# The Relationship between the Sequence Identities of Alpha Helical Proteins in the PDB and the Molecular Similarities of Their Ligands

John B. O. Mitchell†

Unilever Centre for Molecular Informatics, Department of Chemistry, University of Cambridge, Lensfield Road, Cambridge CB2 1EW, U.K.

Received February 17, 2001

This paper considers the relationship between the percentage sequence identities of protein chains and the molecular similarities of the ligands they bind. Among a set of alpha helical proteins from the PDB, it is found that related proteins tend to bind similar ligands. Furthermore, the property of binding similar ligands can be used to define the categories of "like" and "unlike" pairs of protein chains, separated by an approximate cutoff at a sequence identity of, or somewhat above, 45%. Similarly, the property of binding related protein chains can be used to define "low" and "high" similarity pairs of ligand residues, with a cutoff at a Tanimoto score of 0.70. The ligands bound to two "like" protein chains are five times more likely to be of high similarity than would be expected if protein sequence identity and ligand molecular similarity were independent variables. Nonetheless, the nature of the PDB means that it is unclear whether the same conclusions would be reached with a data set representing an unbiased sample of all protein−ligand complexes in a living cell. The construction of an appropriate data set for such a study represents a significant challenge.

## INTRODUCTION

The binding of ligand molecules to proteins is often integral to protein functions such as transport, storage, signaling, regulation, energy transfer, and especially enzyme catalysis. The Protein Data Bank (PDB)[1] now contains a large number of structures of protein−ligand complexes; according to the PDBsum[2,3] database, more than 7700 PDB entries contain some kind of ligand molecule. This work addresses what is called here the similarity question: "To what extent do similar proteins bind similar ligands?". It is often tacitly assumed that related proteins will tend to bind similar ligands, but the present work seeks to test this assumption in a novel and quantifiable way, for a set of alpha helical proteins in the PDB databank. The possibilities of extending the study from this subset, first, to the full structural diversity represented by the PDB and, second, to a data set more representative of protein−ligand interactions in nature will be discussed. This work is intended to be a pathfinder, encouraging groups worldwide to develop effective ways of posing and methods of addressing the similarity question in more diverse and more representative data sets.

Often, the bound ligand is a molecule that the protein has evolved to bind as part of its function. Thus, gene duplication followed by divergent evolution is a likely route to the appearance of structurally and functionally similar proteins that bind related, but nonidentical, ligands. In enzymes, one might expect that the general features of the active site and the essential chemistry of the reaction would be conserved but that the details of the binding and the structures of the substrate and products would be somewhat changed.[4] Previous studies have shown a clear preference for a particular kind of ligand to be bound by one structural class of protein domain (e.g., heme by mainly α domains and nucleotides by $\alpha\beta$ domains), though there is no strong relationship between structural class and the first digit of an enzyme's E.C. number.[5,6] Such correlations do seem to appear when sequence data is added to the analysis.[7] In any case, it is reasonable to hypothesize that evolution may produce some relationships between protein sequence identity and ligand molecular similarity.

Another cause of such a relationship might be protein−ligand binding that is only moderately specific. Thus, a given protein in the PDB will often be capable of binding, for instance, a variety of peptides or several different polysaccharides. Different crystal structures may therefore capture the molecule with different, but related, ligands.

Furthermore, experimentalists are often interested in structures where the complexed molecule is a modified version or analogue, either natural or synthetic, of the biological ligand. Such an analogue might be an inhibitor that mimics an enzyme's substrate and can also fit into the binding site. Similarly, there are solved structures where a given protein is complexed with different members of a series of related pharmaceutical candidates. Thus one expects to find separate PDB structures of identical (or near-identical) proteins binding their natural ligands and very similar related molecules. Much of this is likely to be due to human design and would not be replicated in a data set truly representative of protein−ligand interactions in living organisms. Clearly the PDB is not such a data set.

All the above (divergent evolution, binding specificity and human design) may be considered reasons to expect that, broadly speaking, similar proteins in the PDB will be seen to bind similar ligands.

There are also instances, however, where the ligand molecules complexed to similar (or identical) proteins are

† Corresponding author phone: (U.K.) 01223-762983; phone: (international) +44-1223-762983; fax: (U.K.) 01223-763076; fax: (international) +44-1223-763076; e-mail: jbom1@cam.ac.uk.

