Published online 27 November 2006 *Nucleic Acids Research, 2006, Vol. 34, No. 22* ***e152***
doi:10.1093/nar/gkl788

Rapid detection of similarity in protein structure and function through contact metric distances

Andreas Martin Lisewski and Olivier Lichtarge*

Department of Molecular and Human Genetics, Baylor College of Medicine, One Baylor Plaza, Houston, TX 77030, USA

Received September 4, 2006; Revised September 28, 2006; Accepted September 29, 2006

ABSTRACT

The characterization of biological function among newly determined protein structures is a central challenge in structural genomics. One class of computational solutions to this problem is based on the similarity of protein structure. Here, we implement a simple yet efficient measure of protein structure similarity, the contact metric. Even though its computation avoids structural alignments and is therefore nearly instantaneous, we find that small values correlate with geometrical root mean square deviations obtained from structural alignments. To test whether the contact metric detects functional similarity, as defined by Gene Ontology (GO) terms, it was compared in large-scale computational experiments to four other measures of structural similarity, including alignment algorithms as well as alignment independent approaches. The contact metric was the fastest method and its sensitivity, at any given specificity level, was a close second only to Fast Alignment and Search Tool—a structural alignment method that is slower by three orders of magnitude. Critically, nearly 40% of correct functional inferences by the contact metric were not identified by any other approach, which shows that the contact metric is complementary and computationally efficient in detecting functional relationships between proteins. A public 'Contact Metric Internet Server' is provided.

INTRODUCTION

There are now over 39 000 entries in the Protein Data Bank (PDB) (1), with about 100 more added weekly. A growing fraction are from the protein structure initiative (2), 30–50% of which are listed without a known function (3). Thus the functional annotation gap in the structural proteome may eventually come to mirror that encountered in genomics where, for example, up to 40% of the genes sequenced at NCBI's RefSeq databank lack annotation of biological function (4). In this light, there is an important need for novel methods of protein function prediction that exploit the available structural knowledge in order to go beyond the limitations of sequence analysis (5).

Methods to infer functional similarity from structure currently fall in two broad classes. Those that rely on a global similarity, scoring whole structures likeness, and others that estimate local similarity among structural 'motifs' that embody key functional properties. Eventually, one may expect that both approaches will be complementary (6–8) since similar folds may mediate different biological functions (5,9,10), while conversely different folds may support identical functions (11) based on common local structural motifs (12). This study, however, focuses on the first class motivated by examples where structural similarity points to functional similarity, long after any common evolutionary origin is rubbed out from sequence comparison (13–15). Many such methods compare whole protein structures, or domains, and annotate function (16–33).

Typically, protein similarity is obtained by computing structural alignments (34). These alignments are computationally hard (35,36), however, and thus often require heuristics and approximations. For example, the DALI algorithm (37) looks for common local patterns in residue–residue distance matrices, and then maximizes their size by combining smaller patterns into larger ones using a Monte Carlo method; the CE algorithm (21) searches for the maximum alignment by a combinatorial extension of a path of aligned fragment pairs that satisfy certain similarity criteria; VAST (29,30) uses a graph–theoretical approach to align secondary structure elements based on their type, relative orientation and connectivity.

Alternatively, other algorithms compute similarity much faster by avoiding structural alignment: the method of Gauss integrals applied to the topological curve in space defined by the polypeptide chain's C_α backbone resulting in the so-called Scaled Gauss Metric, SGM (38); PRIDE and PRIDE2 (25,39) describe proteins as distributions of backbone carbon $C_\alpha(i)$–$C_\alpha(i+n)$ geometrical distances, where i is the residue number in the protein chain and n is an integer in the range [3–30], so that structural similarity is evaluated on 28 distance distributions and expressed into a single similarity score.

*To whom correspondence should be addressed. Tel: +1 713 798 5646; Fax: +1 713 798 7773; Email: lichtarge@bcm.edu

Recently, some of these methods have also been evaluated against structural similarity benchmarks (17,40), although their performance for function prediction has not been tested comparatively. These studies focused on the ability to recognize CATH (41) fold classifications of several distantly related proteins, and to detect 'difficult cases' of structural similarity proposed previously (42). However, in contrast to structural similarities based on manually maintained or automatically generated fold classifications, predicted functional relationships can be ultimately tested in experiments, and hence it is desirable to benchmark protein structure similarity measures against existing functional annotations.

