Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Structural Biology

Evolutionary studies of ligand binding sites in proteins

Rafael J Najmanovich

Biological processes at their most fundamental molecular aspects are defined by molecular interactions with ligand–protein interactions in particular at the core of cellular functions such as metabolism and signalling. Divergent and convergent processes shape the evolution of ligand binding sites. The competition between similar ligands and binding sites across protein families create evolutionary pressures that affect the specificity and selectivity of interactions. This short review showcases recent studies of the evolution of ligand binding-sites and methods used to detect binding-site similarities.

Address
Department of Pharmacology and Physiology, Faculty of Medicine, Université de Montreal, Montreal H3T 1J4, Quebec, Canada

Corresponding author: Najmanovich, Rafael J (rafael.najmanovich@umontreal.ca)

Current Opinion in Structural Biology 2017, **45**:85–90

This review comes from a themed issue on **Engineering and design: New trends in designer proteins**

Edited by **Niv Papo** and **Julia Shifman**

For a complete overview see the Issue and the Editorial

Available online 18th December 2016

http://dx.doi.org/10.1016/j.sbi.2016.11.024

Introduction

The study of the evolution of ligand binding-sites in proteins, most notably enzymes has greatly benefited from our understanding of the evolution of protein structure [1–3] and function [4] as well as the availability of methods for the detection of binding-site similarities [5]. In recent years, insights on the evolution of protein function both as a result of divergent and convergent evolutionary processes have been the result of top-down and bottom-up approaches. Top-down approaches using the amassed available data in public repositories and efficient computational methods to perform large scale analyses clarified the relationship and extent of evolutionary connections between enzyme families whereas bottom-up approaches, particularly experimental approaches using directed evolution as a probing tool permitted to probe functional changes and evolutionary pathways in great detail. Recent breakthroughs include the understanding of the evolution of catalytic mechanisms across enzyme superfamilies and the appreciation of the role of convergent evolution. This short review focuses on small-molecule binding-sites but many of the ideas may be applicable to the evolution of protein–protein interactions [6].

Methods for the detection of binding-site similarities

Over the years a large number of methods to detect binding-site similarities have been developed. Only a few are presented here as representatives of the approaches mentioned. Different methods can be categorized according to the level of detail used to represent binding-sites, the methodology used to search for similarities and the scoring scheme (Table 1).

The detection of binding-site similarities depends on the capacity to detect or define binding-sites. Research on the detection of binding-sites aims in general at identifying cavities that are either biologically relevant (e.g., known to bind a ligand or allosterically affecting binding) or 'druggable'. A number of methods exist for such a purpose, from purely geometric such as PASS [7], SURFNET [8] and its modern implementation within the NRGsuite PyMOL plugin [9] to methodologies that consider additional information such as evolutionary conservation [10] or energetic considerations as exemplified by PocketFinder [11]. The resulting cavities detected with any of the methods above (among others) can be used as input for the detection of similarities.

Representation. At a most basic or reduced level of representation, one can map binding-site residues onto the primary sequence and use pairwise or multiple sequence alignments to define a Tanimoto coefficient of binding-site sequence identity [12]. In other words, one can count the number of identical aligned binding-site residues c and normalize that number by the number of residues in either binding-site (a and b respectively) to obtain a Tanimoto score ($c/(a+b-c)$). The eMatchSite algorithm goes a step further and creates sequence order-independent alignments of ligand binding-sites [13]. Considering the Functionalist principle discussed below, different methods representing binding-sites at increasing levels of biological complexity aim at detecting similarities that capture the biological information responsible for higher levels of conservation. Climbing the representation complexity ladder, at the level of structure we find a number of approaches to represent binding-sites, from C-alpha atoms and microenvironments to all-atom representations. PSILO [14], SOIPPA [15] as well as the C-alpha mode of IsoCleft [16,17] represent binding-sites via C-alpha atoms. APoc represents binding-sites utilizing the C-alpha and C-beta atoms as well as a classification of amino-acids into 8 classes [18]. Pre-defined atomic