10.1021/ci010364q CCC: $20.00 
Published on Web 11/26/2001

not expected to be similar. Convergent evolution, where a common fold is "reinvented" to perform a related function, would cause similar ligands to be bound by proteins that, despite structural similarities, had low sequence identities and hence would not be expected produce similarities in a study such as this. Also, a single enzyme active site may bind quite different molecules, such as a cofactor and a substrate. In other cases, molecules are included in the experimental structure simply through accidents of the crystallization conditions.

## METHODS

There are many different ways of measuring molecular similarity.[8] For the present work, a specific implementation was developed, designed to make the best use of the information provided in the PDB's CIF format file[9] of all ligands in the databank. In association with this work, software has been developed to convert the CIF format representation of each ligand residue into a 199 bit binary string. Of the 199 bits, 109 describe the presence or absence of chemical elements or members of groups of elements. A further 50 bits are devoted to the covalent bonds and bond orders. The remaining 40 bits describe the different bonded environments in terms of SATIS codes[10] and the related BLEEP atom types.[11] There are 26 bits in total dedicated to recording multiple occurrences of atom and bond types. This allows molecular size to be taken proper account of and thus deals with one of the major difficulties identified by Flower[12] in the use of bit strings. Further details of the bit strings used here are given on a web page[13] and are also available as Supporting Information.

The data set used in this study was restricted to alpha helical proteins, listed as "mainly alpha" by CATH[14] and having any of the following architectures: bundle (CATH code begins 1.20), horseshoe (1.25), alpha solenoid (1.40), and alpha/alpha barrel (1.50). Thus, among "mainly alpha" proteins, only the "nonbundle" (1.10) architecture was (somewhat arbitrarily) excluded. This limited data set was designed as a relatively tractable test case for developing methods of addressing the similarity question. It was expected to provide a reasonably large proportion of proteins structurally and evolutionarily related to one another, giving an adequate number of related as well as unrelated protein pairs when the data are considered pairwise. More diverse data sets will prove more challenging in this respect. Any complex with ligand bound to a protein domain of one of the specified architectures was a candidate for inclusion; 194 protein−ligand complexes in the PDB had at least one eligible domain.

For consistency with usage in the PDB, some ligands are defined in terms of residues. This particularly affects peptides and polysaccharides. Most ligands, however, are represented as monomers, and the residue is then equivalent to the entire molecule. It is not appropriate to include multiple representatives of identical ligand residues bound to closely related domains in the same homologous superfamily (sharing the first four numbers of the CATH code), as this would lead to an observation of similar proteins binding similar ligands that was clearly an artifact of the data set. Thus, only a single representative was retained of identical ligand residues (as defined by the three letter identifier in the PDB) bound to domains in the same homologous superfamily. Inspection of the PDBsum[2,3] database entries also showed that some ligands bind to other domains in the structure but not to the relevant alpha helical ones, and these were not included in the final data set. A few complexes were also excluded due to the ligand being absent from the CIF format dictionary file.

The resulting data set consisted of 140 {protein domain − ligand residue} interactions, hereafter considered as {protein chain − ligand residue} interactions, where the chain is that containing the given domain. These are listed in full on a web page[15] and are also available as Supporting Information. The 140 {protein domain − ligand residue} interactions were then analyzed pairwise, giving a data set consisting of 9730 combinations of the kind {protein chain 1 − ligand residue 1, protein chain 2 − ligand residue 2}. These combinations compared each of the 140 {protein domain − ligand residue} interactions with every other different {protein domain − ligand residue} interaction. Combinations of an interaction with itself were excluded. Subsequent analysis focuses on the set of 9730 combinations.

Each of these combinations was assigned two scores. The first of these, assigned on the basis of the {protein chain 1, protein chain 2} pairing, was the percentage sequence identity between the protein chains concerned, as calculated by the Web-based software suite SAS.[16,17] A correction factor was introduced where sequence identities over 20% were calculated for alignments of fewer than 25 residues, to prevent large values of sequence identity being assigned where only a small fraction of the residues in the chains were aligned. This was

$$S = 20 + 2(S' - 20)N_a/(L_1 + L_2)$$

where $S$ is the corrected percentage sequence identity, $S'$ is the uncorrected percentage sequence identity, $N_a$ is the number of aligned residues, and $L_1$ and $L_2$ are the lengths of the two protein chains.

The second score, assigned to each combination on the basis of the {ligand residue 1, ligand residue 2} pairing, was the Tanimoto molecular similarity calculated between the bit strings generated by the two ligand residues. The Tanimoto score used here is defined as

$$T = N_c/(N_a + N_b - N_c)$$

where $N_c$ is the number of bits set to 1 which are common to the bit strings of ligand residues A and B, while $N_a$ and $N_b$ are the total numbers of bits set to 1 for ligand residues A and B, respectively.

