): p. 323-30, 307-8.
13. Brylinski, M. and J. Skolnick, *A threading-based method (FINDSITE) for ligand-binding site prediction and functional annotation.* Proc Natl Acad Sci U S A, 2008. 105(1): p. 129-34.
14. Lopez, G., A. Valencia, and M.L. Tress, *firestar - prediction of functionally important residues using structural templates and alignment reliability.* Nucleic Acids Research, 2007. 35: p. W573-W577.
15. Oh, M., K. Joo, and J. Lee, *Protein-binding site prediction based on three-dimensional protein modeling.* Proteins, 2009. 77 Suppl 9: p. 152-6.
16. Pandit, S.B., et al., *PSiFR: an integrated resource for prediction of protein structure and function.* Bioinformatics, 2010. 26(5): p. 687-8.
17. Wass, M.N., L.A. Kelley, and M.J. Sternberg, *3DLigandSite: predicting ligand-binding sites using similar structures.* Nucleic Acids Res, 2010. 38(Web Server issue): p. W469-73.
18. Soro, S. and A. Tramontano, *The prediction of protein function at CASP6.* Proteins, 2005. 61 Suppl 7: p. 201-13.
19. Lopez, G., et al., *Assessment of predictions submitted for the CASP7 function prediction category.* Proteins, 2007. 69 Suppl 8: p. 165-74.
20. Lopez, G., I. Ezkurdia, and M.L. Tress, *Assessment of ligand binding residue predictions in CASP8.* Proteins, 2009. 77 Suppl 9: p. 138-46.
21. Huang, S.Y., S.Z. Grinter, and X. Zou, *Scoring functions and their evaluation methods for protein-ligand docking: recent advances and future directions.* Phys Chem Chem Phys, 2010. 12(40): p. 12899-908.

22. Leach, A.R., B.K. Shoichet, and C.E. Peishoff, *Prediction of protein-ligand interactions. Docking and scoring: successes and gaps.* J Med Chem, 2006. 49(20): p. 5851-5.
23. Shoichet, B.K., *Virtual screening of chemical libraries.* Nature, 2004. 432(7019): p. 862-5.
24. Warren, G.L., et al., *A critical assessment of docking programs and scoring functions.* J Med Chem, 2006. 49(20): p. 5912-31.
25. Berman, H., et al., *The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data.* Nucleic Acids Res, 2007. 35(Database issue): p. D301-303.
26. The UniProt Consortium, *Ongoing and future developments at the Universal Protein Resource.* Nucleic Acids Research, 2011. 39(suppl 1): p. D214-D219.
27. Krissinel, E. and K. Henrick, *Inference of macromolecular assemblies from crystalline state.* J. Mol. Biol., 2007. 372: p. 774-797.
28. Biasini, M., et al., *OpenStructure: a flexible software framework for computational structural biology.* Bioinformatics, 2010. 26: p. 2626-2628.
29. Matthews, B.W., *Comparison of the predicted and observed secondary structure of T4 phage lysozyme.* Biochim Biophys Acta, 1975. 405(2): p. 442-51.
30. Roche, D.B., S.J. Tetchner, and L.J. McGuffin, *The binding site distance test score: a robust method for the assessment of predicted protein binding sites.* Bioinformatics, 2010. 26(22): p. 2920-2921.
31. Wilcoxon, F., *Individual Comparisons by Ranking Methods.* Biometrics Bulletin, 1945. 1(6): p. 80-83.
32. R Development Core Team, *R: A language and environment for statistical computing*. 2011, R Foundation for Statistical Computing: Vienna, Austria.
33. Wass, M.N. and M.J. Sternberg, *Prediction of ligand binding sites using homologous structures and conservation at CASP8.* Proteins, 2009. 77 Suppl 9: p. 147-51.
34. Arnold, K., et al., *The Protein Model Portal.* J Struct Funct Genomics, 2009. 10(1): p. 1-8.

3. Assessment of ligand binding site predictions in CASP10

This chapter has been published as:

"*Assessment of ligand binding site predictions in CASP10*", Tiziano Gallo Cassarino[1,2], Lorenza Bordoli[1,2] and Torsten Schwede[1,2]

[1]. Biozentrum, University of Basel, Switzerland

[2]. SIB Swiss Institute of Bioinformatics, Basel, Switzerland

Contribution: I analysed the targets and their biological function, assessed the biologically relevance of their ligands, evaluated the participants' performances, measured the statistical significance of the assessment and wrote the manuscript.

Abbreviations

MCC: Matthews Correlation Coefficient

TBM: Template based modeling category in CASP

FM: Free-modeling category in CASP

Abstract

The identification of amino acid residues in proteins involved in binding small molecule ligands is an important step for their functional characterization, as the function of a protein often depends on specific interactions with other molecules. The accuracy of computational methods aiming to predict such binding residues was evaluated within the "function prediction (prediction of binding sites, FN)" category of the critical assessment of protein structure prediction (CASP) experiment. In the last edition of the experiment (CASP10), 17 research groups participated in this category, and their predictions were evaluated on 13 prediction targets containing biologically relevant ligands. The results of this experiment indicate that several methods achieved an overall good

performance, showing the usefulness of such methods in predicting ligand binding residues. As in previous years, methods based on a homology transfer approach were dominating. In comparison to CASP9, a larger fraction of the top predictors are automated servers. However, due to the small number of targets and the characteristics of the prediction format, the differences observed among the first ten methods were not statistically significant and it was also not possible to analyze differences in accuracy for different ligand types or overall structure, difficulty. To overcome these limitations and to allow for a more detailed evaluation, in future editions of CASP, methods in the FN category will no longer be evaluated on the “normal” CASP targets, but assessed continuously by CAMEO (continuous automated model evaluation) based on weekly pre-released sequences from the PDB.

Introduction

Proteins interact with a broad range of molecules to perform their function. While the majority of these interactions are unspecific and transient (e.g., with water molecules, ions and other solutes in the cell), others are very specific and essential for the function of the protein. Specific interactions can be stable, for example oligomeric proteins, or transient, for example in signaling networks or motor proteins. Binding partners of a protein are not limited to other proteins, but can include the whole range of other molecule types. Typical examples include complexes of enzymes with substrates and co-factors, receptors and ligands, antibodies and epitopes, transcription factors and cognate DNA, protein–RNA assemblies such as the ribosome, or ligands in a protein structure with a structural role. For the characterization of a new protein, information about ligands, cofactors and binding sites often provides crucial hints about its function. However, when determining the structure of a protein experimentally, ligands with medium to low binding affinities are often lost during the purification procedure, and the resulting structures often do not contain ligands. Additionally, in many cases neither the three-dimensional structure of the protein itself is known, nor the location and identity of possible ligands. To overcome these limitations, computational methods were established to predict from a protein's sequence its three-dimensional structure and possible ligand binding residues. Several computational approaches for predicting ligand binding sites have been developed, which differ with respect to the information they are based on: (1) only the sequence of the protein; (2) its structural properties; (3) both sequence and structure; (4) homologue proteins.[1-12]

The aim of the critical assessment of protein structure prediction (CASP) experiment is to assess the current state of the art of such methods, and to highlight bottlenecks and opportunities for further development. The accuracy of predictions of three-dimensional structures is assessed as part of the template-based (TBM) and free-modeling (FM) categories,[13-16] which includes the accuracy of binding site coordinates.[17, 18] Function prediction (FN) was introduced as a new category in CASP6.[19] In later editions, the definition of function predictions was specified as prediction of residues involved in binding relevant ligands.[20-22] Here, we describe the results of the assessment in the category “prediction of binding sites (function prediction, FN)” of the CASP10 experiment.

Materials and Methods

Prediction format

As in previous CASP experiments, the format for the predictions of binding residues for a given target protein consisted of a list of the amino-acid positions that were predicted to be in contact with a biologically relevant ligand. The CASP format did not include a confidence score, so that residues are classified in a binary way, either binding or not binding to any ligand. Predictors could optionally propose the name and the category of the compounds that could bind to these residues. One consequence of the prediction format is that it is not possible to correctly assess over-predictions; neither in case a target did not include any biologically relevant ligand, nor if a prediction indicated a binding site for a different ligand elsewhere in the target.

Prediction targets

All CASP10 target sequences were sent out as prediction targets in the FN category. [23] For the assessment, a subset of the target structures (coordinates available as of 2013-08-05) were selected, which contained at least one biologically relevant ligand. To define which ligands were considered as “biologically relevant,” we used information coming from scientific literature, Swiss-Prot[24] annotations, sequence conservation of functionally important residues, and information from homologous structures. For the purpose of the assessment, covalently bound ligands in the reference structure were handled the same way as noncovalently bound ones. In case of oligomeric assemblies, the “biological assembly units” as defined by the authors were used as reference target structures.

Binding site definition

A binding site was defined by all protein residues in the target structure having at least one (non-hydrogen) atom within a certain distance ($d_{i,j}$) to biologically relevant ligand atoms:

$$d_{i,j} = r_i + r_j + c$$

where $d_{i,j}$ is the distance between a residue atom *i* and a ligand atom *j*, r_i and r_j are the Van der Waals radii of the involved atoms, while *c* is a tolerance distance of 0.5 Å. In case the biological assembly of the experimental target structure represents a homo-oligomeric protein, or in case of NMR ensembles, residues were included in the binding site definition if they fulfilled the distance criterion in at least half of the reference chains. The binding site definitions used for the assessment are shown in Table 2. Analysis of ligand binding sites was implemented using OpenStructure (version 1.4). [25, 26]

Binding site prediction evaluation

According to the binding site definition in the experimental reference structure, predicted binding residues were classified as true positives (TP: correctly predicted binding site residue), true negatives (TN: correctly predicted nonbinding residue), false negatives (FN: incorrectly not predicted binding site residue), false positives (FP: incorrectly predicted non-binding residue). As in the previous CASP assessment [22]the evaluation of the quality of the binding site predictions was performed using the Matthews correlation coefficient (MCC):

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP) \cdot (TP + FN) \cdot (TN + FP) \cdot (TN + FN)}}$$

MCC is a useful measure when the two classes (in our case binding and non-binding residues) are of very different sizes. MCC ranges from +1 (perfect prediction) to −1 (inverse prediction), where a MCC of 0 corresponds to random prediction. Raw scores and confusion matrices for all groups and all targets are provided in Supporting Information Table SI, SIV, and Figure S1. Finally, the MCCs were standardized by calculating their *Z* scores to allow the combination of scores for targets of different difficulty:

$$Z_{P,T} = \frac{MCC_{P,T} - \overline{MCC_T}}{\sigma_T}$$

where $MCC_{P,T}$ is the MCC of the predictor *P* for target *T*, $\overline{MCC_T}$ is the mean MCC for the target *T* by all predictors *P* and σ_T is the standard deviation of the MCCs for the target *T* by all predictors *P*. The final ranking of the methods was based on the average value of the MCC across all targets. Cumulative confusion matrices are provided in Supporting Information. For completeness, we also assessed the accuracy of the predictions using the distance based method BDT[27] (see Supporting Information Table SIII).

Statistical significance and robustness of the ranking

To measure the statistical significance of the assessment results, we applied two-tailed Student's paired *t* test and Wilcoxon signed-rank test on the MCCs values for each target's predictions. Both tests were performed using the R statistical package (version 2.11.1) [28]. The robustness of the ranking based on MCC values was assessed by 100 rounds of random sampling using 70% of the targets.

Results and discussion

Prediction targets

Although the CASP10 experiment had in total about 100 prediction targets, only very few of them had ligands which were classified as biologically relevant[23, 29]. In total, we identified 13 targets with biologically relevant ligands, as listed in Tables 1 and 2. In eight targets, metal ions (Zn^{2+}, Mg^{2+}, Mn^{2+}, Na^{+}) are present, one contained an iron–sulfur cluster (SF4), one bound an adenine-mono-phosphate (AMP), one had a reduced Flavin mononucleotide (FNR), two had Flavin-adenine-dinucleotide (FAD) ligands, and one had LPP (*N′*-pyridoxyl-lysin-5′-monophosphate) covalently bound at the dimer interface. It is worth emphasizing that the presence or absence of a ligand in a target structure depends on the experimental conditions, that is, the same binding site can be occupied by a ligand under one condition, and be empty or occupied by a different ligand under different conditions. Therefore, target structures without bound ligands can therefore not be considered as reference in the assessment. This issue is especially pronounced in prediction targets solved by high-throughput methods, where the experimental conditions often do not contain the biologically relevant ligands or cofactors. As a consequence, the number of targets that bind a relevant compound and that can be used for further prediction assessment in CASP10 is quite small. The following paragraph provides a short overview of the assessed targets.

Table 1 Targets with biologically relevant ligands used in the FN prediction assessment.

Target	PDB ID	Ligand ID	Type	Interface
T0652	4HG0	AMP	Non-metal	No
T0657	2LUL	ZN	Metal	No
T0659	4ESN	ZN (2)	Metal	No
T0675	2LV2	ZN (2)	Metal	No
T0686	4HQO	MG	Metal	No
T0696	n.a.	NA	Metal	No
T0697	n.a.	LLP (2)	Non-metal	A-A
T0706	n.a.	MG (2)	Metal	A-A
T0720	4IC1	MN(10)/SF4(10)	Metal	No
T0721	4FK1	FAD (2)	Non-metal	No
T0726	4FGM	ZN	Metal	No
T0737	3TD7	FAD	Non-metal	No
T0744	2YMV	FNR	Non-metal	No

Table 2 Definition of ligand binding residues.

Target	Binding site (residue numbers)
T0652	74, 79, 80, 99, 100, 101, 102, 103, 104, 165, 180, 182, 183
T0657	121, 132, 133, 143
T0659	43, 48, 63
T0675	21, 24, 37, 42, 49, 52, 65, 70
T0686	28, 30, 103
T0696	18, 69, 104
T0697	91, 150, 151, 152, 190, 243, 245, 247, 272, 274, 301, 303, 304, 351
T0706	25, 27, 99, 101, 129, 130
T0720	32, 34, 35, 62, 99, 113, 114, 115, 182, 188, 191, 194, 197, 200
T0721	10, 12, 13, 14, 33, 34, 35, 36, 37, 38, 39, 42, 45, 46, 60, 78, 79, 80, 109, 110, 111, 114, 126, 136, 235, 237, 268, 269, 277, 278, 281
T0726	273, 277, 307
T0737	37, 40, 41, 42, 44, 45, 49, 78, 83, 114, 117, 118, 120, 121, 123, 124, 128, 130, 135, 138, 174, 237
T0744	22, 23, 24, 26, 58, 61, 120, 121, 122, 124, 196, 214, 216, 270, 271, 272, 273, 314, 316

A residue in the target structure was defined as binding if it had at least one heavy atom of a biologically relevant ligand within 0.5 Å distance of the sum of the Van der Waals radii of the involved atoms.

Target T0652 (PDB: 4HG0)

The magnesium and cobalt efflux protein CorC contains two CBS (cystathionine-beta-synthase) domains, which bind an Adenosine monophosphate (AMP) [Fig. 1(A)], next to a transporter associated domain (CorC_HlyC) at the C terminus of the protein. CBS is a small intracellular module, mostly found in two or four copies next to a wide range of protein domains in bacteria, *archaea*, and eukaryotes [30] [31]. Pairs of CBS domains can bind adenosyl groups such as AMP, ATP or SAM, thus they could regulate the activity of the attached domains [32] and they may act as sensors of intracellular metabolites [33]. The CorC_HlyC transporter associated domain is found in a family of proteins of unknown function with CBS domain and also in CorC

involved in magnesium and cobalt efflux; it is hypothesized that it could modulate the transport of ion substrates.

Target T0657 (PDB: 2LUL)

The tyrosine-protein kinase Tec is composed by a PH (Pleckstrin homology) domain and a BTK (Btk-type zinc finger) domain that binds a Zn^{2+} cation [Fig. 7(A)]. The first occurs in many proteins involved in intracellular signaling or as part of the cytoskeleton [34], such as the beta/gamma subunits of heterotrimeric G proteins [35]. This domain has specificities for different membrane phosphoinositides phosphorylated at different sites within the inositol ring, so the function of PH-containing proteins is modulated by enzymes that dephosphorylate such rings. PH recruits proteins to different cellular compartments or it allows them to be involved in signal transduction pathways. The structure of this domain consists of two perpendicular antiparallel beta sheets followed by an amphiphatic helix; the loop between the beta strands has a very variable length. The BTK domain contains a conserved zinc-binding motif of one histidine and three cysteine residues, it is very close to the PH domain and it consists in a long loop held together by a zinc ion.

Target T0659 (PDB: 4ESN)

There are no sequence annotations on this target, a homo-dimer that binds a Zn^{2+} ion in both chains at the same position [Fig. 7(B)]. A DELTA-BLAST [36] search revealed a conserved domain of unknown function homologous *to Listeria innocua* Lin0431, a protein similar to the N-utilization substance G (NusG) N terminal (NGN) insert (domain II, DII). Lin0431 has a similar structure and charged surface distribution to *Aquifex aeolicus* NusG DII, indicating a possible role in transcription or translation regulating functions.

Target T0675 (PDB: 2LV2)

The insulinoma-associated protein 1 contains two Zinc finger domains [Fig. 1(B)], which are stable structural motifs that bind DNA, RNA, protein, or lipid substrates [37] [38] [39] [40] [41]. Some types of this domain use zinc, others use iron or form salt-bridges to create the correct fold, which often does not change conformation upon binding the target. Zinc fingers are usually found in groups and they have different binding specificities depending on their amino acid sequence and on the overall structure of the protein containing them. The domains in this target are of the C2H2 type, where two conserved cysteines and histidines coordinate a zinc ion inside

of two short beta strands followed by an alpha helix; this “finger” binds the major groove of the DNA.

Target T0686 (PDB: 4HQO)

The *sporozoite* surface protein 2 is the ectodomain of a thrombospondin repeat anonymous protein (TRAP), a mediator in the infection of mosquito and vertebrate cells and in the gliding motility of *sporozoites*, which is an important target of pre-erythrocytic malaria vaccines. TRAP passes through the plasma membrane and is attached to the actin cytoskeleton by aldolase [42]. This structure has a Von Willebrand factor type A (VWA) domain, binding a Mg^{2+} ion [Fig. 1(C)] which is additionally coordinated by three water molecules, and a thrombospondins (TSP) domain. The first is found in various plasma proteins, for example, complement factors or integrins, and is often involved in protein complexes which participate in various biological process (e.g., signal transduction, cell adhesion, pattern formation, and migration) [43]; it contains a metal ion site at the surface that could represent a general metal ion-dependent adhesion site (MIDAS) for binding protein ligands [44]. This site binds magnesium in the I-domain of integrins CD11b [44] and manganese in CD11a [45] by slightly different coordination of the same conserved residues [45]. TSP is a multimeric multidomain glycoprotein functioning in the extracellular matrix and it regulates cell interactions.

Target T0696 (PDB: n.a.)

A DELTA-BLAST search relates this target with a conserved domain superfamily called “Glyoxalase/fosfomycin resistance/dioxygenase domain,” which is found in a variety of structurally related, but functionally diverse metallo-proteins, including glyoxalase I, type I extradiol dioxygenases and some antibiotic resistance proteins. They use different metal cations for their catalytic activity (e.g., Fe^{2+}, Mn^{2+}, Zn^{2+}, Ni^{2+}, or Mg^{2+}). In this target the binding site is occupied by a Na^{+} [Fig. 1(D)], which substitutes one of mentioned metal ions.

Target T0697 (PDB: n.a.)

It belongs to the pyridoxal phosphate (PLP)-dependent decarboxylase family (EC number 4.1.1) group 2, which includes glutamate, histidine, tyrosine, and aromatic-l-amino-acid decarboxylases. This family is involved in the biosynthesis of amino acids, their derived metabolites, amino sugars and in the synthesis or catabolism of neurotransmitters. The PLP cofactor [Fig. 1(E)] forms a Schiff base with a conserved lysine in the active site, which is

temporarily displaced by the substrate; the resulting aldimine is the common central intermediate for PLP-catalyzed reactions[46].

Target T0706 (PDB: n.a.)

A DELTA-BLAST search indicates that the target belongs to the Von Willebrand factor type A domain family, which contains a metal ion-dependent adhesion site (MIDAS) for binding protein ligands (for details, see also target T0686). In this target, a Mg^{2+} ion is bound to the adhesion site [Fig. 1(F)].

Target T0720 (PDB: 4IC1)

The CRISPR-associated exonuclease Cas4 (EC = 3.1.-.-) protein is involved in the mobile genetic elements immunity of the CRISPR (clustered regularly interspaced short palindromic repeat) system in most bacteria and *archea* [47]. Short DNA sequences from viruses, the "spacers," are flanked by CRISPR repeats in the host genome and transcribed into CRISPR RNAs (crRNAs), which are used by Cas (CRISPR-associated) proteins to recognize and degrade viral cognate sequences. This target in particular belongs to the Cas4 family of proteins, which resembles the RecB family [48] and contains a cysteine-rich motif similar to the AddB family [49]. It is a 5' ssDNA metal-dependent (magnesium or manganese) exonuclease that needs an iron–sulfur cluster for structural stability [Fig. 1(G,H)] [50].

Target T0721 (PDB: 4FK1)

The putative Thioredoxin reductase TrxB contains a FAD-dependent pyridine nucleotide-disulfide oxidoreductase domain with a FAD bound [Fig. 1(I)].

Target T0726 (PDB: 4FGM)

It contains an M61 glycyl aminopeptidase and a PDZ domain. Metalloproteases containing the first domain bind a divalent cation, through His, Glu, Asp, or Lys amino acids, that activates the water molecule; usually a zinc ion is bound by three residues which often can be described by an HEXXH motif (X can be any amino acid) [51]. The target binds a Zn^{2+} ion with the motif's histidines and a different glutamate [Fig. 1(J)]. The second domain is found in eukaryotes [52] and it binds the target protein by extending its beta-sheet with a strand from the partner C-terminus, so acting as a bridge between transmembrane proteins and the cytoskeleton in signaling pathways [53].

Target T0737 (PDB: 3TD7)

It is the probable FAD-linked sulfhydryl oxidase R596 from the *Acanthamoeba polyphaga* mimivirus. Its sequence contains a ERV/ALR sulfydryl oxidase domain which catalyzes disulfide bond formations. This module has a CXXC motif next to a FAD cofactor [in Fig. 1(K)] which is used to transfer electrons from the thiol substrates to the (non-thiol) acceptor. A structure with bound FAD (PDB code: 3GWN) was available at the time of prediction for this target.

Target T0744 (PDB: 2YMV)

It is a homologue of *Mycobacterium tuberculosis* Acg (Rv2032) in the reduced form from *Mycobacterium smegmatis*. The proteins in the Acg family are monomers that resemble the nitroreductase homodimer fold, with a single flavin mononucleotide binding site [Fig. 1(L)] closed by a lid, instead of two open binding sites as in homodimeric nitroreductases. The structure and the lack of reduction by NADPH suggest that this proteins has lost the nitroreductase function and instead they may act as inhibitor of another nitroreductase by storing the flavin cofactor during the dormancy state of the bacteria [54].

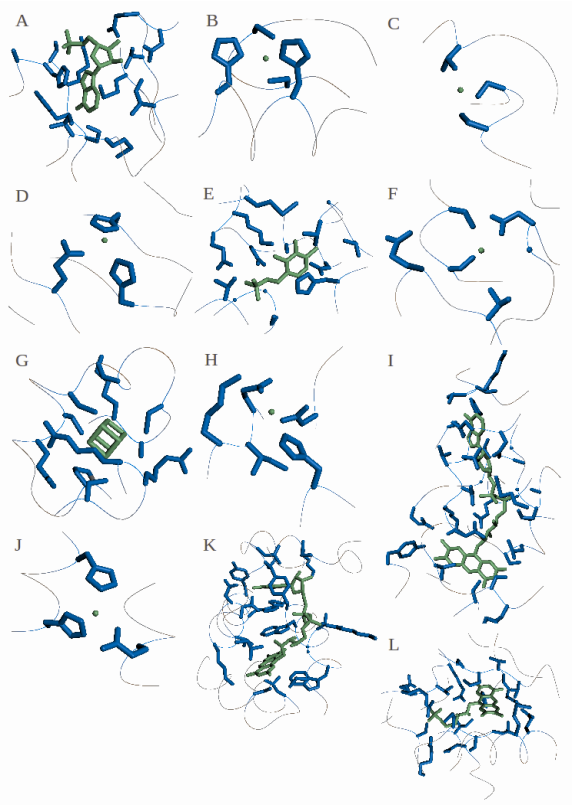

Figure 1 Binding sites and ligands of the assessed targets. Biologically relevant ligands are colored in green and the residues included in the binding site are colored in blue. Targets and ligands included here:

(**A**) T0652: AMP, (**B**) T0675: ZN, (**C**) T0686: MG, (**D**) T0696: NA, (**E**) T0697: LLP, (**F**) T0706: MG, (**G**,**H**) T0720: SF4 and MN, (**I**) T0721: FAD, (**J**) T0726: ZN, (**K**) T0737: FAD, (**L**) T0744: FNR. Targets T0657 and T0659 are in Figure 7.

Overall performance

As in previous years, the evaluation of the binding site prediction accuracy was based on the Matthew Correlation Coefficient. A total of 1817 submissions by 19 groups for the FN category were received by the Prediction Center. In CASP10 only 13 target proteins contained relevant ligands, that is, only a small subset of all submissions could be used for the assessment (Fig. 2). Of the 17 groups[4] in the assessment, most of them submitted predictions for all 13 targets. Missing predictions were assigned a MCC score of zero, corresponding to a random prediction. Figure 3 shows a box plot representing the MCC distributions for each target, which gives a first estimate of the prediction difficulty. On most targets the predictors achieved on average a good performance around an MCC of 0.6, except in three cases, where in two (T0657 and T0659) the median scores were around zero and in one (T0720) was around 0.2.