Here, we introduce a new vector representation of polypeptide structure that is remarkably fast, quantitative, and, as we show, lends itself to global structure comparison and function prediction. The components of the vector are the frequencies of C_α backbone contacts at a given sequence separation, i.e. at a given contact length. This so-called 'contact vector' embodies the histogram of a structure's contact lengths and thus quantifies the topology of the protein fold by taking into account residue–residue contacts from local secondary as well as from non-local tertiary structure. To compare structures, we use a distance metric between vectors, the 'contact metric'.

We aim to confirm the following hypotheses: (i) that this measure is computationally very efficient since, as a consequence of the contact vector representation, it does not require a structural alignment; (ii) that it carries enough information that the similarities it detects match, or are correlated, with the more familiar root mean square deviation (RMSD) between structural alignments—despite the simplified representation through contact vectors; (iii) that it is useful for functional prediction, namely that it detects functional similarity among remote homologs either more accurately than, or in a way that is complementary to, comparable methods. As a result, the contact metric would then efficiently complement current structural similarity tools used for automated functional annotation of proteins.

The following results confirm these hypotheses. First, a similarity search against ~34 000 protein chains in the current PDB runs in a few seconds of single CPU time. Thus across large molecular databases, the contact metric computes similarity nearly instantaneously. Second, for small contact metric distances, randomly chosen pairs of protein structures positively correlate (between 0.48 and 0.64) with maximum alignment RMSD. Finally, receiver operating characteristics (ROC) over 1.38 million pairs of remotely homolog PDB chains distributed among the Gene Ontology (GO) classes 'molecular function', 'biological process' and 'cellular component' (43) show that the contact metric sensitivity is higher, across all specificity levels tested, than those obtained with Basic Local Alignment and Search Tool (BLAST) (44), SGM, PRIDE2, and nearly as high as Fast Alignment and Search Tool (FAST) (33), a detailed 3D alignment method shown to be computationally more efficient and accurate for structural recognition than standard alignment algorithms, such as DALI and CE. Critically, up to 44% of functional relationships in GO detected by the contact metric could not be found by any of the other methods tested, including FAST, which itself missed nearly 60% of the contact metric hits. The contact metric therefore provides complementary information to more established approaches in structure-based functional annotation.

A public 'Contact metric internet server' for similarity searches against the PDB is maintained and available at the internet URL http://mammoth.bcm.tmc.edu/cm/.

RESULTS

The contact metric as a similarity measure for protein folds

The contact metric uses a histogram representation of protein structures. These histograms record the number of residue–residue contacts in a structure, as a function of their separation along the sequence. These values are then ordered with a so-called 'contact vector', which can then be compared directly between proteins by the contact metric. Figure 1A illustrates this representation in human ubiquitin (PDB 1ubq). In this 76 residue-long protein, every pair of residues with C_α backbone atoms closer than 9 Å is recorded in a contact matrix (45). Summing up, diagonally in the contact matrix, all the contacts among residues i and j, such that $i - j = k - 1$ with $k \geqslant 3$, leads to a histogram that enumerates all structural contacts among residues that are $k-1$ positions apart in the sequence. Typically many contact lengths k are short ($k = 3$, 4 and 5), consistent with the local secondary structure constraints of α-helices and turns. But other contact lengths are longer, with even some comparable to the length of the chain, and these carry information about the entire fold. Thus the contact vector representation of a tertiary structure is $(q_3, \ldots, q_k, \ldots, q_{400})$, where q_k is the absolute number of contact lengths $3 \leqslant k \leqslant 400$. The cut-off at $k = 400$ reflects the few contribution above this characteristic limit (Supplementary Figure S1).