The distribution of Tanimoto scores for the 5460 possible unique pairs of different ligand residues, taken from the 105 ligand residues in the data set, is shown in Figure 1. This gives a guide to the interpretation of the scores. Of the pairs in this biased sample of chemical space, 5.2% have Tanimoto scores above 0.70. The Tanimoto scores for some representative pairs of ligand residues are given in Table 1, and inspection of these gives some insight into the level of chemical similarity associated with a given Tanimoto score. A high Tanimoto coefficient is often taken as an indication that two molecules have similar biological activity. The value of 0.70 is similar to that for which Xue et al.[18,19] obtained

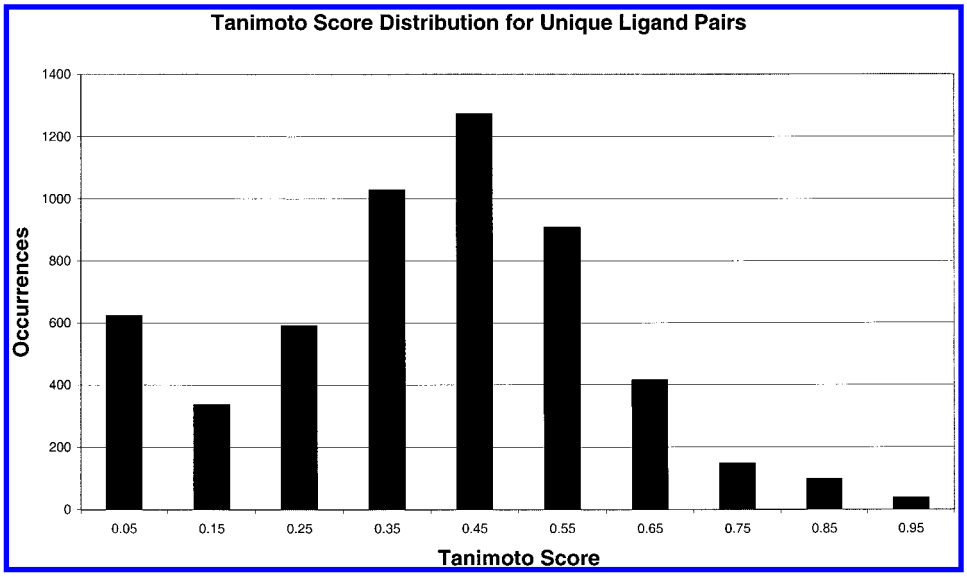


**Figure 1.** Distribution of Tanimoto molecular similarity scores for the 5460 unique pairs taken from the 105 different ligand residues in the data set. These data correspond to all possible unique pairs of ligand residues and not to {protein chain 1 − ligand residue 1, protein chain 2 − ligand residue 2} combinations. The mean Tanimoto score is 0.39 and the median is 0.41.

**Table 1:** Some Examples of Tanimoto Scores for Ligand Residue Pairs[a]

| | ARG | GLY | LEU | VAL | GOL | MAN | ACR | GAC | BOG | SO4 | HEM |
|---|---|---|---|---|---|---|---|---|---|---|---|
| ARG | **1.00** | 0.69 | **0.80** | **0.79** | 0.38 | 0.39 | 0.49 | 0.55 | 0.45 | 0.06 | 0.49 |
| GLY | 0.69 | **1.00** | **0.80** | **0.87** | 0.46 | 0.41 | 0.45 | 0.47 | 0.35 | 0.08 | 0.39 |
| LEU | 0.80 | **0.80** | **1.00** | **0.92** | 0.43 | 0.48 | 0.54 | 0.61 | 0.50 | 0.07 | 0.50 |
| VAL | **0.79** | **0.87** | **0.92** | **1.00** | 0.46 | 0.47 | 0.53 | 0.56 | 0.44 | 0.07 | 0.46 |
| GOL | 0.38 | 0.46 | 0.43 | 0.46 | **1.00** | 0.64 | 0.39 | 0.41 | 0.52 | 0.10 | 0.27 |
| MAN | 0.39 | 0.41 | 0.48 | 0.47 | 0.64 | **1.00** | 0.60 | 0.64 | **0.74** | 0.07 | 0.32 |
| ACR | 0.49 | 0.45 | 0.54 | 0.53 | 0.39 | 0.60 | **1.00** | **0.89** | 0.56 | 0.05 | 0.52 |
| GAC | 0.55 | 0.47 | 0.61 | 0.56 | 0.41 | 0.64 | **0.89** | **1.00** | 0.64 | 0.05 | 0.48 |
| BOG | 0.45 | 0.35 | 0.50 | 0.44 | 0.52 | **0.74** | 0.56 | 0.64 | **1.00** | 0.07 | 0.46 |
| SO4 | 0.06 | 0.08 | 0.07 | 0.07 | 0.10 | 0.07 | 0.05 | 0.05 | 0.07 | **1.00** | 0.04 |
| HEM | 0.49 | 0.39 | 0.50 | 0.46 | 0.27 | 0.32 | 0.52 | 0.48 | 0.46 | 0.04 | **1.00** |

