Interplay of physics and evolution in the likely origin of protein biochemical function

Jeffrey Skolnick[1] and Mu Gao

Center for the Study of Systems Biology, School of Biology, Georgia Institute of Technology, Atlanta, GA 30318

Edited by Nick V. Grishin, Howard Hughes Medical Institute, Dallas, TX, and accepted by the Editorial Board April 23, 2013 (received for review January 2, 2013)

The intrinsic ability of protein structures to exhibit the geometric and sequence properties required for ligand binding without evolutionary selection is shown by the coincidence of the properties of pockets in native, single domain proteins with those in computationally generated, compact homopolypeptide, artificial (ART) structures. The library of native pockets is covered by a remarkably small number of representative pockets (~400), with virtually every native pocket having a statistically significant match in the ART library, suggesting that the library is complete. When sequences are selected for ART structures based on fold stability, pocket sequence conservation is coincident to native. The fact that structurally and sequentially similar pockets occur across fold classes combined with the small number of representative pockets in native proteins implies that promiscuous interactions are inherent to proteins. Based on comparison of PDB (real, single domain protein structures found in the Protein Data Bank) and ART structures and pockets, the widespread assumption that the co-occurrence of global structure, pocket similarity, and amino acid conservation demands an evolutionary relationship between proteins is shown to significantly underestimate the random background probability. Indeed, many features of biochemical function arise from the physical properties of proteins that evolution likely fine-tunes to achieve specificity. Finally, our study suggests that a repertoire of thermodynamically (marginally) stable proteins could engage in many of the biochemical reactions needed for living systems without selection for function, a conclusion with significant implications for the origin of life.

protein evolution | protein pocket space | protein–ligand interactions

One of the remarkable features of proteins is their ability to bind a variety of small molecules such as metabolites and drugs, with most binding surfaces formed by concave shapes or "pockets" on the protein's surface (1). The response of proteins to ligand binding gives rise to a plethora of biological processes that are essential for life (2). The ability to engage in these interactions reflects the convolution of the fundamental geometric, physical, and chemical properties of proteins with selection due to evolution. How could the ability to bind ligands have arisen? Consider two extreme views: The "inherent functionality model" asserts that the ability to engage in such interactions is just a physical chemical property of proteins, with the formation of ligand-binding cavities arising from defects in the packing of secondary structural elements comprised of hydrophobic and hydrophilic residues (3, 4), which evolution then exploits and amplifies. At the other extreme, in the "acquired functionality model," proteins were spherical objects without ligand-binding pockets. Evolution selected for function by literally sculpting out pockets to create functionally competent proteins. The inherent functionality model has the appeal that it a priori provides a nonzero background probability on which evolutionary selection operates. Because it is intrinsic, proteins could engage in a large variety of functions. These low-level, ligand–protein interactions should be highly promiscuous and act like biochemical noise that would be difficult, if not impossible, to eliminate. In contrast, the acquired functionality model implies that selection for function is very rare. Promiscuous interactions, while possible, would not be inherent and could be readily eliminated by functional selection. The goal of this contribution is to provide insights into the interplay of physics and evolution in dictating protein ligand-binding properties by eliminating for representative protein models any selection for protein function and then exploring the overlap of the structure and sequence properties of the resulting pockets with those in native proteins.

To test which of these two views of the origin of protein biochemical function is likely more correct, one must remove the effects of evolutionary selection for protein function. Protein design studies are one way of achieving this experimentally. Support for the inherent functionality model is provided by Hecht et al., who created a combinatorial library of designed four-helix bundle proteins expressed in *Escherichia coli*, through binary patterning of six polar and five nonpolar residues (5). The resulting superfamily, neither designed nor selected for function, was screened for a variety of functions including heme binding, peroxidase, and lipase activities. The majority of proteins bound heme, with a sizeable fraction showing activity in all assays. This suggests that protein structure and sequence composition provide rudimentary activity that "serve(s) as a feedstock for evolution." It also agrees with Jensen's conjecture (6) that "primitive enzymes possessed a broad range of specificity" that would allow an early cell to carry on the chemistry of life. This idea is compatible with observations that enzymes routinely catalyze other, sometimes barely related, chemical reactions (7–10). Such latent functions could then evolve without interfering with the original catalytic activity (11–13). Tawfik et al. argue that catalytic promiscuity is inherent to enzymes and suggest that contemporary enzymes diverged from ancestral proteins that catalyzed a plethora of low-level reactions (9, 10). Consistent with the notion that such activity is ubiquitous, protein design studies often find the desired low-level function after a remarkably small number of generations (14–16). These experimental studies suggest that promiscuous, low-level protein function is inherent.

How can one demonstrate that the ability to engage in a variety of low-level biochemical functions without selection is an intrinsic property of proteins that holds not just for a limited number of designed proteins (5, 8, 17, 18), but is likely true in general? Here, computational models of proteins can play a significant role. They offer the advantages of being comprehensive and could allow us to tease out which features of protein structures/sequences likely give rise to which functional properties. In earlier computational studies (3, 4, 19–21), the ability of protein structures to exhibit some of the geometric features required for molecular function sans evolution was examined in three representative protein structure libraries: the PDB library, real, single domain protein structures found in the Protein Data

Author contributions: J.S. and M.G. designed research, performed research, analyzed data, and wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission. N.V.G. is a guest editor invited by the Editorial Board.

Data deposition: All datasets reported in the paper have been made available through Center for the Study of Systems Biology, http://cssb.biology.gatech.edu/pocketlib.

[1]To whom correspondence should be addressed. E-mail: skolnick@gatech.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1300011110/-/DCSupplemental.