The contact metric between any two protein chains (X, Y) is the absolute distance between two contact vectors, i.e.,

$$d(X,\ Y) = \sum_{k=3}^{400} \mid q_k(X) - q_k(Y) \mid .$$

Even though short contact lengths, say with $k \leqslant 10$, dominate in contact vectors, summation over long-range contacts contributes almost equally, and the contact metric reflects both secondary and tertiary structure. For example, in PDB 1 ubq the sum of contacts with $k \leqslant 10$ is 176 while all remaining contact give 140, a comparable number (Figure 1A). This balance between short and long range contributions also holds on a large scale. The Pearson correlation between the contact metric d and $d_{<10}$, which is a short range contact metric that takes into account only close contacts up to $k = 10$, is 0.51 over a randomly selected set of 32 525 PDB pairs (Supplementary Figure S2A). Similarly, the correlation between d and $d_{>10}$, i.e. the long-range contact metric accounting contacts larger than $k = 10$, is 0.79 (Supplementary Figure S2B). This shows that contact metric values take into account local structure to some extent, but long range contacts contribute most. Hence, it cannot be attributed to local features alone, such as to secondary structures.

Differences in protein chains length bias (raise) the contact metric distances of longer chains. This follows because in native structures chain lengths are proportional to the number

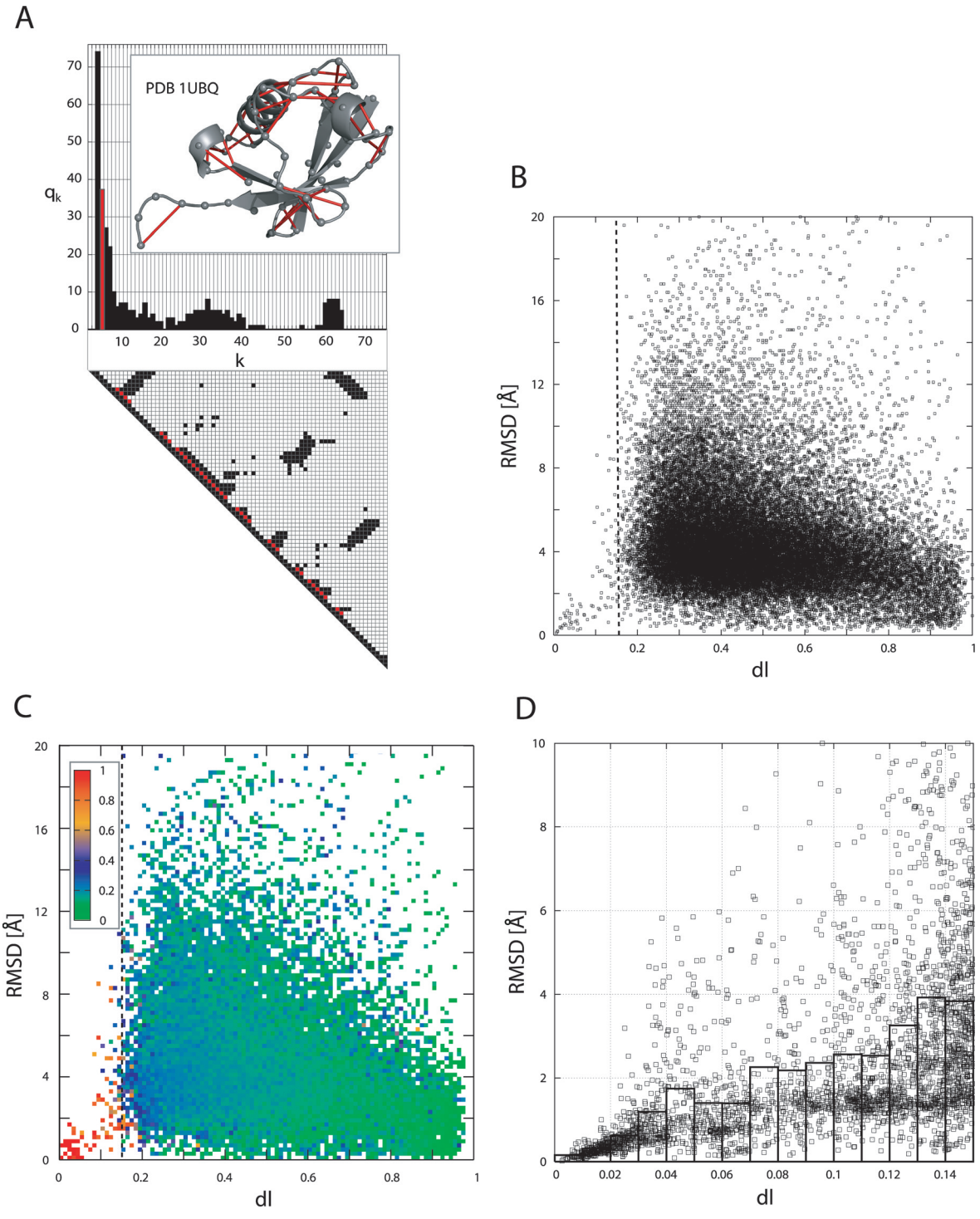

Figure 1. (**A**) Contact vector representation of human ubiquitin (PDB 1ubq). Contact lengths are given as integer values k, and their frequencies are counted by q_k. Red lines indicate contacts between C_α-atoms (spheres) with sequence separation $k - 1 = 3$. A contact matrix C is derived from the spatial coordinates of C_α-atoms documented in the PDB. For a single-chain protein of length L, one defines a contact by using a Euclidean distance threshold of 9 Å between C_α-atoms. In a protein sequence consecutive residues are in contact, thus $C(i, j) = 1$ for all i, j in $\{1,\ldots, L\}$ with $|i - j| = 1$. Then the frequency q_k is the number of contact pairs (i, j) for any given sequence separation $k - 1 = j - i$ with $2 < k - 1 < L - 1$ and with $j > i$. (**B**) Scatter plot of length corrected contact metric (LCM) and the maximum geometrical alignment RMSD calculated with FAST for 32 525 pairs of PDB structures. (**C**) Distribution of alignments fraction f_{avg} for the same data as in (B); pairs of structures which can be aligned close to perfect ($f_{avg} > 0.8$) cluster at small contact metric and RMSD values. (**D**) LCM-RMSD scatter plot for 3074 pairs of PDB chains, with 1838 pairs having sequence identity below 25%. The average alignment fraction for all pairs is 0.82 (SD 0.19). Solid lines give the average RMSD by steps of 0.01 in dl. The Pearson correlation coefficient between RMSD and dl is 0.48, between RMSD and sequence identity it is 0.55, and between LCM and sequence identity it is 0.67.

