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

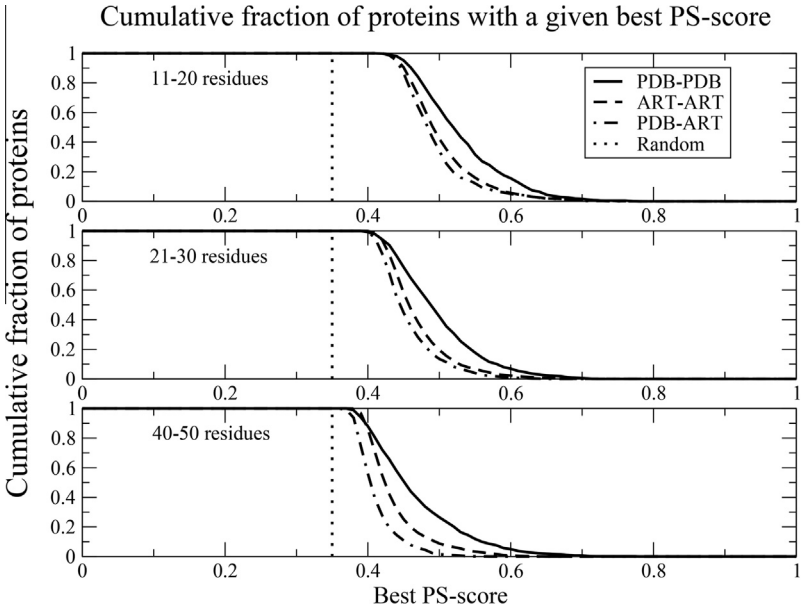

Figure 1. For different size pockets, cumulative fraction of proteins whose best PS-score to a pocket in the given structural library $\geqslant$ the specified PS-score threshold.

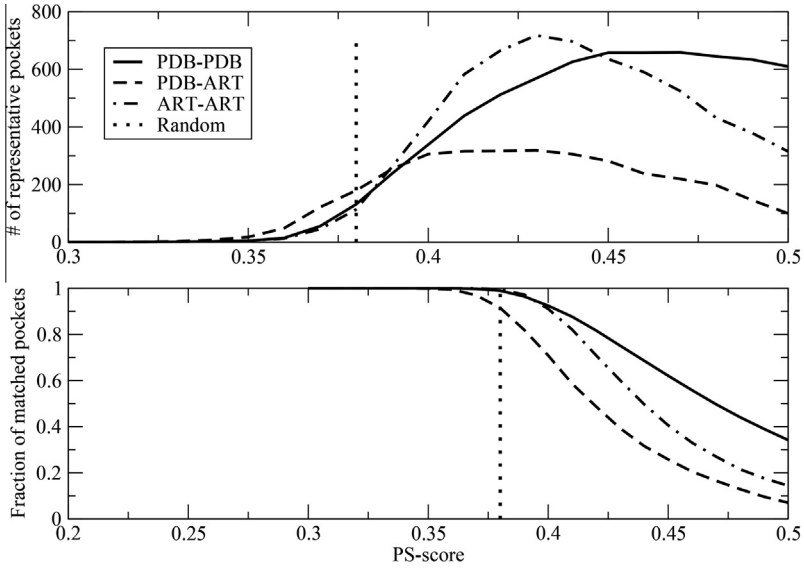

Figure 2. (A) Number of representative pockets for a given PS-score threshold versus PS-score in the PDB–PDB, PDB–ART and ART–ART pocket libraries. (B) Fraction of matched pockets in the PDB–PDB, PDB–ART and ART–ART libraries.

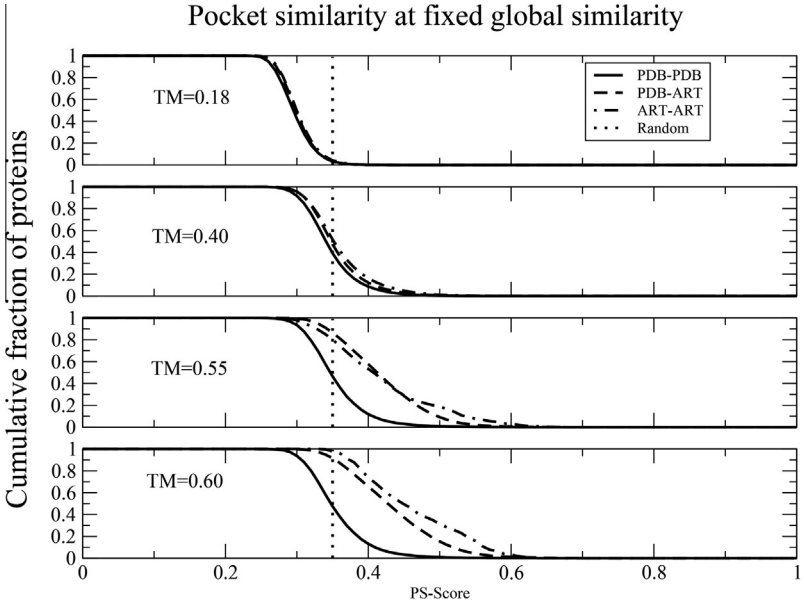

Figure 3. At a given level of global structural similarity as assessed by the TM-score, the cumulative fraction of proteins whose best PS-score to a pocket in the given structural library $\geqslant$ the specified PS-score threshold.