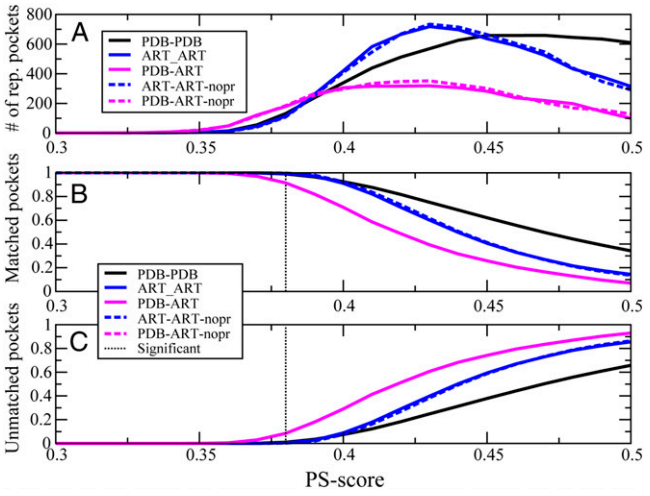

Fig. 1. (*A*) Number of representative, distinct pocket structures in the pocket template library that match pockets in the given target library at the given PS-score. (*B*) Fraction of pockets in the target library that are matched to representative pocket templates at the specified PS-score. (*C*) Fraction of target pockets that do not have a match to any other pocket structure in the pocket template library at the specified PS-score. That is, self-matching pockets are excluded from the PDB pocket template library. Solid (dashed) lines are protein sequences selected for global stability in ART structures generated using burial, secondary structure, and (without) pair potentials.

Bank (22); the ART library, computationally generated, compact homopolypeptide, artificial, structures with protein-like secondary structure; and the QS library, quasi-spherical, random protein structures packed in the same average spherical volume as proteins but lacking backbone secondary structure and hydrogen bonding. Without evolutionary selection, the library of artificial structures has statistically significant structural matches to the global structures of native proteins and due to defects in packing secondary structural elements, native like pocket volumes. However, the similarity of their pockets to native proteins was unexplored. While QS structures have a statistically significant match to the global structures of native proteins, lacking secondary structure, they are more densely packed and contain pockets that are too tiny to bind small molecules. Thus, backbone hydrogen bonding is likely one important underlying cause of protein function.

Since QS proteins lack the inherent features required for molecular interactions, while ART proteins have cavities of the requisite volume, the ART library is the reference against which we will compare its features with native pockets. To explore the dependence of the ART results on the potential used to evaluate stability, we generated sequences whose stability is determined by burial, secondary structure, and residue-based pair preferences (ART), or just by burial and secondary structure preferences in "no pair potential" ("no-pr") (3). Comparison will be made between pockets in native protein structures, PDB–PDB; pockets in the PDB to those in artificial structures, PDB–ART; and among the artificial structures, ART–ART. We shall determine the number of structurally distinct pockets needed to represent all pocket structures and examine the completeness of the space of protein pocket structures—that is, whether all representative pockets are already present in the library of solved protein structures (22). Next, we explore the relationship between pocket geometry and global structural similarity. Does high global fold similarity demand that the protein pair always have similar ligand-binding pockets? Conversely, given a pair of structurally similar pockets, how similar are the global folds of the proteins where they reside? The importance of this study is to examine how often structurally similar pockets occur in proteins of completely different global fold and whether this is common to both native and artificial proteins. If so, this has important implications for both ligand-binding promiscuity and the use of global structural similarity to infer evolutionary relationships and/or functional similarity. We next delineate the interrelationship between global structure, pocket geometry, and amino acid sequence. For ART sequences selected to maximize stability in structures with the same backbone structure, we will explore the similarity of their protein pockets and whether they show the same extent of sequence conservation at structurally equivalent positions as a function of pocket similarity. Finally, we highlight the implications of this work.

Results

In what follows, we will compare the global similarity of protein structures and the structural similarity of their pockets. For global comparison, the structure alignment algorithm that uses the template modeling score (TM-score) as the comparison metric is used (23, 24). The TM-score ranges from 0 for unrelated structures to 1.0 for identical structures, with a mean of 0.30 for the best structural alignment of a random protein structure pair. For pocket comparison, the sequence-order-independent APoc algorithm with similarity assessed by the Pocket Similarity score (PS-score) is used (25). The PS-score ranges from 0 for entirely dissimilar pockets to 1.0 for identical pockets. A table of P values for a given PS-score is given in Table S1. For pocket pairs of similar length, a PS-score above 0.38 is significant with a P value < 0.03.

How Many Representative Pockets Are There, and Is Pocket Space Complete? If there were a small number of distinct ligand-binding pockets that shared significant sequence conservation, this would have profound consequences. Most importantly, it would imply that similar ligand–protein interactions should occur across different protein folds and rationalizes the large number of off-target interactions of drugs (26). In Fig. 1, for a representative set of native, single domain proteins between 40 and 250 residues in length (whose pairwise sequence identity $< 35\%$; *Methods*) and corresponding ART proteins, we plot the number of representative distinct pockets, the fraction of target pockets that match these representative pockets (i.e., which have at least one matching pocket), and the fraction of unmatched (singleton) pockets versus PS-score without a template size restriction (see Fig. S1 for results when the template pockets are restricted to be smaller than 80 residues in length). What is striking is that at a PS-score of 0.38, a small number of pockets, 132 (180), in the PDB (ART) library are required to cover 99.0% (91.5%) of the space of PDB pockets. PDB–PDB and PDB–ART pockets behave