of contacts with $k \geqslant 3$ (Supplementary Figure S3), thus we have $dl(X,Y) \leqslant c\ (L_X + L_Y)$, with the proportionality constant $c \approx 5.8$ and protein chain lengths L_X, L_Y. This can be corrected by normalizing the contact metric with the factor $1/[c(L_X + L_Y)]$ to yield the length corrected contact metric (LCM), used henceforth. We note that although the contact metric is a true metric mathematically, LCM is not: it is still positive, non-degenerate and symmetric, but does not necessarily satisfy the triangle inequality. The length corrected contact metric can then be written by the simple formula

$$dl(X,Y) = \frac{\sum_{k=3}^{400} \mid q_k(X) - q_k(Y) \mid}{\sum_{k=3}^{400} q_k(X) + q_k(Y)}$$

With this expression all *dl*-values are limited between 0 and 1, where 0 signalizes maximum similarity, and 1 the minimum.

LCM distances are always well defined for any two polypeptide structures, regardless of their geometrical similarity. Because they can be computed rapidly, it was possible to estimate the statistical significance of any number $dl(X, Y)$, by assigning a *P*-value from the total distribution of contact metric distances. This distribution was randomly sampled choosing $N_s = 2.5 \times 10^6$ single-chain protein pairs (X, Y) taken from the PDB (Supplementary Figure S4), and it suggested a level of statistical significance at values below $dl_c = 0.15$, because for larger distances their distribution is characterized by a rapid 'blow-up' in relative frequency indicating the onset of a random regime. Hence we considered structures to be significantly similar if their LCM distance was below $dl_c = 0.15$, which corresponded to 99.7% significance level ($P = 0.003$).

To demonstrate that LCM is not wholly distinct from the intuitive, standard geometric similarity measure between molecules, it was compared to the RMSD of a maximum geometrical alignment of two chains. We used the same set of 32 525 PDB pairs, and for each pair (X, Y) calculated a maximum alignment RMSD with FAST, as shown in the scatter plot of Figure 1B. Over the entire LCM domain, there is no correlation between the length corrected contact metric and the alignment RMSD (the Pearson correlation coefficient is -0.11). But, for statistically significant values of LCM ($dl \leqslant 0.15$), the RMSD between aligned structures falls toward 1 Å, as shown in Figure 1D, and these values are positively correlated with maximum alignment RMSD: for the 79 pairs in Figure 1B the correlation coefficient is 0.64, and for the randomly chosen 3074 PDB pairs in Figure 1D we measured a correlation of 0.48, again in the range $0 \leqslant dl \leqslant 0.15$.