[a] Key: arginine (ARG), glycine (GLY), leucine (LEU), valine (VAL), glycerol (GOL), mannose (MAN), acarbose (ACR), dihydroacarbose (GAC), $\beta$-octylglucoside (BOG), sulfate (SO4), heme (HEM). High similarity pairs are shown in bold.

their best discrimination of such molecules, though somewhat lower than the threshold of 0.85 suggested by the work of Matter[20] and of Patterson et al.[21]

If protein−ligand complex A is similar, in terms of its two scores, to both complex B and complex C, it is very likely that complexes B and C will also be similar to one another. Thus a cluster of $N$ similar complexes will give rise to $N(N-1)/2$ pairwise observations of similarity. This o($N^2$) effect serves to amplify the signal of similar proteins binding similar ligands.

## RESULTS

A plot of protein−protein sequence identity against the ligand−ligand Tanimoto similarity score, using the full data set, is shown in Figure 2. Chi-squared analysis based on a division of this plot into 35 regions shows that the two variables are clearly not independent ($\chi^2 = 472$ with 24 degrees of freedom; $p = 1 \times 10^{-84}$). Nonetheless, the overall correlation between the two variables is very modest indeed ($r = 0.146$).

The data are analyzed by plotting the average Tanimoto ligand similarity scores for all combinations where the protein sequence identity is above and below each given value (Figure 3a). Thus the average Tanimoto score for sequence identities above 30.0% is 0.38 and that for sequence identities above 50.0% is 0.54. It is clear that one can use these data to draw a distinction between "like" pairs of protein chains (those with high percentage sequence identities) and "unlike" pairs. Combinations with like protein chains tend to have ligands with higher Tanimoto similarities than do combinations with unlike protein chains. Only a central range of sequence identities is shown in Figure 3a, because outside it one or other of the averages has substantial statistical uncertainty. One can use the property of binding a higher proportion of similar ligands to define the minimum sequence identity corresponding to like protein chains. From Figure 3a, it is clear that a cutoff at a sequence identity of around 45% gives rise to a sensible distinction between the like and unlike pairs of protein chains. The relative paucity of combinations with sequence identities between 45 and 99%, however, means that this cutoff is very much an approximate value, and a substantially higher value would give almost identical results for this data set.

Similarly, one can also plot the average protein sequence identities for all combinations where the Tanimoto ligand similarity score is above and below each given value (Figure 3b). This figure shows that pairs of ligands with high molecular similarity tend to bind to proteins of higher