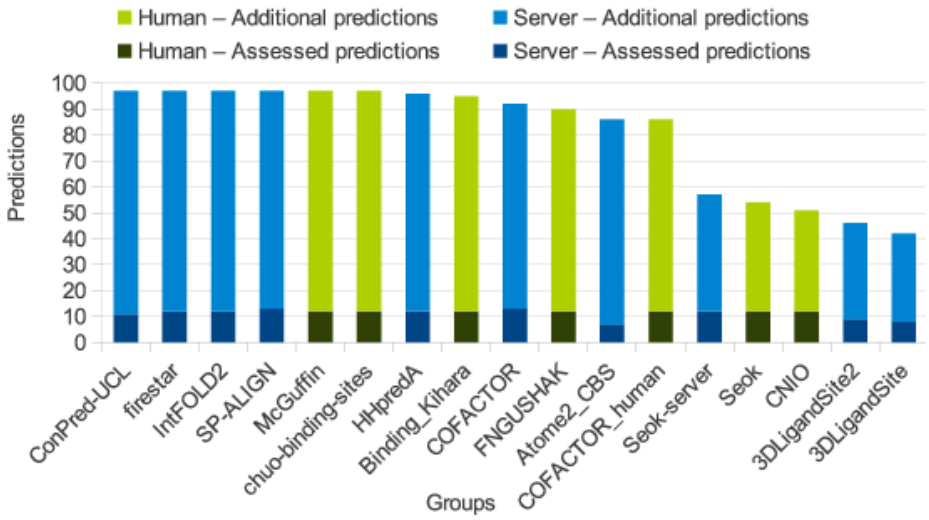

Figure 2. Number of predictions per group. Because only a small number of all CASP10 targets contained relevant ligands, only a few predictions could be used for the assessment (dark blue and dark green), while the majority of the predictions could not be evaluated (blue and green).

[4] Predictions by two groups were excluded from the assessment by the CASP organizers.

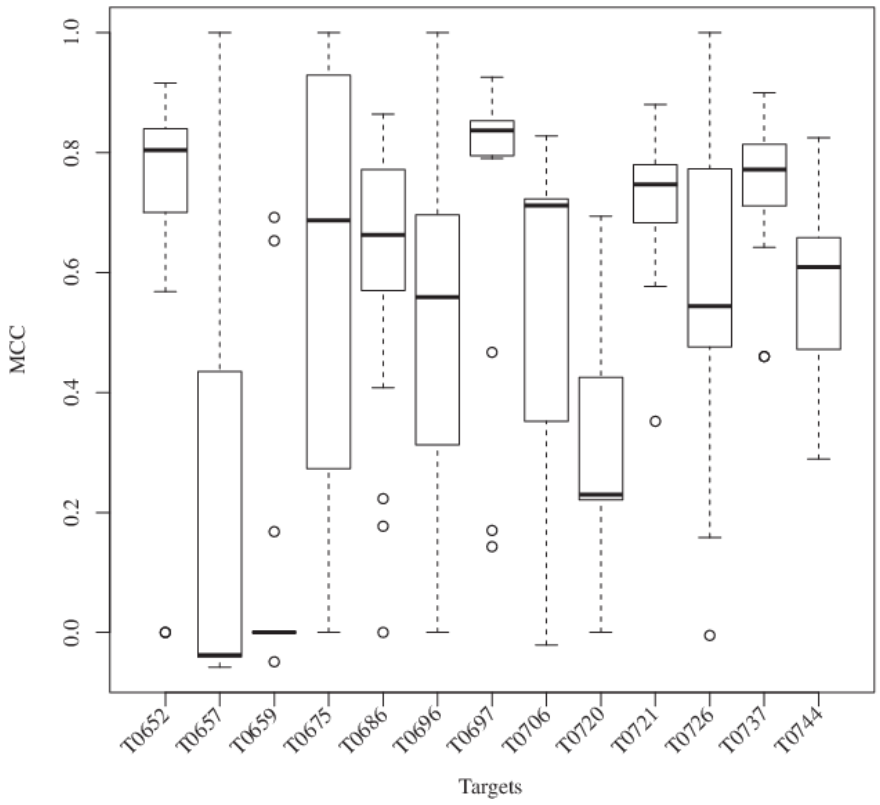

Figure 3 Target difficulty. Distribution of the predictor's MCC for each target shown as Box plot (1st quartile, median and 3rd quartile), indicating the difficulty in the prediction of the various binding sites.

For comparison of the method's overall performance, groups were ranked according to the average value of their MCCs normalized on all prediction targets (Fig. 4; Supporting Information Table SI). Within the first ten groups, there were more “servers” at CASP10 than in CASP9, six instead of two, with an average MCC of 0.62. Their performance was indistinguishable from the “human” predictors, which is an improvement with respect to the results obtained in CASP9. The main differences between “human” and “server” methods is that the former could access human-only readable data (e.g., literature or databases) to identify relevant ligands, and have access to the pool of 3D structure predictions by servers due to the late submission deadline. While there was only a difference of 0.15 between the top ten groups based on average MCC, group FN119 (Firestar [10]) and FN326 (SP-ALIGN [11]) achieved the best scores of 0.715 and 0.707, respectively. These two methods had an overall different behaviour: Firestar was one of the two predictors, together with HHPredA, with the highest number of top scores; it had the best MCC in three targets (T0696, T0726, T0744) while SP-ALIGN only for T0659, which was the most difficult target.

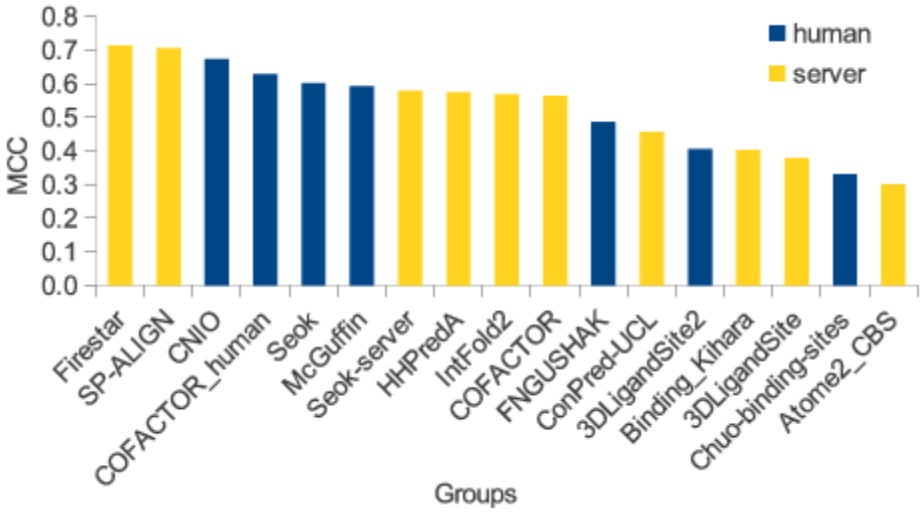

Figure 4 Groups ranking by MCC. The predictors are ranked in decreasing order by the average value of the MCC, calculated on all the evaluated targets. Human predictors are shown in blue and server predictors in yellow.

We also evaluated the predictors' performances based on the distance-based BDT measure (Supporting Information Table SIII), which gave a ranking very similar to the MCC averages that was within deviations expected from the robustness test described below. To better understand to which extend these methods were useful in practice, we compared their performance with a baseline predictor that inferred the target's binding sites using the first ten templates with ligands found by DELTA-BLAST and collecting all the residues in contact with them. The resulting average MCC was 0.339, which is only half of the performance obtained by the top predictors, and only two methods in the experiment performed worse than this baseline. This result indicates that most of the methods assessed in CASP10 give advantages in the ligand binding site prediction compared to a naïve homology search approach and could positively support the characterization of a protein's function.

Assessment robustness

Because the number of prediction targets was extremely small, we assessed the robustness of the ranking by calculating MCC distributions with 100 cycles of random sampling using 70% of the targets (Fig. 5). Although, the median values confirm the order of the top groups ranked by MCC, the rank spread is rather large and fluctuations by 10 positions are not unusual, that is, the ranking is strongly influenced by the composition of the data set and does therefore not necessarily correctly reflect the differences in prediction accuracy of the individual methods. When calculating the statistical significance of the overall ranking by applying Student's *t* test (data not shown) and Wilcoxon signed-rank test (Supporting Information Table SII), the results

indicated that the ranking was not robust and the differences between the top ten groups were not statistically significant. Both results are not surprising, considering the fact that the assessment had to be based on a very small number of target structures.

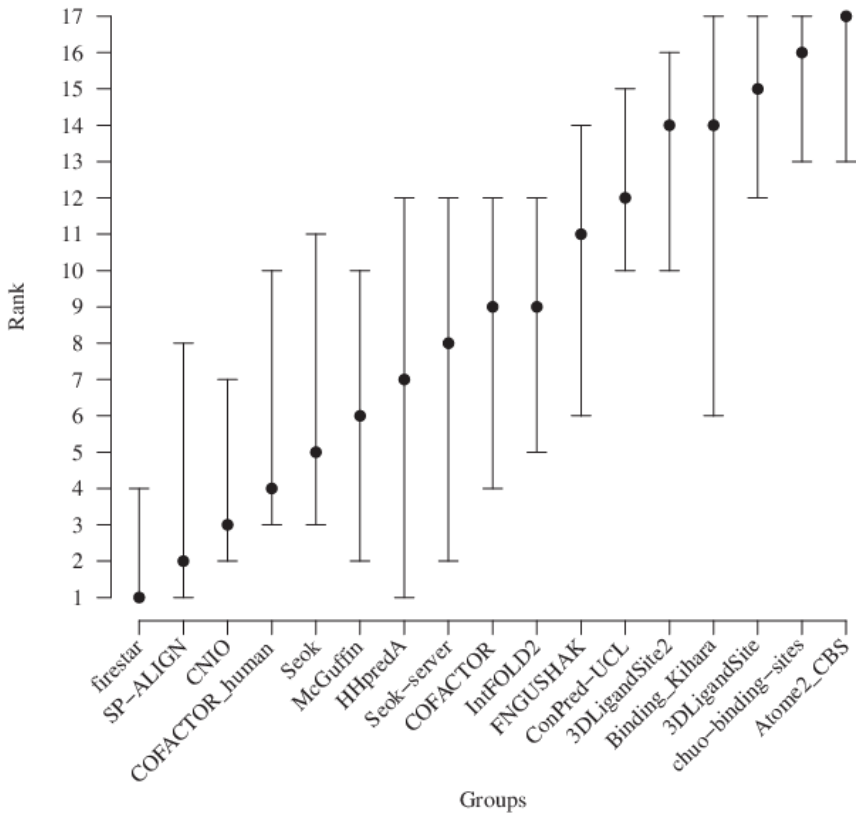

Figure 5 Groups ranking robustness. Methods were ranked using the median value of the MCC distributions after 100 cycles of random sampling using 70% of the targets. Bars indicate best, median and worst ranking for each group.

Top predictors' methods are based on homology transfer

Let's take a closer look at the groups ranked highest by MCC: Firestar (FN119), SP-ALIGN (FN326), CNIO (FN475), and Cofactor_human (FN208); the first two were registered to CASP10 as "server," while the last two as "human" predictor groups. Firestar [10] bases its predictions on homology transfer of functionally important residues, found by local evolutionary sequence conservation; SP-ALIGN, an update to FINDSITE [11], is a threading based method to detect ligand binding sites by the employment of remote template identification and superimposition, structure-pocket alignment and binding site clustering guided by the template ligands; CNIO combines predictions from Firestar and 3DLigandSite [12], which clusters superimposed ligands from homologous structures to identify the binding residues; Cofactor_human requires human assistance to validate the binding residues found by the Cofactor algorithm [55], which employs

a local superimposition of conserved residues taken from the target's templates. The common theme among these methods is that they are all based on the analysis of the ligands bound to homologous structures. Firestar and CNIO use FireDB [56], Cofactor employs structures from BioLip [57], while SP-ALIGN uses an *ad hoc* template library. As a consequence, the performance of these methods is tied to the availability of annotated protein structures and the ability of finding homologue templates. Nevertheless, it has to be noted that the protocol to transfer the information on binding residues is different among those methods.

In recent years, homology based methods for structure prediction have started to reach a substantial coverage for proteins of interest: today some form of structural information—either experimental or computational—is available for the majority of amino acids encoded by common model organism genomes[58]. For almost all known protein-protein interactions for which the individual components are structurally characterized, structures of complexes can be identified in the PDB which can be used for template-based prediction approaches[59-61]. The overall good performances of methods such as Firestar and SP-ALIGN in CASP10, and their ability to identify ligand binding sites in different families of proteins in the absence of close homologue targets indicates that the field of ligand binding site prediction shows a similar trend.

It should be noted that in previous editions of CASP, almost all FN targets were classified as “template based modeling” (TBM) and only very few as “free modeling” (FM). In this round of CASP10, none of the relevant ligand binding sites were located in FM targets. Although target T0737 is classified as “free modeling,” the part of the protein to which the ligand FAD is bound has experimental structure information (Fig. 6). This directly implies that the CASP assessment is mainly suitable to evaluate methods based on homology transfer to predict binding residues, but unable to measure the performance on harder targets, for which template structure information is not a useful source of information.

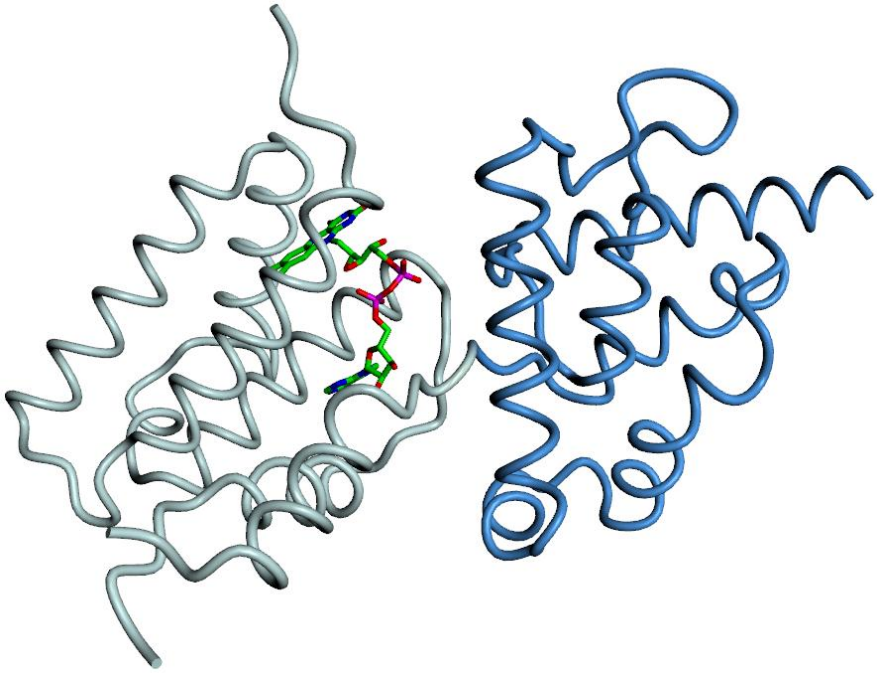

Figure 6 FAD binding site in target T0737. Residues 176–292 (D1, blue) have been classified as "free modeling." However, the N-terminal domain (grey) where the ligand FAD is bound, is covered by experimental structures. (Image generated with OpenStructure).

Prediction examples

Two targets, T0657 and T0659, appeared to be most challenging as predictors obtained on average the lowest scores. The first (PDB: 2LUL) was a solution NMR structure of the PH domain from the "Tyrosine-protein kinase Tec," bound to a Zn^{2+} ion in a Btk-type zinc finger [Fig. 7(A)]. On a first view, this appears to be a simple template-based modeling target, since at least one template with the correct ion bound (e.g., PDB:1B55) is easily detectable with BLAST. However, the median MCC achieved for this target was −0.05, where the best predictor ("Binding_Kihara", FN231) achieved an MMC of 1 (Supporting Information Table SI). Other predictors achieved a lower MCC of about 0.3, mainly because they predicted more binding residues than were present in the reference structure, some of which have been assigned to other ligands than Zinc as indicated in the comments field. This example illustrates one of the limitations of the current binary prediction format.

The second target, T0659 (PDB: 4ESN), was a crystal structure of a hypothetical protein that bound a Zn^{2+} ion by three conserved Cysteines [Fig. 7(B)]. The median MCC was zero, while the best score, an MCC of 0.69 (Supporting Information Table SI), was obtained for a prediction by SP-ALIGN (FN326), which is shown in Figure 7(B). Easily detectable homologous structures of this protein did not contain any ligand, which explains the overall weak performance on this

target. Interestingly, SP-ALIGN predicted an iron ion bound at this position; potentially, this could be due to the employment of its threading based method that detected a remote homologue bound to that ion.

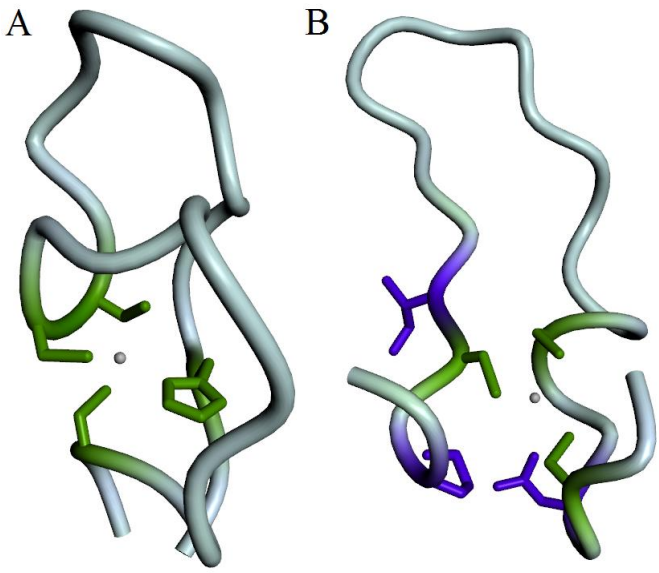

Figure 7 Binding site prediction examples. Residues are colored as “correctly predicted” (true positive, green) and “wrongly predicted” (false positive, violet). (**A**) The Zn^{2+} binding site in a Btk-type zinc finger in the PH domain from the Tyrosine-protein kinase Tec (T0657) is formed by His 121, Cys 132, Cys 133, and Cys 143. Coloring according to prediction by group “Binding_Kihara” (FN231). (**B**) Structure of a hypothetical protein T0659 with a Zn^{2+} ion bound by three conserved Cysteine residues (Cys 43,48,63). Coloring according to predictions by group “SP-ALIGN” (FN326).

Conclusions

Predicting a protein's binding site is an important step toward understanding its function, and has implications for gene product characterization, drug design and enzyme engineering. The 13 targets evaluated in the assessment include proteins with interesting functions. For example T0686, which contains a metal ion-dependent adhesion site (MIDAS) which mediates the invasion of vertebrate cells by malaria Sporozoites; or T0720—a CRISPR-associated (Cas) protein involved in the genetic mobile elements defense and it contains a catalytic magnesium ion plus a structural iron–sulfur cluster. As in previous years, homology transfer approaches, in which the target binding residues are inferred from homologous proteins, have scored best with an average MCC of 0.71.

As in previous rounds of the CASP experiment, only a very limited number of targets with biologically relevant ligands (13 out of 97 targets) were available. Consequently, the assessment

did not lead to a stable ranking of the participating methods, and it was not possible to differentiate methods by their performance on different types of targets or ligands. Another limitation originates from the current binary prediction format (“binding” or “not binding”), which does not include any information on the type of compounds or a level of confidence for the prediction. For a more detailed discussion, see assessment of ligand binding predictions in CASP9 [22].

During the CASP10 predictors meeting in Gaeta, it was recognized that the current procedure is not appropriate to assess the state of the art in ligand binding site predictions, and therefore does not stimulate the development of new approaches. To overcome these limitations, the following improvements should be implemented: (a) Binary predictions should be replaced by predicting continuous probability values. (b) The prediction format should include the specification of ligand type/ligand identity. (c) The number of prediction targets, specifically those without trivial templates, needs to be increased substantially.

Based on these considerations, prediction methods in the FN category in future editions of CASP will no longer be evaluated based on the regular set of CASP target proteins. Instead, ligand binding site prediction servers will be evaluated continuously using an automated system called continuous automated model evaluation (CAMEO, http://www.cameo3d.org/), which is based on weekly pre-released sequences from the PDB. Continuous evaluation allows developers to constantly monitor the performance of new developments. Thanks to the larger number of targets, continuous evaluation also provides statistically robust assessment of ligand binding site predictions and allows for a more detailed assessment of methods, for example, by ligand type or target difficulty. We hope that these new developments will stimulate new methods and approaches in this important area of structural bioinformatics.

Acknowledgements

The authors thank the previous CASP assessors Tobias Schmidt and Florian Kiefer for fruitful discussions and their contributions to the assessment methods for the ligand binding category. They are grateful to Konstantin Arnold (SIB Swiss Institute of Bioinformatics and Biozentrum University of Basel) for professional systems support. Computational resources were provided by the $[BC]^2$ Basel Computational Biology Center.

Supporting information

Table S1 Raw scores. Raw scores for each group that provided a prediction for a FN target. TP: True Positive, FP: False Positive, FN: False Negative, TN: True Negative, MCC: Matthews Correlation Coefficient.

Target	Group	TP	FP	FN	TN	MCC	MCC Z-score
T0652	CONPRED-UCL	11	7	2	212	0.7	0.010
T0652	FIRESTAR	12	4	1	215	0.821	0.447
T0652	3DLIGANDSITE2	12	4	1	215	0.821	0.447
T0652	COFACTOR_HUMAN	11	4	2	215	0.774	0.277
T0652	ATOME2_CBS	13	5	0	214	0.84	0.516
T0652	COFACTOR	11	4	2	215	0.774	0.277
T0652	BINDING_KIHARA	8	6	5	213	0.568	-0.468
T0652	SEOK-SERVER	11	0	2	219	0.916	0.791
T0652	INTFOLD2	12	3	1	216	0.85	0.552
T0652	MCGUFFIN	12	3	1	216	0.85	0.552
T0652	SP-ALIGN	11	3	2	216	0.804	0.386
T0652	HHPREDA	12	4	1	215	0.821	0.447
T0652	CHUO-BINDING-SITES	9	4	4	215	0.674	-0.084
T0652	SEOK	11	0	2	219	0.916	0.791
T0652	CNIO	12	8	1	211	0.726	0.104
T0657	CONPRED-UCL	0	17	4	133	-0.058	-0.744
T0657	FNGUSHAK	0	13	4	137	-0.05	-0.721
T0657	FIRESTAR	4	16	0	134	0.423	0.636
T0657	3DLIGANDSITE2	0	9	4	141	-0.041	-0.696
T0657	COFACTOR_HUMAN	4	15	0	135	0.435	0.670
T0657	ATOME2_CBS	0	14	4	136	-0.052	-0.727
T0657	COFACTOR	4	15	0	135	0.435	0.670
T0657	BINDING_KIHARA	4	0	0	150	1	2.292
T0657	3DLIGANDSITE	0	9	4	141	-0.041	-0.696
T0657	SEOK-SERVER	0	7	4	143	-0.036	-0.681
T0657	INTFOLD2	0	9	4	141	-0.041	-0.696
T0657	MCGUFFIN	0	9	4	141	-0.041	-0.696
T0657	SP-ALIGN	4	1	0	149	0.891	1.979
T0657	HHPREDA	0	15	4	135	-0.054	-0.733
T0657	CHUO-BINDING-SITES	4	44	0	106	0.243	0.119
T0657	SEOK	0	8	4	142	-0.038	-0.687
T0657	CNIO	4	14	0	136	0.449	0.710
T0659	CONPRED-UCL	0	0	3	71	0	-0.383
T0659	FNGUSHAK	1	6	2	65	0.168	0.364
T0659	FIRESTAR	0	0	3	71	0	-0.383
T0659	COFACTOR	0	4	3	67	-0.049	-0.601

T0659	BINDING_KIHARA	0	0	3	71	0	-0.383
T0659	INTFOLD2	0	0	3	71	0	-0.383
T0659	MCGUFFIN	0	0	3	71	0	-0.383
T0659	SP-ALIGN	3	3	0	68	0.692	2.693
T0659	HHPREDA	2	1	1	70	0.653	2.520
T0659	CHUO-BINDING-SITES	0	0	3	71	0	-0.383
T0675	CONPRED-UCL	6	4	2	63	0.627	0.169
T0675	FNGUSHAK	3	1	5	66	0.495	-0.179
T0675	FIRESTAR	7	0	1	67	0.929	0.964
T0675	COFACTOR_HUMAN	8	0	0	67	1	1.151
T0675	ATOME2_CBS	0	0	8	67	0	-1.482
T0675	COFACTOR	5	6	3	61	0.467	-0.252
T0675	BINDING_KIHARA	8	1	0	66	0.936	0.982
T0675	SEOK-SERVER	6	1	2	66	0.78	0.572
T0675	INTFOLD2	4	0	4	67	0.687	0.327
T0675	MCGUFFIN	4	0	4	67	0.687	0.327
T0675	SP-ALIGN	4	0	4	67	0.687	0.327
T0675	CHUO-BINDING-SITES	7	29	1	38	0.273	-0.763
T0675	SEOK	8	0	0	67	1	1.151
T0675	CNIO	8	0	0	67	1	1.151
T0686	CONPRED-UCL	3	1	0	250	0.864	1.032
T0686	FNGUSHAK	3	6	0	245	0.57	-0.111
T0686	FIRESTAR	3	2	0	249	0.772	0.674
T0686	3DLIGANDSITE2	2	1	1	250	0.663	0.250
T0686	COFACTOR_HUMAN	3	2	0	249	0.772	0.674
T0686	ATOME2_CBS	0	0	3	251	0	-2.328
T0686	COFACTOR	3	14	0	237	0.408	-0.741
T0686	BINDING_KIHARA	1	5	2	246	0.223	-1.461
T0686	3DLIGANDSITE	3	1	0	250	0.864	1.032
T0686	SEOK-SERVER	2	2	1	249	0.572	-0.104
T0686	INTFOLD2	3	2	0	249	0.772	0.674
T0686	MCGUFFIN	3	1	0	250	0.864	1.032
T0686	SP-ALIGN	2	1	1	250	0.663	0.250
T0686	HHPREDA	3	4	0	247	0.649	0.196
T0686	CHUO-BINDING-SITES	3	67	0	184	0.177	-1.640
T0686	SEOK	3	2	0	249	0.772	0.674
T0686	CNIO	2	2	1	249	0.572	-0.104
T0696	CONPRED-UCL	3	6	0	91	0.559	0.099
T0696	FNGUSHAK	3	3	0	94	0.696	0.549
T0696	FIRESTAR	3	0	0	97	1	1.547
T0696	COFACTOR_HUMAN	2	1	1	96	0.656	0.418
T0696	ATOME2_CBS	0	0	3	97	0	-1.736
T0696	COFACTOR	2	1	1	96	0.656	0.418