Higher LCM values do not maintain this correlation because alignment RMSD can be restricted to very different alignment fractions (or, alignment coverage). In contrast, the contact metric always relates entire protein chains and a loss of correlation is expected as structures become more dissimilar, which was confirmed by the following analysis.

Given any two protein chains *X* and *Y* with lengths L_X and L_Y, their maximum number of geometrically aligned residue pairs is $L_{max} = (L_X + L_Y)/2$, which is reached if both chains have the same number of residues and if geometrical alignment is perfect (i.e. every residue has exactly one aligned partner residue). We then defined the alignment fraction *f* as the ratio of actually aligned residue pairs $L \leqslant L_{max}$ over the maximum number L_{max}. Figure 1C gives a 100×100 array of alignment fractions f_{avg} derived from the data in Figure 1B (Each pixel in Figure 1C covers an area of (0.1 Å $\times$ 0.01) in RMSD/contact metric space and coloring represents the average value f_{avg} of all points within a single pixel.) High alignment values of f_{avg} ($f_{avg} > 0.8$, red pixels) cluster at small contact metric distances ($dl \leqslant dl_c$) and RMSD values (RMSD < 8 Å), while for $dl > dl_c$ well aligned pairs with $f_{avg} > 0.6$ were rarely observed, and the correlation coefficient was -0.10. Only statistically significant LCM values signalized both small RMSD and high alignment coverage. This convergence between the contact metric and the geometrical alignment RMSD suggests that significant contact metric distances predict that two polypeptide structures can be aligned with a small geometrical error—thus they correspond to our intuitive definition of geometrical similarity.

Large-scale annotation with Gene Ontology terms

To demonstrate that LCM distances carry functional information, we performed large-scale functional recognition experiments and compared the results to several other methods for structural comparison. We mimicked realistic conditions and selected a test set of 1662 non-redundant protein structures with <25% mutual sequence identity from PDBselect25 (46), March 2006, that also had at least one available GO annotation term recorded (version GOA 28.0). Starting with 2372 in the PDBselect25, there were $N = 1662$ proteins with GO terms representing 261 molecular functions, 216 biological processes and 75 cellular components. The other methods were BLAST, the standard for sequence alignment and similarity measurements; FAST, which performs structural alignments; PRIDE2 and SGM, which do not require structural alignment.

The algorithms were evaluated and compared in terms of their sensitivity and specificity combined into ROC curves. First, all $N(N-1)/2 = 1.38 \times 10^6$ pairwise similarity scores were calculated for each method. Then, if needed, scores were converted into a distance measure between all structures. For example, since FAST measures structural similarity with a normalized similarity measure *Sn*, which is larger when similarity is greater, we used the formula $1/(1 + Sn)$ to obtain a distance value. Finally, for a variable distance cutoff r_c, true positives were defined as any pair of proteins (X, Y) within distance r_c that shared at least one GO term. False positives pairs also had distances smaller than r_c but were without a common GO term; true negatives had a distance greater than r_c and also no common GO term; and false negatives had a distance greater than r_c but with a common GO term. The test was limited to pairs with relatively few false positives, and thus to specificities of no less than 90%.

For all three GO categories, the ROC graphs in Figure 2A–C show that LCM and FAST have an advantage over BLAST, SGM and PRIDE2 of at most 5% better sensitivity, for specificities in the range 90–100%. The difference between FAST and LCM is smaller at most 1.0, 1.5 and 0.7% for molecular function, molecular process and cellular