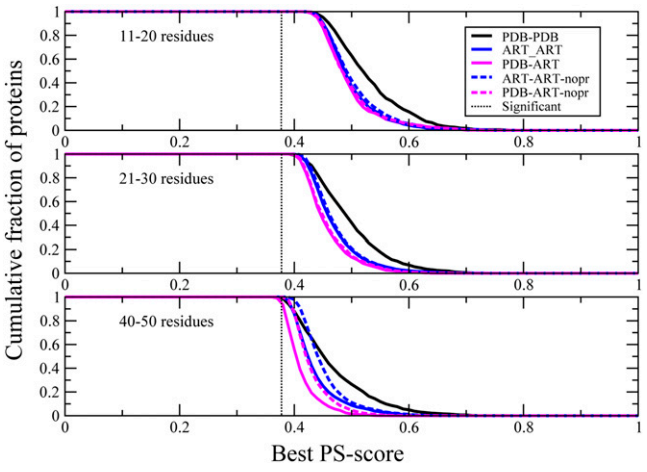

Fig. 2. Best pocket match between protein structures. For different size pockets, cumulative fraction of proteins whose best PS-score to a pocket in the given structural library is greater than or equal to the PS-score on the abscissa. The mean PS-score for a significantly related pair of pockets ($P <$ 0.03) is indicated in the dotted black line.

similarly to ART–ART structures that require 114 pockets to cover 99.5% of all pockets. The reason for this relatively small number is that a few large pockets cover the space of many small pockets. The average number of residues/template-pocket is 106.0 (91.4) for PDB (ART) templates that cover all PDB pockets. At a PS-score of 0.40, as shown in Fig. 1*B*, 92.5% (91.1%) of PDB (ART) pockets match one of the 339 (420) representative PDB (ART) structures, with an average number of residues/template-pocket of 87.8 (80.8). Thus, the number of distinct pocket structures found in either PDB or ART proteins is remarkably small. As the PS-score rises, as in Fig. 1*C*, the number of unmatched pockets increases. Little dependence is seen on which potential is used to generate the ART pocket sequences, suggestive that these conclusions are quite robust.

Further evidence that pocket space is likely complete is provided in Fig. 2, where the distribution of best PS-scores of pairs of pockets is between native proteins, between ART proteins, and when PDB proteins are compared with ART templates. Consistent with Fig. 1, virtually every native ligand-binding pocket has a statistically significant structure match in the ART library and does not require evolution for them to occur; rather, they are a geometric effect arising from defects in the packing of secondary structural elements. The most distinct differences are seen when artificial pockets are compared with PDB pockets, with larger discrepancies seen for larger pockets. As pockets increase in size, structural differences are amplified, as more residues must be aligned to give the same PS-score.

Global Structural Similarity Versus Pocket Similarity. The global similarity of protein structures is often used to infer functional similarity (27, 28). The underlying justifications are the assumptions that a similar global fold is sufficient to guarantee that the pair of proteins have structurally similar pockets and that the coincidence of such pockets (implicitly assumed to be very rare) implies an evolutionary and, therefore by implication, functional relationship. To explore the validity of these assumptions, in Fig. 3, for the largest pocket in the protein of interest, we examine, for a given extent of global structural similarity as assessed by their TM-score, the cumulative fraction of proteins whose best PS-score to a pocket in a protein in the library of interest is greater than or equal to the abscissa. For structurally unrelated proteins (TM-score = 0.18), most pocket structures are dissimilar; yet, even here, ~0.5% of their pockets are structurally similar, a consistent fraction for both native and artificial proteins, independent of how the ART sequences are generated. For globally similar proteins with a TM-score = 0.40, 16% of PDB–PDB proteins have structurally similar pockets, with similar behavior

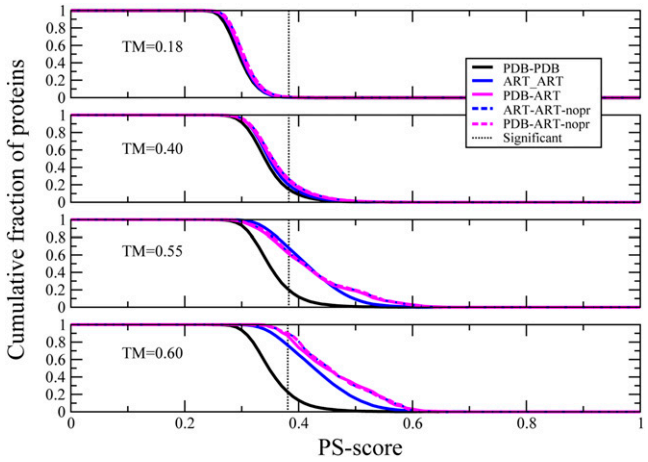

Fig. 3. For fixed TM-score, the cumulative fraction of proteins whose best match PS-score is greater than or equal to abscissa. The PS-score for a significantly related pair of pockets is indicated in the dotted black line.

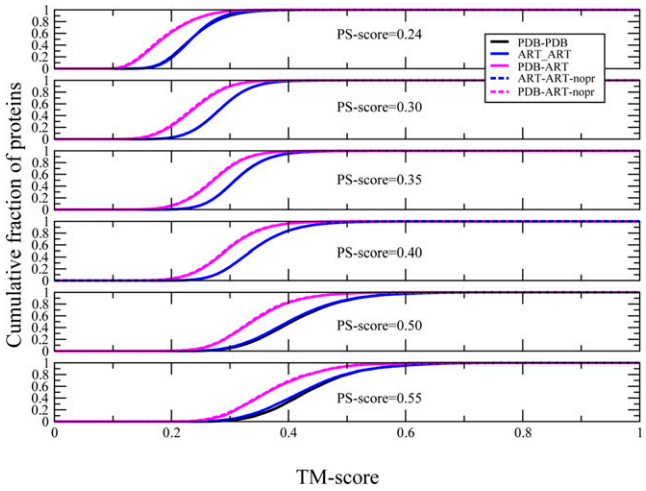

Fig. 4. For a given PS-score, cumulative fraction of proteins with TM-score less than or equal to the abscissa. Solid (dashed) lines are protein sequences generated with burial, secondary structure, and (without) pair potentials.