T0696	BINDING_KIHARA	3	0	0	97	1	1.547
T0696	3DLIGANDSITE	1	2	2	95	0.313	-0.708
T0696	SEOK-SERVER	2	1	1	96	0.656	0.418
T0696	INTFOLD2	1	2	2	95	0.313	-0.708
T0696	MCGUFFIN	1	2	2	95	0.313	-0.708
T0696	SP-ALIGN	2	0	1	97	0.812	0.930
T0696	HHPREDA	3	11	0	86	0.436	-0.304
T0696	CHUO-BINDING-SITES	3	22	0	75	0.305	-0.734
T0696	SEOK	2	5	1	92	0.411	-0.386
T0696	CNIO	3	1	0	96	0.862	1.094
T0697	CONPRED-UCL	1	2	13	446	0.143	-2.477
T0697	FNGUSHAK	13	4	1	444	0.837	0.374
T0697	FIRESTAR	12	3	2	445	0.823	0.316
T0697	3DLIGANDSITE2	12	3	2	445	0.823	0.316
T0697	COFACTOR_HUMAN	13	1	1	447	0.926	0.739
T0697	ATOME2_CBS	12	1	2	447	0.886	0.575
T0697	COFACTOR	13	1	1	447	0.926	0.739
T0697	BINDING_KIHARA	2	6	12	442	0.17	-2.366
T0697	3DLIGANDSITE	12	3	2	445	0.823	0.316
T0697	SEOK-SERVER	11	0	3	448	0.883	0.562
T0697	INTFOLD2	12	2	2	446	0.853	0.439
T0697	MCGUFFIN	12	2	2	446	0.853	0.439
T0697	SP-ALIGN	12	4	2	444	0.795	0.201
T0697	HHPREDA	11	1	3	447	0.844	0.402
T0697	CHUO-BINDING-SITES	14	44	0	404	0.467	-1.146
T0697	SEOK	10	0	4	448	0.841	0.390
T0697	CNIO	13	6	1	442	0.79	0.180
T0706	CONPRED-UCL	4	1	2	197	0.723	0.592
T0706	FNGUSHAK	3	1	3	197	0.603	0.222
T0706	FIRESTAR	4	1	2	197	0.723	0.592
T0706	COFACTOR_HUMAN	5	3	1	195	0.712	0.558
T0706	ATOME2_CBS	0	0	6	198	0	-1.637
T0706	COFACTOR	4	0	2	198	0.812	0.866
T0706	BINDING_KIHARA	0	3	6	195	-0.021	-1.702
T0706	SEOK-SERVER	4	1	2	197	0.723	0.592
T0706	INTFOLD2	4	2	2	196	0.657	0.389
T0706	MCGUFFIN	4	2	2	196	0.657	0.389
T0706	SP-ALIGN	4	0	2	198	0.812	0.866
T0706	HHPREDA	5	1	1	197	0.828	0.916
T0706	CHUO-BINDING-SITES	5	23	1	175	0.352	-0.552
T0706	SEOK	4	1	2	197	0.723	0.592
T0706	CNIO	4	1	2	197	0.723	0.592
T0720	CONPRED-UCL	0	0	16	186	0	-1.371

T0720	FNGUSHAK	4	7	12	179	0.253	-0.184
T0720	FIRESTAR	4	0	12	186	0.485	0.904
T0720	3DLIGANDSITE2	2	2	14	184	0.221	-0.334
T0720	COFACTOR_HUMAN	2	2	14	184	0.221	-0.334
T0720	ATOME2_CBS	0	0	16	186	0	-1.371
T0720	COFACTOR	2	2	14	184	0.221	-0.334
T0720	BINDING_KIHARA	2	1	14	185	0.267	-0.118
T0720	SEOK-SERVER	2	2	14	184	0.221	-0.334
T0720	INTFOLD2	3	5	13	181	0.222	-0.330
T0720	MCGUFFIN	3	3	13	183	0.273	-0.090
T0720	SP-ALIGN	5	0	11	186	0.543	1.177
T0720	HHPREDA	9	1	7	185	0.694	1.885
T0720	CHUO-BINDING-SITES	5	13	11	173	0.23	-0.292
T0720	SEOK	4	1	12	185	0.425	0.623
T0720	CNIO	8	0	8	186	0.692	1.876
T0721	CONPRED-UCL	20	11	11	257	0.604	-0.894
T0721	FNGUSHAK	26	7	5	261	0.791	0.679
T0721	FIRESTAR	23	7	8	261	0.726	0.132
T0721	3DLIGANDSITE2	26	16	5	252	0.683	-0.230
T0721	COFACTOR_HUMAN	23	5	8	263	0.757	0.393
T0721	ATOME2_CBS	22	8	9	260	0.69	-0.171
T0721	COFACTOR	23	5	8	263	0.757	0.393
T0721	BINDING_KIHARA	6	2	25	266	0.352	-3.015
T0721	3DLIGANDSITE	29	16	2	252	0.747	0.309
T0721	SEOK-SERVER	21	7	10	261	0.681	-0.246
T0721	INTFOLD2	25	4	6	264	0.815	0.881
T0721	MCGUFFIN	21	3	10	265	0.747	0.309
T0721	SP-ALIGN	28	11	3	257	0.78	0.587
T0721	HHPREDA	29	5	2	263	0.88	1.428
T0721	CHUO-BINDING-SITES	31	46	0	222	0.577	-1.122
T0721	SEOK	22	8	9	260	0.69	-0.171
T0721	CNIO	27	8	4	260	0.798	0.738
T0726	CONPRED-UCL	3	5	0	579	0.61	0.089
T0726	FNGUSHAK	3	10	0	574	0.476	-0.389
T0726	FIRESTAR	3	0	0	584	1	1.479
T0726	3DLIGANDSITE2	3	5	0	579	0.61	0.089
T0726	COFACTOR_HUMAN	3	7	0	577	0.544	-0.146
T0726	ATOME2_CBS	3	21	0	563	0.347	-0.848
T0726	COFACTOR	3	7	0	577	0.544	-0.146
T0726	BINDING_KIHARA	0	3	3	581	-0.005	-2.103
T0726	3DLIGANDSITE	3	3	0	581	0.705	0.428
T0726	SEOK-SERVER	3	0	0	584	1	1.479
T0726	INTFOLD2	3	2	0	582	0.773	0.670

T0726	MCGUFFIN	3	0	0	584	1	1.479
T0726	SP-ALIGN	3	18	0	566	0.372	-0.759
T0726	HHPREDA	3	8	0	576	0.519	-0.235
T0726	CHUO-BINDING-SITES	3	97	0	487	0.158	-1.522
T0726	SEOK	3	2	0	582	0.773	0.670
T0726	CNIO	3	8	0	576	0.519	-0.235
T0737	CONPRED-UCL	18	14	4	217	0.642	-0.803
T0737	FNGUSHAK	18	9	4	222	0.711	-0.251
T0737	FIRESTAR	16	3	6	228	0.764	0.173
T0737	3DLIGANDSITE2	17	4	5	227	0.772	0.237
T0737	COFACTOR_HUMAN	17	4	5	227	0.772	0.237
T0737	ATOME2_CBS	20	10	2	221	0.755	0.101
T0737	COFACTOR	17	4	5	227	0.772	0.237
T0737	BINDING_KIHARA	5	0	17	231	0.46	-2.260
T0737	3DLIGANDSITE	20	2	2	229	0.9	1.262
T0737	SEOK-SERVER	16	1	6	230	0.814	0.573
T0737	INTFOLD2	18	2	4	229	0.845	0.822
T0737	MCGUFFIN	18	1	4	230	0.87	1.022
T0737	SP-ALIGN	16	7	6	224	0.683	-0.475
T0737	HHPREDA	16	4	6	227	0.741	-0.011
T0737	CHUO-BINDING-SITES	19	40	3	191	0.46	-2.260
T0737	SEOK	16	1	6	230	0.814	0.573
T0737	CNIO	18	2	4	229	0.845	0.822
T0744	CONPRED-UCL	13	15	6	293	0.531	-0.287
T0744	FNGUSHAK	16	6	3	302	0.768	1.213
T0744	FIRESTAR	15	2	4	306	0.825	1.573
T0744	3DLIGANDSITE2	16	9	3	299	0.716	0.884
T0744	COFACTOR_HUMAN	12	7	7	301	0.609	0.207
T0744	ATOME2_CBS	9	10	10	298	0.441	-0.856
T0744	COFACTOR	12	7	7	301	0.609	0.207
T0744	BINDING_KIHARA	3	2	16	306	0.289	-1.818
T0744	3DLIGANDSITE	10	3	9	305	0.619	0.270
T0744	SEOK-SERVER	7	13	12	295	0.318	-1.635
T0744	INTFOLD2	10	2	9	306	0.647	0.447
T0744	MCGUFFIN	9	1	10	307	0.639	0.396
T0744	SP-ALIGN	16	13	3	295	0.658	0.517
T0744	HHPREDA	9	8	10	300	0.472	-0.660
T0744	CHUO-BINDING-SITES	14	38	5	270	0.392	-1.166
T0744	SEOK	9	7	10	301	0.489	-0.553
T0744	CNIO	15	4	4	304	0.776	1.263

Table S2 Statistical significance and robustness of the ranking. Differences in accuracy of binding residue predictions was assessed using the Wilcoxon signed-rank test, indicating that the differences between the top ten groups are not significant. The first 11 groups are shown.

		FN119	FN326	FN475	FN208	FN473	FN285	FN261	FN430	FN273	FN227	FN082
Firestar	**FN119**	NA	0.59	0.56	0.31	0.22	0.17	0.07	0.33	0.07	0.07	0.01
SP-ALIGN	**FN326**	0.59	NA	0.79	0.54	0.54	0.5	0.27	0.74	0.5	0.08	0.05
CNIO	**FN475**	0.56	0.79	NA	0.45	0.31	0.33	0.22	0.67	0.27	0.09	0.03
COFACTOR_human	**FN208**	0.31	0.54	0.45	NA	0.76	0.91	0.31	0.55	0.62	0.36	0.13
Seok	**FN473**	0.22	0.54	0.31	0.76	NA	0.84	0.91	0.95	0.48	0.74	0.24
McGuffin	**FN285**	0.17	0.5	0.33	0.91	0.84	NA	0.93	0.74	0.29	0.59	0.27
Seok-server	**FN261**	0.07	0.27	0.22	0.31	0.91	0.93	NA	0.95	0.67	0.82	0.45
HHPredA	**FN430**	0.33	0.74	0.67	0.55	0.95	0.74	0.95	NA	0.81	0.95	0.31
IntFold2	**FN273**	0.07	0.5	0.27	0.62	0.48	0.29	0.67	0.81	NA	0.64	0.31
COFACTOR	**FN227**	0.07	0.08	0.09	0.36	0.74	0.59	0.82	0.95	0.64	NA	0.64
FNGUSHAK	**FN082**	0.01	0.05	0.03	0.13	0.24	0.27	0.45	0.31	0.31	0.64	NA

Table S3 Group BDT scores for each target.

Group / Target	T0652	T0657	T0659	T0675	T0686	T0696	T0697	T0706	T0720	T0721	T0726	T0737	T0744
Firestar	0.77	0.2	0	0.9	0.6	1	0.82	0.77	0.27	0.81	1	0.77	0.83
SP-ALIGN	0.84	0.8	0.5	0.51	0.75	0.7	0.78	0.77	0.43	0.74	0.14	0.77	0.57
CNIO	0.62	0.22	0	1	0.56	0.75	0.69	0.77	0.63	0.79	0.27	0.86	0.83
Seok-server	0.91	0.02	0	0.82	0.56	0.7	0.83	0.77	0.18	0.77	1	0.79	0.43
COFACTOR_human	0.79	0.21	0	1	0.6	0.7	0.93	0.67	0.19	0.81	0.3	0.83	0.71
McGuffin	0.83	0.01	0	0.51	0.75	0.42	0.87	0.77	0.24	0.75	1	0.86	0.55
Seok	0.91	0.02	0	1	0.6	0.31	0.78	0.77	0.27	0.79	0.6	0.79	0.54
IntFold2	0.83	0.01	0	0.51	0.6	0.42	0.87	0.77	0.25	0.85	0.6	0.86	0.58
HHPredA	0.77	0.01	0.73	0	0.43	0.21	0.82	0.88	0.62	0.86	0.27	0.8	0.58
COFACTOR	0.79	0.21	0.06	0.5	0.18	0.7	0.93	0.77	0.19	0.81	0.3	0.83	0.71
FNGUSHAK	0	0.01	0.18	0.43	0.33	0.5	0.77	0.57	0.29	0.81	0.23	0.71	0.75
ConPred-UCL	0.65	0.01	0	0.66	0.75	0.33	0.17	0.77	0	0.71	0.38	0.58	0.51
Binding_Kihara	0.65	1	0	0.89	0.24	1	0.26	0.02	0.26	0.31	0.04	0.37	0.32
3DLigandSite2	0.77	0.01	0	0	0.75	0	0.83	0	0.31	0.66	0.38	0.83	0.66
3DLigandSite	0	0.01	0	0	0.75	0.45	0.82	0	0	0.65	0.5	0.92	0.59
Atome2_CBS	0.72	0.01	0	0	0	0	0.89	0	0	0.78	0.13	0.68	0.54
Chuo-binding-sites	0.81	0.08	0	0.21	0.04	0.12	0.24	0.19	0.31	0.4	0.03	0.34	0.29

Table S4 Cumulative confusion matrices for all groups.

Group name	TP	FP	FN	TN	TPR	ACC
ConPred-UCL	82	83	63	2995	0,566	0,955
FNGUSHAK	93	73	39	2786	0,705	0,963
Firestar	106	38	39	3040	0,731	0,976
3DLigandSite2	90	53	35	2592	0,720	0,968
COFACTOR_human	103	51	39	2956	0,725	0,971
Atome2_CBS	79	69	63	2938	0,556	0,958
COFACTOR	99	70	46	3008	0,683	0,964
Binding_Kihara	42	29	103	3049	0,290	0,959
3DLigandSite	78	39	21	2298	0,788	0,975
Seok-server	85	35	57	2972	0,599	0,971
IntFOLD2	95	35	50	3043	0,655	0,974
McGuffin	90	27	55	3051	0,621	0,975
SP-ALIGN	110	61	35	3017	0,759	0,970
HHpredA	102	63	35	2948	0,745	0,969
chuo-binding-sites	117	467	28	2611	0,807	0,846
Seok	92	35	50	2972	0,648	0,973
CNIO	117	54	25	2953	0,824	0,975

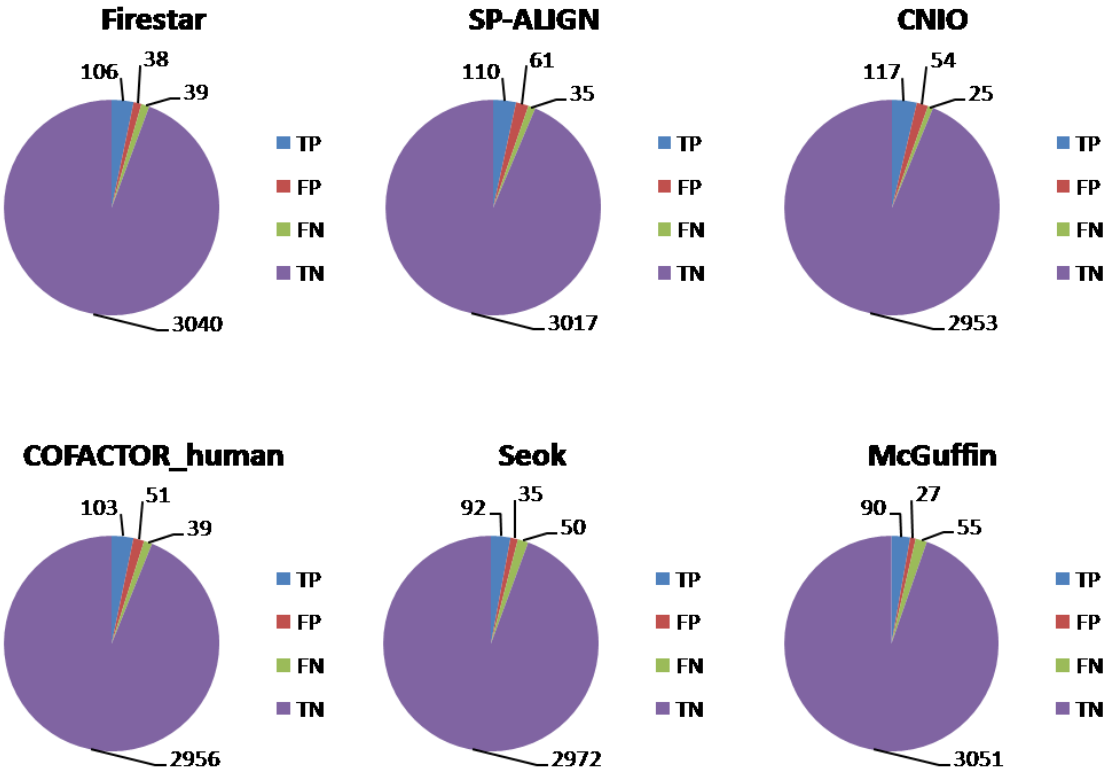

Figure S1: Cumulative confusion matrices for the top 6 groups as pie charts.

References

1. Pupko, T., et al., *Rate4Site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues.* Bioinformatics, 2002. 18 Suppl 1: p. S71-7.
2. Capra, J.A. and M. Singh, *Predicting functionally important residues from sequence conservation.* Bioinformatics, 2007. 23(15): p. 1875-82.
3. Fischer, J.D., C.E. Mayer, and J. Soding, *Prediction of protein functional residues from sequence by probability density estimation.* Bioinformatics, 2008. 24(5): p. 613-20.
4. Laurie, A.T. and R.M. Jackson, *Q-SiteFinder: an energy-based method for the prediction of protein-ligand binding sites.* Bioinformatics, 2005. 21(9): p. 1908-16.
5. Binkowski, T.A. and A. Joachimiak, *Protein functional surfaces: global shape matching and local spatial alignments of ligand binding sites.* BMC Struct Biol, 2008. 8: p. 45.
6. Ghersi, D. and R. Sanchez, *EasyMIFS and SiteHound: a toolkit for the identification of ligand-binding sites in protein structures.* Bioinformatics, 2009. 25(23): p. 3185-6.
7. Huang, B. and M. Schroeder, *LIGSITEcsc: predicting ligand binding sites using the Connolly surface and degree of conservation.* BMC Struct Biol, 2006. 6: p. 19.
8. Glaser, F., et al., *A method for localizing ligand binding pockets in protein structures.* Proteins, 2006. 62(2): p. 479-88.
9. Capra, J.A., et al., *Predicting protein ligand binding sites by combining evolutionary sequence conservation and 3D structure.* PLoS Comput Biol, 2009. 5(12): p. e1000585.
10. Lopez, G., A. Valencia, and M.L. Tress, *firestar--prediction of functionally important residues using structural templates and alignment reliability.* Nucleic Acids Res, 2007. 35(Web Server issue): p. W573-7.
11. Brylinski, M. and J. Skolnick, *A threading-based method (FINDSITE) for ligand-binding site prediction and functional annotation.* Proc Natl Acad Sci U S A, 2008. 105(1): p. 129-34.
12. Wass, M.N., L.A. Kelley, and M.J. Sternberg, *3DLigandSite: predicting ligand-binding sites using similar structures.* Nucleic Acids Res, 2010. 38(Web Server issue): p. W469-73.
13. Mariani, V., et al., *Assessment of template based protein structure predictions in CASP9.* Proteins, 2011. 79 Suppl 10: p. 37-58.
14. Kinch, L., et al., *CASP9 assessment of free modeling target predictions.* Proteins, 2011. 79 Suppl 10: p. 59-73.
15. Huang, Y.J., et al., *Assessment of template-based protein structure predictions in CASP10.* Proteins, 2014. 82 Suppl 2: p. 43-56.
16. Tai, C.H., et al., *Assessment of template-free modeling in CASP10 and ROLL.* Proteins, 2014. 82 Suppl 2: p. 57-83.
17. Kopp, J., et al., *Assessment of CASP7 predictions for template-based modeling targets.* Proteins, 2007. 69 Suppl 8: p. 38-56.
18. Battey, J.N., et al., *Automated server predictions in CASP7.* Proteins, 2007. 69 Suppl 8: p. 68-82.
19. Soro, S. and A. Tramontano, *The prediction of protein function at CASP6.* Proteins, 2005. 61 Suppl 7: p. 201-13.
20. Lopez, G., et al., *Assessment of predictions submitted for the CASP7 function prediction category.* Proteins, 2007. 69 Suppl 8: p. 165-74.
21. Lopez, G., I. Ezkurdia, and M.L. Tress, *Assessment of ligand binding residue predictions in CASP8.* Proteins, 2009. 77 Suppl 9: p. 138-46.
22. Schmidt, T., et al., *Assessment of ligand-binding residue predictions in CASP9.* Proteins, 2011. 79 Suppl 10: p. 126-36.

23. Taylor, T.J., et al., *Definition and classification of evaluation units for CASP10.* Proteins, 2014. 82 Suppl 2: p. 14-25.
24. Magrane, M. and U. Consortium, *UniProt Knowledgebase: a hub of integrated protein data.* Database (Oxford), 2011. 2011: p. bar009.
25. Biasini, M., et al., *OpenStructure: a flexible software framework for computational structural biology.* Bioinformatics, 2010. 26: p. 2626-2628.
26. Biasini, M., et al., *OpenStructure: an integrated software framework for computational structural biology.* Acta Crystallogr D Biol Crystallogr, 2013. 69(Pt 5): p. 701-9.
27. Roche, D.B., S.J. Tetchner, and L.J. McGuffin, *The binding site distance test score: a robust method for the assessment of predicted protein binding sites.* Bioinformatics, 2010. 26(22): p. 2920-2921.
28. R Development Core Team, *R: A language and environment for statistical computing*. 2011, R Foundation for Statistical Computing: Vienna, Austria.
29. Kryshtafovych, A., et al., *Challenging the state of the art in protein structure prediction: Highlights of experimental target structures for the 10th Critical Assessment of Techniques for Protein Structure Prediction Experiment CASP10.* Proteins, 2014. 82 Suppl 2: p. 26-42.
30. Bateman, A., *The structure of a domain common to archaebacteria and the homocystinuria disease protein.* Trends Biochem Sci, 1997. 22(1): p. 12-3.
31. Ignoul, S. and J. Eggermont, *CBS domains: structure, function, and pathology in human proteins.* Am J Physiol Cell Physiol, 2005. 289(6): p. C1369-78.
32. Scott, J.W., et al., *CBS domains form energy-sensing modules whose binding of adenosine ligands is disrupted by disease mutations.* J Clin Invest, 2004. 113(2): p. 274-84.
33. Kemp, B.E., *Bateman domains and adenosine derivatives form a binding contract.* J Clin Invest, 2004. 113(2): p. 182-4.
34. Mayer, B.J., et al., *A putative modular domain present in diverse signaling proteins.* Cell, 1993. 73(4): p. 629-30.
35. Wang, D.S., et al., *Binding of PH domains of beta-adrenergic receptor kinase and beta-spectrin to WD40/beta-transducin repeat containing regions of the beta-subunit of trimeric G-proteins.* Biochem Biophys Res Commun, 1994. 203(1): p. 29-35.
36. Boratyn, G.M., et al., *Domain enhanced lookup time accelerated BLAST.* Biol Direct, 2012. 7: p. 12.
37. Klug, A., *Zinc finger peptides for the regulation of gene expression.* J Mol Biol, 1999. 293(2): p. 215-8.
38. Hall, T.M., *Multiple modes of RNA recognition by zinc finger proteins.* Curr Opin Struct Biol, 2005. 15(3): p. 367-73.
39. Brown, R.S., *Zinc finger proteins: getting a grip on RNA.* Curr Opin Struct Biol, 2005. 15(1): p. 94-8.
40. Gamsjaeger, R., et al., *Sticky fingers: zinc-fingers as protein-recognition motifs.* Trends Biochem Sci, 2007. 32(2): p. 63-70.
41. Matthews, J.M. and M. Sunde, *Zinc fingers--folds for many occasions.* IUBMB Life, 2002. 54(6): p. 351-5.
42. Song, G., et al., *Shape change in the receptor for gliding motility in Plasmodium sporozoites.* Proc Natl Acad Sci U S A, 2012. 109(52): p. 21420-5.
43. Colombatti, A., P. Bonaldo, and R. Doliana, *Type A modules: interacting domains found in several non-fibrillar collagens and in other extracellular matrix proteins.* Matrix, 1993. 13(4): p. 297-306.
44. Lee, J.O., et al., *Crystal structure of the A domain from the alpha subunit of integrin CR3 (CD11b/CD18).* Cell, 1995. 80(4): p. 631-8.
45. Qu, A. and D.J. Leahy, *Crystal structure of the I-domain from the CD11a/CD18 (LFA-1, alpha L beta 2) integrin.* Proc Natl Acad Sci U S A, 1995. 92(22): p. 10277-81.