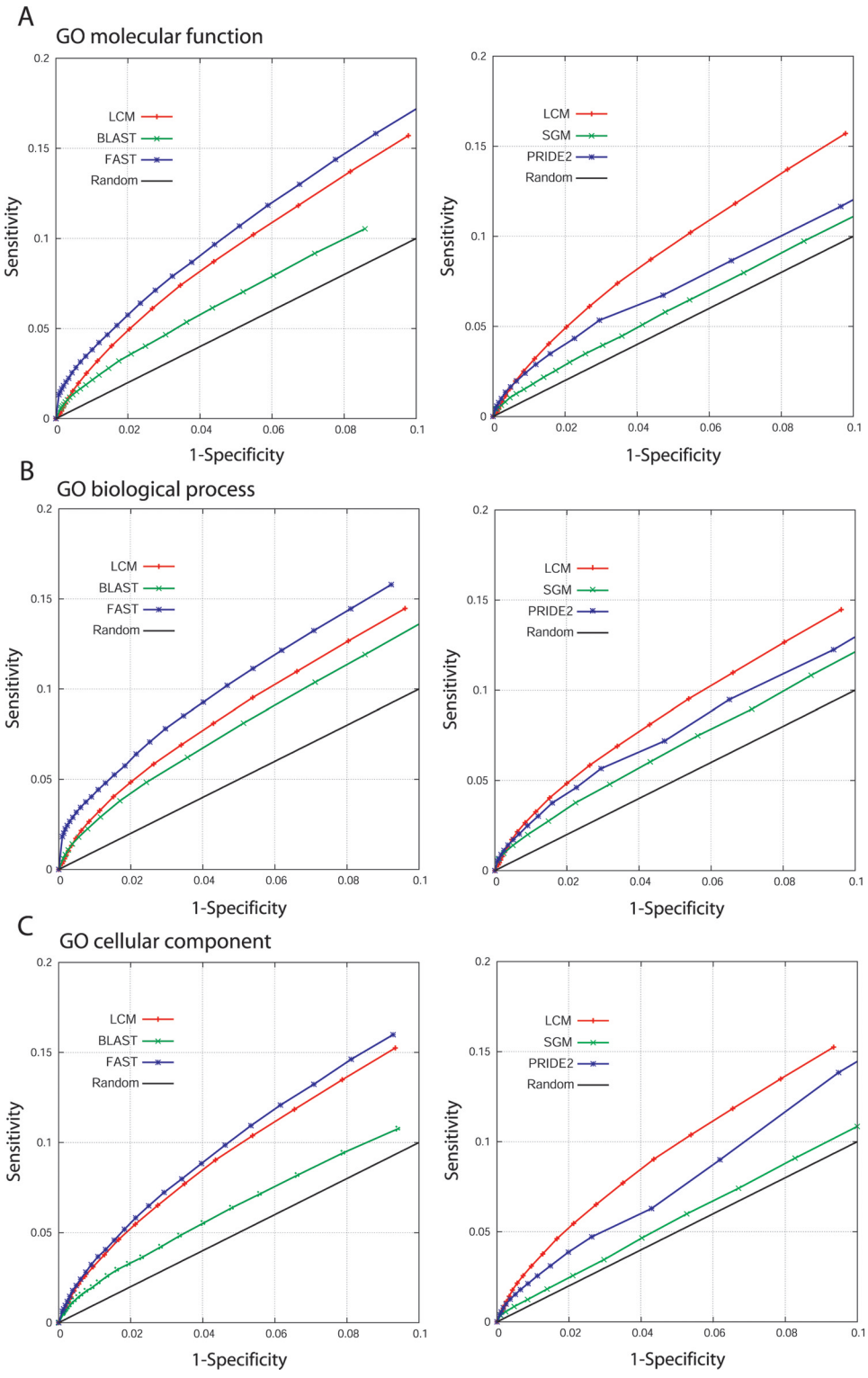

Figure 2. (**A–C**) Gene Ontology (GO) standardized Receiver Operating Characteristic (ROC) curves for 1.38 million pairs of PDB chains with <25% sequence identity. For the three GO categories molecular function, biological process, and cellular component the length-corrected contact metric (LCM) is more sensitive than all method tested, except FAST.

component GO categories, respectively. Thus, overall, LCM performed nearly on par, but not better than FAST. The advantages of LCM stem from two other factors, however, its algorithmic speed and its complementarity to current methods. Complementarity, a desirable feature of any novel protein similarity measure, is the capacity to identify functional similarities which cannot be detected by known methods.