exhibited by all three types of compared structures (PDB–PDB, PDB–ART, ART–ART). This does not mean that there are few PDB proteins with this level of pocket similarity. From Fig. 2, greater than 90% of proteins have a pocket that matches another protein in the PDB, but most of these matched pockets are in proteins at different levels of global fold similarity or even with dissimilar folds. For structurally very similar proteins having TM-scores of 0.55 and 0.60 (where for native proteins the pair could be evolutionarily related), pairs of PDB–PDB proteins have significant pocket matches in ~22% of the cases, a fraction quite insensitive to TM-score. Since for all TM-score values, there are significant matches to structurally similar pockets in both the PDB–ART and ART–ART libraries, having a significant pocket match does not guarantee that the pair of proteins shares any evolutionary relationship whatsoever. Interestingly, ART–ART and PDB–ART pairs have a larger fraction of significant pocket structural matches for globally similar structures (those with TM-scores >0.4) than PDB–PDB pairs. The implication is that evolution seems to have acted to increase the specificity for particular ligands for a given backbone by making pockets in globally similar structures more dissimilar than would be expected on the basis of their global structure alone. As shown below, the observed pocket variability in similar global folds is readily achieved by appropriate sequence variability. However, even for a TM-score of 0.60, global fold similarity is insufficient to guarantee the presence of structurally similar pockets, as ~77% of pairs of PDB proteins have dissimilar pockets. Thus, global structural similarity alone cannot guarantee that a pair of proteins has a similar ligand-binding pocket.

We next consider the converse and examine the extent of global structural similarity for a given extent of pocket similarity. In Fig. 4, for a given PS-score, we show the fraction of proteins whose TM-score is less than or equal to the abscissa. On comparing PDB–PDB or ART–ART structures, both curves are very similar even up to a PS-score of 0.55. Thus, the fraction of globally similar structures that have structurally similar pockets up to quite significant PS-scores is independent of any evolutionary selection. When PDB structures are compared with ART structures, proportionally more similar pockets are found in more globally dissimilar proteins. There is also the trend that as the pockets become more similar so does the global structural similarity of the protein pair. However, even for a PS-score of 0.5, about 12% of similar pockets are in unrelated structures (TM-score = 0.30). This is consistent with previous observations that similar pockets occur across different folds and different proteins (29, 30), a conclusion now shown to hold more comprehensively and rationalized as to why it happens. Thus, the

major conclusion from Fig. 4 is that pocket similarity and global structural similarity are only weakly correlated. In fact, we find that there is very little overlap between residues aligned on the basis of the global structural alignment and the pocket alignment (Table S2). Pockets often reorient and move about within the same global fold. This reflects a significant interplay between the backbone structure and the specific amino acids that form the pocket. Given that similar pockets can occur in proteins of completely different global structures and dissimilar pockets can occur in proteins with the same global structures, the issue is to establish what aspects of pocket similarity are required to infer functional similarity.

Global Fold Versus Protein Sequence in Determining Pocket Similarity. Given structurally similar pockets, we next explore the interrelationship between global structure and protein sequence in dictating pocket similarity. If one has an open clam-shell-like structure, on average, then global fold should dictate pocket location. However, when the sequence is composed entirely of small residues, then the resulting pocket will be much larger than if bulky amino acids dominate. Thus, binding pocket size/location should depend on the interplay of global structure and protein sequence. By way of illustration for ART target and templates whose backbone structure is exactly the same, in Fig. 5, for two sets of 20 protein stability-selected sequences (with average sequence identity of ~7%), one set selected with and the other without using the pair potential, we plot the resulting PS-score distribution. These results suggest that in native proteins, pocket geometric similarity as a function of sequence could vary dramatically even for proteins with globally similar structures; *viz*. as indicated above, pockets are very plastic, with global fold and pocket geometry weakly coupled. This is exactly what happens in native ligand-binding pockets that have virtually the identical span of PS-scores in globally very similar structures (TM-score ≥ 0.6) as ART protein sequences (see *SI Methods* for additional details). Overall, the space of native pocket shapes that nature explores is remarkably similar to that found in the library of artificial pocket structures. Thus, protein pocket structure is likely strongly driven by selection for fold stability rather than by function.

We also considered in Figs. S2 and S3 the effect of thermal fluctuations on PS-score in a set of 2,801 nonhomologous proteins, whose distorted structures were clustered by their global Cα rmsd from native. For example, in structures with a 1.0 < rmsd < 1.5 Å, greater than 95% of the pockets have a significant match to the native pocket. For 21–30 residue pockets in structures whose 0.0 < rmsd < 0.5 (1.0 < rmsd < 1.5) Å, the mean PS-score is 0.68 ± 0.12 (0.57 ± 0.11) for pockets 21–30 residues in length. For additional details, see Table S3. Thus, the majority of thermal fluctuations has a marginal effect on pocket identification and overlap with the native, with the effect diminishing with increasing pocket size.

