Bioorganic & Medicinal Chemistry Letters 25 (2015) 1163–1170

Contents lists available at ScienceDirect

Bioorganic & Medicinal Chemistry Letters

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

BMCL Digest

Implications of the small number of distinct ligand binding pockets in proteins for drug discovery, evolution and biochemical function

Jeffrey Skolnick *, Mu Gao, Ambrish Roy, Bharath Srinivasan, Hongyi Zhou

Center for the Study of Systems Biology, Georgia Institute of Technology, 250 14th St NW, Atlanta, GA 30318, USA

ARTICLE INFO

Article history:
Received 14 November 2014
Revised 23 January 2015
Accepted 24 January 2015
Available online 3 February 2015

Keywords:
Ligand binding pockets
Virtual ligand screening
Kinases, proteases, phosphatases
Drug repurposing
Drug side effects

ABSTRACT

Coincidence of the properties of ligand binding pockets in native proteins with those in proteins generated by computer simulations without selection for function shows that pockets are a generic protein feature and the number of distinct pockets is small. Similar pockets occur in unrelated protein structures, an observation successfully employed in pocket-based virtual ligand screening. The small number of pockets suggests that off-target interactions among diverse proteins are inherent; kinases, proteases and phosphatases show this prototypical behavior. The ability to repurpose FDA approved drugs is general, and minor side effects cannot be avoided. Finally, the implications to drug discovery are explored.

Introduction: Despite the tremendous effort that goes into designing a small molecule drug that uniquely binds to a specific protein, often that drug binds other, sometimes evolutionarily unrelated, proteins.[1,2] This interaction promiscuity leads to unexpected side effects, which depending on their nature, could result in the drug failing a clinical trial or being repurposed to treat other diseases.[3] These results imply that the number of distinct small molecule binding sites or pockets must be reasonably small; otherwise, the likelihood that two evolutionarily unrelated proteins would share similar stereochemical shapes and environments would be inconsequential. Indeed, the widespread prevalence of drug side effects raises a plethora of questions: (1) how special are the observed small molecule ligand binding pockets? Are they just a byproduct of protein structure and amino acid composition or do they require evolutionary selection for them to occur? (2) How many distinct ligand binding pockets are there? (3) Is the space of ligand binding pockets complete; that is, are all small molecule binding pockets known? (4) What is the relationship between the global fold of a protein and the structure of their ligand binding pockets? Must two proteins have the same global fold for them to share similar pockets or is pocket geometry weakly coupled to global fold? (5) Conversely, if two proteins have high sequence and structural similarity, must they have very similar pockets? (6) To what extent can one infer similar protein–ligand interactions by the similarity of their ligand binding pockets? (7) Can one use these insights to design better virtual screening algorithms based on ligand binding pocket similarity?[4–7] (8) For possible off target interactions of the major classes of drug targets, kinases, proteases and phosphatases,[8] how often do their pockets match those in other protein families? (9) What are the consequences of such promiscuity for the development of better drug discovery paradigms? In what follows, we address each of these questions and suggest possible answers.

Simulations to tease out inherent protein properties: To separate out the intrinsic properties of proteins from those due to evolution, in principle one could design proteins without any selection for function, solve their structures, assay them for ligand binding and explore the similarity between their pockets and those in native proteins.[9–12] To cover all representative protein folds and pocket geometries would be a long, expensive process, that is, at present, impractical. Rather, we chose to perform a series of computer experiments where a library of compact homopolypeptides from 40 to 250 residues in length were generated using the TASSER structure prediction algorithm.[13] Then, sequences with protein-like composition are selected by optimizing their thermodynamic stability (using potentials describing secondary structure, burial and pair interactions) in the putative fold of interest.[14] We then compare the properties of the pockets found in these artificial, ART, proteins with those found in the PDB.[15] The qualitative results that emerge are independent of the particular potential used to select the sequences, thereby suggesting that the results are robust. Parenthetically, we note that the set of folds in the ART library matches those in the PDB,[16] as does its set of protein–

* Corresponding author. Tel.: +1 404 407 8975; fax: +1 404 385 7478.
E-mail address: skolnick@gatech.edu (J. Skolnick).

http://dx.doi.org/10.1016/j.bmcl.2015.01.059
0960-894X/

protein interfaces.[17] Thus, with regards to a variety of other structural features, the PDB and ART libraries are very similar. The recapitulation of many native like protein properties lends credence that ART proteins might also recapitulate many features of pockets in native proteins.

Pocket comparison algorithm: We first address the requirements to generate native-like protein pockets in single domain protein structures. To do so, one needs an algorithm that can compare the structures of protein pockets. Here, we employ the APoc pocket structural alignment algorithm.[18] Pockets are ranked using a pocket structural similarity, PS-score that goes from 0 to 1 (identical pockets). A PS-score of 0.38 has a *P*-value of 2.6×10^{-3} We use this as the threshold that a pair of pockets is structurally related. The PS-score offers the advantages that its mean is pocket size independent and its statistical significance is provided. We further note that structural fluctuations have a marginal effect on pocket identification and the resulting overlap between a pair of pockets.[14]

Software tools: For the convenience of the reader, Table 1 provides a summary of the computational tools used to generate the data in this review as well as the URL where the software can be obtained.