(PS-score >0.5), the proteins need not be evolutionarily be related. Hence, care has to be taken to infer evolutionary similarity even when their global fold and pockets are structurally similar. Conversely, for structurally very similar proteins, with TM-scores of 0.55 and 0.60, there is still a significant fraction of unrelated pockets for PDB–PDB (>60%), PDB–ART (>20%) and ART–ART (>10%) structures. Thus, while pocket similarity tends to increase with increasing global similarity, one can always find pockets in globally similar proteins that are very different. This is a feature of protein structure. Indeed, for native proteins, their pockets are even less coupled to the global fold than are the ART proteins.

Pocket similarity in globally very similar structures: In Figure 4, for each of the ART structures, we generated a set of 20 sequences selected for stability (whose backbone structure is the same) and examined the resulting PS-score distribution. Clearly, in these proteins with virtually identical global protein structures, pockets range from being highly similar to completely unrelated. This effect can be understood by the following: Imagine, for a same fixed backbone structure, one had nothing but GLY residues. Then, the entire protein interior is a pocket. Conversely, consider the case of nothing but TRP residues. Now, because TRP is a large amino acid, the pockets will be very tiny. For sequences with protein like composition, pockets can grow when a bulky amino acid is replaced by a small amino acid. Conversely, pockets can shrink or even bifurcate if a bulky amino acid is added. Interestingly, both ART and PDB structures explore the same range of pocket similarity for very similar global structures. This once again clearly indicates that there is a rather weak coupling between global fold and pocket similarity and the general features of protein pockets do not require evolution for them to be reproduced.

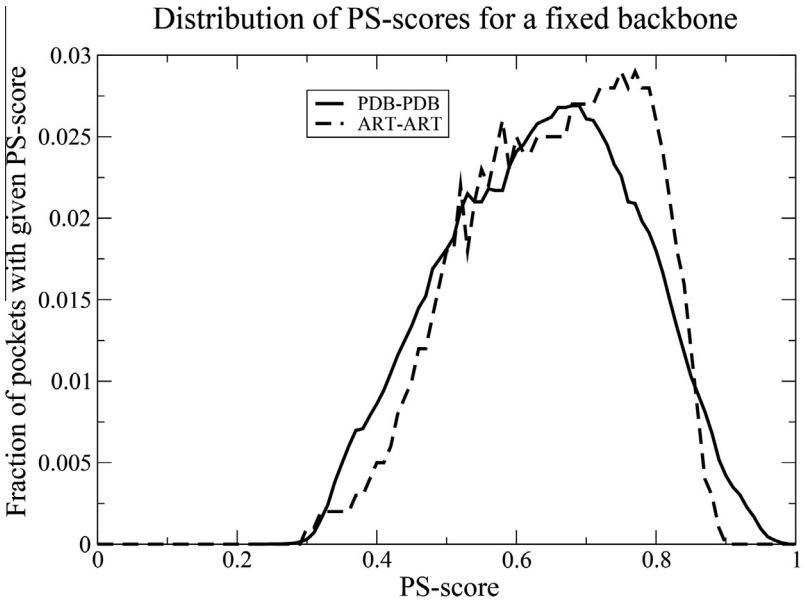

Figure 4. Relationship of protein sequence and PS-score for a fixed protein backbone structure for ART proteins and closely related PDB structures, whose TM-score $\geqslant$ 0.6.

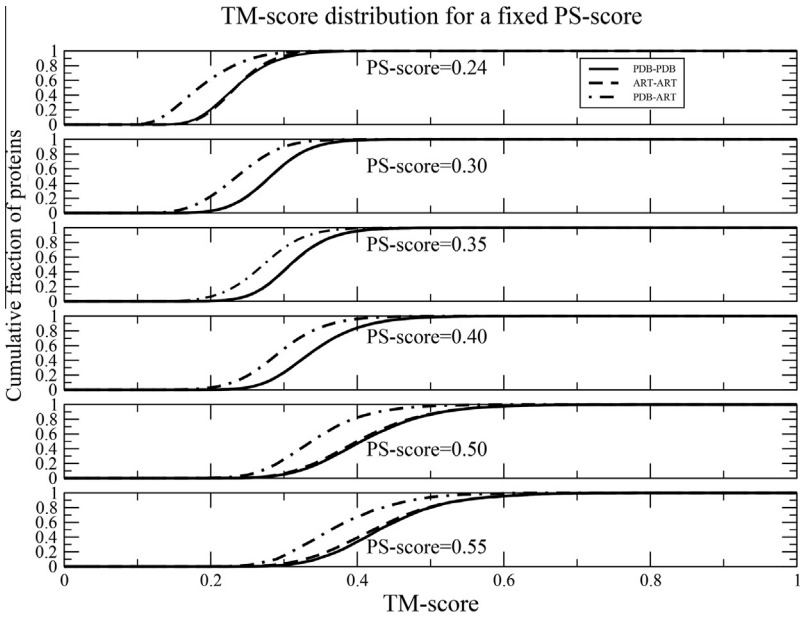

Figure 5. For a given PS-score, the cumulative fraction of proteins whose TM-score $\leqslant$ the given threshold.