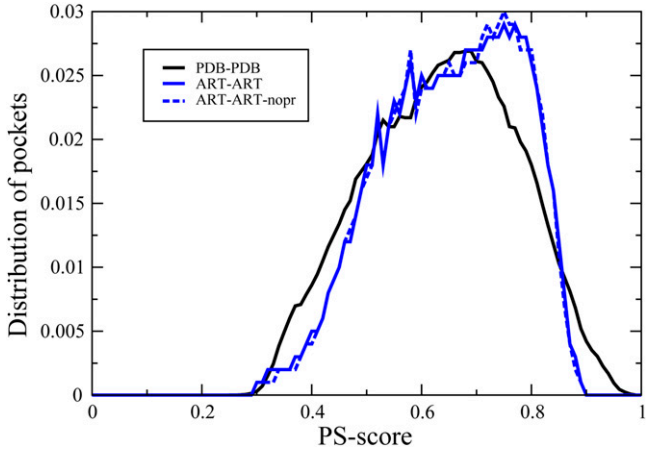

Fig. 5. Relationship of protein sequence and PS-score for a fixed protein backbone in the artificial structure library in blue and in PDB structures in black. Solid (dashed) blue lines are protein sequences generated using burial, secondary structure, and (without) pair potentials.

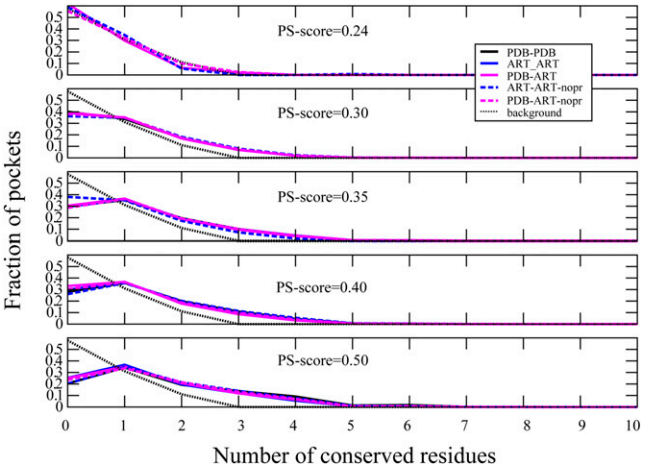

Fig. 6. For a given PS-score and pockets that are 31–40 residues in size, the fraction of proteins with a given number of conserved residues at structurally aligned positions. The black dotted line is the background for randomly related pockets.

How conserved are the residues at a given position in structurally similar pockets? Do pockets have to be related by evolution for such sequence conservation to occur, or can it just result from the selection of sequences that are stable in the fold of interest? Since native protein pockets are the convolution of physical interactions and evolution, to tease out these effects, in Fig. 6, we compare the fraction of proteins that have a given number of pocket residues conserved at structurally aligned positions in the pocket as a function of PS-score for both ART and PDB proteins. Even up to a PS-score of 0.5, the sequence conservation behaviors of PDB–PDB, PDB–ART, and ART–ART sets of pockets are very similar. Once again, as in Fig. 4, similar behavior does not demand that the pair of pockets be evolutionarily related. This is suggestive of the origin of functional promiscuity conjectured by Jensen (6) and Tawfik (8, 9, 13). We find on comparing PDB–PDB and PDB–ART pocket pairs (Table S4, columns 2 and 3) that ALA, VAL, ILE, LEU, and GLU are conserved independent of pocket structural similarity; these are residues that are just likely to be at structurally equivalence pocket positions. ARG (SER) becomes more (less) conserved as the pocket similarity increases. Thus, ART proteins recapitulate the native results for which residues are conserved, suggesting that to a significant extent the degree of sequence conservation in pockets is driven by the selection for thermodynamic stability.

Discussion

Overall, there is a remarkable coincidence of the properties of native protein pockets with those of artificial proteins, whose sequences are selected purely for fold stability and not function. We find that the structural space of pockets is remarkably small and likely complete. Combined with the fact that similar pockets occur across many different types of protein structures with similar patterns of amino acid conservation, we conclude that ligand-binding promiscuity is likely an inherent feature resulting from the geometric and physical–chemical properties of proteins. This promiscuity implies that the notion of one molecule–one protein target that underlies many aspects of drug discovery is likely incorrect, a conclusion consistent with recent studies (26, 30, 31). Moreover, within a cell, a given endogenous ligand likely interacts at low levels with multiple proteins that may have different global structures.

This view is confirmed in Fig. S4, where for nonhomologous proteins in the human and all proteomes, we examine the histogram of the distribution of the number of known interactions a given ligand makes with nonhomologous receptor proteins. The results for 1,284,577 entries extracted from the ChEMBL15 (32) and BindingDB (33) databases are reported. We note that the reported number of ligand–receptor interactions is a lower bound as many such interactions are currently uncharacterized. Even so, in more than 1,400 ligands, each binds to 40 or more nonhomologous proteins. Thus, there is considerable experimental evidence that a given ligand interacts with many proteins in a proteome; *viz.* such interactions are quite promiscuous.

While it is very likely that evolution acted to selectively enhance this low-level function to provide specificity, we suspect that background biochemical noise reflective of a soup of functions is always present. On the one hand, this produces robustness, yet it makes regulation and control more difficult. How nature achieves the collective behavior needed for living cells is not fully understood. Furthermore, even when protein structures are highly similar, there is remarkable plasticity in pocket geometry. Protein pocket shape is just partly coupled to global fold. The coincidence of structurally similar pockets in structurally similar proteins with similar patterns of residue conservation cannot in and of itself be used to infer an evolutionary relationship between proteins. Thus, the implicit and widespread assumption that the co-occurrence of global structure, pocket similarity, and amino acid conservation demands an evolutionary relationship between proteins significantly underestimates the background probability for the random coincidence of such properties. The clear implication is that the fundamental physical–chemical properties of proteins are sufficient to explain many of their structural and molecular functional properties, with evolution acting to fine-tune/amplify their biochemical function, as in the inherent functionality model. This suggests that a collection of stability-selected proteins should have the inherent ability to engage in many of the biochemical reactions needed by living systems without any selection for function; this work has significant implications for the origin of the biochemical processes needed for life.