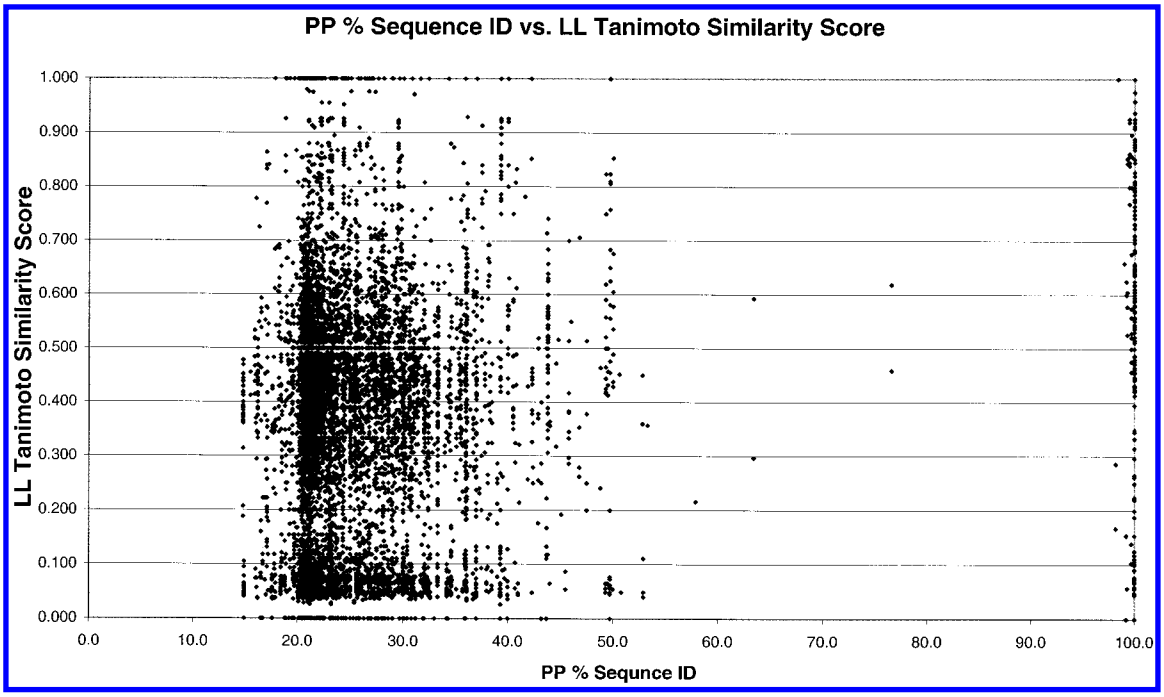


**Figure 2.** Protein−protein percentage sequence identity plotted against ligand−ligand Tanimoto molecular similarity score for the 9730 combinations of {protein chain 1 − ligand residue 1, protein chain 2 − ligand residue 2}.

sequence identity than do ligands of low similarity. One can use the property of tending to bind to protein chains of greater sequence identity to define a range of Tanimoto scores corresponding to high similarity ligand residues. A cutoff at a Tanimoto score of 0.70 gives rise to a clear distinction between the properties of "low" and "high" similarity ligand residue pairs. There are sufficient combinations with Tanimoto ligand similarity scores between 0.70 and 0.99 to allow a much more confident assignment of this cutoff than is the case for protein sequence identities. This cutoff admits 5.6% of the 9730 combinations (and 5.2% of the 5460 unique pairs of different ligand residues, see Figure 1) as having high similarity ligands.

Table 2 shows the distribution of low and high similarity ligands binding to unlike and like protein chains across the 5460 {protein chain 1 − ligand residue 1, protein chain 2 − ligand residue 2} combinations. Only 4.8% of combinations with unlike protein chains have high similarity ligands, compared with 27.4% of combinations with like protein chains. Thus like protein chains are substantially more likely than unlike ones to bind ligands of high similarity. Only 2.6% of combinations with low similarity ligand residues have like protein chains, compared with 16.6% of combinations with high similarity ligand residues. Thus high similarity ligands are substantially more likely than low similarity ligands to bind like protein chains. An additional chi-squared test using the newly defined categories confirms the nonindependence of the variables ($\chi^2 = 309$ with 1 degree of freedom; $p = 4 \times 10^{-69}$).

The enrichment factor is calculated for each given combination of sequence identity and ligand similarity categories. It is defined here as the ratio of the actual number of observations to the number that would be expected if protein sequence identity and ligand molecular similarity were independent variables. For like protein chains in combination with high similarity ligand residues, this factor is 4.93.

As a validation of the methods used, a control experiment was performed by randomly shuffling the list of protein−ligand interactions and repeating the analysis. Chi squared analyses for this randomized data set, in which "interactions" were dictated by chance rather than by real binding, indicated no grounds to reject the hypothesis that protein sequence identity and ligand molecular similarity were now independent variables. The resulting enrichment factor for like protein chains in combination with high similarity ligand residues was not significantly different from 1.00. Versions of parts a and b of Figure 3 using the randomized data give lines that are essentially flat, except for the expected statistical fluctuations among the very small number of data at extremely high values of ligand-ligand molecular similarity. The overall correlation coefficient is close to zero (−0.0172).

## DISCUSSION

In distinguishing unlike from like protein chains on the basis of their tendency to bind similar ligands, the suggested cutoff is at, or somewhat above, 45% sequence identity. This is sufficiently large a value that the proteins are highly likely to be closely related in structural and evolutionary terms.[22,23]

Carrying out a similar analysis of ligand residues, to distinguish high from low similarity pairs on the basis of their preferentially binding like protein chains, a reasonable cutoff is at a Tanimoto score of 0.70. Such a value indicates a high degree of chemical relatedness. This is consistent with molecular similarity being essentially local in nature; a molecule may behave similarly to its near neighbors in chemical space, but it is hard to make any meaningful predictions on the basis of more distant relationships.