Distribution of global structures at fixed pocket similarity: We now further explore the coupling of pocket similarity to global protein structure. On comparing PDB–PDB or ART–ART structures (Fig. 5), at fixed pocket similarity, the cumulative TM-score curves are very similar up to a PS-score of 0.55. Thus, the ART library of pockets recapitulates the coupling (or more accurately the lack thereof) of a given pocket with its occurrence across both related and unrelated global folds. Even for a highly significant PS-score of 0.5, about 12% of similar pockets are in unrelated structures (TM-score = 0.30). However, in this high pocket similarity regime, most pockets are found in globally related structures (even if they are unrelated by evolution). Overall, on comparing PDB pockets that match those in the ART library, for ART proteins, matching pockets are more likely found in globally dissimilar proteins. This further reinforces the idea that pocket geometry and global protein fold are weakly coupled.

Sequence conservation in similar pockets: Thus far, we focused on protein structural properties and did not consider the explicit role of protein sequence. However, even if two pockets have very similar geometric shapes, whether or not they interact with a given ligand also depends on the sequence of residues that line the pocket. Often, sequence conservation of pocket residues is used to infer an evolutionary, or more precisely, functional relationship.[22,23] Since native protein pockets are the convolution of physical interactions and evolution, to tease out these effects, in Figure 6, we compare the fraction of proteins that have a given number of pocket residues conserved at a given equivalenced position in the pocket as a function of PS-score for ART and PDB proteins. Interestingly, up to a PS-score of 0.4, the sequence conservation behaviors of PDB–PDB, PDB–ART and ART–ART sets of pockets are virtually identical. For PDB–PDB pockets, a slight echo of evolutionary conservation possibly occurs around a

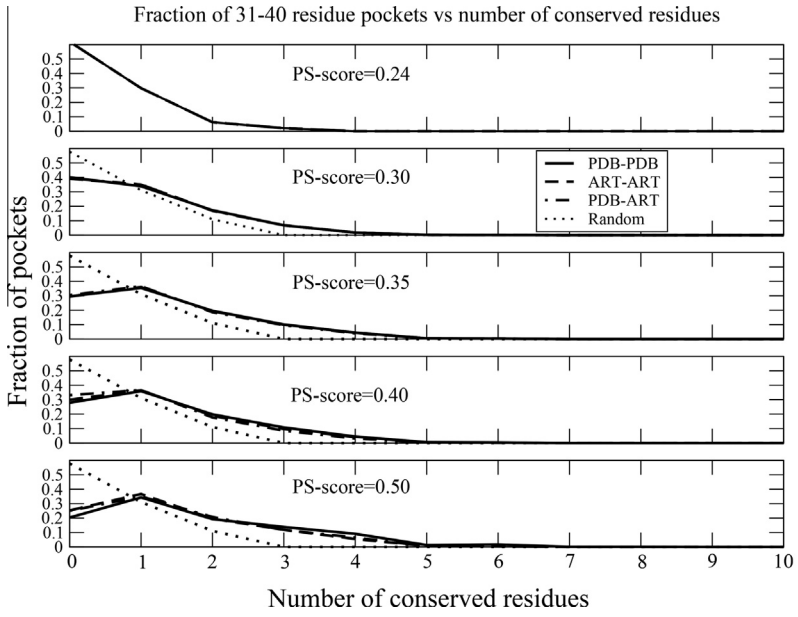

Figure 6. For a given PS-score, the fraction of proteins with a given number of conserved residues when PDB–PDB, PDB–ART and ART–ART structures are compared.

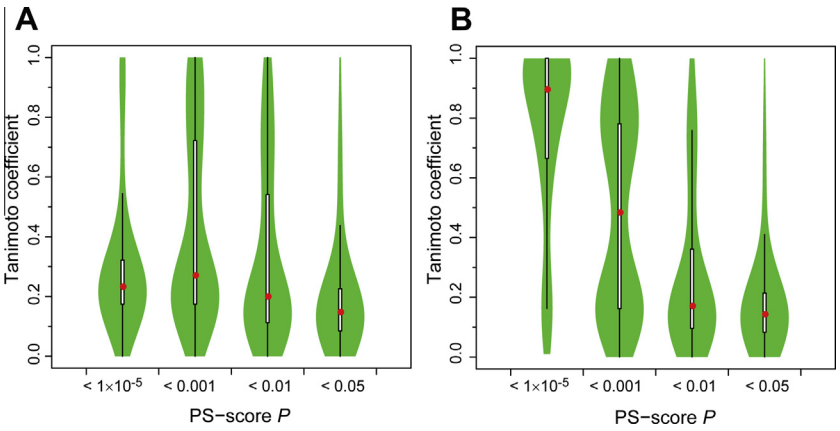

Figure 7. Tanimoto coefficient between pairs of ligands for a given *P*-value of the PS-score for pairs of proteins with similar (A), and different (B), global folds.

PS-score of 0.5, where higher sequence conservation is just barely seen in PDB–PDB pairs of pockets. Thus, pockets that are entirely unrelated by evolution (PDB–ART or ART–ART) but that are structurally similar (PS-score = 0.40, *P*-value $\sim 10^{-3}$) have virtually the same degree of sequence conservation at fixed pocket positions as native proteins. This is suggestive of the origin of functional promiscuity conjectured by Jensen[24] and Tawfik.[25–27] The ability to bind similar ligands perhaps arises from protein sequences that have thermodynamically stable native protein structures. Once again, this reinforces the point that care needs to be taken to infer whether a pair of proteins is evolutionarily related. They can have similar structure, similar pockets and similar pattern of conserved residues, and yet need not be related by evolution.