Methods

Pocket Detection. Two pocket detection methods are used in this study. For comparative studies between the artificial and native protein pocket structures, an in-house pocket detection method CAVITATOR was used (25). CAVITATOR is a geometry-based method similar to LIGSITE (1), but it is designed to be less sensitive to minor structural distortions. Each protein-heavy atom is mapped to a cubic lattice grid with a spacing of 1 Å and occupies the central grid point and all adjacent grid points within $\sqrt{2}$ Å. Thus, a single heavy atom occupies 27 grid points. For an unoccupied point to be part of a pocket, it must be bounded on both sides by occupied points along the X, Y, or Z directions. For ligand/protein complex structures collected from the PDB, the program LPC (34) was applied to extract protein residues that make direct contact with each ligand, yielding observed ligand-bound protein pockets.

Pocket Comparison. For the structural alignment and comparison of protein pockets, the APoc algorithm was developed, with the details of the method described elsewhere (25). Here, we recapitulate the main ideas. Given two input pockets, a template and a target, APoc evaluates their PS-score, which measures the similarity in their backbone geometries, side-chain orientations, and the chemical similarities between the aligned pocket-lining residues. The length of a pocket is the number of Cα atoms of the pocket residues. Suppose an alignment is obtained between a query (target) of length L_Q and a template of length L_T. The PS-score of the alignment is

$$\text{PS - score} = (S + s_0)/(1 + s_0), \quad [1]$$

$$S = \frac{1}{L_Q} \max_{sup}\left[\sum_{i=1}^{N_a} p_i r_i / \left(1 + d_i^2/d_0^2\right)\right], \quad [2]$$

$$P_i = \begin{cases} 1 & \text{if } \theta_i \le \pi/3 \\ \max(0.1, 0.5 + \cos\theta_i) & \text{if } \theta_i > \pi/3 \end{cases}, \quad [3]$$

$$r_i = \max\left(0.8, \delta\left(a_i^Q, a_i^T\right)\right). \quad [4]$$

Here, N_a is the number of aligned residue pairs, d_i is the distance in Å between the Cα atoms of the ith aligned residue pair, and the empirical scaling factor $d_0 \equiv 0.70(L_Q - 5)^{1/4} - 0.2$. The constants in d_0 were obtained by fitting the distribution of Cα distances in random alignments of pockets. p_i measures the directional similarity between two Cα to Cβ vectors in the two pockets, which span an angle θ_i at the ith alignment position of two non-Glycine residues. For Glycine, the value of p_i is assigned 1 if both amino acids are Glycine and 0.77 if only one residue is Glycine. The latter is the mean p_i derived from random alignments. r_i measures the chemical similarity of the two aligned amino acids. $\delta(a_i^Q, a_i^T)$ has a value of 1 if the two amino acids a_i^Q, a_i^T belong to the same group (I–VIII) defined as: I (LVIMC), II (AG), III (ST), IV (P), V (FYW), VI (EDNQ), VII (KR), VIII (H) (29), and 0 otherwise. The scaling factor $s_0 = 0.23 - 12/L_Q^{1.88}$ ensures that the mean score of two aligned random pockets is independent of their length. To calculate the distances used in d_i and p_i, aligned residues are superimposed using the Kabsch algorithm (35) to minimize the rmsd of the full or subset of aligned residues. In principle, the number of all possible superpositions exponentially increases as the alignment length grows. The notation "max" in Eq. **2** indicates that the PS-score corresponds to the superposition that gives the maximum of all scores. In practice, a heuristic iterative extension algorithm is used to calculate the PS-score, similar to that used for calculating the TM-score (36). Note that identical pocket structures have a PS-score of 1.0, which is the upper bound of the PS-score.

To obtain the "optimal" alignment, APoc consists of three phases: In the first phase, several guessed solutions are generated from gapless alignments, secondary structure comparisons, fragment alignments, and local contact pattern alignments. Starting from these guessed "seed" alignments, dynamic programming is iteratively applied in the second phase. This yields optimal sequential (*viz.* protein sequence order–dependent) alignments between two pocket structures. In the third phase, an iterative procedure searches for the best nonsequential alignment between two pockets, which is then selected if this alignment has a better PS-score than the optimal sequential alignment. The problem of finding an optimal nonsequential alignment (or match) is converted to the Linear Sum Assignment Problem (LSAP), which is a special case of integer programming and is also equivalent to the problem of finding a maximum weight matching in a weighted bipartite graph. To efficiently solve LSAP, we implemented the shortest augmenting path algorithm (37), which has a polynomial time complexity of $O(N^3)$, where $n = \max(L_T, L_Q)$.

Statistical Significance. The statistical significance of the PS-score is estimated by comparing millions of randomly selected pocket pairs (25). The distribution of PS-scores are fitted to the type I extreme value distribution (Gumbel distribution):

$$f(z) = \exp[-z - \exp(-z)], \quad [5]$$

where z denotes the Z-score given by $z = (s - \mu)/\sigma$. The variable s denotes the PS-score, μ is the location parameter, and σ is the scale parameter. The corresponding P value of the score can be calculated according to the formula:

$$P = 1 - \exp[-\exp(-z)]. \quad [6]$$

The location and scale parameters can be estimated through linear regression fits of

$$\begin{aligned} \mu &= a + b\ln(L_Q) + c\ln(L_T) \\ \sigma &= d + e\ln(L_Q) + f\ln(L_T), \end{aligned} \quad [7]$$

to obtain the parameters a to f through maximum likelihood estimates using the Extreme Value Distribution (EVD) package in the statistical platform R (www.r-project.org). The values of a to f are 0.3117, 0.0277, –0.029, 0.0366, 0.0025, and –0.0084, respectively. The P value of PS-scores for different size cavities is given in Table S1.