Table 1

A compilation of methods for the detection of binding-site similarities

Method	Representation	Search	Scoring
eMatchSite[a] [13]	Sequence	Geometric hashing	Correlation
PSILO[b] [14]	C-alpha atoms	Exhaustive	RMSD
SOIPPA[c] [15]	C-alpha atoms	Graph matching	Profile distance
IsoCleft[d] [16,17]	All-atoms	Graph matching	Tanimoto/volume
APoc[e] [18]	C-alpha C-beta atoms and sequence	Structural alignment	PS-score
CavBase[f] [19]	Pseudo-centres	Graph matching	Surface overlap
SiteEngine[g] [20]	Pseudo-centres	Geometric hashing	Surface overlap
PocketFEATURE[h] [21]	Microenvironments	Exhaustive	Tanimoto score
GRID-FLAP[i] [22]	MIFs	Exhaustive	Volume overlap
IsoMIF[j] [23•,24]	MIFs	Graph matching	Tanimoto/volume

[a] Free source code and web-accessible at http://www.brylinski.org/ematchsite.
[b] Commercially available, Chemical Computing Group.
[c] Source code available as part of SMAP at www.compsci.hunter.cuny.edu/~leixie/smap/smap and web interface: www.bioinfo.cs.pu.edu.tw/cloud-PLBS.
[d] Free source code and web-accessible at www.bcb.med.usherbrooke.ca/icfi.
[e] Source code freely available at http://cssb.biology.gatech.edu/APoc.
[f] Unknown availability.
[g] Free source code for non-commercial users and web-accessible at www.bioinfo3d.cs.tau.ac.il/SiteEngine.
[h] Free source available at www.simtk.org/projects/pocketfeature.
[i] Commercially available, Molecular Discovery.
[j] Free source code and web-accessible at www.bcb.med.usherbrooke.ca/isomif.

pseudo-centres aim at capturing the presence of important interacting groups while decreasing the number of objects that need to be compared. Representative methods of this approach are CavBase [19] and SiteEngine [20]. PocketFEATURE [21] defines microenvironments at specific geometric centres of residues and calculated physico-chemical properties associated with the atoms present in concentric shells at different radii. The all-atom mode of IsoCleft [16,17] uses all non-hydrogen atoms to represent binding-sites. Further still in the complexity-representation ladder, we find methods that represent binding-sites by the potential interactions that could be made with particular chemical probes at different positions within the volume of the cavity using potential energy functions to define molecular interaction fields (MIFs). Notable methods in this category are GRID-FLAP [22] and IsoMIF [23•,24]. The potential advantage of using MIFs rather than the specific positions of atoms or associated properties at the molecular surface is to account for different binding-site residue configurations that do not affect binding or cases where small differences can have drastic effects.

Search and scoring. In addition to representation, the other two pillars of any optimization problem are the method used for searching for solutions (in this case similarities between the binding-sites) and the scoring scheme. Predominantly used search algorithms are geometric hashing [25], graph matching [26] and exhaustive enumeration. The choice of method and its implementation depends on the type of representation used as that generally dictates the size of the search space. Different distance measures can be used to quantify similarity such as the Tanimoto score, root mean square distance (RMSD) of the identified similarities after superimposition, and surface overlap (Table 1).

Performance. Despite the same purpose of detecting binding-site similarities, the different methods in Table 1 were developed with slightly different applications in mind and therefore were evaluated using tailor-made datasets. It would be interesting to add to this list of 'benchmark' datasets the Shoichet dataset discussed above [27••]. As reported in [23•], different methods perform well on particular datasets but poorly on others, with eMatchSite and IsoMIF having the largest average Area Under the receiver-operator Curve (AUC) across datasets at around 0.80. It is interesting to note that the two methods at the extremes of the scale of biological complexity representation discussed above have the best performance. However, unlike eMatchSite, IsoMIF displays a very low AUC variance, thus its performance is more robust across datasets. The wide variance in performance across datasets suggests that using multiple datasets is beneficial. Thus, the lack of any single ultimate benchmark dataset is a situation that should be maintained. Instead of a single benchmark dataset, even more benchmarking datasets should be used as part of a benchmark dataset pool. Whereas the advantages of such an approach are clear within the realm of methods for the detection of binding-site similarities, another field that would drastically benefit from such an approach is that of small-molecule docking simulation methods where benchmarking is dominated by the use of the Astex datasets [28,29] but different methods clearly vary in their performance when tested on different datasets [30].