46. Toney, M.D., *Reaction specificity in pyridoxal phosphate enzymes.* Arch Biochem Biophys, 2005. 433(1): p. 279-87.
47. Kunin, V., R. Sorek, and P. Hugenholtz, *Evolutionary conservation of sequence and secondary structures in CRISPR repeats.* Genome Biol, 2007. 8(4): p. R61.
48. Jansen, R., et al., *Identification of genes that are associated with DNA repeats in prokaryotes.* Mol Microbiol, 2002. 43(6): p. 1565-75.
49. Yeeles, J.T., R. Cammack, and M.S. Dillingham, *An iron-sulfur cluster is essential for the binding of broken DNA by AddAB-type helicase-nucleases.* J Biol Chem, 2009. 284(12): p. 7746-55.
50. Zhang, J., T. Kasciukovic, and M.F. White, *The CRISPR associated protein Cas4 Is a 5' to 3' DNA exonuclease with an iron-sulfur cluster.* PLoS One, 2012. 7(10): p. e47232.
51. Rawlings, N.D. and A.J. Barrett, *Evolutionary families of metallopeptidases.* Methods Enzymol, 1995. 248: p. 183-228.
52. Ponting, C.P., *Evidence for PDZ domains in bacteria, yeast, and plants.* Protein Sci, 1997. 6(2): p. 464-8.
53. Ranganathan, R. and E.M. Ross, *PDZ domain proteins: scaffolds for signaling complexes.* Curr Biol, 1997. 7(12): p. R770-3.
54. Chauviac, F.X., et al., *Crystal structure of reduced MsAcg, a putative nitroreductase from Mycobacterium smegmatis and a close homologue of Mycobacterium tuberculosis Acg.* J Biol Chem, 2012. 287(53): p. 44372-83.
55. Roy, A., J. Yang, and Y. Zhang, *COFACTOR: an accurate comparative algorithm for structure-based protein function annotation.* Nucleic Acids Res, 2012. 40(Web Server issue): p. W471-7.
56. Lopez, G., A. Valencia, and M. Tress, *FireDB--a database of functionally important residues from proteins of known structure.* Nucleic Acids Res, 2007. 35(Database issue): p. D219-23.
57. Yang, J., A. Roy, and Y. Zhang, *BioLiP: a semi-manually curated database for biologically relevant ligand-protein interactions.* Nucleic Acids Res, 2013. 41(Database issue): p. D1096-103.
58. Schwede, T., *Protein Modeling: What Happened to the "Protein Structure Gap"?* Structure, 2013. 21(9): p. 1531-1540.
59. Kundrotas, P.J., et al., *Templates are available to model nearly all complexes of structurally characterized proteins.* Proc Natl Acad Sci U S A, 2012. 109(24): p. 9438-41.
60. Stein, A., R. Mosca, and P. Aloy, *Three-dimensional modeling of protein interactions and complexes is going 'omics.* Curr Opin Struct Biol, 2011. 21(2): p. 200-8.
61. Xu, Q. and R.L. Dunbrack, Jr., *The protein common interface database (ProtCID)--a comprehensive database of interactions of homologous proteins in multiple crystal forms.* Nucleic Acids Res, 2011. 39(Database issue): p. D761-70.

4. CAMEO Ligand binding

Introduction

The recent developments of high-throughput sequencing techniques and the setup of structural genomic initiatives have increased the number of available protein sequences and structures, however mostly without functional characterization. To investigate the role of these proteins, several computational methods have been developed (for some examples, see [1-3]); in particular, the annotation of protein binding sites and of their ligands provided a fundamental step in the discovery of protein functional details at the molecular level. This piece of information is, in fact, essential for important applications such as drug design and enzyme engineering. For this reason, computational methods for ligand binding prediction were assessed each two years starting from the 7th edition of CASP in 2007 [4]. These evaluations provided a valuable tool for comparing the performances of different methods, but two major limitations in the assessment emerged during the last CASP editions (see [5] in chapter 2 and [6] in chapter 3). Essentially, these drawbacks consisted in the low number of target structures bound to biologically relevant ligands and the classification of the target residues in either “binding” or “non-binding”, without taking into account the affinities for different potential ligands.

The Continuous Automated Model EvaluatiOn (CAMEO) Ligand Binding framework was developed to solve these issues and to provide a constant assessment of the state-of-the-art prediction methods. More in detail, participants are evaluated on the weekly PDB releases, in order to assess their server performances, in the long term, on a larger number of targets than in CASP. Additionally, the binary classification has been substituted by a continuous score that reflects the binding likelihood. Finally, the predictions are evaluated in a separate way for each chemical type of ligand.

Methods

Targets

To each registered server, every week CAMEO sends a group of pre-released PDB sequences with a minimum length of 30 amino acids. All the received predictions are collected by CAMEO

until the PDB publishes the structures of the pre-released sequences. Then, the Ligand Binding section of CAMEO selects all the assembly units of only those structures with biologically relevant ligands and evaluates the performances of each participant on the selected targets.

Ligands

All the small molecules present in the PDB target structures are categorized in four classes – Ion (I), Organic(O), poly-Nucleotide(N), poly-Peptide(P) (as described in http://www.cameo3d.org/cameong_help/lb/) – which are derived from the scheme adopted in the chemical component dictionary of the PDB. Then, each ligand is classified as "biologically relevant", "irrelevant" or "covalently bound" according to the following different criteria: (i) the distances between the ligand atoms and the protein, (ii) the annotation of the ligand as a commonly observed buffer or crystallization molecule, (iii) the presence/absence of covalent bonds between the ligand and the protein. Additionally, a web-based annotation platform allows users to manually change these classifications for any ligand bound to a CAMEO target (a tutorial page is available at http://www.cameo3d.org/annotation/support).

Prediction Format

We developed a new format for the binding site prediction that overcomes the limitations observed in the last CASP assessments [5, 6]; however, servers still formatting the predictions in the CASP style are allowed to send them to CAMEO, which will automatically perform the conversion to the new format. In the CAMEO format, a probability can be assigned either to each atom, to each residue or to a mixture of both, for every target chain. In the prediction, each entry contains two mandatory and one optional blocks of data, separated by the symbol "|", that contain a set of key-value pairs. In the following description of the block structure, the names within the symbols "<" and ">" indicate a value, while the data between the symbols "[" and "]" is optional. The first section uniquely indicates a residue or an atom; it is mandatory to specify the residue name and number, while it is optional to indicate the chain name or the atom name. The syntax is:

"r=<residue name>; n=<residue number>; [c=<chain name>;] [a=<atom name>;] |"

The second section contains the probabilities assigned to each ligand category, which reflect the likelihood of a ligand to be in contact with the residue, or with the atom, of the entry; the data consists of:

“I=<score>; O=<score>; N=<score>; P=<score>; |”

In the third section it is possible to indicate a score for each predicted compound using the three-letter code of the PDB:

[<compound *i* name>=<score>;].

Moreover, it is also possible to skip residue or atom entries for which all the probabilities were predicted to be zero.

Baseline homology predictor server

We implemented three servers to use as baseline methods for the comparison to the participant servers. Each one of the three reference server uses a different approach, which is based on sequence conservation, geometric binding pocket identification and homology transfer, respectively.

In particular, the server employing the latter approach collects the small molecules present in the protein template and places them within the target model. More in detail, the server superimposes the template onto the model built by SWISS-MODEL and identifies as members of the binding site all those residues that are within 3 Angstroms from a ligand. Finally, the server transfers to the model only those ligands which are included in a list of biologically relevant molecules (see Table 1 in the Supplementary information chapter) and which fulfil different criteria: (i) the ligand must bind at least 3 residues, (ii) the model binding residues must be completely conserved, (iii) none of the ligand atoms should be within 1.5 Angstroms from any of the protein atoms, and (iv) the RMSD of the binding residues between the template and the model must be less than 2 Angstroms. These strict rules allow a high confidence in the correctness of the predicted ligands and of their pose within the model.

The server calculates a score that represents the likelihood of each atom to bind a ligand, using a linear function that depends on the distance d between the protein’s atom and the nearest ligand atom. The score $s(d)$ is calculated for each ligand category as:

$$s(d) = \begin{cases} 1, & d < 3 \\ 2 - \frac{1}{3}d, & 3 \leq d \leq 6 \\ 0, & d > 6 \end{cases}$$

and its value is 0.5 with a distance *d* of 4 Angstroms.

Assessment

A reference probability for each atom is calculated for all the chains in a biological assembly of a given target using a sigmoid function, defined as:

$$p(d) = \frac{1}{1 + e^{(1.5d - 7.5)}}$$

where d is the distance between a protein atom and a ligand atom. The function parameters were optimized to result in a probability of:

- *p(d) = 1* for distances less or equal than 3 Angstroms
- *p(d) = 0* for distances greater or equal than 7 Angstroms
- *p(d)* = 0.5 at a distance of 5 Angstroms.

This probability is calculated for each ligand category. Additionally, if a compound is specified in either the organic or ion category, the probability is also computed for that particular compound.

Each week CAMEO assigns a set of scores to every server using four different methods: the Area Under the Curve (AUC), the Pearson's correlation coefficient, Spearman's rank correlation coefficient and Matthew's correlation coefficient. The first score can be interpreted as the probability that the server assigns a higher score to a binding residue than to a non-binding residue. CAMEO defines this score as follows: if an atom has a probability above 0.5, the atom is defined as "binding"; otherwise, if the probability is below 0.5, it is defined as "non-binding". The second score measures the linear correlation between the prediction and the reference probabilities, while the third measures the monotonic relationship between the same two sets of probabilities. The last score is a correlation coefficient between the predicted and the reference classifications of the residues and is used for comparing the server performances with the CASP sessions.

In case multiple chains are present in the reference, in the prediction or in both, the overall score assigned to the server is the average over all the best scores calculated for each chain in the prediction. Moreover, if the prediction contains all the correct ligands belonging to a certain category, the comparison will be based on the probabilities of these single compounds, rather than on the value assigned to the whole category.

Results and discussions

The main limitations in CASP ligand binding assessments were the prediction of the binding residues without considering the type of the bound ligand and the small number of target structures binding biologically relevant compounds. Moreover, only a subset of these proteins was considered to be “difficult targets”, that is, proteins for which there were no structures or ligand annotations associated to close homologs. CAMEO was developed to overcome these issues by evaluating the registered servers on the weekly released PDB structures. To date (2014-04-11), the number of targets evaluated by CAMEO reached 5260 proteins, containing 2473 ions, 3877 organics, 351 poly-nucleotides and 119 poly-peptides. An overview of the performances obtained over the last 3 months (from 2014-01-10 to 2014-04-11) in the ion and organic categories by the registered servers is shown in Figure 4.1.

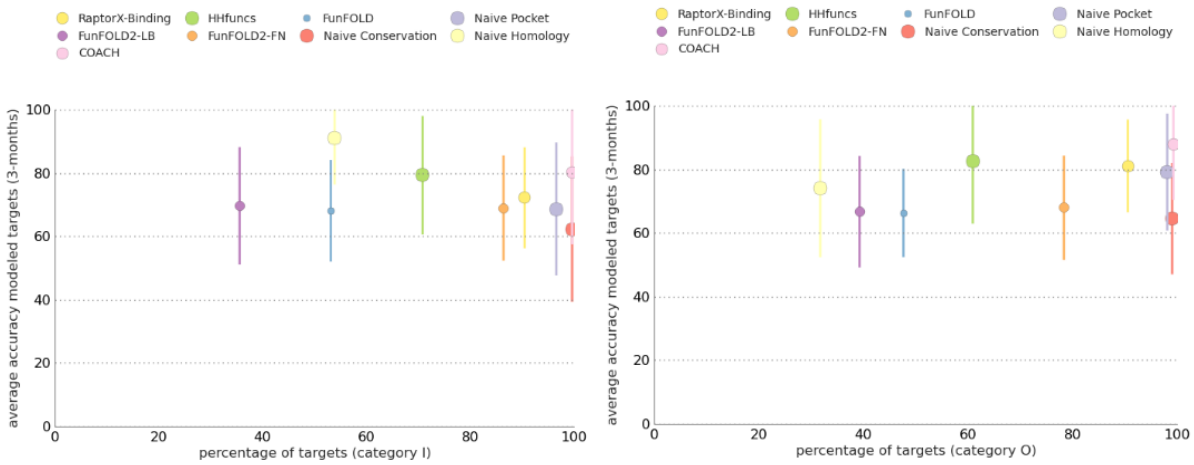

Figure 4.1 The performances over the last 3 months (from 2014-01-10 to 2014-04-11) of the registered servers in CAMEO Ligand Binding within the Ion (left) and Organic (right) categories. Our baseline homology predictor server is indicated as “Naive homology” by the light yellow dot. The x axis represents the percentage of targets for which a prediction was calculated, while the y axis represents the average accuracy calculated on the predicted targets.

The plots in Figure 4.1 show that our baseline predictor is very good at predicting binding sites for ions, but only for about half of the targets received; on the other hand, in the organic category our predictor had an average accuracy, but on a low number of targets.

These results indicate that our baseline method can be used as a first choice for predicting ion binding sites, since it showed the best performance in finding the correct residues. However, the performances of this server are limited by the small number of templates that bind biologically relevant ligands. Therefore, the first improvement to be made would be the analysis of more

homologs for a given target, in order to enlarge the available pool of ligands for predicting the correct model binding sites. To improve the functional prediction of a model by modelling ligands in the predicted binding sites, we developed a new method that takes into account ligands and binding sites from the multiple homologs of a target (see chapter 6).

Supplementary information

Table 1 Ligands evaluated by SWISS-MODEL.

Type	PDB codes
Ions	CA, CO, CU, CU2, FE, FE2, MG, MN, MO, NA, NI, ZN
Organic molecules	ADP, AMP, ATP, BTN, COA, BGC, GLC, GDP, GMP, GTP, GSH, FAD, FMN, HEM, HEA,HEB, NAD, NAP, NDP, NAI, PLP, SAM, THG, TPP, UDP, CDP, SF4, FES

References

1. Capra, J.A. and M. Singh, *Predicting functionally important residues from sequence conservation.* Bioinformatics, 2007. 23(15): p. 1875-82.
2. Pazos, F., A. Rausell, and A. Valencia, *Phylogeny-independent detection of functional residues.* Bioinformatics, 2006. 22(12): p. 1440-8.
3. Lopez, G., et al., *firestar--advances in the prediction of functionally important residues.* Nucleic Acids Res, 2011. 39(Web Server issue): p. W235-41.
4. Lopez, G., et al., *Assessment of predictions submitted for the CASP7 function prediction category.* Proteins, 2007. 69 Suppl 8: p. 165-74.
5. Schmidt, T., et al., *Assessment of ligand-binding residue predictions in CASP9.* Proteins, 2011. 79 Suppl 10: p. 126-36.
6. Gallo Cassarino, T., L. Bordoli, and T. Schwede, *Assessment of ligand binding site predictions in CASP10.* Proteins, 2014. 82 Suppl 2: p. 154-63.

5. SWISS-MODEL: modelling protein tertiary and quaternary structure using evolutionary information

This chapter has been accepted for publication as:

"SWISS-MODEL: modelling protein tertiary and quaternary structure using evolutionary information", Marco Biasini[1,2], Stefan Bienert[1,2], Andrew Waterhouse[1,2], Konstantin Arnold[1,2], Gabriel Studer[1,2], Tobias Schmidt[1,2], Florian Kiefer[1,2], Tiziano Gallo Cassarino[1,2], Martino Bertoni[1,2], Lorenza Bordoli[1,2] and Torsten Schwede[1,2],*. Nucleic Acids Research.

[1] Biozentrum, University of Basel, Basel, 4056, Switzerland

[2] SIB Swiss Institute of Bioinformatics, Basel, 4056, Switzerland

Contribution: I developed and implemented the part of the SWISS-MODEL pipeline that models ligands into the protein model structures.

Abstract

Protein structure homology modelling has become a routine technique to generate three-dimensional models for proteins when experimental structures are not available. Fully automated servers such as SWISS-MODEL with user-friendly web interfaces generate reliable models without the need for complex software packages or downloading large databases. Here, we describe the latest version of the SWISS-MODEL expert system for protein structure modelling. The SWISS-MODEL template library provides annotation of quaternary structure and essential ligands and co-factors to allow for building of complete structural models, including their oligomeric structure. The improved SWISSMODEL pipeline makes extensive use of model quality estimation for selection of the most suitable templates and provides estimates of the expected accuracy of the resulting models. The accuracy of the models generated by SWISS-MODEL is continuously evaluated by the CAMEO system. The new web site allows users to interactively search for templates, cluster them by sequence similarity, structurally compare alternative templates, and select the ones to be used for model building. In cases where multiple alternative template structures are available for a protein of interest, a user-guided template

selection step allows building models in different functional states. SWISS-MODEL is available at http://swissmodel.expasy.org/.

Intoduction

SWISS-MODEL (http://swissmodel.expasy.org/) is an automated system for modelling the three-dimensional structure of a protein from its amino acid sequence using homology modelling techniques. SWISS-MODEL has been established 20 years ago as the first fully automated server for protein structure homology modelling, and has been continuously developed and improved since then [1, 2] [3] [4]. The server features a user-friendly web interface, which allows also non-specialists to generate three-dimensional models for their protein of interests from a simple web-browser without the need to install and learn complex molecular modelling software, or to download large databases [5]. Today, SWISS**-**MODEL is one of the most widely used structure modelling web servers world-wide, with more than 0.9 million requests for protein models annually (i.e. approximately one model per minute). Recently, its functionality has been greatly extended: SWISS-MODEL now models oligomeric structures of target proteins, and includes evolutionary conserved ligands such as essential cofactors or metal ions in the model. A newly developed interactive web interface allows users to conveniently search for suitable templates using sensitive HMM searches against the SWISS-MODEL Template Library (SMTL), analyse alternative templates and alignments, perform structural superposition and comparison, explore ligands and cofactors in templates, and compare the resulting models using mean force potential based model quality estimation tools. Model quality estimation is an essential component of protein structure predictions, as the accuracy of a model determines its usefulness for practical applications. SWISS-MODEL provides model quality estimates (visually in the web page and numerically for download) based on a QMEAN potential [6] [7] specifically re-parameterized for models build by SWISS-MODEL. The accuracy of the SWISS-MODEL server is independently evaluated in comparison with other state-of-the-art methods by the CAMEO project (http://cameo3d.org/; Continuous Automated Model EvaluatiOn) [8] based on target sequences weekly pre-released by the PDB [9].

Materials and Methods

Overview

Homology modelling (or comparative modelling) relies on evolutionarily related structures (templates) to generate a structural model of a protein of interest (target). The process typically comprises the following steps: (I) Template identification; (II) Template selection; (III) Model building; and (IV) Model quality estimation [10] [11]. In brief, a library of experimentally determined protein structures is searched with sensitive sequence search tools to identify proteins which are evolutionary related to the target protein. If one or more templates are identified, the information of the alignment of the target and the template sequences together with the 3-dimensional coordinates of the template(s), are used to build a structural model for the protein of interest. Finally the quality of the computed model is estimated to indicate the expected quality and suggest possible application of the obtained model.

The SWISS-MODEL Template Library (SMTL)

Comparative modelling methods make use of information from experimentally determined protein structures to generate models for a target protein. A well-curated and annotated template library which supports efficient queries is therefore a crucial component of a modelling server. The SMTL aggregates information of experimental structures from the PDB (9) and augments it with derived information. When a new structure is released by the PDB, the coordinates and accompanying information are processed and imported into the template library. SMTL entries are organized by likely quaternary structure assemblies, termed “bio units”, which are created according to the author- and software annotated oligomeric states listed in the PDB deposition. Template amino acid sequences are indexed in a searchable databases for BLAST [12], and added to a HMM library that can be searched by HHblits [13]. Sequence Profiles, predicted secondary structure (SSpro [14], PSIPRED [15]), predicted solvent accessibility (ACCpro [14]), per-residue solvent accessibility, (NACCESS (S. Hubbard and J. Thornton)), secondary structure (DSSP [16]) are calculated and stored alongside the structure. In addition, protein purification tags, such as HIS or TAP tags are detected in the sequences and marked as such. The implementation of computational routines in SMTL is based on OpenStructure [17].

Annotation of Ligands in SMTL

In most crystal structures low molecular weight ligands are observed, but only some of those are functionally or structurally relevant for the protein. Instead of their natural ligands, some structures contain synthetic analogues or inhibitors which occupy competitively the same binding site. Often, buffer or precipitant molecules are encountered, which are added by experimentalists to facilitate crystallization. SMTL implements a two-stage process to annotate

biologically relevant ligands and synthetic analogues. The first stage uses a list of rules to automatically categorize the ligands based on their chemical identity. For example, all potassium ions are classified as solvent at this stage. In a second stage, the SMTL web interface provides a way to change the ligand classification manually. For example, in case of a potassium channel structure some of the before-mentioned potassium ions may be re-annotated as biologically relevant. While re-annotations can be suggested by any SWISSMODEL user, before taking effect in SMTL, the annotations are reviewed by a curator to guarantee high quality of annotations.

Template Search and Selection

The SWISS-MODEL Template Library is searched in parallel both with BLAST and HHblits to identify templates and to obtain target-template alignments. The combined usage of these two methods guarantees good alignments at high and low sequence identity levels [18]. In order to select the most suitable templates, the procedure implemented in SWISS-MODEL uses properties of the target-template alignment (sequence identity, sequence similarity, HHblits score, agreement between predicted secondary structure of target and template, agreement between predicted solvent accessibility between target and template; all normalized by alignment length) to predict the expected quality of the resulting model (Biasini, M., *et al, manuscript in preparation*). In brief, each of the alignment properties is modelled as probability density function (PDF) of the estimate for a resulting model having a certain structural similarity to the target. The use of PDFs has the advantage of at once including the expectation value as well as the accuracy of the estimate for each property. It also takes into account, that some properties are better (more accurate) at predicting the quality at high levels of sequence identity, whereas others are more accurate in the twilight zone of sequence alignments. For each property the most likely structural similarity of the template to the target is the value at which the PDF is maximal. Properties are combined based on their relevance, which has been determined from large sets of target/template alignments with known target structures. When combining the estimates of each property, the most likely structural similarity is the value at which the joint distribution is maximized, termed the global quality estimation score (GMQE).

Model Building and Scoring

After templates are selected for model building, either by using the automated or manual selection mode, the target/template alignment is used as input for generating an all-atom model for the target sequence using ProMod-II [19]. In case loop modelling with ProMod-II does not

give satisfactory results, an alternative model is built with MODELLER [10]. By default, models are built using the homo-oligomeric structure of the template as annotated in SMTL, provided the oligomeric structure is predicted as conserved (see Oligomeric Structure Prediction below). An indispensable part of every modelling procedure is the estimation of a protein model's accuracy, directly providing the user with information regarding the range of its possible applications [11, 20, 21]. Here, model quality is assessed with the local composite scoring function QMEAN, which uses several statistical descriptors expressed as potentials of mean force: geometrical features of the model (pairwise atomic distances, torsion angles, solvent accessibility) are compared to statistical distributions obtained from experimental structures and scored. Each residue is assigned a reliability score between 0 and 1, describing the expected similarity to the native structure. Higher numbers indicate higher reliability of the residues. The weights of QMEAN have been specifically retrained for SWISS-MODEL, leading to more accurate local quality predictions for single models (Studer, G., *et al.*, *manuscript in preparation*). In addition, global QMEAN scores are calculated as indicators for the overall model quality. Global QMEAN estimates are provided as a Z-score which relates the obtained values to scores calculated from a set of high-resolution X-ray structures [7]. Additionally, a combined quality estimate is provided, which combines the QMEAN estimate with the GMQE obtained from the target-template alignment as described before. The resulting GMQE is again expressed as a number between zero and one, where higher numbers indicate higher reliability.

Oligomeric Structure Prediction

The majority of proteins in a living cell exist as part of complexes and quaternary structure assemblies, monomeric proteins being the exception rather than the rule [22]. Frequently, ligand binding sites and enzyme active sites are located at protein chain interfaces, and modelling of the oligomeric structure of a protein is therefore essential to build models which are useful in biomedical applications [23]. Here, the homo-oligomeric structure of a target protein is modelled based on the hypothesis that the quaternary structure is conserved in one of the templates. To test this hypothesis, conservation of the oligomeric structure is predicted by analysing properties of interfaces between polypeptide chains such as sequence identity, sequence similarity, interface hydrophobicity, and consensus occurrence of the same interface in the set of identified templates. A random forest is generated using these features as input parameters to predict the probability of conservation for each interface. When the size-weighted average of interface conservation is higher than a defined threshold, the oligomeric structure of the target is predicted to be the same as in the template.

Modelling of Ligands

For predicting essential ligands and cofactors for a given target protein, we apply a conservative homology transfer approach to small molecules which are observed in the templates identified in the SMTL. Ligands in SMTL are annotated either as: a) relevant, non-covalently bound ligand, b) covalent modifications, or c) non-functional binders (e.g. buffer or solvent). A non-covalently bound ligand is considered for the model if the coordinating residues are conserved in the target-template alignment. The relative coordinates of the ligand are transferred from the template, if the resulting atomic interactions in the model are within the expected range for van der Waals interactions and water mediated contacts.