PDB Dataset. This set, composed of 5,371 nonredundant monomeric, single domain protein structures from 40 to 250 residues in length, serves as the reference set of PDB structures and is also the primary random background for statistical significance analysis. All proteins in this set have <35% global pairwise sequence identity. We applied CAVITATOR to each structure and restricted our analysis to the largest pocket (the target) that contains at least 10 and no more than 60 residues and has a volume >100 Å^3 (100 grid points). PDB template pockets have no size restriction.

Artificial Structures. For a randomly chosen subset of PDB comprised of 1,259 proteins, we extract the corresponding secondary structure preferences and, following the previously described procedure (38), performed TASSER structure predictions of the tertiary structure of the corresponding polyleucine homopolypeptide. More precisely, this is a homopolypeptide whose side chain excluded volume envelope matches that of leucine but whose secondary structure preferences were taken from the corresponding native template structures. Leucine is selected because it gives the same average pocket volume as the pockets in native sequences for the same backbone configuration. We note that the average TM-score to the native template is 0.33, which is very close to the average value of the best random structural alignment of 0.30. Despite the fact that these proteins have similar secondary structural preferences and native structures, in general, they are globally unrelated to the native structure from which these preferences are excised. For each homopolypeptide, the top centroid of the most populated structural cluster is selected.

To estimate the stability of a given sequence in the given ART structure, we use a knowledge-based potential having a centrosymmetric amino acid burial term (39) and secondary structure preferences generated using a neural network and a statistical pair potential (40). Sequences were also generated without the pair potential; these results are indicated by no-pr. For a given polyleucine structure, a randomized sequence is generated based on the average amino acid composition in the PDB. The sequence is shuffled to give a low-energy structure for the ART structural template of interest (3). Then, the corresponding all-atom conformation is built from the Cα trace by Pulchra (41). For each polyleucine template, 20 randomly related sequences are generated (mean sequence identity of ∼7%). This procedure generates a total of 25,180 structures and constitutes the ART template library. CAVITATOR was applied to identify all pockets in the resulting structures. Using PDB structures as the target, ART template pockets that are more than 10 residues in length and have more than 100 grid points in volume are considered. Then, the best of top 10 pockets with the highest PS-score to the corresponding native pocket are selected. When comparing ART target pockets to themselves, we consider all ART target pockets whose volume is more than 100 grid points and whose lengths are more than 10 and fewer than or equal to 60 residues. This gives a total of 11,094 ART target pockets and 96,090 ART template pockets.

Representative Pockets. We seek to find the smallest set of representative pockets that are sufficient to cover the full set of pockets at a desired level of similarity. In terms of graph theory, pocket similarity relationships can be viewed as a directed graph *G*, wherein each node represents a pocket, and an edge from pocket A to the pocket B indicates that A as a pocket structural template provides significant similarity to target pocket B above a specified PS-score threshold. Thus, the size *N* of the sought-after representative set is the domination number for *G*, which is defined as the cardinality of the smallest dominating set of the graph (42). An approximation to the domination number of the set of protein pockets is constructed as follows (42): For a given template library, the pocket with the largest number of matching targets at the specified PS-score is selected. Then, the pocket with the next highest number of matching targets (after all matching targets to the first pocket are removed) is calculated. The process is iterated until all target pockets that can be matched to a template pocket at the specified score threshold are selected. The resulting number of distinct pockets is reported. The fraction of matching pockets is the ratio of the number of pockets assigned to the dominating set divided by the total number of pockets.

ACKNOWLEDGMENTS. We thank Dr. Michal Brylinski for preparing the ensemble of unfolded structures and Ambrish Roy for performing the ligand–protein interaction analysis. This research was supported in part by Grant GM-48835 from the Institute of General Medical Sciences, National Institutes of Health.