First, LCM achieves its results far quicker than FAST, or the other methods. On a 1 GHz Athlon AMD processor, these computations took 0.3 h for LCM, 0.3 h for SGM, 0.4 h for PRIDE2, 10 h for BLAST and 74 h for FAST. LCM, GI and PRIDE2 use a precompiled vector representation of the N protein set calculated before the run, and these compilations took between 1 and 3 min of computer runtime. Distance calculations for LCM and SGM were computationally essentially equivalent, and required for any two protein chains, the calculation of a metric value between two vectors (SGM uses a 29 vector components in 8 bit precision, and LCM subtracts 397 integer vector components). The total runtime of PRIDE2 was slightly longer since it searched for similarities in sequence domains by applying a sliding window approach, whereas LCM and SGM consider only entire chains. BLAST and FAST were slower because they performed, in contrast to the other algorithms, alignment calculations for every pair.

Second, LCM detects—while being sensitive—novel functional relationships. This is seen in Figure 3A–C, at fixed specificity of 95%, in all three GO categories. Green bars indicate the absolute numbers of correct functional assignments (true positives) for every method tested, and red bars show the number of those true positives that were not identified by any of the other methods combined. In the GO category molecular function, FAST had 6347 true positives (10.8% sensitivity). Of these 2772 went undetected by any of the other methods. By comparison, BLAST uniquely identified 2425 and LCM 2154, while PRIDE2 and SGM reached lower numbers with 1867 and 1449, respectively. Likewise, for GO biological process, BLAST exclusively identified 1486 correct pairs, FAST 1411, PRIDE2 1030, LCM 955, and SGM 877. And for GO cellular component the ranking was FAST with 1954, BLAST with 1750, LCM with 1614, PRIDE2 with 1129, and SGM with 1077. These data show that LCM contributed the highest level of complementary information among the alignment independent methods tested here.

Since FAST was the most accurate while LCM is both accurate and efficient, we specifically tested the complementarity of these two methods. At 95% specificity, 3360 (60%) of the correct molecular function pairs found by LCM were not found by FAST, while, vice-versa, only 2987 (47%) of FAST true positives were not detected by LCM (Figure 4A). For GO biological process, the extent to which the pairs identified by each method are disjoint is the following: 1626 (58%) pairs were found by LCM and not by FAST against 2141 (64%) found by FAST and not LCM, and for the GO cellular component the numbers were 2528 (63%) against 2791 (41%), respectively. As shown in Figure 4A–C, no other method reached this absolute level of complementarity to FAST.

Public server

A publicly available software tool for structural and functional similarity detection across the PDB with the contact metric was implemented (the 'Contact Metric Server': http://mammoth.bcm.tmc.edu/cm/). As of August 2006, it contained contact vectors of 33 687 PDB chains, and for any chosen query structure X, it outputs an ordered table of most similar PDB entries, each characterized by LCM distance and sequence identity to the query, by fold CATH characterization, and by available GO terms. To use this tool for functional annotation, a simple a nearest-neighbor rule can be followed by taking the first non-trivial (a match with $dl = 0$ is trivial) value $dl(X,Y)$. Consequently, protein chain Y annotates the function of X if the function of Y is known, and if $dl(X,Y)$ is statistically significant, i.e. if $dl(X,Y) < dl_c = 0.15$. If the closest match Y has unknown function the nearest neighbor rule can be applied to the second closest match $dl(X,Z)$, and repeated until any functionally characterized protein within a significant distance is met. Conversely, if no statistically significant and functionally classified nearest-neighbor is found, functional annotation with this strategy fails.

We tested this annotation strategy through 5-fold cross validation among our set of 1662 chains with FAST, LCM, BLAST, PRIDE2 and SGM. Since for LCM a significance level at $dl_c = 0.15$ ($P = 0.003$) corresponded to 0.4% specificity, we set this specificity level for all methods, which were then evaluated according to their accuracy, i.e. their fraction of correct GO annotations. For all three GO categories combined, FAST again ranked best with 41% accuracy, LCM second with 27%, PRIDE2 with 25%, SGM with 19% and BLAST with 18% (random nearest neighbor selection had 5% accuracy).