Examination of pocket and ligand similarity: The preceding analysis was pocket centric and entirely ignored the identity and chemistry of the ligands that bind in the protein's pocket. We now examine the extent of pocket similarity in PDB proteins required to infer that they bind similar ligands. This is a first step towards the development of a pocket/ligand similarity based virtual screening algorithm.[28] Figure 7A shows a violin plot of the Tanimoto coefficient versus *P*-value of the PS-score for a representative set of PDB pockets and associated ligands.[28] About 13% of ligand pairs share significant chemical similarity when their Tanimoto coefficients (Tc) > 0.4 and $0.01 \leqslant P$-value < 0.05. This value increases to 37% on increasing the pocket similarity to a *P*-value of 1×10^{-5}. However, the percent of similar ligands drops to 18% when pocket *P*-value $<1 \times 10^{-5}$. This is due to the fact that many pockets are promiscuous and interact with chemically different ligands (at least as assessed by their Tc); however, a subset of the ligands might adopt similar stereochemical shapes, this is not apparent in their Tc scores.

Pocket similarity can imply binding ligand similarity: Next, we examine how often protein-ligand interactions found in the PDB can be matched to a template having a similar pocket that binds a similar ligand? By way of illustration, we search for the best pocket structural hit in a template protein whose global sequence identity <30% and whose chemical similarity Tc of bound ligands is above a specified value. As shown in Figure 8, at a significant Tc >0.4, ~86% of pockets find a template that binds similar ligands whose pocket similarity PS-score *P*-value <0.05. More conservatively, 72%, 60%, and 54% of ligands have pocket matches whose *P*-values <0.01, 0.001, and 0.0001, respectively. At a highly significant Tc >0.7, ~50% of pockets match a template whose *P*-value <0.01. Thus, structural comparison of pockets could be useful for

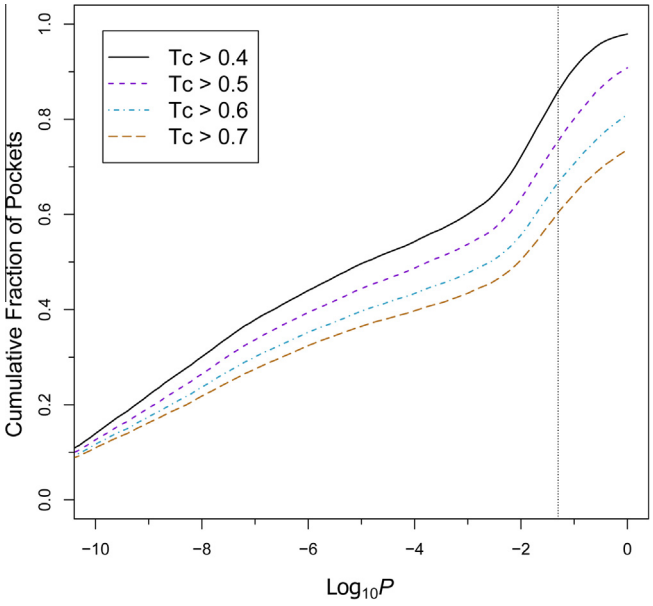

Figure 8. Cumulative fraction of pockets with a given Tc versus the log of the *P*-value of the PS-score.

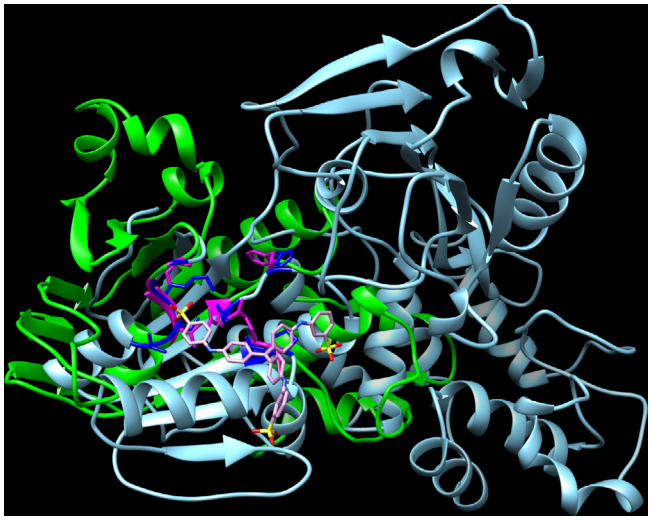

Figure 9. Binding site match between CFTR (green) and NS3 helicase (cyan). Their TM-score is 0.35 and pairwise sequence identity is 4%; yet, the *P*-value of the pocket similarity score is 1.1×10^{-3}.[18]

inferring ligand binding. Furthermore, many top structural hits come from proteins with different global folds. If Tc >0.4 (0.7) and the pocket *P*-value <0.05, about 35% (19%) of the top template hits are from proteins whose TM-score <0.4; viz. they have different global structures. Note that for all Tc values, there is a significant fraction of pockets that are structurally unrelated, and yet, they bind similar ligands. This could reflect the fact that ligands have internal degrees of freedom and can adopt different shapes that can bind to different pocket structures.