1. Huang BD, Schroeder M (2006) LIGSITE^{csc}: Predicting ligand binding sites using the Connolly surface and degree of conservation. *BMC Struct Biol* 6:19.
2. Alberts B (2008) *Molecular Biology of the Cell* (Garland Science, New York), 5th Ed.
3. Brylinski M, Gao M, Skolnick J (2011) Why not consider a spherical protein? Implications of backbone hydrogen bonding for protein structure and function. *Phys Chem Chem Phys* 13(38):17044–17055.
4. Gao M, Skolnick J (2012) The distribution of ligand-binding pockets around protein-protein interfaces suggests a general mechanism for pocket formation. *Proc Natl Acad Sci USA* 109(10):3784–3789.
5. Patel SC, Bradley LH, Jinadasa SP, Hecht MH (2009) Cofactor binding and enzymatic activity in an unevolved superfamily of de novo designed 4-helix bundle proteins. *Protein Sci* 18(7):1388–1400.
6. Jensen RA (1976) Enzyme recruitment in evolution of new function. *Annu Rev Microbiol* 30:409–425.
7. Khersonsky O, Malitsky S, Rogachev I, Tawfik DS (2011) Role of chemistry versus substrate binding in recruiting promiscuous enzyme functions. *Biochemistry* 50(13):2683–2690.
8. Tawfik DS (2010) Messy biology and the origins of evolutionary innovations. *Nat Chem Biol* 6(10):692–696.
9. Khersonsky O, Tawfik DS (2010) Enzyme promiscuity: A mechanistic and evolutionary perspective. *Annu Rev Biochem* 79:471–505.
10. Khersonsky O, Roodveldt C, Tawfik DS (2006) Enzyme promiscuity: Evolutionary and mechanistic aspects. *Curr Opin Chem Biol* 10(5):498–508.
11. Khersonsky O, et al. (2012) Bridging the gaps in design methodologies by evolutionary optimization of the stability and proficiency of designed Kemp eliminase KE59. *Proc Natl Acad Sci USA* 109(26):10358–10363.
12. Ben-David M, et al. (2012) Catalytic versatility and backups in enzyme active sites: The case of serum paraoxonase 1. *J Mol Biol* 418(3-4):181–196.
13. Bar-Even A, et al. (2011) The moderately efficient enzyme: Evolutionary and physicochemical trends shaping enzyme parameters. *Biochemistry* 50(21):4402–4410.
14. Jürgens C, et al. (2000) Directed evolution of a $(\beta\ \alpha)_8$-barrel enzyme to catalyze related reactions in two different metabolic pathways. *Proc Natl Acad Sci USA* 97(18):9925–9930.
15. Song G, et al. (2006) Rational design of intercellular adhesion molecule-1 (ICAM-1) variants for antagonizing integrin lymphocyte function-associated antigen-1-dependent adhesion. *J Biol Chem* 281(8):5042–5049.
16. Pande J, Szewczyk MM, Grover AK (2010) Phage display: Concept, innovations, applications and future. *Biotechnol Adv* 28(6):849–858.
17. Khersonsky O, et al. (2011) Optimization of the in-silico-designed kemp eliminase KE70 by computational design and directed evolution. *J Mol Biol* 407(3):391–412.
18. Khare SD, et al. (2012) Computational redesign of a mononuclear zinc metalloenzyme for organophosphate hydrolysis. *Nat Chem Biol* 8(3):294–300.
19. Skolnick J, Arakaki AK, Lee SY, Brylinski M (2009) The continuity of protein structure space is an intrinsic property of proteins. *Proc Natl Acad Sci USA* 106(37):15690–15695.
20. Skolnick J, Zhou HY, Brylinski M (2012) Further evidence for the likely completeness of the library of solved single domain protein structures. *J Phys Chem B* 116(23): 6654–6664.
21. Hildebrand A, Remmert M, Biegert A, Söding J (2009) Fast and accurate automatic structure prediction with HHpred. *Proteins* 77(Suppl 9):128–132.
22. Berman H, Henrick K, Nakamura H, Markley JL (2007) The worldwide Protein Data Bank (wwPDB): Ensuring a single, uniform archive of PDB data. *Nucleic Acids Res* 35(Database issue):D301–D303.
23. Pandit SB, Skolnick J (2008) Fr-TM-align: A new protein structural alignment method based on fragment alignments and the TM-score. *BMC Bioinformatics* 9:531.
24. Xu J, Zhang Y (2010) How significant is a protein structure similarity with TM-score = 0.5? *Bioinformatics* 26(7):889–895.
25. Gao M, Skolnick J (2013) APoc: Large-scale identification of similar protein pockets. *Bioinformatics* 29(5):597–604.
26. von Eichborn J, et al. (2011) PROMISCUOUS: A database for network-based drug-repositioning. *Nucleic Acids Res* 39(Database issue):D1060–D1066.
27. Petrey D, Fischer M, Honig B (2009) Structural relationships among proteins with different global topologies and their implications for function annotation strategies. *Proc Natl Acad Sci USA* 106(41):17377–17382.
28. Brylinski M, Skolnick J (2010) Comparison of structure-based and threading-based approaches to protein functional annotation. *Proteins* 78(1):118–134.
29. Zhang ZD, Grigorov MG (2006) Similarity networks of protein binding sites. *Proteins* 62(2):470–478.
30. Sturm N, Desaphy J, Quinn RJ, Rognan D, Kellenberger E (2012) Structural insights into the molecular basis of the ligand promiscuity. *J Chem Inf Model* 52(9):2410–2421.
31. Lim E, et al. (2010) T3DB: A comprehensively annotated database of common toxins and their targets. *Nucleic Acids Res* 38(Database issue):D781–D786.
32. Gaulton A, et al. (2012) ChEMBL: A large-scale bioactivity database for drug discovery. *Nucleic Acids Res* 40(Database issue):D1100–D1107.
33. Liu TQ, Lin YM, Wen X, Jorissen RN, Gilson MK (2007) BindingDB: A web-accessible database of experimentally determined protein-ligand binding affinities. *Nucleic Acids Res* 35(Database issue):D198–D201.
34. Sobolev V, Sorokine A, Prilusky J, Abola EE, Edelman M (1999) Automated analysis of interatomic contacts in proteins. *Bioinformatics* 15(4):327–332.
35. Kabsch W (1976) Solution for best rotation to relate two sets of vectors. *Acta Crystallogr Sect A* 32(SEP1):922–923.
36. Zhang Y, Skolnick J (2004) Scoring function for automated assessment of protein structure template quality. *Proteins* 57(4):702–710.
37. Derigs U (1985) The shortest augumenting path method for solving assignment problems—Motivation and computational experience. *Algorithms and Software for Optimization*, ed Monma CL (Baltzer, Basel), Vol 4, pp 57–102.
38. Zhou H, Skolnick J (2009) Protein structure prediction by pro-Sp3-TASSER. *Biophys J* 96(6):2119–2127.
39. Zhang Y, Skolnick J (2004) Automated structure prediction of weakly homologous proteins on a genomic scale. *Proc Natl Acad Sci USA* 101(20):7594–7599.
40. Skolnick J, Kolinski A, Ortiz A (2000) Derivation of protein-specific pair potentials based on weak sequence fragment similarity. *Proteins* 38(1):3–16.
41. Rotkiewicz P, Skolnick J (2008) Fast procedure for reconstruction of full-atom protein models from reduced representations. *J Comput Chem* 29(9):1460–1465.
42. Fomin F, Grandoni G, Pyatkin A, Stepanov A (2008) Combinatorial bounds via measure and conquer: Bounding minimal dominating sets and applications. *ACM Trans Algorithms* 5(1):9.