Performance of the Method (CAMEO)

The performance and reliability of the SWISS-MODEL server is continuously evaluated by the CAMEO project (Continuous Automated Model EvaluatiOn) [8]. Modelling servers are blindly assessed based on sequences pre-released by the PDB for proteins which structure will be published in the next release. Servers have four days to predict the 3-dimensional structure of the target proteins before models are evaluated against the protein structure coordinates released by the PDB using superposition-independent scoring methods such as CAD score [24] and lDDT [25]. The current CAMEO evaluation for this version of SWISS-MODEL consist of 6424 predictions for 599 target proteins collected over 52 weeks (i.e. from 2013-03-01 to 2014-02-28; data available at http://cameo3d.org). SWISS-MODEL accuracy is compared to other state-of-the-art protein structure prediction servers [26-32] and to previous version of the server [5].

Webserver Implementation

The web frontend to SWISS-MODEL follows the typical design of modern websites where business logic is implemented in JavaScript and executed directly in the browser. For improved user-interaction, data is fetched asynchronously from the server, without the need to reload the complete page. The front-end uses jQuery (jquery.com) to guarantee cross-browser compatibility. For 3D structure visualization, the user can chose between a modified version of OpenAstexViewer (openastexviewer.net) Java plugin, and the WebGL-based PV (https://biasmv.github.io/pv). The frontend communicates with a Django (www.djangoproject.com) backend that handles all incoming requests. Computationally demanding calculations, e.g. template search and modelling, are submitted via a queuing system to a dedicated compute cluster.

SWISS-MODEL web interface

Input. Model building with SWISS-MODEL can be initiated from different starting information: In the simplest case, a protein amino acid sequence can be specified directly (raw one letter sequence or FASTA format) or by referring to its UniProt accession code, in which case SWISS-MODEL will automatically retrieve the corresponding entry from UniProt [33]. Alternatively, a target-template sequence alignment can be specified in the form of a multiple sequence alignment containing the target, the template, and eventually other homologous sequences, or in the form of a DeepView project file [3, 19]. At this point, the user can initiate the template selection step, which allows to manually select specific templates, or directly invoke the fully automated modelling pipeline.

Output template search results and manual template selection. Thanks to tremendous technical advances in experimental structure determination, for an increasing number of protein families there is not only one template, but multiple alternative template structures available. For some well-studied protein families, finding hundreds of possible templates for a target protein is not unusual. Often, these represent different functional states or structures in complex with different ligands. Depending on the intended application of a model, selecting a different template than the top-ranked one might be necessary, e.g. to build a model of a protein in complex with a ligand – rather than its apo form – for applications in drug design when induced fit movements are expected [34]. We have therefore developed a manual template selection mode to make template selection available to a larger user base. All the steps of manual template selection can be performed directly in the web-interface without the need to leave the browser environment (**Figure 1**).

Suitable templates identified for the target sequence are listed in a tabular form, sorted by their predicted global quality estimation score (GMQE). Each template lists biologically relevant ligands, the predicted oligomeric structure conservation and the target-template alignment. The tabular view allows quickly gaining an overview on the identified templates. The user can directly select one or more templates and initiate model building. Apart from comparing template properties in tabular form, two graphical comparison views help to better understand the landscape of available templates. An interactive 3D view of superposed templates shows the aligned part of selected template structures (**Figure 1C**), at the user's choice using a WebGL-based (PV), or Java-based (OpenAstex) viewer. The second view shows the evolutionary distance between templates on 2D plot (**Figure 1A**).

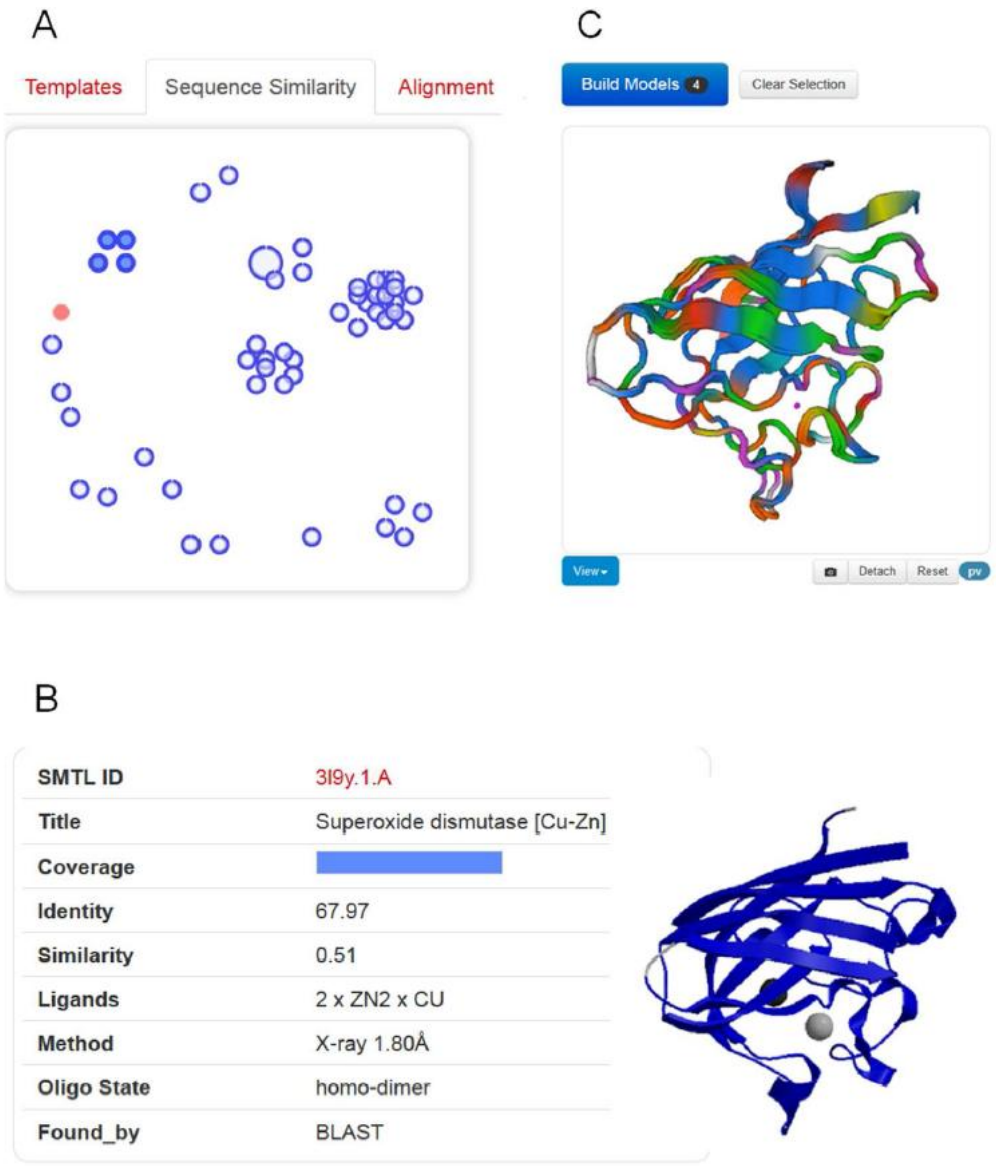

Figure 1. Templates Selection and Visualization. (A) An interactive chart shows the relationship of detected templates in sequence-similarity space. The target protein is represented as filled red circle. Each template is displayed as a blue circle, where the thick blue arc indicates target coverage (the N-terminus of the target protein starts at the top of the circle, and ends in clockwise direction with the C-terminus to close the circle). The distance between different templates is proportional to the pairwise sequence similarity, i.e. evolutionarily closely related templates will clustered together. (B) Clicking on a circle will display template-specific information. A group of similar templates can be also visualized and selected by hovering over a cluster of templates. (C) The superposed structures of the selected templates will be instantaneously displayed in 3D to visually inspect structural differences.

Groups of high-sequence identity templates cluster together, whereas more distant proteins are separated. The interactive graph allows marking groups of templates for structural superposition by selecting them with the cursor. The sequence similarity cluster view in combination with template superposition allows identifying functionally relevant states of the templates (“open / closed”). It also supports defining structurally conserved cores in the identified template

structures, and such regions where template which are not closely related share common structural features, are most likely well modelled in the target, while segments of structural variation in templates typically correlate with errors in the model [30, 35].

Output modelling. For each model generated based on the selected templates (either by the fully automated pipeline or interactively by the user), SWISS-MODEL provides the model coordinates along with relevant information to assess the modelling process and expected accuracy of the model (**Figure 2**): the target-template alignment, a step-by-step modelling log, information about the oligomeric state, ligands and cofactors in the model, as well as QMEAN model quality estimation. Models can be displayed interactively, initially coloured by model quality estimates assigned by QMEAN to highlight regions of the model which are well or poorly modelled. If several alternative models have been built for a target sequence, these can be interactively superposed and visualized. Model coordinates and information displayed on the website can be downloaded for later reference.

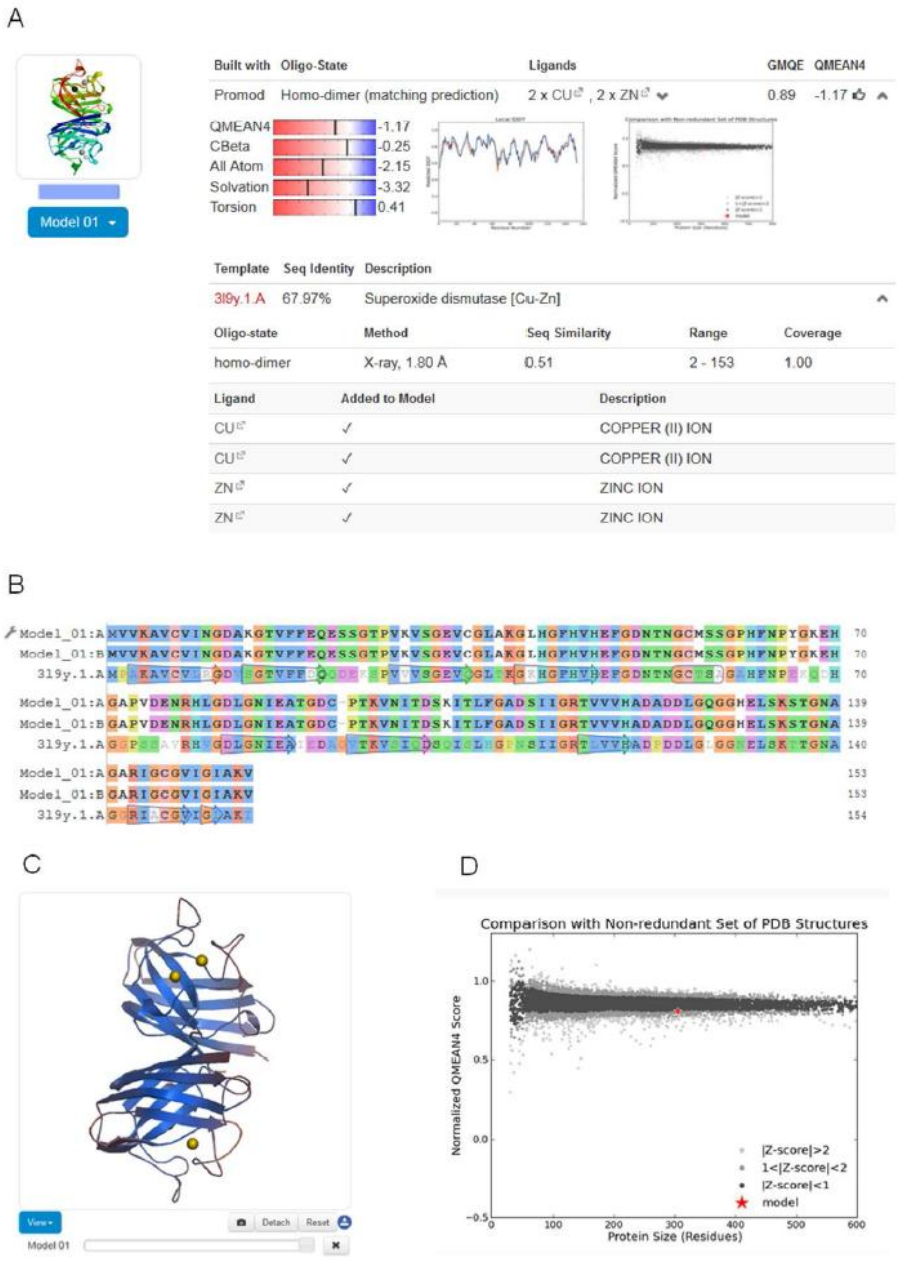

Figure 2. Modelling Results. (A) For each model, coordinates, target-template alignment, modelling log, and quality estimation information are provided. Information about the oligomeric structure, ligands and cofactors is also provided. (B) The colouring of the target-template sequence alignment can be changed to another scheme by clicking on the option button (adjustable spanner icon). Changes are simultaneously reflected in the structural representation of the model. (C) Models displayed in the interactive viewer are initially coloured by model quality estimates assigned by QMEAN. This allows instantly discriminating regions of the model which are well or poorly modelled. Local estimates of the model quality based on the QMEAN scoring function are shown as per-reside plot (A) and global score (GMQE) in relation to a set of high-resolution PDB structures (Z-score) (D).

Discussion and conclusions

Protein structure homology modelling has become a routine method to provide structural models on life science research in cases where no experimental structures are available. However, in order to support the understanding a protein's function in its biological context, realistic structural models should not only correctly represent the overall fold of a single protein chain, but also its quaternary structure, as well as the atomic details of interactions with essential cofactors and ligands. Modelling and assessment procedures must also be able to account for structural flexibility since proteins are not static entities, but may exist in structurally distinct functional states. With the new version of SWISS-MODEL presented here, we aimed to address these aspects by introducing a new augmented SWISS-MODEL Template Library, which includes information on quaternary structures and the role of ligands bound to the template. At the same time, we have significantly improved the accuracy of the fully automated SWISS-MODEL pipeline, aiming to reliably provide accurate models which are useful for applications in biomedical research. The expected accuracy of each specific model is communicated to the user in the form of QMEAN score, and the overall accuracy of SWISS-MODEL is continuously monitored in CAMEO. The implementation of the new web interface allows users to interactively compare alternative templates and select those which are more suitable for the intended application of the model (e.g. based on the presence / absence of specific ligands or structurally different functional states). The interactivity of the new web site required the usage of innovative programming techniques for the web front end, as well as speed optimization and hardware upgrades of the backend in order to provide a satisfying user experience.

Acknowledgements

We would like to thank all SWISS-MODEL users who participated in the user survey or individually sent us their feed-back, which greatly helped us in developing the new version. The authors gratefully acknowledge the computational resources provided by the sciCORE / $[BC]^2$ Basel Computational Biology Center at the University of Basel and the support by the system administration team.

Funding

This work was supported by SIB Swiss Institute of Bioinformatics. G. Studer has been supported by a PhD fellowship of the SIB by the “Swiss Foundation for Excellence and Talent in Biomedical Research”. M. Bertoni has been supported by a fellowship for the Biozentrum University of Basel international PhD program by the Werner von Siemens Foundation. Funding for open access charge: SIB Swiss Institute of Bioinformatics.

References

1. Guex, N., M.C. Peitsch, and T. Schwede, *Automated comparative protein structure modeling with SWISS-MODEL and Swiss-PdbViewer: a historical perspective.* Electrophoresis, 2009. 30 Suppl 1: p. S162-73.
2. Schwede, T., et al., *SWISS-MODEL: An automated protein homology-modeling server.* Nucleic Acids Res, 2003. 31(13): p. 3381-5.
3. Bordoli, L., et al., *Protein structure homology modeling using SWISS-MODEL workspace.* Nat Protoc, 2009. 4(1): p. 1-13.
4. Kiefer, F., et al., *The SWISS-MODEL Repository and associated resources.* Nucleic Acids Res, 2009. 37(Database issue): p. D387-92.
5. Arnold, K., et al., *The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling.* Bioinformatics, 2006. 22(2): p. 195-201.
6. Benkert, P., M. Kunzli, and T. Schwede, *QMEAN server for protein model quality estimation.* Nucleic Acids Res, 2009. 37(Web Server issue): p. W510-4.
7. Benkert, P., M. Biasini, and T. Schwede, *Toward the estimation of the absolute quality of individual protein structure models.* Bioinformatics, 2011. 27(3): p. 343-50.
8. Haas, J., et al., *The Protein Model Portal--a comprehensive resource for protein structure and model information.* Database (Oxford), 2013. 2013: p. bat031.
9. Berman, H., et al., *The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data.* Nucleic Acids Res, 2007. 35(Database issue): p. D301-303.
10. Sali, A. and T.L. Blundell, *Comparative protein modelling by satisfaction of spatial restraints.* J Mol Biol, 1993. 234(3): p. 779-815.
11. Schwede, T., et al., *Protein Structure Modeling*. 2008: World Scientific Publishing.
12. Altschul, S.F., et al., *Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.* Nucleic Acids Res, 1997. 25(17): p. 3389-402.
13. Remmert, M., et al., *HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment.* Nat Methods, 2012. 9(2): p. 173-5.
14. Cheng, J., et al., *SCRATCH: a protein structure and structural feature prediction server.* Nucleic Acids Res, 2005. 33(Web Server issue): p. W72-6.
15. Jones, D.T., *Protein secondary structure prediction based on position-specific scoring matrices.* J Mol Biol, 1999. 292(2): p. 195-202.
16. Kabsch, W. and C. Sander, *Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features.* Biopolymers, 1983. 22(12): p. 2577-637.
17. Biasini, M., et al., *OpenStructure: an integrated software framework for computational structural biology.* Acta Crystallogr D Biol Crystallogr, 2013. 69(Pt 5): p. 701-9.
18. Sadowski, M.I. and D.T. Jones, *Benchmarking template selection and model quality assessment for high-resolution comparative modeling.* Proteins, 2007. 69(3): p. 476-85.
19. Guex, N. and M.C. Peitsch, *SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling.* Electrophoresis, 1997. 18(15): p. 2714-23.
20. Baker, D. and A. Sali, *Protein structure prediction and structural genomics.* Science, 2001. 294(5540): p. 93-6.
21. Schwede, T., et al., *Outcome of a workshop on applications of protein models in biomedical research.* Structure, 2009. 17(2): p. 151-9.
22. Poupon, A. and J. Janin, *Analysis and prediction of protein quaternary structure.* Methods Mol Biol, 2010. 609: p. 349-64.
23. Mariani, V., et al., *Assessment of template based protein structure predictions in CASP9.* Proteins, 2011. 79 Suppl 10: p. 37-58.

24. Olechnovic, K., E. Kulberkyte, and C. Venclovas, *CAD-score: a new contact area difference-based function for evaluation of protein structural models.* Proteins, 2013. 81(1): p. 149-62.
25. Mariani, V., et al., *lDDT: a local superposition-free score for comparing protein structures and models using distance difference tests.* Bioinformatics, 2013. 29(21): p. 2722-8.
26. Kallberg, M., et al., *Template-based protein structure modeling using the RaptorX web server.* Nat Protoc, 2012. 7(8): p. 1511-22.
27. Yang, Y., et al., *Improving protein fold recognition and template-based modeling by employing probabilistic-based matching between predicted one-dimensional structural properties of query and corresponding native properties of templates.* Bioinformatics, 2011. 27(15): p. 2076-82.
28. Kelley, L.A. and M.J. Sternberg, *Protein structure prediction on the Web: a case study using the Phyre server.* Nat Protoc, 2009. 4(3): p. 363-71.
29. Rykunov, D., et al., *Improved scoring function for comparative modeling using the M4T method.* J Struct Funct Genomics, 2009. 10(1): p. 95-9.
30. Buenavista, M.T., D.B. Roche, and L.J. McGuffin, *Improvement of 3D protein models using multiple templates guided by single-template model quality assessment.* Bioinformatics, 2012. 28(14): p. 1851-7.
31. Kim, D.E., D. Chivian, and D. Baker, *Protein structure prediction and analysis using the Robetta server.* Nucleic Acids Res, 2004. 32(Web Server issue): p. W526-31.
32. Soding, J., A. Biegert, and A.N. Lupas, *The HHpred interactive server for protein homology detection and structure prediction.* Nucleic Acids Res, 2005. 33(Web Server issue): p. W244-8.
33. The UniProt Consortium, *Ongoing and future developments at the Universal Protein Resource.* Nucleic Acids Research, 2011. 39(suppl 1): p. D214-D219.
34. Schmidt, T., A. Bergner, and T. Schwede, *Modelling three-dimensional protein structures for applications in drug design.* Drug Discov Today, 2013.
35. Kryshtafovych, A., et al., *Assessment of the assessment: evaluation of the model quality estimates in CASP10.* Proteins, 2014. 82 Suppl 2: p. 112-26.

6. Modelling cofactors in homology models

This chapter describes the work in preparation for the manuscript:

"*Modelling cofactors in homology models by evolutionary inference*", Tiziano Gallo Cassarino[1,2] and Torsten Schwede[1,2]

[1]. Biozentrum, University of Basel, Switzerland

[2]. SIB Swiss Institute of Bioinformatics, Basel, Switzerland

Abstract

Recent developments in the field of ligand binding site prediction, assessed during the last CASP editions, indicate that the currently best performing methods make use of the close homologs of a target protein to infer its binding residues. Our aim is to extend this approach to the modelling of ligands, in particular ions and organic cofactors, in homology modelled protein structures. By comparing the target with a set of its homologous templates, we analyse several properties of their binding sites and find the best similarity descriptor to identify the most likely ligands that should be placed in the target model. To verify the quality of this approach, we assessed the performances of our method against the two leading prediction servers, COACH and RaptorX-Binding, from the CAMEO Ligand Binding category. Using a blind-test approach on a dataset consisting of several hundreds of protein structures, we show that our method performs clearly better than the other two servers, with the best precision-recall for ions and the highest sensitivity for organic cofactors.

Introduction

One of the biggest challenges in Biology is to reduce the gap between the ever-growing number of protein sequences deposited in public databases and, on the other hand, the relatively small fraction of these for which a biological function is known. A fundamental help in understanding protein functions is provided both by their 3D structure and by the interactions they are involved in with other molecules. To this purpose, it is necessary to identify which protein residues

participate in these interactions and, most interestingly, which molecules can serve as ligands within a protein binding pocket. Moreover, a deeper knowledge of protein binding preferences has proved to be essential to identify novel therapeutic targets [1], but also to discover natural ligands by structure-based drug design [2]. Current methods providing functional annotation at the residue level can be classified according to the main approach they adopt, which can be based either on the protein sequence [3-5] or on its structure [6-11].

Methods belonging to the first group usually measure each residue's conservation within homologous proteins and define as "binding residues" the most conserved amino acids in the sequence. However, although a clear advantage of this approach is that it can be used even when the protein structure is unknown, the main drawback is that some residues considered to be functional might have actually been evolutionarily conserved for other reasons (for instance, they might be involved in protein-protein interfaces or in maintaining protein stability).

On the other hand, the methods belonging to the second group, which adopt protein structure to identify binding residues, can be further distinguished in those that recognize protein surface cavities (for an example, see [8]) and those that infer the target binding residues from its homologous proteins (for an example, see [9]). Although pocket detection algorithms can be more successful in case the target has only distant homologs, these methods are outperformed in the opposite scenario - that is, when close homologs are available - as shown in the last edition of the community-wide Critical Assessment of protein Structure Prediction (CASP) competition under the "FN category" [12] and, more recently, in the Ligand Binding category of the Continuous Automated Model Evaluation (CAMEO) server (http://cameo3d.org/lb/).

Overall, the main objective of the above mentioned approaches is focussed on the identification of the specific residues which might be in contact with a ligand. However, so far only few attempts have been done to predict, in addition, the precise ligand conformation in the model (for an example, see [13]) and, thus, to provide a complete functional annotation for the protein of interest.

In this study, we contribute to improve the knowledge about a target protein by inferring its natural, i.e. biologically relevant, small molecule ligands and by placing these in the most likely conformation within the modelled structure. Our method analyses the similarities between the target structure and the ligand binding sites of its homologous proteins; a range of features are tested and compared, the best of which is used in the final implementation of the ligand modelling method. Moreover, we also extract and integrate the annotations from UniProt [14]

and from the SWISS-MODEL Template Library (SMTL) (see chapter 5), regarding all ligands interacting with the target's homologous proteins. Finally, the model binding site(s) are identified and filled (when possible) with the small molecules they are most compatible with. In a blind test using 364 targets provided by CAMEO, we compare the performances of our method to two state-of-the-art ligand binding predictors and we demonstrate that our method performs significantly better than the other two.

Methods

Datasets

To create a training dataset containing as many correctly bound ligands as possible, we built a non-redundant training set of high quality PDB protein chains that are experimentally annotated as binding ions (Cobalt, Calcium, Copper, Iron, Magnesium, Manganese, Nickel, Zinc) or organic cofactors, like S-adenosyl-L-methionine, Biotin or Flavin adenine dinucleotide (the full list can be found at CoFactor database: http://www.ebi.ac.uk/thornton-srv/databases/CoFactor/).

We used the EMBOSS suite 6.2.0 [15] to fetch all those UniProt-SwissProt (UniProt release 2013_05) entries which are annotated to bind one of the above mentioned molecules in the Sequence annotation (Features) field. Next, we retrieved the associated PDB code by using the SIFTS service [16]. From this pool of proteins, we only kept the high quality chains (X-Ray resolution <= 2Angstrom and R-Free <=0.25) bound to a ligand which was not located in an interface and which was in contact with the residues indicated by SwissProt. Following previous indications in the literature, for example in [17], this filtering step ensured that: (i) the ion-protein distances were in agreement with the Cambridge Structural Database (CSD) [18] statistics and (ii) the ion in the PDB structure was reliably identified.

For each small molecule, we obtained a non-redundant set of protein domains by clustering all protein chains bound to the selected ligand on the basis of the PFAM classification [19] (as of 2013-06-25) and by keeping only one member of each cluster as representative of the whole protein family domain. The resulting set was composed by 434 monomers bound to 495 small molecules, of which 352 are ions and 143 are organic cofactors. Afterwards, we built a model of each protein using the SWISS-MODEL server and we filtered out the templates belonging to our training set.