Application to virtual ligand screening: Can these ideas be translated into a practical, pocket similarity based virtual ligand screening algorithm? To demonstrate that pocket matching can successfully predict ligands that bind to similar pockets that come from globally dissimilar proteins, we tested our pocket matching algorithm, on the nucleotide binding domain of CFTR, a Cystic Fibrosis associated protein.[29,30] From its experimentally measured thermal shift, NSC 93739, a small molecule from the NCI diversity set,[31] identified on the basis of pocket similarity to NS3 helicase, has a K_D = 60 nM. As shown in Figure 9, the two proteins clearly lack global fold similarity, with a TM-score = 0.35; yet, the *P*-value of their pockets PS-score is 1.1×10^{-3}. The best binder, NSC10447 with a K_D = 16.5 nM, is from a similar pocket in cyclodextrin transferase. In total, 7/18 predicted ligands identified on the basis of pocket similarity have a $K_D \leqslant 100$ μM. Thus, in practice a pocket based virtual screening methodology finds a number of promising leads, but far more validation is needed.

Uniqueness of pockets in drug targets: Here, we focus on three classes of enzymes: kinases, phosphatases, and proteases, which provide a rich list of potential drug targets.[8] For each enzyme class, we used enzyme EC numbers to search for protein structures in a non-redundant set curated from the PDB.[28,32] This yields 2052 kinases, 188 phosphatases, and 545 proteases. We further group these proteins according to their EC numbers, resulting in 94, 16, and 23 groups for kinases, phosphatases, and proteases, respectively, at the four EC digit level (see Supplementary Tables 1–3). Next, ligand-bound pockets in these protein structures were identified. For each target pocket, we then search for similar pockets in other proteins, whose sequence identity with respect to the target protein is less than 30% according to the global structural alignment. Here, we define a pocket as a hit if it shares with the target pocket a significant PS-score >0.36 and a *P*-value <0.05 according to APoc.[18] As shown in Figure 10, for each EC group, the mean number of hits is 678 for kinases, 411 for phosphatases, and 398 for proteases. Although it is expected that remote homologs may share highly similar pocket as these enzymes have many paralogs, many hits are from apparently unrelated proteins with different structural folds. About 263 (39%), 369 (90%), 154 (39%) hits are from proteins which have TM-score <0.4 in the global structure alignment for kinases, phosphatases, and proteases, respectively. The most significant hits to target pockets usually have a highly similar pocket shape, with a mean PS-score of 0.60/0.52/0.59, mean *P*-value of $3 \times 10^{-7}/4 \times 10^{-5}/1 \times 10^{-6}$, mean pocket RMSD of 1.58/1.93/1.64 Å, mean sequence identity (in the pocket alignment) of 42%/32%/36%, and alignment coverage of 88%/90%/89%, for kinases/phosphatases/proteases, respectively.

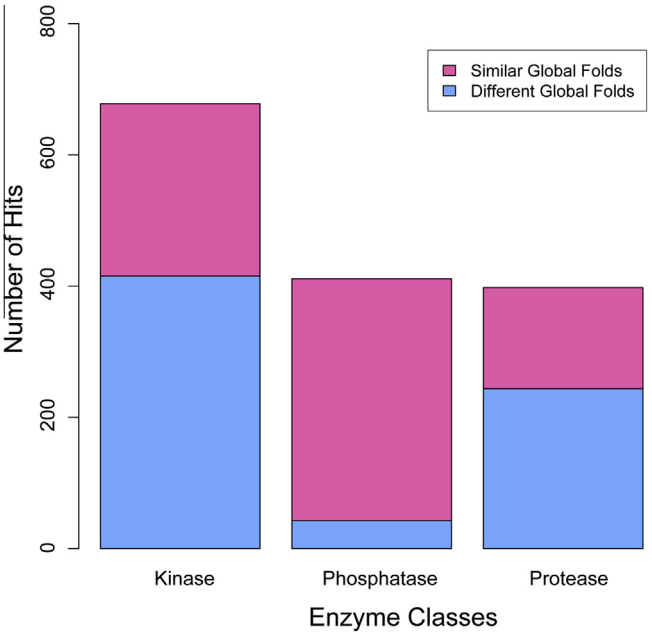

Figure 10. Number of similar pocket hits to evolutionary distant, if not unrelated, proteins in kinases, proteases and phosphatases with similar (pink) and different (blue) folds.