Matching of pockets: Figure 1 plots the cumulative fraction of proteins whose best PS-score matches a pocket that exceeds the given threshold. Every native pocket has a statistically significant match in the ART library and vice versa. Both ART–PDB and ART–ART libraries have somewhat lower quality matching pockets than those found in the PDB–PDB comparison. This is partly because PDB structures have a somewhat greater number of larger pockets than are found in ART proteins. Large pockets can be a source of many matches to small pockets. The fact that all PDB pockets up to 60 residues in size have a statistically significant match to pockets in the ART library suggests that the library of native pockets is likely complete. Since ART pockets are generated without any functional selection or evolution, this implies that the space of protein pockets is mainly determined by the compact packing of secondary structural elements, as the volume of pockets is very tiny in compact proteins lacking secondary structure.[16] This is an important conclusion with implications for the origin of the biochemistry of life.

Number of pockets: Next, in Figure 2, we compute the number of representative pockets as a function of PS-score. For PDB–PDB, PDB–ART and ART–ART pocket pairs above the random threshold (PS-score = 0.38), there are roughly 200–300 representative pockets that cover the entire pocket space. PDB or ART pockets tend to find a larger similarity among themselves than to each other. Again, this reflects the fact that the current ART library has fewer large pockets that can cover many smaller pockets than are present in the PDB. Thus, there is a larger fraction of PDB pockets matched at higher PS-scores. This deficit of larger pockets is likely an artifact of the way the ART library was prepared. Nevertheless, the ART library covers all PDB pockets at a statistically significant level. From Figures 1 and 2, we conclude that the library of PDB pockets is likely complete and covered by a rather small set of distinct pockets.

Relationship between global fold similarity and pocket similarity: To assess global protein structural similarity, we employ the TM-score,[19–21] whose value ranges from 0 to 1.0; proteins with globally related structures have a TM-score $\geqslant 0.4$ (a statistically significant score with a *P*-value of 3.4×10^{-5}).[21] Figure 3 shows the distribution of PS-scores for a given extent of global structure similarity. For globally unrelated proteins, with a TM-score = 0.18, their best matching pocket structures are mostly unrelated; yet, even here, 3.5% of pockets are structurally similar. For globally similar proteins with a TM-score = 0.40, 39% of proteins have structurally similar pockets, with virtually identical behavior when all three sets (PDB–PDB, PDB–ART, ART–ART) are compared. Comparison of PDB–ART structures clearly shows that even when one has high global structural similarity (TM-score = 0.6) and high pocket similarity

Table 1
Computational tools used in this review

Protein 3D structure prediction
http://cssb.biology.gatech.edu/skolnick/webservice/TASSER-VMT/index.html
http://cssb.biology.gatech.edu/TASSER-VMT-Lite/index.html
Comparison of protein global structural similarity
http://cssb.biology.gatech.edu/fr-tm-align
Comparison of protein pocket/local structural similarity
http://cssb.biology.gatech.edu/APoc
Comparison of ligand 3D structural similarity
http://cssb.biology.gatech.edu/LIGSIFT

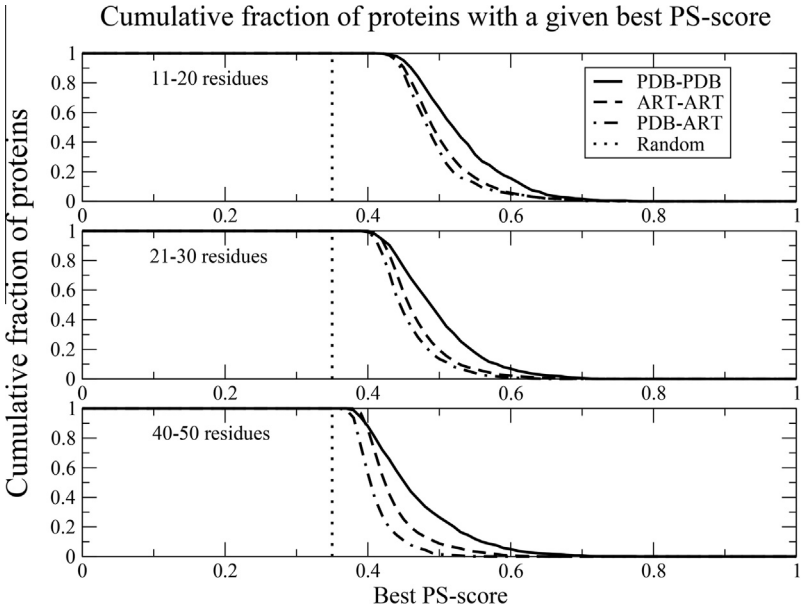

Figure 1. For different size pockets, cumulative fraction of proteins whose best PS-score to a pocket in the given structural library $\geqslant$ the specified PS-score threshold.