There are a number of possible reasons why related proteins in the PDB might be found to bind similar ligands. One is the occurrence of divergent evolution. Two proteins with a relatively recent common ancestor are likely to have

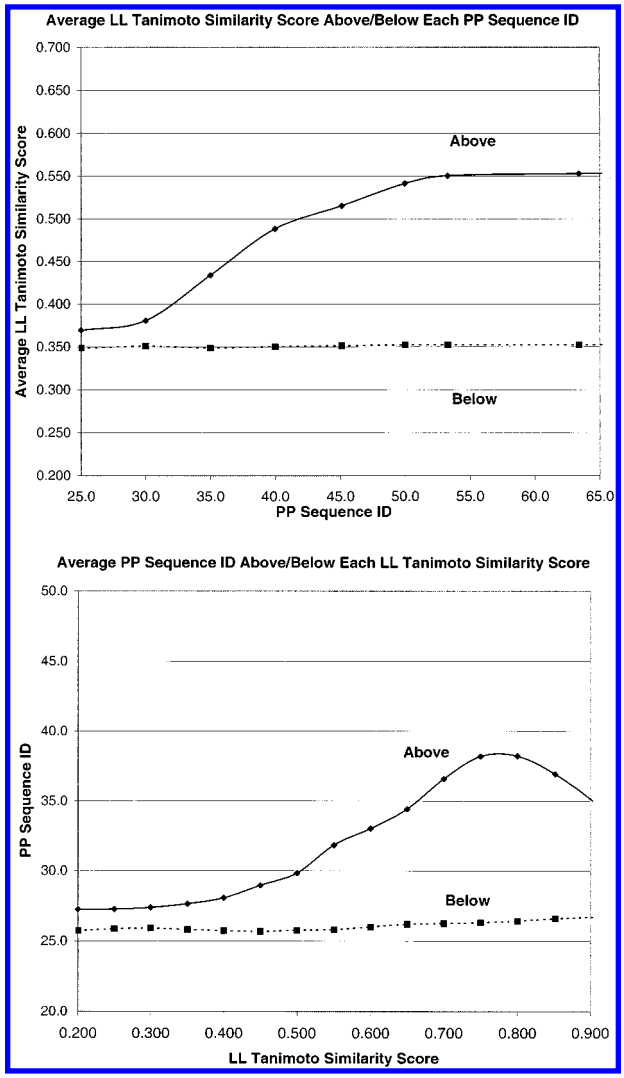


**Figure 3.** (a) Average ligand−ligand Tanimoto molecular similarity score for all combinations with protein−protein percentage sequence identity above (solid line) and below (broken line) each given value. This plot is used to define the approximate cutoff between unlike and like protein chains at a sequence identity of 45%. (b) Average protein−protein percentage sequence identity for all combinations with ligand−ligand Tanimoto molecular similarity score above (solid line) and below (broken line) each given value. This plot is used to define the cutoff between low and high similarity ligand residues at a Tanimoto score of 0.70.

large sequence identity values and only small changes in ligand specificity. The comparison of these two complexes will give a combination with like protein chains and high similarity ligand residues.

A second reason is that some proteins have only moderately specific binding properties and are thus capable of binding a variety of related ligands. This will also lead to identical or near-identical protein chains binding ligands of high similarity.

A third possible contribution to the observed trend is that the data set may contain examples where experimentalists have deliberately solved separate structures of a given protein complexed with its natural ligand and with one or more (natural or synthetic) analogues thereof.

A quite reasonable reaction to the results presented here might be a mild surprise that the correlation between protein and ligand similarity is not stronger. There seem to be several