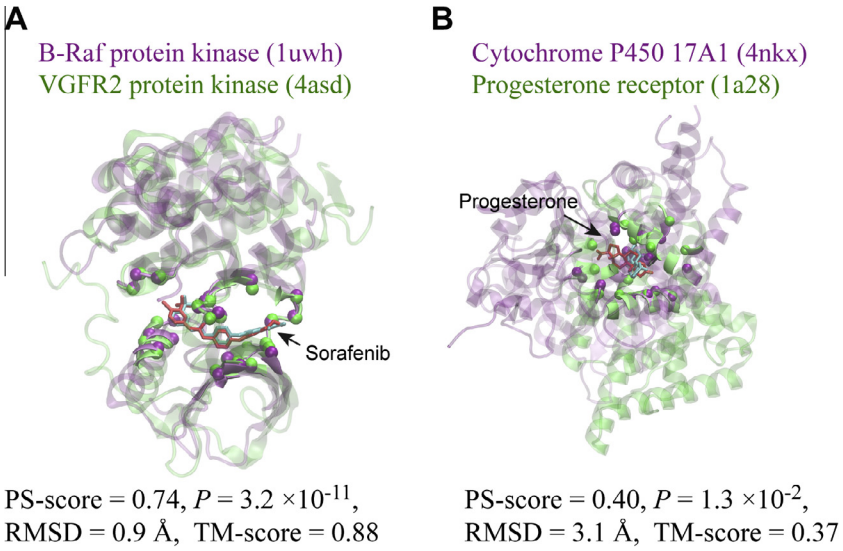

Figure 11. (A) Example of Sorafenib that binds to two different kinases, B-Raf and VGFR2, with low sequence identity. (B) Example of progesterone that binds to two proteins whose pockets are similar despite the fact that their global structural similarity is low. The RMSD of the pocket residues is indicated, as is the global TM-score.

As an example, sorafenib is one of FDA approved anti-cancer drugs that have multiple targets in the family of protein kinases. Two of these kinases, B-Raf[33] and VGFR2,[34] have been crystallized in complex with the drug molecule. Figure 11A, shows the superposition of these two structurally similar complexes, which demonstrate a high similarity in their drug-binding pockets at a PS-score of 0.74, despite a low global sequence similarity of 30%. Perhaps this result is not that surprising, given that protein kinases share high level structural similarity in its catalytic domain.

We next turn to the example of a drug, progesterone, a steroid that has many known targets, binding to a pocket found in two proteins with globally different structures. As shown in Figure 11B, two of its targets, the human progesterone receptor[35] and the human steroidogenic cytochrome P450 17A1,[36] are resolved in complex with progesterone. However, significant structural similarity is detected in the progesterone-binding pockets at a PS-score of 0.40. These two examples illustrate that the same drug molecule can bind to similar pockets located in multiple targets, whether they are structurally related or not at the global structural level.

More generally, protein tyrosine kinases (EC 2.7.10.2) and protein serine–threonine kinases (2.7.11.1) are two groups of kinases frequently targeted by drugs to treat cancer.[37] As shown in Table S1, these two groups of kinases contain pockets that have similar counterparts in over 1200 proteins, with ~300 or so hits to pockets in globally unrelated structures. Many have a highly significant PS-score >0.60.

Recently, tyrosine phosphatases (EC 3.1.3.48) (see Table S2) have been proposed as potential therapeutic targets for diabetes, obesity, and cancer.[38] We found fewer significant pockets hits for these targets compared to kinases and proteases, but there are still over 200 hits in evolutionary very distant if not unrelated proteins whose global fold is different. Finally, the HIV protease family (EC 3.4.23.16) is another drug target.[39] From Table S3, the main pocket of this target can be found in more than 700 protein structures, 536 of which have unrelated global folds. The clear implication of this analysis is that drugs, which target these pharmaceutically relevant classes of proteins, likely have many off-target interactions. Many might belong to proteins of entirely unrelated structures. This clearly shows that promiscuous off target drug interactions must be dealt with in drug discovery.

Conclusions: Many biological activities originate from interactions between small-molecule ligands and their protein targets in pockets in the protein's structure.[28] In this work, the intrinsic ability of protein structures to exhibit the geometric/sequence properties required for ligand binding without evolutionary selection was shown by the coincidence of properties of pockets in native, single domain proteins with computationally generated, compact homopolypeptide, artificial structures, ART.[14] The library of native pockets is covered by a very small number of representatives (<400); almost all native pockets have a statistically significant match to the ART pocket library, suggesting that the library is complete. The facts that structurally and sequentially similar pockets occur across fold classes and the number of representative native pockets is small imply that promiscuous interactions are inherent to proteins. These results are consistent with a large-scale study on a non-redundant set of ~20,000 known ligand-binding pockets that finds that their structural space is crowded, likely complete, and represented by a similar number of pockets.[28] Moreover, proteins have pockets that can interact with diverse ligand scaffolds.[28] Thus, large scale drug repurposing, where FDA approved drugs are applied to new uses,[3,40–54] should be general. While pocket matching is straightforward, specific ligand selection can be problematic and requires the development of better tools.

This work rationalizes why drug side effects occur despite focused attempts to design drugs that interact with a specific protein target. There are simply many similar pockets in the cellular proteome, a fact clearly demonstrated for kinases, proteases and phosphatases. While one might have expected cross reactivity within a protein family, (although due to possible pocket divergence within the family, this might not be as crucial as naively expected, see Fig. 4), the fact that similar pockets occur in evolutionarily unrelated proteins (see Fig. 11B) suggests that the best drug design strategy would be to identify all putative pockets across a proteome, identify proteins that are more likely to give rise to side effects and then design inhibitors towards the protein of interest and destabilize their interactions towards the other, most deleterious protein targets. Of course, in reality one does not always have the structures of these off target proteins. Here, structure prediction can play a role to identify at least some off-target proteins, as ~86% of the human proteome can be modeled at the requisite level of resolution.