A large scale blind-test was used to assess the performances of our approach in comparison to other two state-of-the-art methods. The testing set used to this purpose was built by collecting the target proteins sent weekly by CAMEO to our and to the other two participants with the best performances, RaptorX-Binding (http://raptorx.uchicago.edu/BindingSite/) and COACH [11], in the Ligand Binding category of the Continuous Automated Model EvaluatiOn server (CAMEO). The models produced by these two participants were retrieved through the data sent by each participant to CAMEO Ligand Binding. For COACH, we considered the representative ligand of the best scoring cluster, while for RaptorX-Binding we used the compound indicated for each pocket.

CAMEO tries to validate the biologically relevance of ligands with a crowd-sourcing approach through an annotation platform; therefore, CAMEO targets could contain both biologically and non-biologically relevant compounds. To exclude irrelevant small molecules that have not been manually curated, we kept in the structures only those ligands included in a list of known natural occurring, or "cognate", compounds (as in Table 2 within the Supplementary information chapter), that we retrieved from the FireDB database [20]. Moreover, we excluded those “cognate” ligands that could be irrelevant for a specific target structure by removing all the molecules not bound to at least three protein residues, or present more than 15 times. All the targets that did not bind at least one relevant molecule were removed from the testing set.

A total of 614 target structures, and the corresponding server models, were collected between 2014-01-10 and 2014-04-04; among these, there were 364 proteins bound to 1004 biologically relevant ligands, of which 555 were ions, 436 organic, 7 nucleotides and 6 polypeptides, according to the CAMEO classification of hetero-compounds. The complete list of ligands and of their frequencies can be found in Table 1 within the Supplementary information chapter. The 614 targets correspond to about 300 different protein domains, which are homogeneously distributed (Fig. 6.1).

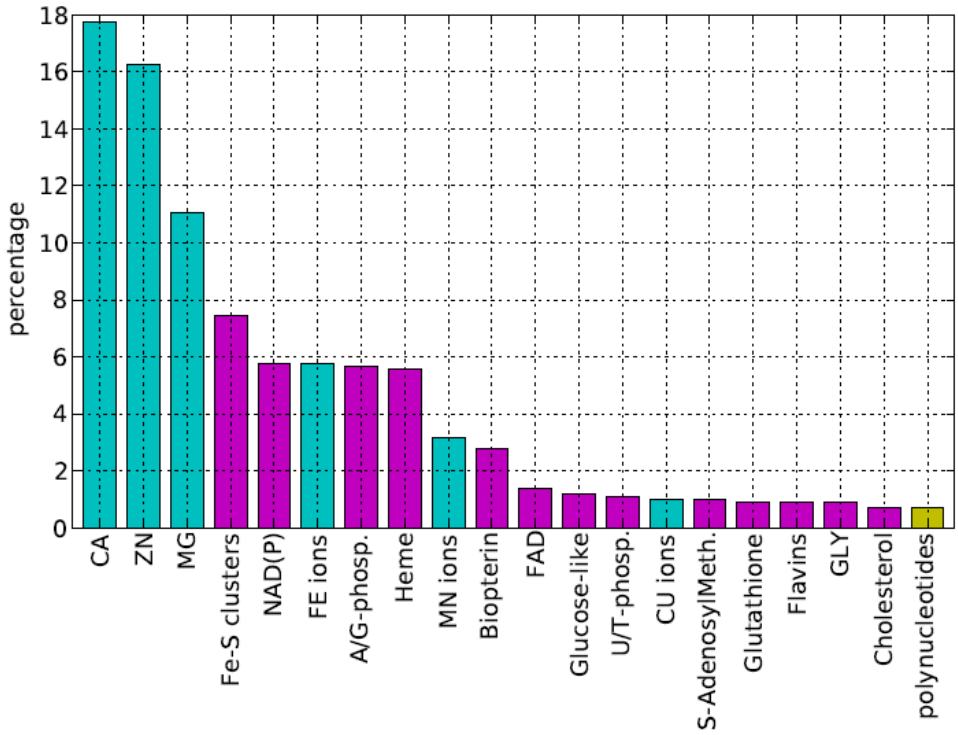

Figure 6.1 Distribution of the small molecules within the CAMEO targets used in the testing set. In cyan are indicated ions, in magenta organic ligands and in yellow the category of polynucleotides. FAD indicates Flavin-adenine dinucleotide and GLY is Glycine. For clarity, only the ligands representing at least 1% of the dataset are shown.

Training

During the training step, we predicted ligands by using seven different scores based on various properties of template and target proteins. In addition, we implemented an annotation score that takes into account UniProt annotations, in order to assign a higher weight to those ligands for which there is an experimental evidence of binding. This score, described in the previous section as the fraction $10\,u/b$, was summed to the following seven scores:

(i) the fraction score, named *freq*, which is the ratio between the number of ligands of a given type and the total number of ligands in the cluster;

(ii) the sequence conservation score, named *cons,* of a ligand binding site; *cons* is the normalized average entropy of the ligand binding residues and is calculated using the Jensen-Shannon Divergence (JSD) implemented in [13];

(iii) the functional specificity score of a binding site, named *sdp_bres*, defined as the average score obtained by implementing a modified version of the method described in [14]. In that study, the aim was to identify subgroups in protein families through the Specificity Determinant Positions (SDP) and each residue's score ranged from 0 for

conserved residues to minus infinite for a very specific position. In our case, both conserved and specific residues are considered relevant, such that the score assigned to a ligand is the average of the standardized absolute SDP values of the binding residues;

(iv) the local structural similarity score, named *rmsd*, derived from the RMSD between the template ligand site and the aligned model site, using a logistic function:

$$p = 1 + \frac{-1}{1 + e^{-3(RMSD-2)}}$$

where a RMSD of 2 Angstroms results in a probability of 0.5. The RMSD is calculated by TMAlign [15] in the superimposition between the template binding site – defined as all the residues within 10 Angstroms from the ligand – and the corresponding model residues;

(v) the structural sequence identity, named *str_seqid*, considers the fraction of annotated binding residues and the sequence identity between a template binding sites and a model pocket, as described in the following Algorithm section;

(vi) the sequence identity, named *seqid*, calculated performing a local alignment between the subset of the template residues which are in contact with the ligand and the corresponding model residues found in the target-template alignment from SWISS-MODEL;

(vii) the aggregated score, named *rank*, defined as the sum of the annotation score, the RMSD-based probability, the structural sequence identity and, only for ions, the fraction score.

Algorithm

Overall, our algorithm collects and assigns a score to all the small molecules present in the homologous structures (the "templates") found by SWISS-MODEL for a target sequence; then, the ligands with the best scores are transferred to the model structure and a report for the SWISS-MODEL website is provided.

More in detail, the initial step consists in the identification of candidate ligands, which should be: (i) bound to the template structure and (ii) considered as "biologically relevant". To satisfy the first criterion, a molecule must have at least 3 protein residues within 4 Angstroms (3.2 in case of ions and Fe-S clusters). From now on, the ligand binding site will be composed by the residues within this distance. Regarding the second criterion, we give priority to molecules that are listed in the SWISS-MODEL Template Library (SMTL) Ligand Annotation system (see chapter 5),

which is a semi-manually curated database of ligand ontologies. In case the molecule is not annotated in SMTL or it is annotated as "Non-covalent", the algorithm checks whether it is contained in: (i) a list of known natural compounds or (ii) a list of biological molecules that can be used as buffer or solvent (for example sulfate ions) and, at the same time, among the UniProt ligands. These two manually curated lists are adapted from the FireDB [20] "Cognate" (see Table 2 in the Supplementary information) and "Ambiguous" ligands (see Table 3 in the Supplementary information), respectively.

Ligands defined as "biologically relevant" are subsequently clustered and scored. Any two ligands belonging to different templates are considered to be in the same group when at least one third of the ligand binding residues for one of them (or simply two residues, for ions), are aligned in the merged pair-wise alignment produced by SWISS-MODEL for the target and its homologous protein sequences.

After the clustering step, each ligand receives a score corresponding to the sum of two terms: the first is the weighted fraction of the template binding residues annotated in UniProt; the second is the sequence identity, calculated from the structural alignment between the ligand binding residues and the corresponding residues in the model. These residues are found by a sequence independent superimposition – calculated by TMAlign [21] – of the full structure of the template containing the ligand on the model structure. A second superimposition of only the binding residues is performed to produce a refined local structural alignment, allowing a refined placement of the ligand into the model. The score s assigned to a ligand is finally calculated as:

$$s = 10\frac{u}{b} + \frac{m}{b},$$

where u is the number of binding residues annotated in UniProt, m is the number of matching residues (i.e. with the same one letter code) between the template and the model binding site, while b is the number of ligand binding residues. In order to rank the ligands primarily on the UniProt annotation, the fraction of annotated binding residues (that is, the term u/b) is multiplied by 10.

After a close inspection of the training set, we decided to set the binding site sequence identity (m/b) cutoff to 0.25; in this way, we avoid to transfer ligands into a model binding site which is unrelated to the template ligand binding site. For each cluster where at least one ligand has a score s greater than zero, the best scoring ligand that fits the model pocket is transferred into it.

Assessment

To assess the accuracy of our method, we need to identify which modelled ions and cofactors correspond to the molecules bound to the target structure. The first step is to superimpose the model to the target; then, for each of ligand in the target, we look for small molecules in the model within 3 Angstroms of the ligand centre. The similarity between the found molecules and the target ligand is compared using the Tanimoto score (calculated using the SMSD software [22]) in case of organic ligands; instead, in case of ions, the similarity is measured only using the atom's element, without considering the oxidation state. A model ligand is evaluated as wrong if it belongs to a category (ions or organics) different from the target ligand, while it is considered correct either if it has a chemical similarity greater than 0.77, or, for ions, if it has the same element as the target ligand. The ability of our predictor to place the correct ligands into the model is assessed using the recall versus the precision plot. The precision, also called Positive Predicted Value (PPV), and the recall, i.e. the sensitivity or True Positive Rate (TPR), are calculated as:

$$precision = \frac{TP}{TP + FP}, \qquad recall = \frac{TP}{TP + FN}$$

where *TP* is the number of correct ligands that are placed in the right binding site of the model, *FP* is the number of ligands in the models that are not present in the target structures or that are in the wrong binding site and *FN* is the number of target ligands that were not placed in the models. A high precision means that most of the ligands placed in the models are correct, while a high recall indicates that most of the target ligands are included in the models. These measures are employed first in the training and later in the testing phase to compare our method with the performances of COACH and RaptorX-Binding.

Whenever a ligand is correctly predicted, we measure how much the ligand conformation is similar to the corresponding molecule present in the target structure by superimposing their binding sites. In case of organic ligands we use the RMSD between the ligand atoms coordinates in the target structure and the corresponding ligand within the model. In case of ions, we measure the distance between the atom in the target structure and the corresponding ion atom in the model. As we are interested only in the ligand conformations and positions respect to the target, we do not exclude correct ligands that overlap badly modelled binding sites.

Results

In order to check which of the seven scores used by our algorithm was best performing on ions and organic ligands, we measured their precision-recall values on the training set, as shown in the Figures 6.2 and 6.3. Overall, for each scoring method, the recall was higher for the ion category than for the organic ligands, while the precision had the opposite trend, being greater for organics than for the ions. Two clusters can be identified in the precision-recall plots: the first group, on the left side of the plot, composed by scores (*cons* and *sdp_b-res*) based on the evaluation of the merged pairwise alignment between the target and all the templates; the second group, on the right side of the plot, including the remaining scores, which take into account the similarity between each single pair of target-template binding site. A special case is constituted by the fraction score *freq*, which considers all the ligands in all the templates and, thus, is more similar to the first than to the second group. The structural sequence identity score, *str_seqid*, achieved the best precision both for the prediction of ions (0.67) and of organic ligands (0.83), while the recall was similar to that of the other features (0.88), meaning that this type of score allows the identification of more correct than incorrect ligands with respect to the other measures.

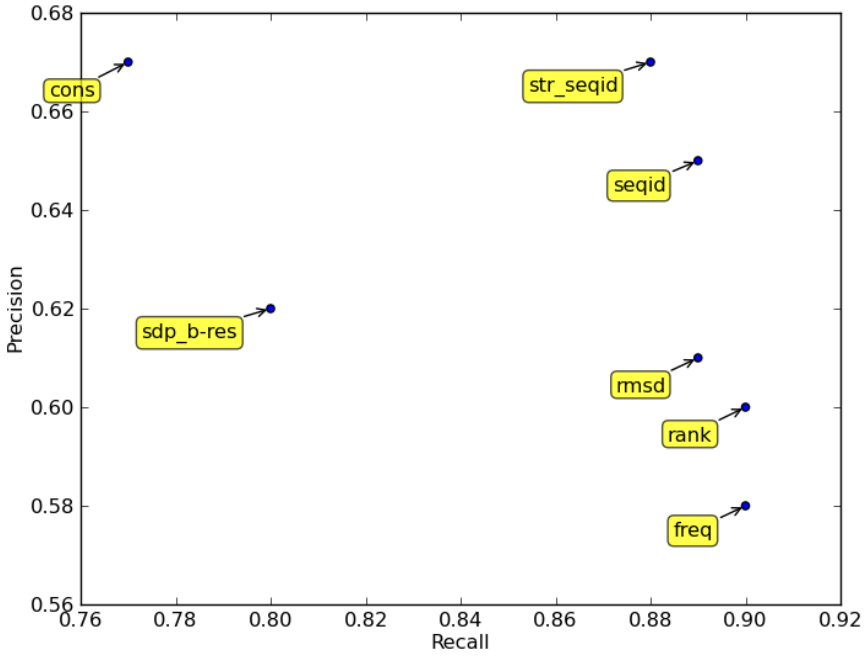

Figure 6.2 Precision-recall plot of the performances for ion predictions achieved by using different scores. *Str_seqid* is the structural sequence identity score, *seqid* is the sequence identity measure, *rmsd* is the local structural similarity, *rank* is the rank score, *freq* is the fraction score, *cons* is the conservation score and *sdp_b-res* is the specificity score.

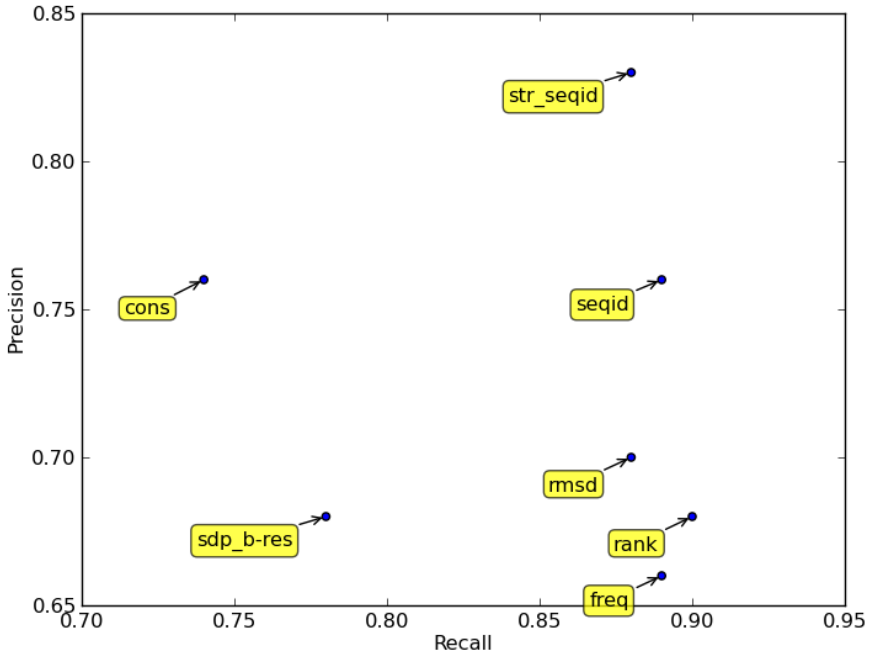

Figure 6.3 Precision-recall plot of the performances for organic ligand predictions achieved by using different scores. *Str_seqid* is the structural sequence identity score, *seqid* is the sequence identity measure, *rmsd* is the local structural similarity, *rank* is the rank score, *freq* is the fraction score, *cons* is the conservation score and *sdp_b-res* is the specificity score.

From the point of view of the ligand conformation accuracy, shown in Figures 6.4 and 6.5, a first observation is that, in each plot, the distribution of the RMSDs was very similar across all the different types of scores. For ions prediction, the median was around 0.2 Angstroms and the upper quartile was below 0.5 Angstroms, while for the organics prediction the median was less than 0.7 Angstroms and the upper quartile around 1 Angstrom. Only in the organic category, we noticed that the *cons* and *sdp_bres* scores showed a broader distribution than the other scoring methods.

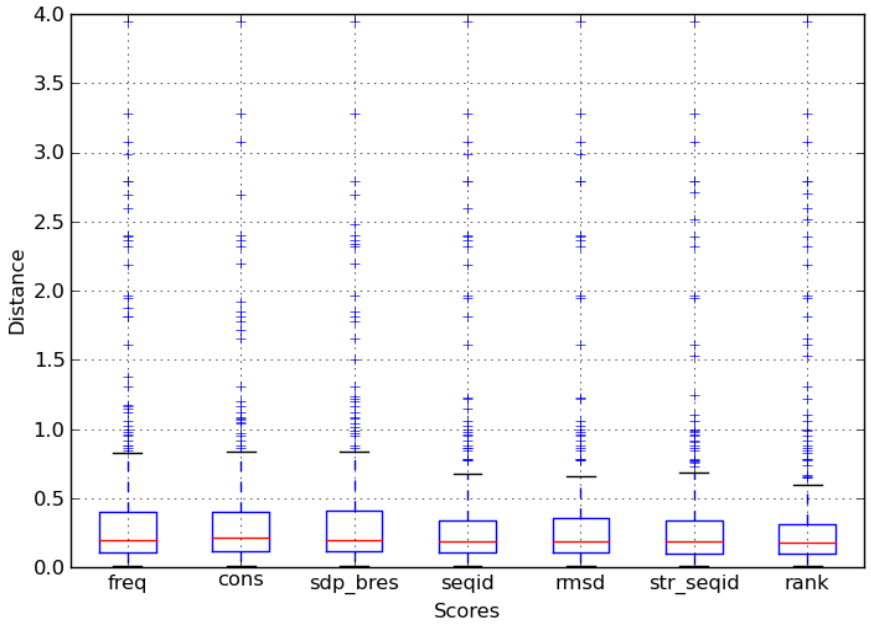

Figure 6.4 Distributions of the distances, for each ion, between its position within the model and the target, achieved by using different scores. Red lines indicate medians; blue boxes show the upper (75%) and lower (25%) quartile; whiskers are 1.5 times the upper and lower quartile; blue crosses correspond to the outliers. *Str_seqid* is the structural sequence identity score, *seqid* is the sequence identity measure, *rmsd* is the local structural similarity, *rank* is the rank score, *freq* is the fraction score, *cons* is the conservation score and *sdp_bres* is the specificity score.

Since the score *str_seqid* achieved the best precision, while maintaining recall and RMSD distributions with very similar values to the other approaches, we decided to employ it for the testing phase. To assess the performances of our method against COACH and RaptorX-Binding servers, we compared the precision-recall values and the RMSD distributions between the target and the subset of ligands which were correctly predicted. Table 6.1 shows the values of True Positives (TP), False Positives (FP) and False Negatives (FN) together with precision and recall, grouped by the ligand category and the server.

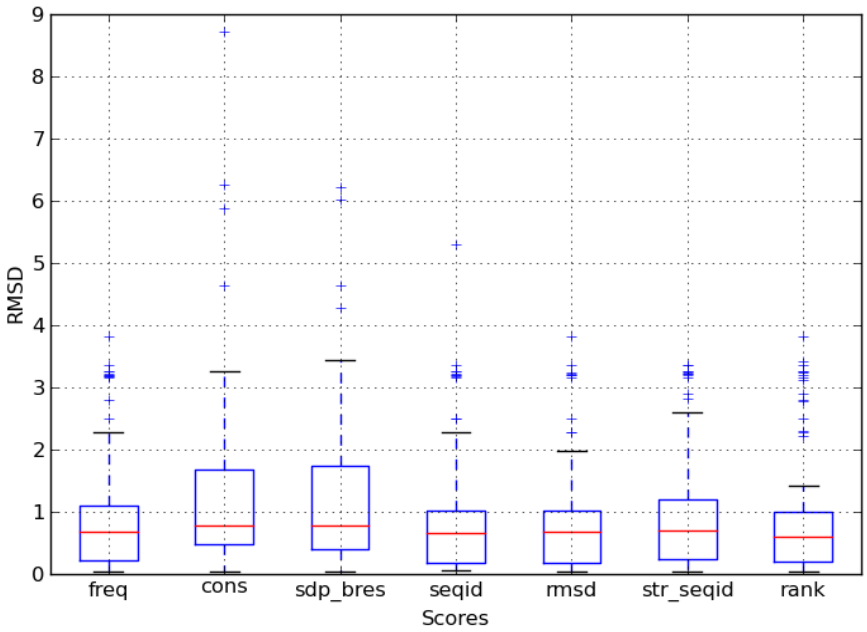

Figure 6.5 Distributions of the RMSDs, for each organic cofactor, between its position within the model and the target, achieved by using different scores. Red lines indicate medians; blue boxes show the upper (75%) and lower (25%) quartile; whiskers are 1.5 times the upper and lower quartile; blue crosses correspond to the outliers. *Str_seqid* is the structural sequence identity score, *seqid* is the sequence identity measure, *rmsd* is the local structural similarity, *rank* is the rank score, *freq* is the fraction score, *cons* is the conservation score and *sdp_bres* is the specificity score.

Table 6.1 Counts of FN, FP, TP with precision and recall values in the ions and organics category obtained by our method, using the *str_seqid* score, COACH and RaptorX-Binding.

Ions	**str_seqid**	**COACH**	**RaptorX**
FN	191	400	302
FP	172	73	160
TP	237	29	83
precision	0.580	0.284	0.342
recall	0.554	0.068	0.216

Organics	**str_seqid**	**COACH**	**RaptorX**
FN	214	287	230
FP	208	89	179
TP	167	88	90
precision	0.445	0.497	0.335
recall	0.438	0.235	0.281

In Figure 6.6 and Figure 6.7 we show the precision-recall values, while Figure 6.8 and 6.9 display the RMSD distributions of COACH, RaptorX-Binding and our method. In the prediction of ions, our method performed clearly better than the other two servers, with a precision of 0.58 and a recall of 0.55. In the organics category, our method showed a slightly lower precision (0.45) than COACH (0.50), although it reached the best recall (0.44).

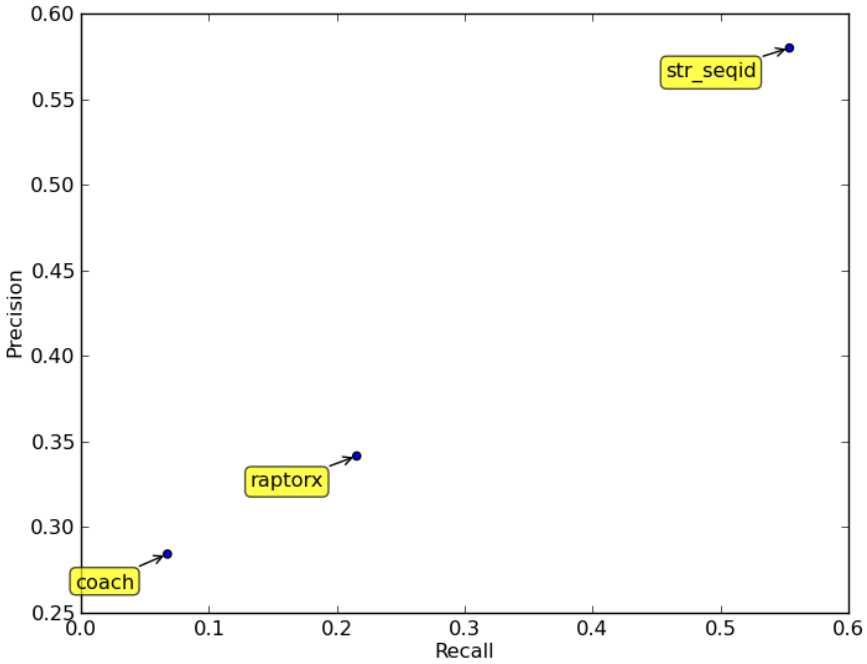

Figure 6.6. Precision-Recall achieved by our method (using the score *str_seqid*), COACH and RaptorX-Binding in the ion predictions.

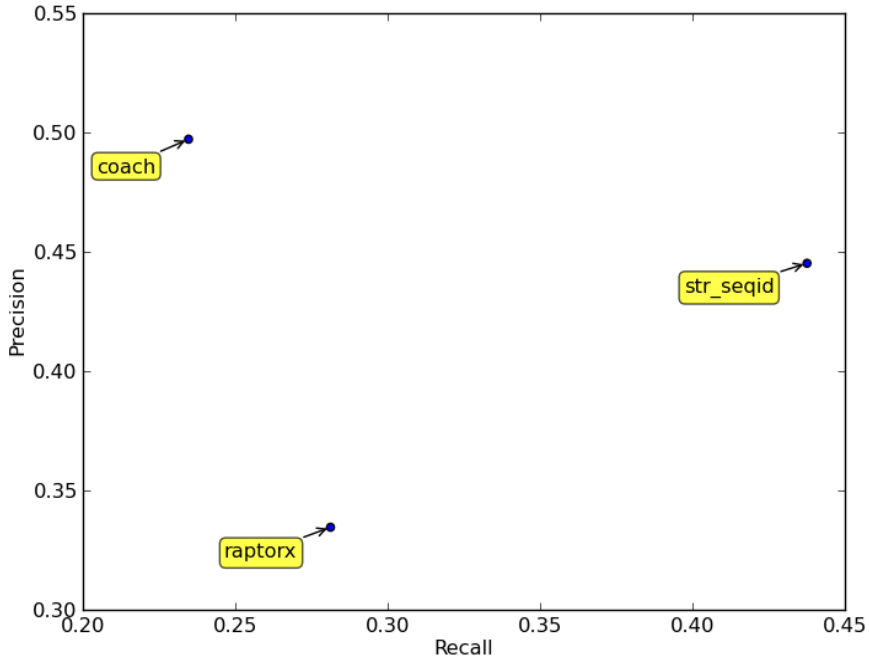

Figure 6.7. Precision-Recall achieved by our method (using the score *str_seqid*), COACH and RaptorX-Binding in the organic ligand predictions.