**Table 2:** Results for Full Data Set[a]

| | ligand molecular similarity | | |
|---|---|---|---|
| | low | high | total |
| a. Actual Numbers of Observations | | | |
| unlike sequences | 8950 | 452 | 9402 |
| like sequences | 238 | 90 | 328 |
| total | 9188 | 542 | 9730 |
| b. Expected Number of Observations | | | |
| unlike sequences | 8878.3 | 523.7 | 9402 |
| like sequences | 309.7 | 18.3 | 328 |
| total | 9188 | 542 | 9730 |
| c. Enrichment Factors | | | |
| unlike sequences | 1.01 | 0.86 | |
| like sequences | 0.77 | 4.93 | |
| d. Percentages of Each LL Type within Each PP Type | | | |
| unlike sequences | 95.2% | 4.8% | 100.0% |
| like sequences | 72.6% | 27.4% | 100.0% |
| e. Percentages of Each PP Category within Each LL Category | | | |
| unlike sequences | 97.4% | 83.4% | |
| like sequences | 2.6% | 16.6% | |
| total | 100.0% | 100.0% | |

[a] The expected numbers of observations are calculated on the assumption that protein−protein sequence identity and ligand−ligand molecular similarity are independent. The enrichment factor is the ratio of the actual to the expected number of observations.

factors acting to reduce the correlation. One such factor is that many enzymes and other proteins necessarily bind a plurality of unrelated molecules, such as a substrate and a cofactor. Another is that nature has solved the problem of binding certain classes of molecule on more than one occasion. Even if convergent evolution were to reinvent a given fold for binding related compounds, the lack of sequence similarity would lead to an observation here of highly similar ligands binding unlike proteins. In addition, the PDB data set will contain noise from ligands that are found as accidents of the crystallization process.

This work is designed to test the relationship between protein−protein and ligand−ligand similarity in complexes taken from a structurally defined subset of the PDB. Its conclusions cannot necessarily be transferred either to the whole structural diversity of the PDB or to a representative set of in vivo protein−ligand complexes, such as all those found in a living cell. Extension of this work to encompass all structural classes in the PDB would in principle be feasible, but the proportion of combinations exhibiting like sequences would be substantially smaller and the relationship with ligand similarity harder to detect. However, it would potentially allow more accurate determination of the cutoff between like and unlike sequences.

Many of the criteria determining presence in the PDB, such as pharmaceutical interest, ease of forming diffracting crystals, and accidents of the crystallization conditions, do not reflect an unbiased sampling of natural protein−ligand complexes. Indeed, many protein−ligand interactions in the PDB are simply not found in nature. Several contributing factors to the relationship between protein−protein and ligand−ligand similarity within the PDB, such as sets of related synthetic inhibitors, proteins deliberately complexed with various noncognate ligands, and the over-representation of complexes that have been considered scientifically interesting, would not operate in a sampling procedure designed to mirror natural occurrence.

A useful future modification would be to define a data set including only ligands that bind the given proteins in vivo, eliminating all noncognate interactions and synthetic ligands, and to carry out a similar analysis on this. To establish a truly representative data set for protein−ligand interactions occurring in the cell, however, would be a substantial task requiring great care and accurate definition. It is likely that a large number of relevant protein−ligand interactions could be found by a thorough search of the literature and available databases. Analysis of such a well-constructed data set would be required in order fully to understand the relationship between protein−protein and ligand−ligand similarity in nature.

## CONCLUSIONS

Within this data set of alpha helical proteins from the PDB, similar proteins do indeed tend to bind similar ligands. One can use the property of binding a higher proportion of similar ligands to define like and unlike pairs of protein chains, separated by a rather approximate percentage sequence identity cutoff. Similarly, one can use the property of tending to bind to protein chains of greater sequence identity to define high and low similarity pairs of ligand residues, with a reasonable cutoff at a Tanimoto similarity score of 0.70. Observations of like protein chains in combination with high similarity ligand residues are enriched by a factor of 4.93 relative to what would be expected if protein sequence identity and ligand molecular similarity were independent.

In the future, it would be desirable to extend this kind of study, first, to the full structural diversity of the PDB and, second, to a data set reflecting the natural occurrence of protein−ligand interactions, for instance a truly representative sample of the protein−ligand complexes found within a living cell. It is far from clear that the patterns found for the alpha helical PDB subset would be reflected in the results from such data sets. It is hoped that the present work will encourage others to address the similarity question using other methodologies.

## ACKNOWLEDGMENT

Unilever are thanked for their financial support for the Centre for Molecular Informatics. Dr. Irene Nobeli and Dr. Roman Laskowski (University College London) are acknowledged for helpful discussions.

**Supporting Information Available:** Criteria for defining bit strings for ligand residues and data set of 140 pairs used in the study of the relationship between protein−protein sequence identity and ligand−ligand molecular similarity. This material is available free of charge via the Internet at http://pubs.acs.org.