If the one small molecule-one protein target paradigm of Paul Ehrlich were mostly true, then cellular regulation as well as drug discovery would be relatively easy to achieve. Rather, the small

number of ligand binding pockets suggests that the one small molecule-many protein targets (a given ligand likely has different specificity for its protein targets) and the one protein-many small molecule binders viewpoint is a closer approximation to reality. Redundancy could provide for biological robustness and could rationalize how biochemical processes evolved; yet, how to regulate and control such promiscuity is not readily apparent.

On average, a given single domain protein contains from 3 to 5 pockets whose volume is large enough to bind to a typical small molecule ligand. For a given protein target, if one wants to identify drugs with a novel mode of action, a possibility is to select another pocket different from the one targeted by extant drugs. Of course, one would still have to demonstrate that by modifying a secondary pocket, the desired physiological response is achieved.

Overall, this work strongly suggests that a pocket centric view of protein-ligand interactions is a powerful way of repurposing FDA approved drugs, understanding minor side effects and finding molecules with novel modes of interaction. We presented a promising example of how such an approach can predict novel nM binding molecules of the nucleotide binding domain of CFTR, a Cystic Fibrosis associated protein.[29,30] We also showed that drugs that target kinases, proteases and phosphatases likely have off-target interactions across a large set of diverse proteins. Because of the rather small number of ligand binding pockets, a more effective approach to drug discovery should consider these alternative interactions in the design strategy. Here, one can view this as an optimization procedure to target the specific features of the pocket in the protein of interest and design away from the other pockets. Such types of approaches are currently being pursued, as well as questions concerning the completeness of the space of stereochemical shapes adopted by drug like small molecules and how much of that space is represented by FDA approved drugs. We further note that the notion that binding sites with similar structures bind similar ligands is embodied in the idea of similar ligand sensing cores in the PSSC approach;[55] however in PSSC, there is the assumption such ligand sensing cores and their cognate natural products are the product of evolution. Here, we show that such protein binding sites likely arise from the intrinsic properties of proteins without evolution. Thus, fundamental consideration of the physical chemical properties of proteins and their small molecule binding partners can have practical applications to drug discovery.

Acknowledgements

This research was supported in part by grants GM-37408 and GM-48835 of the Division of General Medical Sciences of the National Institutes of Health.

Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.bmcl.2015.01.059.