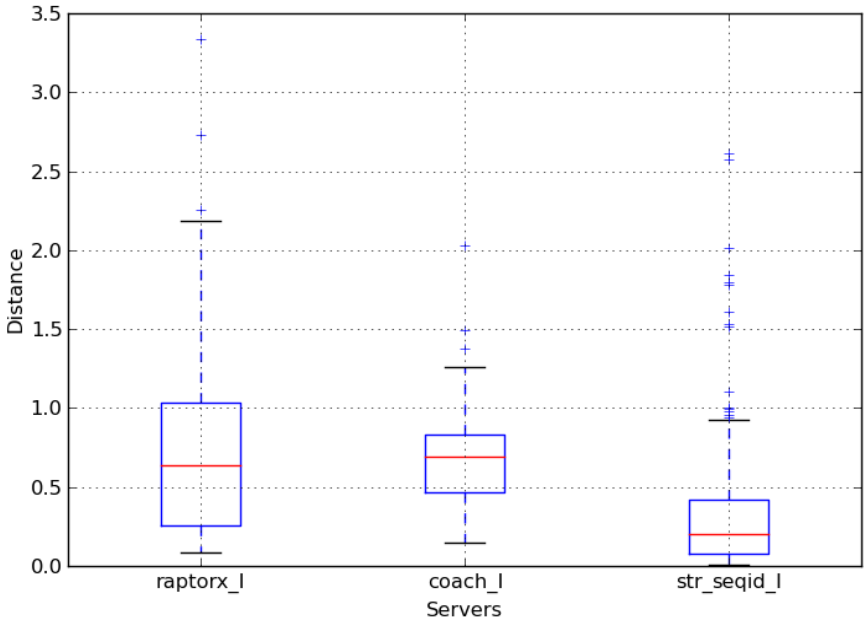

Figure 6.8. Distance distributions of the correctly predicted ligand conformations produced by our method (*str_seqid*), COACH and RaptorX-Binding.

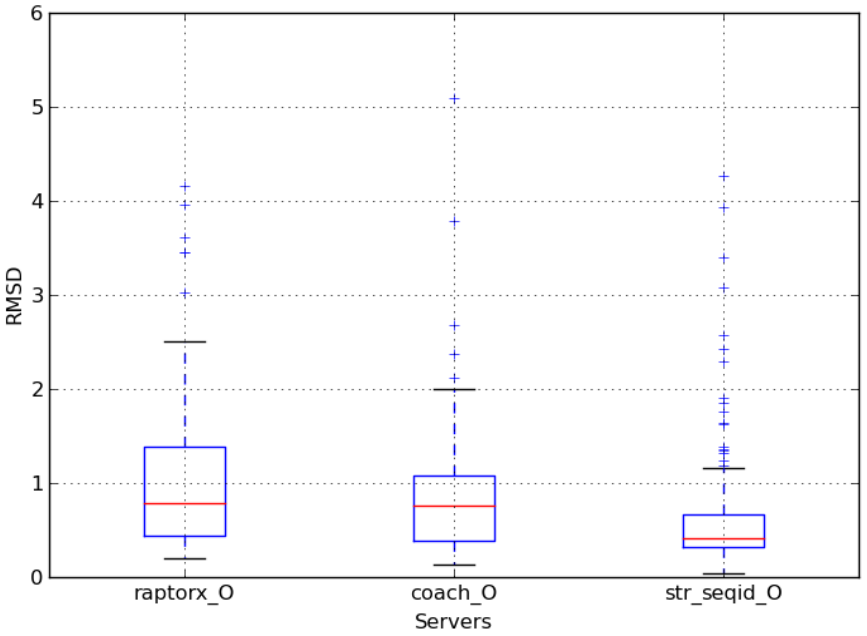

Figure 6.9. RMSD distributions of the correctly predicted ligand conformations produced by our method (*str_seqid*), COACH and RaptorX-Binding.

The distributions of the distances calculated between the coordinates of target and model ligands shows that, in most of the cases, our method placed ions in the models within a distance of only 0.5 Angstroms from the corresponding ion positions in the target. For organic ligands, the

difference between target and model cofactor conformation was usually below 0.8 Angstroms. Both for ions and for organics, the other two methods showed wider distributions, higher medians and higher upper quartiles. In particular, COACH had a median conformation difference of 0.7 Angstroms for ions and an upper quartile above 1 Angstrom for organics, while RaptorX-Binding had the upper quartile above 1 Angstrom for ions and around 1.5 Angstroms for organics.

Considering the fact that it might be difficult to identify the exact ion element from the density map of the protein crystal, we also assessed the three methods by ignoring the precise type of the target ion(s), as shown in Figure 6.10. By comparing the performances displayed in Figures 6.6 and 6.10, we noticed that COACH gained more precision than before, while the ranking was not significantly affected.

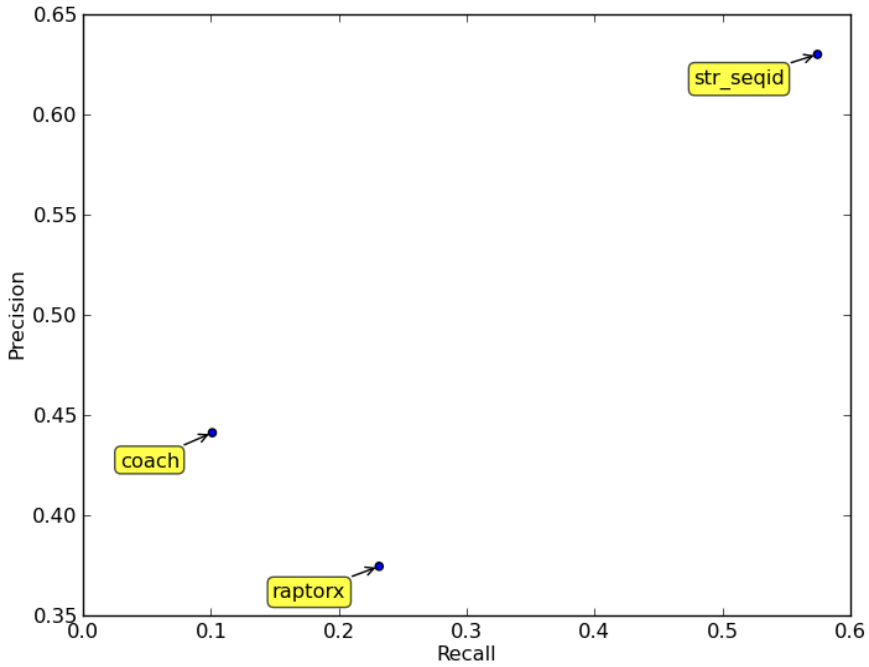

Figure 6.10. Precision-Recall achieved by our method, COACH and RaptorX-Binding in the ion ligand predictions without considering the ion atom element.

Two examples of the accuracy of our method are shown in Figure 6.11 and Figure 6.12. In the first case we transferred the correct ions, two Mn^{2+}, in the target 4BMU_1, a ribonucleotide reductase di-manganese(II), while COACH and RaptorX-Binding predicted only a single iron ion. In the second example we modelled the correct substrate, a triiodothyronine, on a thyroid hormone receptor alpha protein, while COACH predicted a drug-like ligand and RaptorX-Binding placed a different thyroid hormone.

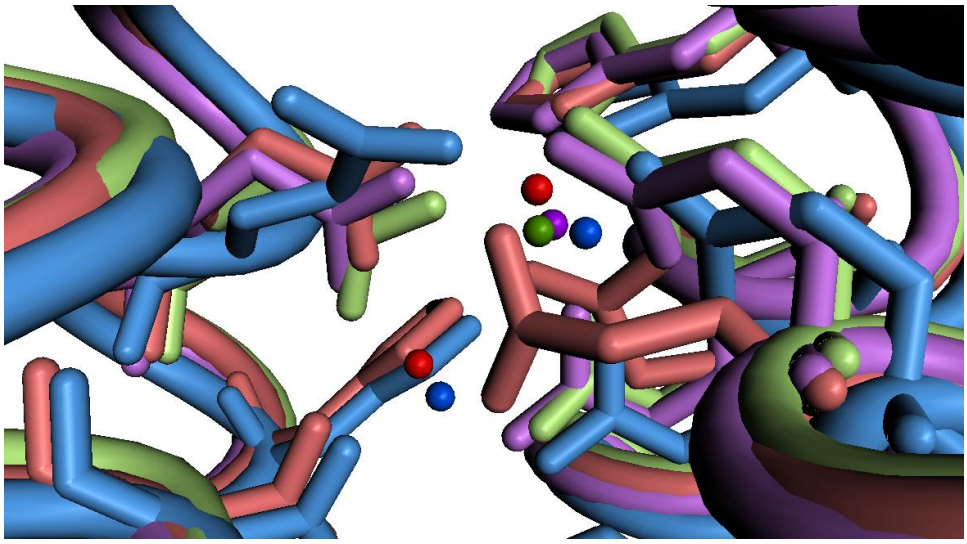

Figure 6.11 Comparison of the position of ions (depicted as spheres) modelled for target 4BMU_1 (red). Method *str_seqid* (in blue) models both manganese ions, while RaptorX-Binding (in violet) and COACH (in green) model only a single iron ion. The binding site residues of the target and the modelled proteins are displayed in lighter colours.

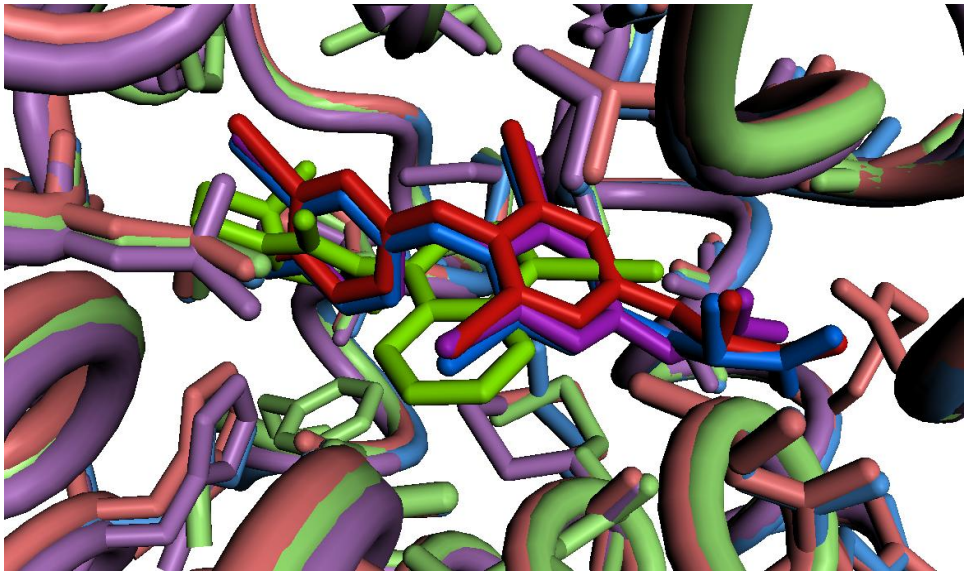

Figure 6.12 Comparison of the cofactors conformations (rendered in sticks) modelled for the target 4LNW_1 (in red). Our method *str_seqid* models the exact target ligand (in blue), while RaptorX-Binding (in violet) models an analogous hormone and COACH (in green) a wrong molecule. The binding site residues of the target and the modelled proteins are displayed in lighter colours.

The precision of our method in identifying a specific cofactor can be seen in Figure 6.13, where we correctly transferred a Nicotinamide Adenine Dinucleotide (NAD) while COACH predicted a

Nicotinamide Adenine Dinucleotide Phosphate (NADP) for the target 4O1M_1, an Enoyl acyl-carrier (ENR) enzyme, which uses NAD+/NADH as cofactor.

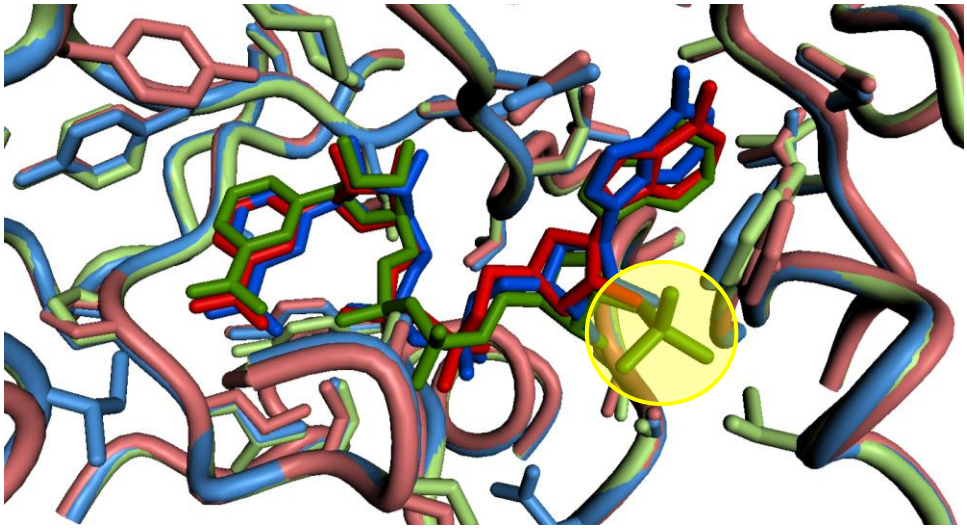

Figure 6.13 Comparison of the cofactors (rendered in sticks) modelled for the target 4O1M_1, which binds a NAD (in red). Our method *str_seqid* correctly modelled a NAD (in blue) while COACH wrongly modelled a NADP (in green), which has an additional phosphate group highlighted by the yellow circle.

Discussion

The work presented herein describes a new method for the identification and modelling of ions and organic cofactors within homology models, in this case created by SWISS-MODEL. As shown in the last CASP assessments [12, 23], the most successful methods for ligand binding site prediction apply a homology transfer approach. Moreover, this strategy is already successfully applied within the field of the protein structure prediction; therefore, we decided to further investigate this method and to employ it for the additional step of modelling ligands in their binding sites. We compared a set of models against their homologous proteins using different properties based on their binding site sequences and structures. We found that the best performances in terms of precision, recall and ligand conformation could be obtained using the structural sequence identity, combined to the fraction of ligand binding residues annotated in UniProt. In this approach, we used TM-Align to superimpose the model to the template in order to identify the binding site within the model; afterwards, we repeated the same step only using the binding site region (i.e. all the residues within 10 Angstroms from a ligand), to obtain an optimal local structural alignment between the model and the binding residues. The distance of 10 Angstroms was chosen to include enough residues for the superimposition step, in case a

ligand was in a relatively flat site. Moreover, as TM-Align utilizes only the C-alpha atoms, the superimposition is not affected by the atom coordinates accuracy of the model binding side-chains. Finally, a ligand is transferred (i.e. the atom coordinates are copied) to the model if it does not overlap, within a distance of 1.6 Angstrom, with any atom of the model. This constrain avoids creating unreliable model-ligand complexes. We decided to not apply a local docking procedure for the optimization of a candidate ligand conformation because it has been shown that the best pose in a model can be achieved by preserving the ligand native conformation [24].

Employing the local structural alignment allows to take into account the effects of the structural variability between two sites. Moreover, the evaluation of the resulting sequence identity allows accounting for the chemical similarity between the model and the target binding residues. In addition to these aspects, the use of the UniProt sequence features enhances the reliability of a ligand to be biologically relevant and of its corresponding templates binding residues to be the correct pocket site. We decided to use UniProt as a source of annotation because it contains information about all known protein sequences and is updated every 4 weeks; in contrast, the Catalytic Site Atlas database [25], which is used by other predicting methods [5, 26], only stores the catalytic residues of enzymes reported in literature and is updated every few years. In addition, the score we selected for the comparison against other methods (*str_seqid*) only depends on one parameter, that is, the minimum sequence identity for which a binding site is considered enough similar to the one in the model. The results of the training step indicate that the correct ligand for a target should be selected, in most of the cases, from the template having the highest binding site similarity, independently from the remaining residues of the template structure.

To assess the performances of our method, we carried out a blind test on a large set of proteins representing a real scenario. We chose targets and models provided by the CAMEO Ligand Binding server. Our approach was compared with two state-of-the-art methods in the ligand binding predictions that place ligands in their model: COACH [11] and RaptorX-Binding (http://raptorx.uchicago.edu/BindingSite/). Briefly, COACH adopts a support vector machine to score binding sites predictions made by different methods that use sequence and structural information. RaptorX-Binding infers the likelihood of a pocket, and of the bound ligand, to be correct using the number of occurrences in the templates of the target.

As clearly shown by the results of the precision and recall, our method was the best performing in the prediction of ions and had the highest recall among the three tested methods for the prediction of organic ligands. In this latter category, COACH showed the best precision, slightly

higher than our method. The higher recall obtained by our method in the organic category is due to the fact that we focus on biologically relevant ions and cofactors, both in the clustering and in the scoring step; moreover, the evaluation of the structural local sequence identity, which was used to select the most similar template binding site to the model, resulted in a gain in sensitivity. Additionally, our method is able to model ligands even in structures composed by multiple chains. On the other hand, although we obtained the highest recall of the three methods, the number of false negatives (FN) is still quite high (191 for ions and 214 for organic cofactors). The main reason for that is the lack of biologically relevant ions and cofactors in the evaluated targets. Likewise, the higher false positives (FP) number of our method with respect to COACH, within the organic category, is due to the extra amount of ligands that were not present in the target structures, but were wrongly transferred into the models.

Despite our method showed the best performances on the testing set, there are still a few limitations that remain to be addressed. First of all, since we infer the target ligands by a homology transfer approach, we are not able to model ligands when there are no cofactors in any of the protein template structures. Secondly, if the model was not correctly built or the binding site was modelled using a template which did not have any ligand, it might happen that the binding site cannot accommodate the selected ligand. Finally, the templates of a target may not be annotated yet in UniProt and, thus, the relevance of their cofactors would remain uncertain.

In the future developments of our method, we will try to address these issues using several strategies. For example, if there are no ligands in the templates, a pocket detection algorithm could be used to predict binding sites in the target, which would be later compared to a set of protein sites binding experimentally annotated compounds. Otherwise, a Potential of Mean Force approach could be used to create a set of statistics for ligand-binding site complexes, from which the conformation of the candidate ligand in the model could be inferred. In the case of a model with a bad quality binding site, one solution (alternative to ligand docking) would be to rebuild the model using the template from which the candidate ligand was selected by our scoring function. Finally, if the experimental annotation of a protein cofactor is not included in UniProt, it could be retrieved from additional sources, like the Kyoto Encyclopedia of Genes and Genomes (KEGG) COMPOUND database (http://www.kegg.jp/kegg/compound/). However, this type of databases includes only the ligand interacting with a protein and does not indicate the binding site, which should be validated by an additional source of annotations.

Supplementary information

Table 1 Complete list of ligands and their frequency within the CAMEO targets composing testing set. The percentage column indicates the fraction of relevant ligand instances over the total number of relevant compounds.

Ligand name or PDB code	CAMEO category	relevant	not_relevant	percentage
CA	I	178	55	17.72908367
ZN	I	163	28	16.23505976
MG	I	111	116	11.05577689
Fe-S clusters	O	75	28	7.470119522
NAD(P)	O	58	2	5.77689243
FE ions	I	58	2	5.77689243
A/G-phosp.	O	57	1	5.677290837
Heme	O	56	0	5.577689243
MN ions	I	32	3	3.187250996
Biopterin	O	28	0	2.788844622
FAD	O	14	0	1.394422311
Glucose-like	O	12	1	1.195219124
U/T-phosp.	O	11	0	1.09561753
CU ions	I	10	11	0.996015936
S-AdenosylMeth.	O	10	0	0.996015936
Glutathione	O	9	0	0.896414343
Flavins	O	9	0	0.896414343
GLY	O	9	0	0.896414343
Cholesterol	O	7	0	0.697211155
polynucleotides	N	7	5	0.697211155
Molybdopterin	O	6	0	0.597609562
polypeptides	P	6	9	0.597609562
ASP	O	6	0	0.597609562
Triiodothyronine	O	5	1	0.498007968
OXY	O	4	1	0.398406375
IMD	O	4	5	0.398406375
Pyridoxal-phosp.	O	4	0	0.398406375
Coenzyme A	O	3	0	0.298804781
Sugar alcohols	O	3	0	0.298804781
ARG	O	3	0	0.298804781
LBT	O	3	0	0.298804781
22B	O	3	0	0.298804781
NI	I	3	5	0.298804781
CAA	O	3	0	0.298804781
HIS	O	2	0	0.199203187
ADE	O	2	0	0.199203187
TRP	O	2	0	0.199203187

UPG	O	2	0	0.199203187
BET	O	2	0	0.199203187
MO ions	O	2	0	0.199203187
FOL	O	2	0	0.199203187
PAU	O	2	0	0.199203187
ICS	O	2	0	0.199203187
POP	O	1	0	0.099601594
BG6	O	1	0	0.099601594
ILE	O	1	0	0.099601594
3PG	O	1	0	0.099601594
STR	O	1	0	0.099601594
TES	O	1	0	0.099601594
BXP	O	1	0	0.099601594
ORO	O	1	0	0.099601594
STL	O	1	0	0.099601594
LEU	O	1	0	0.099601594
CE6	O	1	0	0.099601594
CE5	O	1	0	0.099601594
CYS	O	1	0	0.099601594
CTR	O	1	0	0.099601594
Arabinofunarose	O	1	0	0.099601594
ASD	O	1	0	0.099601594

Table 2 PDB codes of the “cognate”, or biologically relevant, small molecules taken from FireDB.

00A, 00C, 01A, 01B, 03F, 03W, 045, 06C, 13P, 149, 152, 15L, 16G, 17Z, 188, 191, 1AL, 1CA, 1CL, 1CP, 1CU, 1GN, 1GP, 22B, 2AM, 2DG, 2FP, 2GP, 2HA, 2HP, 2MC, 2MO, 2MR, 2OB, 2OM, 35G, 3CO, 3CP, 3GC, 3GP, 3H9, 3HC, 3ML, 3PG, 46D, 46M, 488, 4IP, 4ML, 4MO, 4PS, 5GP, 5RP, 6MO, 8OG, A, A3P, A4D, A5P, A8S, ABF, ABU, ACD, ACH, ACO, AD0, ADA, ADE, ADP, ADQ, ADX, AFP, AG2, AGC, AHR, AIR, AKG, ALA, ALE, ALL, ALO, AMP, AMZ, AND, ANE, ANR, AOR, AOS, ARA, ARB, ARG, AS1, AS4, ASD, ASN, ASP, ATP, B12, B1M, B1Z, B2G, B4G, BAL, BCA, BCL, BCO, BCR, BCT, BDP, BEM, BET, BG6, BGC, BGP, BIO, BLA, BLD, BMA, BPB, BPD, BPH, BT5, BTN, BXP, BYC, BZX, C, C0R, C1O, C2F, C5P, CA, CAA, CAO, CAP, CAQ, CBI, CBU, CBY, CCQ, CDL, CDN, CDP, CE5, CE6, CE8, CEG, CFM, CFN, CFO, CGL, CH, CHL, CIR, CIS, CL1, CL7, CLA, CLF, CLL, CLP, CLR, CM1, CM2, CMO, CMP, CN1, CNB, CNC, CNF, CO2, CO6, CO8, COA, COB, COD, COH, COJ, COO, COS, COW, COZ, CP2, CP3, CRN, CSE, CTP, CTR, CU, CU1, CU3, CUA, CUB, CUK, CUM, CUN, CUO, CXR, CYC, CYS, CYT, CZL, D2V, DA, DA2, DAC, DAK, DAL, DC, DCC, DEF, DFL, DFV, DG, DGL, DHB, DHC, DHE, DHT, DI, DLZ, DN, DNO, DOC, DPM, DPN, DQR, DT, DTP, DU, DXC, E2P, EA2, EB4, ECH, EDC, EFE, EIC, EMU, F3S, F42, F43, F4S, F6P, F6R, FA, FAD, FAQ, FBP, FCA, FCB, FCI, FCO, FDA, FDC, FDP, FE,

FE2, FEL, FEO, FER, FES, FMN, FOC, FOL, FPC, FPP, FRE, FRU, FS1, FS2, FS3, FS4, FSF, FSO, FUB, FUC, FUD, FUM, G16, G1P, G1R, G2Q, G2R, G3H, G3P, G4P, G6P, G6Q, GA3, GA4, GAE, GAL, GAR, GCD, GCS, GCU, GCV, GDC, GDD, GDP, GDR, GDU, GLA, GLC, GLN, GLO, GLP, GLU, GLV, GLY, GMP, GNP, GP5, GRA, GRG, GSH, GTP, GTR, GTS, GTT, GUD, GUN, GZL, H35, H4B, H4M, HAM, HAS, HBI, HC4, HCB, HCC, HCN, HCO, HDC, HDD, HDE, HEA, HEB, HEC, HEG, HEM, HEQ, HGS, HIF, HIS, HMG, HPA, HSC, HSE, HSM, HSO, HTL, HXC, I0P, I2A, I2P, I3P, I3S, I4P, I5P, I6P, IAC, ICA, ICS, ICT, IDR, IGP, IHP, ILE, IMD, IMI, IMP, IND, INS, IP1, IP2, IPL, IPR, ISC, ISD, ITM, ITT, JB2, JB3, JN3, KDG, KDP, KOJ, LAI, LAN, LAT, LBT, LBV, LDP, LEU, LFC, LFR, LGU, LMG, LNL, LNR, LPA, LUM, LYS, M1P, M2P, M43, M6P, MAB, MAN, MAX, MC4, MCA, MCN, MDO, MET, MEV, MG, MH2, ML1, MLC, MLR, MM4, MMP, MN, MN3, MNH, MO, MOM, MOO, MOS, MOW, MP1, MQ7, MQ8, MQ9, MRR, MRS, MSS, MTA, MTL, MTQ, MTT, MTV, MXY, MXZ, MYA, MYR, NAD, NAI, NAP, NBC, NCA, NDP, NFC, NFE, NFO, NFR, NFS, NFV, NG1, NGA, NI, NIO, NLG, NMN, NO, NOS, NTM, NTN, OAA, OC1, OC2, OC3, OC4, OC5, OC6, OC7, OC8, OCR, OFE, OMO, OMP, OPC, ORN, ORO, OXK, OXS, OXY, P5P, P7I, PAB, PAU, PC, PCA, PCD, PCG, PDP, PEB, PEE, PEP, PG2, PGP, PHE, PIE, PLP, PNS, POP, POR, PP9, PPR, PQN, PQQ, PRO, PTE, PTR, PTT, PUB, PVL, PVN, PXL, PXM, PXP, PYG, PYH, PYM, PYQ, PYR, QDK, QUE, R1P, R5P, RAF, RB5, RBF, RCC, REA, RED, RET, RG1, RIB, RIP, RNS, RNT, ROA, ROM, RTL, RUB, RUT, S0N, S3P, S6P, SAC, SAH, SAM, SAR, SCA, SCG, SER, SF3, SF4, SFT, SMO, SOR, SPF, SPH, SPN, SPO, SRM, SRO, ST9, STE, STL, STR, SUG, SUO, SUP, T1G, T3, T3P, T6P, TC6, TCH, TDP, TDR, TES, TGC, TH3, THG, THM, THP, THR, TMP, TP7, TPO, TPP, TPQ, TPS, TRA, TRP, TRQ, TS5, TSS, TTP, TTQ, TYD, TYR, U, U10, U5P, UAG, UD1, UD2, UDP, UMA, UMP, UP2, UP3, UPG, UQ, UQ1, UQ2, UQ7, UQ8, URA, URC, URI, UTP, VAL, VBN, VD3, VDX, VDY, VER, VIB, VIT, VIV, VK3, WCC, XAN, XCC, XMP, XX2, XX3, XXP, XYL, XYP, XYQ, XYS, ZEA, ZIR, ZN, ZNH.