## REFERENCES AND NOTES

(1) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235−242
(2) Laskowski, R. A.; Hutchinson, E. G.; Michie; A. D.; Wallace, A. C.; Jones, M. L.; Thornton, J. M. PDBsum: A Web-based Database of Summaries and Analyses of all PDB Structures. *Trends Biochem. Sci.* **1997**, *22*, 488−490
(3) http://www.biochem.ucl.ac.uk/bsm/pdbsum/index.html.
(4) Gerlt, J. A.; Babbitt, P. C. Mechanistically Diverse Enzyme Superfamilies: The Impotance of Chemistry in the Evolution of Catalysis. *Curr. Opin. Chem. Biol.* **1998**, *2*, 607−612.
(5) Martin, A. C. R.; Orengo, C. A.; Hutchinson, E. G.; Jones, S.; Karmirantzou, M.; Laskowski, R. A.; Mitchell, J. B. O.; Taroni, C.; Thornton, J. M. Protein Folds and Functions. *Structure* **1998**, *6*, 875−884.
(6) Todd, A. E.; Orengo, C. A.; Thornton, J. M. Evolution of Protein Function, from a Structural Perspective. *Curr. Opin. Chem. Biol.* **1999**, *3*, 548−556.
(7) Hegyi, H.; Gerstein, M. The Relationship Between Protein Structure and Function: A Comprehensive Survey with Application to the Yeast Genome. *J. Mol. Biol.* **1999**, *288*, 147−164.
(8) Barnard, J. M.; Downs, G. M.; Willett, P. *Descriptor-Based Similarity Measures for Screening Chemical Databases, in Virtual Screening for Bioactive Molecules*; Böhm, H.-J., Schneider, G., Ed.; Wiley-VCH: Weinheim, 2000.
(9) ftp://ftp.rcsb.org/pub/pdb/data/monomers/components.cif.
(10) Mitchell, J. B. O.; Alex, A.; Snarey, M. SATIS: Atom Typing from Chemical Connectivity. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 751−757.
(11) Mitchell, J. B. O.; Laskowski, R. A.; Alex, A.; Thornton, J. M. BLEEP − Potential of Mean Force Describing Protein−Ligand Interactions: I. Generating Potential. *J. Comput. Chem.* **1999**, *20*, 1165−1176.
(12) Flower, D. R. On the Properties of Bit String-Based Measures of Chemical Similarity. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 379−386.
(13) http://www-mitchell.ch.cam.ac.uk/bitstring.html.
(14) Orengo, C. A.; Michie, A. D.; Jones, S.; Jones, D. T.; Swindells, M. B.; Thornton, J. M. CATH − A Hierarchic Classification of Protein Domain Structures. *Structure* **1997**, *5*, 1093−1108.
(15) http://www-mitchell.ch.cam.ac.uk/proligdata.html.
(16) Milburn, D.; Laskowski, R. A.; Thornton, J. M. Sequences Annotated by Structure: A Tool to Facilitate the Use of Structural Information in Sequence Analysis. *Prot. Eng.* **1998**, *11*, 855−859.
(17) http://www.biochem.ucl.ac.uk/bsm/sas/.
(18) Xue, L.; Godden, J. W.; Bajorath, J., Database Searching for Compounds with Similar Biological Activity Using Short Binary Bit String Representations of Molecules. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 881−886.
(19) Xue, L.; Stahura, F. L.; Godden, J. W.; Bajorath, J. Mini-fingerprints detect similar activity of receptor ligands previously recognized only by three-dimensional pharmacophore-based methods. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 394−401.
(20) Matter, H. Selecting Optimally Diverse Compounds from Structure Databases: A Validation Study of Two-Dimensional and Three-Dimensional Molecular Descriptors. *J. Med. Chem.* **1997**, *40*, 1219−1229.
(21) Patterson, D. E.; Cramer, R. D.; Ferguson, A. M.; Clark, R. D.; Weinberger, L. E. Neighborhood Behavior: A Useful Concept for Validation of Molecular Diversity Descriptors. *J. Med. Chem.* **1996**, *39*, 3049−3059.
(22) Orengo, C. A.; Flores, T. P.; Taylor, W. R.; Thornton, J. M. Identification and Classification of Protein Fold Families. *Prot. Eng.* **1993**, *6*, 485−500.
(23) Teichmann, S. A.; Chothia, C.; Church, G. M.; Park, J. Fast Assignment of Protein Structures to Sequences Using the Intermediate Sequence Library PDB-ISL. *Bioinformatics* **2000**, *16*, 117−124.

CI010364Q