References and notes

1. Gaulton, A.; Bellis, L.; Bento, A.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; Overington, J. *Nucleic Acids Res.* **2012**, *40*, D1100.
2. Liu, T.; Lin, Y.; Wen, X.; Jorissen, R. N.; Gilson, M. K. *Nucleic Acids Res.* **2007**, *35*, D198.
3. Ye, H.; Liu, Q.; Wei, J. *PloS one* **2014**, *9*, e87864.
4. Srinivasan, B.; Zhou, H.; Kubanek, J.; Skolnick, J. *J. Cheminform.* **2014**, *6*, 16.
5. Zhou, H.; Skolnick, J. *J. Chem. Inf. Model.* **2013**, *53*, 230.
6. Fan, H.; Irwin, J. J.; Sali, A. *Methods Mol. Biol.* **2012**, *819*, 105.
7. Brylinski, M.; Skolnick, J. *J. Comput. Chem.* **2010**, *31*, 1093.
8. Imming, P.; Sinning, C.; Meyer, A. *Nat. Rev. Drug Disc.* **2006**, *5*, 821.
9. Korendovych, I. V.; Kim, Y. H.; Ryan, A. H.; Lear, J. D.; Degrado, W. F.; Shandler, S. J. *Org. Lett.* **2010**, *12*, 5142.
10. Saunders, C. T.; Baker, D. *J. Mol. Biol.* **2005**, *346*, 631.
11. Sheffler, W.; Baker, D. *Protein Sci.* **2009**, *18*, 229.
12. Shu, J. Y.; Tan, C.; DeGrado, W. F.; Xu, T. *Biomacromolecules* **2008**, *9*, 2111.
13. Zhang, Y.; Skolnick, J. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 7594.
14. Skolnick, J.; Gao, M. *Proc. Natl. Acad. Sci. U.S.A.* **2013**, *110*, 9344.
15. Rose, P. W.; Beran, B.; Bi, C.; Bluhm, W. F.; Dimitropoulos, D.; Goodsell, D. S.; Prlic, A.; Quesada, M.; Quinn, G. B.; Westbrook, J. D.; Young, J.; Yukich, B.; Zardecki, C.; Berman, H. M.; Bourne, P. E. *Nucleic Acids Res.* **2011**, *39*, D392.
16. Brylinski, M.; Gao, M.; Skolnick, J. *Phys. Chem. Chem. Phys.* **2011**, *13*, 17044.
17. Gao, M.; Skolnick, J. *Proc. Natl. Acad. Sci. U.S.A.* **2010**, *107*, 22517.
18. Gao, M.; Skolnick, J. *Bioinformatics* **2013**, *29*, 597.
19. Zhang, Y.; Skolnick, J. *Nucleic Acids Res.* **2005**, *33*, 2302.
20. Pandit, S. B.; Skolnick, J. *BMC Bioinformatics* **2008**, *9*, 531.
21. Xu, J.; Zhang, Y. *Bioinformatics* **2010**, *26*, 889.
22. Arakaki, A. K.; Huang, Y.; Skolnick, J. *BMC Bioinformatics* **2009**, *10*, 107.
23. Furnham, N.; Holliday, G. L.; de Beer, T. A.; Jacobsen, J. O.; Pearson, W. R.; Thornton, J. M. *Nucleic Acids Res.* **2014**, *42*, D485.
24. Jensen, R. A. *Annu. Rev. Microbiol.* **1976**, *30*, 409.
25. Tawfik, D. S. *Nat. Chem. Biol.* **2010**, *6*, 692.
26. Khersonsky, O.; Tawfik, D. S. *Annu. Rev. Biochem.* **2010**, *79*, 471.
27. Bar-Even, A.; Noor, E.; Savir, Y.; Liebermeister, W.; Davidi, D.; Tawfik, D. S.; Milo, R. *Biochemistry* **2011**, *50*, 4402.
28. Gao, M.; Skolnick, J. *PLoS Comput. Biol.* **2013**, *9*, e1003302.
29. Ehre, C.; Ridley, C.; Thornton, D. J. *Int. J. Biochem. Cell Biol.* **2014**, *52*, 136.
30. Broadbent, S. D.; Wang, W.; Linsdell, P. *Biochem. Cell Biol.* **2014**, 1.
31. In *NCI diversity set IV*.
32. Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. *Nucleic Acids Res.* **2000**, *28*, 235–242.
33. Wan, P. T.; Garnett, M. J.; Roe, S. M.; Lee, S.; Niculescu-Duvaz, D.; Good, V. M.; Jones, C. M.; Marshall, C. J.; Springer, C. J.; Barford, D.; Marais, R. *Cell* **2004**, *116*, 855.
34. McTigue, M.; Murray, B. W.; Chen, J. H.; Deng, Y. L.; Solowiej, J.; Kania, R. S. *Proc. Natl. Acad. Sci. U.S.A.* **2012**, *109*, 18281.
35. Williams, S. P.; Sigler, P. B. *Nature* **1998**, *393*, 392.
36. Petrunak, E. M.; DeVore, N. M.; Porubsky, P. R.; Scott, E. E. *J. Biol. Chem.* **2014**, *289*, 32952.
37. Zhang, J.; Yang, P. L.; Gray, N. S. *Nat. Rev. Cancer* **2009**, *9*, 28.
38. Barr, A. J. *Future Med. Chem.* **2010**, *2*, 1563.
39. Brik, A.; Wong, C. H. *Org. Biomol. Chem.* **2003**, *1*, 5.
40. Xie, L.; Li, J.; Bourne, P. E. *PLoS Comput. Biol.* **2009**, *5*, e1000387.
41. von Eichborn, J.; Murgueitio, M. S.; Dunkel, M.; Koerner, S.; Bourne, P. E.; Preissner, R. *Nucleic Acids Res.* **2011**, *39*, D1060.
42. Gottlieb, A.; Stein, G. Y.; Ruppin, E.; Sharan, R. *Mol. Syst. Biol.* **2011**, *7*, 496.
43. Luo, H.; Chen, J.; Shi, L.; Mikailov, M.; Zhu, H.; Wang, K.; He, L.; Yang, L. *Nucleic Acids Res.* **2011**, *39*, W492.
44. Cobanoglu, M. C.; Liu, C.; Hu, F.; Oltvai, Z. N.; Bahar, I. *J. Chem. Inform. Model.* **2013**, *53*, 3399.
45. Xu, K.; Cote, T. R. *Brief. Bioinform.* **2011**, *12*, 341.
46. Amelio, I.; Gostev, M.; Knight, R. A.; Willis, A. E.; Melino, G.; Antonov, A. V. *Cell Death Dis.* **2014**, *5*, e1051.
47. Bastos, L. F.; Coelho, M. M. *CNS Drugs* **2014**, *28*, 45.
48. Lotharius, J.; Gamo-Benito, F. J.; Angulo-Barturen, I.; Clark, J.; Connelly, M.; Ferrer-Bazaga, S.; Parkinson, T.; Viswanath, P.; Bandodkar, B.; Rautela, N.; Bharath, S.; Duffy, S.; Avery, V. M.; Mohrle, J. J.; Guy, R. K.; Wells, T. *Malar. J.* **2014**, *13*, 143.
49. Nygren, P.; Larsson, R. *Acta Oncol.* **2014**, *53*, 427.
50. Parkkinen, J. A.; Kaski, S. *BMC Bioinform.* **2014**, *15*, 113.
51. Salentin, S.; Haupt, V. J.; Daminelli, S.; Schroeder, M. *Prog. Biophys. Mol. Biol.* **2014**, *12*, 1.
52. Sateriale, A.; Bessoff, K.; Sarkar, I. N.; Huston, C. D. *J. Am. Med. Inform. Assoc.* **2014**, *21*, 238.
53. Xu, R.; Wang, Q. *J. Biomed. Inform.* **2014**, *51*, 191.
54. Yang, J.; Li, Z.; Fan, X.; Cheng, Y. *J. Chem. Inform. Model.* **2014**, *54*, 2562.
55. Bon, R. S.; Waldmann, H. *Acc. Chem. Res.* **2010**, *43*, 1103.