Table 3 PDB codes of the “ambiguous” ligands adapted from FireDB.

1BO, 3GR, ABA, ADN, AG, AME, APR, ASC, ASE, BES, BR, CAC, CD, CFF, CIT, CL, CLM, CO, CO3, CYN, DCE, DSN, FLC, FUL, IOD, K, LAC, MAL, MLA, MLI, MLT, NA, NH2, NH4, NO3, OLA, OXL, PAM, PLM, PO4, RAM, SCN, SEP, SIA, SIN, SMX, SO4, SPD, SPM, SUC, TLA, TRE, URE, XD2, V, W, WO4, WO5, VO4.

References

1. Chang, C.E., et al., *Homology modeling of cannabinoid receptors: discovery of cannabinoid analogues for therapeutic use.* Methods Mol Biol, 2012. 819: p. 595-613.
2. Congreve, M., et al., *Progress in structure based drug design for G protein-coupled receptors.* J Med Chem, 2011. 54(13): p. 4283-311.
3. Capra, J.A. and M. Singh, *Predicting functionally important residues from sequence conservation.* Bioinformatics, 2007. 23(15): p. 1875-82.
4. Fischer, J.D., C.E. Mayer, and J. Soding, *Prediction of protein functional residues from sequence by probability density estimation.* Bioinformatics, 2008. 24(5): p. 613-20.
5. Lopez, G., et al., *firestar--advances in the prediction of functionally important residues.* Nucleic Acids Res, 2011. 39(Web Server issue): p. W235-41.
6. An, J., M. Totrov, and R. Abagyan, *Pocketome via comprehensive identification and classification of ligand binding envelopes.* Mol Cell Proteomics, 2005. 4(6): p. 752-61.
7. Brylinski, M. and J. Skolnick, *A threading-based method (FINDSITE) for ligand-binding site prediction and functional annotation.* Proc Natl Acad Sci U S A, 2008. 105(1): p. 129-34.
8. Capra, J.A., et al., *Predicting protein ligand binding sites by combining evolutionary sequence conservation and 3D structure.* PLoS Comput Biol, 2009. 5(12): p. e1000585.
9. Wass, M.N., L.A. Kelley, and M.J. Sternberg, *3DLigandSite: predicting ligand-binding sites using similar structures.* Nucleic Acids Res, 2010. 38(Web Server issue): p. W469-73.
10. Roche, D.B., S.J. Tetchner, and L.J. McGuffin, *FunFOLD: an improved automated method for the prediction of ligand binding residues using 3D models of proteins.* BMC Bioinformatics, 2011. 12: p. 160.
11. Yang, J., A. Roy, and Y. Zhang, *Protein-ligand binding site recognition using complementary binding-specific substructure comparison and sequence profile alignment.* Bioinformatics, 2013. 29(20): p. 2588-95.
12. Gallo Cassarino, T., L. Bordoli, and T. Schwede, *Assessment of ligand binding site predictions in CASP10.* Proteins, 2014. 82 Suppl 2: p. 154-63.
13. Shin, W.H., et al., *LigDockCSA: protein-ligand docking using conformational space annealing.* J Comput Chem, 2011. 32(15): p. 3226-32.
14. The UniProt Consortium, *Activities at the Universal Protein Resource (UniProt).* Nucleic Acids Res, 2014. 42(Database issue): p. D191-8.
15. Rice, P., I. Longden, and A. Bleasby, *EMBOSS: the European Molecular Biology Open Software Suite.* Trends Genet, 2000. 16(6): p. 276-7.
16. Velankar, S., et al., *SIFTS: Structure Integration with Function, Taxonomy and Sequences resource.* Nucleic Acids Res, 2013. 41(Database issue): p. D483-9.
17. Zheng, H., et al., *Data mining of metal ion environments present in protein structures.* J Inorg Biochem, 2008. 102(9): p. 1765-76.
18. Allen, F.H., *The Cambridge Structural Database: a quarter of a million crystal structures and rising.* Acta Crystallogr B, 2002. 58(Pt 3 Pt 1): p. 380-8.
19. Finn, R.D., et al., *Pfam: the protein families database.* Nucleic Acids Res, 2014. 42(Database issue): p. D222-30.
20. Maietta, P., et al., *FireDB: a compendium of biological and pharmacologically relevant ligands.* Nucleic Acids Res, 2014. 42(Database issue): p. D267-72.
21. Zhang, Y. and J. Skolnick, *TM-align: a protein structure alignment algorithm based on the TM-score.* Nucleic Acids Res, 2005. 33(7): p. 2302-9.
22. Rahman, S.A., et al., *Small Molecule Subgraph Detector (SMSD) toolkit.* J Cheminform, 2009. 1(1): p. 12.

23. Schmidt, T., et al., *Assessment of ligand-binding residue predictions in CASP9.* Proteins, 2011. 79 Suppl 10: p. 126-36.
24. Dalton, J.A. and R.M. Jackson, *Homology-modelling protein-ligand interactions: allowing for ligand-induced conformational change.* J Mol Biol, 2010. 399(4): p. 645-61.
25. Furnham, N., et al., *The Catalytic Site Atlas 2.0: cataloging catalytic sites and residues identified in enzymes.* Nucleic Acids Res, 2014. 42(Database issue): p. D485-9.
26. Roy, A., J. Yang, and Y. Zhang, *COFACTOR: an accurate comparative algorithm for structure-based protein function annotation.* Nucleic Acids Res, 2012. 40(Web Server issue): p. W471-7.

7. Moment invariants for binding sites description

Introduction

The shape of any distribution can be mathematically described by a set of quantitative measures called “moments”. The moments which are most commonly used to characterize a given distribution are mean, variance, skewness and kurtosis. Additional types of moments, those beyond the 4th-order, involve non-linear combinations of the data and are useful to describe or estimate further shape parameters; however, since these moments are harder to estimate and subtle to interpret, they are in general less used.

Moments can be further distinguished in “central”, when they are computed in terms of the deviations from the mean, and “ordinary”, in case the reference point is the zero; the former type of moments is usually preferred, as it is only dependent on the spread and shape of the distribution, rather than on its location.

The first moment is the mean, which is the average value of a distribution. The second moment is the variance, which is a non-negative quantity defined as the square mean of the distances of the values from their mean and represents the spread of the data. The third moment is the skewness, which measures how much the distribution is asymmetrical; it equals zero if the distribution is perfectly symmetrical. The fourth moment is the kurtosis, which describes whether the distribution is peaked and narrow or flat and wide, with respect to a Gaussian distribution having the same variance. Any central moment can be normalized and is independent on the scale if divided by the variance elevated to the order of the moment.

The “moment invariants” are functions of the moments such that they do not change their value when the distribution is transformed. Moment invariants can be used to describe the shape of a three-dimensional distribution of points, independently from their position and orientation. Although they are mostly employed in image analysis, these shape descriptors have been also used in the structural biology field as feature vectors to efficiently represent and compare protein-protein interfaces [1].

Our aim is to adopt moment invariants as protein pocket descriptors in the context of the de novo ion binding site prediction.

Methods

We collected from the PDB database [2] a set of about 13000 X-ray protein structures that had a minimum resolution of 2.5 Angstroms, a R-value ≤ 0.25 and that were bound to divalent cations (Calcium, Copper, Iron, Magnesium, Manganese, Zinc). Next, we only kept those structures whose bound ion was annotated in UniProt/SwissProt [3] in order to create a set of proteins which had experimental evidence of the interaction with a cation. This filtering step produced 1675 proteins, in which 325 were bound to Calcium, 65 to Copper, 82 to Iron, 119 to Manganese, 274 to Magnesium and 435 to Zinc. Ignoring the ions that were bound to less than 3 protein residues resulted in a total number of 1187 binding sites.

In each binding site, we divided the residues in clusters according to the type of atom that was in contact with the ion. For simplicity, every atom was mathematically represented by a Gaussian density function in the three dimensions. For each cluster, the first three moment invariants (mean, variance, skewness) were calculated using three different approaches: (i) considering only the atoms in direct contact with the ion, (ii) using all the atoms belonging to the whole residues and (iii) utilizing only the C-alpha carbons. Accordingly, each cluster of residues was represented by 9 values, corresponding to the sum of the residue moments (1st, 2nd and 3rd) along each axis (x, y and z).

In particular, the first moment was calculated as:

$$\mu_{1,X} = \sum_k (a_k - \bar{x})$$

$$\mu_{1,Y} = \sum_k (a_k - \bar{x})$$

$$\mu_{1,Z} = \sum_k (a_k - \bar{x})$$

the second moment was calculated as:

$$\mu_{2,X} = \sum_k ((a_k - \bar{x})^2 + \sigma^2)$$

$$\mu_{2,Y} = \sum_k ((a_k - \bar{x})^2 + \sigma^2)$$

$$\mu_{2,Z} = \sum_k ((a_k - \bar{x})^2 + \sigma^2)$$

and the third moment was calculated as:

$$\mu_{3,X} = \sum_k ((a_k - \bar{x})^3 + 3(a_k - \bar{x})\sigma^2)$$

$$\mu_{3,Y} = \sum_k ((a_k - \bar{x})^3 + 3(a_k - \bar{x})\sigma^2)$$

$$\mu_{3,Z} = \sum_k ((a_k - \bar{x})^3 + 3(a_k - \bar{x})\sigma^2)$$

where a indicates the position of the kth atom and $\bar{x}$ the centre of mass of the residue cluster, which is used as origin, or central point, for the shape density of the residue cluster. These moments were subsequently normalized and transformed in order to make them “invariant” to their position and orientation in the three-dimensional space, as described in [1]. The moment invariants were finally grouped in a feature vector to represent the binding site; in this way, the binding sites could be compared to each other by their distance in the Euclidean space. To verify that the moment invariants could be sufficiently accurate in describing the binding sites of our dataset and, more importantly, in classifying them according to the bound ion, we used a two-step procedure: first, we employed the R implementation of the Partition Around Medoid (PAM) algorithm [4] to cluster the moment invariants; secondly, we measured the level of correct clusters partitioning through the “silhouette”, a graphical aid for the interpretation and validation of cluster analysis [5]. The PAM method, which belongs to the k-means family of algorithms, divides a dataset in an a priori defined number of groups and clusters the points in the dataset by minimizing the sum of their pairwise dissimilarities. This approach is more robust against outliers and noise with respect to the other k-means methods, which employ the sum of squared Euclidean distances to create the clusters. In our case, the PAM method provides a useful way to group the moment invariants, particularly because we already know the expected number of clusters, which corresponds to the number of different ions in the dataset. The main advantage of the silhouette is to provide a graphical evaluation of the clustering validity, that is, to estimate whether the points in a dataset were correctly clustered in the appropriate group. An average value close to 1 indicates that the number of clusters accurately reproduced the classification of the underling members, while an average value close to -1 means that most of the data points (in our case, the binding sites) were assigned to the wrong cluster.

Results and discussion

The aim of this project was to investigate whether the moment invariants could be used to efficiently describe the geometry of ion binding sites, and therefore, for the de novo identification of ion binding sites in a protein model. To verify the ability of moment invariants in representing binding sites, we clustered them and evaluated whether the resulting number of groups was coherent with the expected number of clusters, corresponding to the 6 different types of ions which were present in our dataset. Accordingly, the maximum average silhouette value should have been measured when dividing the moment invariants in exactly 6 groups. We clustered the moment invariants with a variable number of groups, ranging from 2 to 10, and for each round we calculated the silhouette. Figure 7.1 shows the silhouettes measured with 2 clusters, while Figure 7.2 illustrates the silhouettes obtained using 6 clusters. In both cases all the atoms of the binding residues were used to calculate the moment invariants.

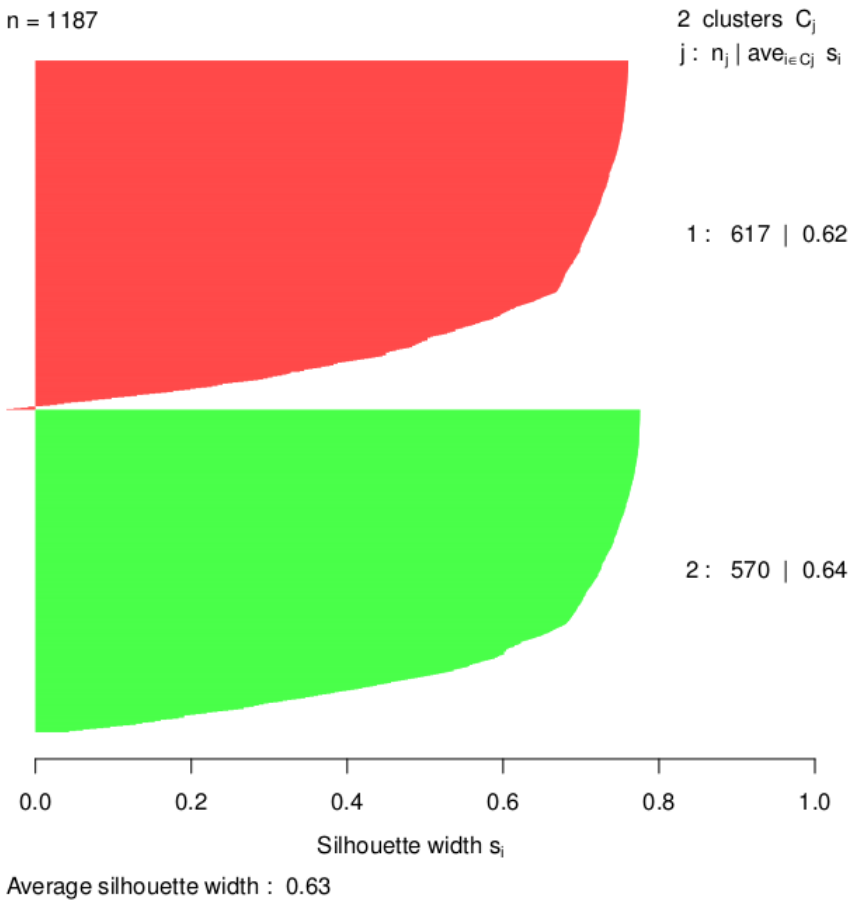

Figure 7.1 Silhouette plot of the moment invariants clustered in 2 groups. On the top left corner, “n” indicates the number of binding sites; at the right side of each silhouette, it is indicated the number of

points in the cluster and the silhouette width; at the bottom of the figure, it is shown the average silhouette width of the 2 clusters.

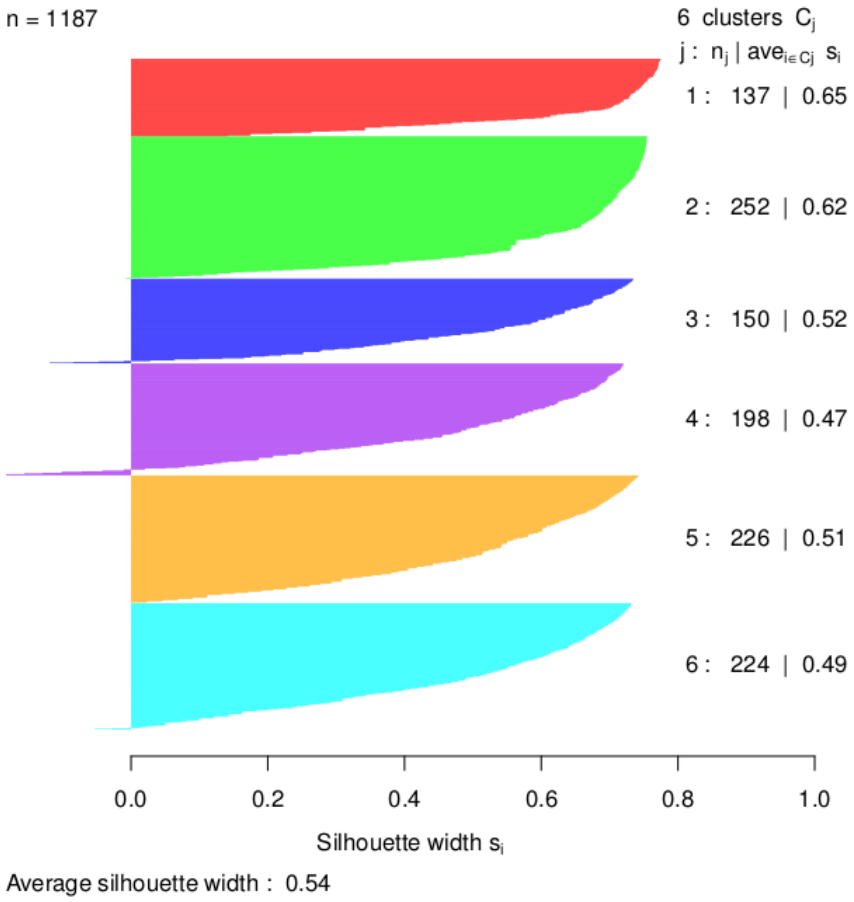

Figure 7.2 Silhouette plot of the moment invariants clustered in 6 groups. On the top left corner, "n" indicates the number of binding sites; at the right side of each silhouette, it is indicated the number of points in the cluster and the silhouette width; at the bottom of the figure, it is shown the average silhouette width of the 6 clusters.

The maximum average silhouette was 0.63 and was obtained by setting to 2 the number of clusters; using different numbers of clusters, the maximum average silhouette was always lower than this value. For example, as shown in Figure 7.2, using 6 clusters produced a value of 0.54. The highest average silhouette was found using 2 clusters also when we calculated the moment invariants by selecting only the atoms in contact with the ions, as well as when we selected only the C-alpha carbons of the binding residues.

Since the expected number of clusters with the best silhouette was 6 instead of 2 (as shown by the silhouette), these results suggest that the moment invariants, even when calculated separately for each type of residue in contact with the ion, are not a suitable method to accurately describe in detail the ion binding sites. We performed the same analysis for the

comparison of the different coordination geometries created by each ion. The silhouette values of the clustered sites followed the same behaviour as in the comparison of all ion sites, meaning that the binding sites of each ion were optimally clustered in two groups, independently of the actual number of coordination geometries. These outcomes can be better understood by taking into account the fact that the distributions of the ionic distances and of the coordination numbers – i.e. the number of atoms making a bond with the ion – are very similar among the ions found in proteins [6]. Moreover, protein ion binding sites can be arranged in a distorted shape that makes its classification into a well defined geometry more difficult. Hence, these features are not enough different from one ion binding site to another and, therefore, moment invariants cannot be efficiently used to distinguish ion binding sites.

However, organic ligand binding sites are more diverse in shape with respect to ion binding sites; therefore, an interesting development for future studies would be to investigate whether the moment invariants could be successfully applied for their identification and classification.

References

1. Sommer, I., et al., *Moment invariants as shape recognition technique for comparing protein binding sites.* Bioinformatics, 2007. 23(23): p. 3139-46.
2. Berman, H.M., et al., *The Protein Data Bank.* Nucleic Acids Res, 2000. 28(1): p. 235-42.
3. The UniProt Consortium, *Activities at the Universal Protein Resource (UniProt).* Nucleic Acids Res, 2014. 42(Database issue): p. D191-8.
4. R Development Core Team, *R: A language and environment for statistical computing*. 2011, R Foundation for Statistical Computing: Vienna, Austria.
5. Rousseeuw, P.J., *Silhouettes: A graphical aid to the interpretation and validation of cluster analysis.* J Comp Applied Math, 1987. 20(11): p. 53-65.
6. Zheng, H., et al., *Data mining of metal ion environments present in protein structures.* J Inorg Biochem, 2008. 102(9): p. 1765-76.

8. Conclusion

In this thesis we presented the assessment of the current methods for the prediction of ligand binding sites in protein models, during the last two rounds of the CASP experiment. In addition, we described the development of a pipeline for modelling small molecules in homology models

and the prediction of their binding sites. Finally, we tested a novel approach for the description of ion binding sites.

In the recent CASP9 and CASP10 editions, the evaluation of the ligand binding site predictors for proteins without a known structure indicated the strengths and weakness of the current state-of-the-art prediction methods. The results showed that the most successful participants employed approaches based on the transfer of binding residue annotations from the target homologous proteins to the model. The main limitation of these methods resides in the variable availability of binding information, either retrieved from the protein sequences or from the structures with bound ligands. Both in CASP9 and CASP10, the assessment method showed some limitations. The first was the low number of targets bound to relevant ligands; the second consisted in the binary classification of the target residues in either “binding” or “non-binding”, without any measure of confidence; the third limitation was the absence of any information regarding the type of bound ligand. The CAMEO server was developed to address these weaknesses, in order to provide a fast and accurate assessment of the current methods and to guide the development of the binding site prediction field towards new directions.

We implemented a baseline homology transfer predictor in SWISS-MODEL to provide a reference performing method of ligand binding site prediction for comparison with more advanced methods within CAMEO. The method consisted in the transfer of the template ligands into the model and in the identification of the conserved residues that were in contact with small molecules. Despite being a baseline approach, our method achieved a very good performance, especially in the ion category. However, since these good results were obtained only for a limited number of targets, this prompted us to develop an improved version for the new SWISS-MODEL server, based on a multi-template approach and focussed on the modelling of biologically relevant ligands. We compared, in a blind test, our new method with the two best public servers that provided models with bound ligands in the CAMEO Ligand Binding section. We showed that our performances, in the ion and organic categories, were overall more accurate with respect to the other two servers, both in terms of ligand type and of conformation within the model. These results indicate that the identification of biologically relevant ligands in templates plays an important role in the prediction of the correct ligand for a given target and, moreover, suggest that the ligand conformation found in the template is enough precise to not require an additional refinement step. Furthermore, these results indicate that while the binding prediction field is mature enough to produce accurate results, new methods should focus on the precise modelling of the ligand itself and of its binding residues. A further improvement in the

direction of ligand and binding site modelling could be the selection of the template based on the best candidate ligands for the target. This approach should allow building models with better quality in the binding site region and a more precise ligand annotation than current homology methods, although the quality in other parts of the models may become worse. This consideration would promote the development of modelling pipelines that employ a template for the binding site and another for the rest of the target. On the other hand, the limitation imposed by available ligands in protein structures suggested that the development of de novo ligand modelling methods might represent a valuable alternative.

Finally, we tested the moment invariants as a new approach for an efficient description and comparison of ligand binding sites in a de novo predictor; however, we concluded that this representation of ion binding sites was not a suitable option, mostly because it was not enough accurate in discriminating different binding sites. In the future steps of the analysis, it will be interesting to verify whether these descriptors could be successfully used to represent and identify the binding sites of organic ligands.

Acknowledgments

I would like to thank Prof. Torsten Schwede for giving me the opportunity to conduct my PhD project in his research group; I am very grateful to him for the helpful advices and interesting discussions. I would further like to thank all the past and present members of my group, for the nice atmosphere and constructive exchange of ideas; a special thanks to Valerio, Marco and Tobias for interesting discussions and scientific feedbacks. Also, I would like to thank the IT team for their assistance on cluster infrastructure and the secretaries for the vital help with the administrative tasks. Finally, I thank my girlfriend, family and friends for their encouragement and support during these years.