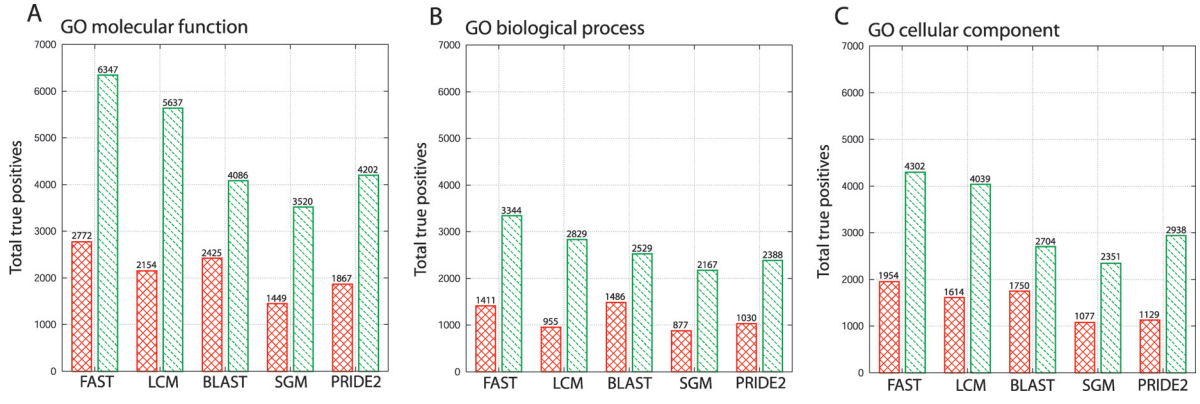

Figure 3. (**A–C**) Match overlap analysis of the different methods at 95% specificity. Green bars represent the absolute number of true positive hits; red indicate the subset of hits which were exclusively found by any of these methods. In each Gene Ontology category the length corrected contact metric contributes a significant number of matches not identified by any of the other methods.

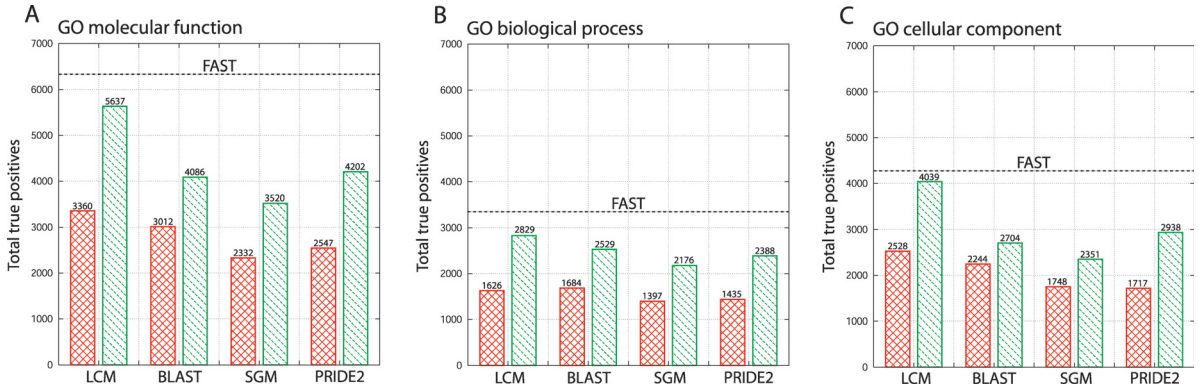

Figure 4. (**A–C**) Overlap analysis relative to the FAST method (at 95% specificity) in three GO categories. Green bars represent the absolute number of true positive numbers; red indicate the subset of hits which were not found by FAST. True positive hits of FAST and indicated by dashed horizontal lines.

DISCUSSION

This study presents a novel measure of protein structural similarity, based on a simple one-dimensional contact vector representation that tallies, by extent of sequence separation, the contact frequency of a given structure. Although other one-dimensional contact patterns have been, along with further protein attributes, included for tertiary structure representation (47–49), their general use for structural and functional recognition has not been assessed previously. Direct comparison of contact vectors is tantamount to rapid structure comparisons but it circumvents the need to compute structural alignment. The results show that LCM is correlated with RMSD measures of structural similarity and adds novel information to discover functional annotations based on structure comparison. Not only does its ROC nearly match FAST, but also it is much faster, and it identifies biological relationships that go undetected by other methods.

An evident limitation of a contact vector representation is the loss of information about the order or composition of structural domains. Therefore, a pair of structurally divergent proteins with a common subdomain may, in general, not result in a small contact metric value, and thus may not be recognized as being similar. A sliding window approach, such as implemented in the PRIDE2 algorithm may, however, be able on sequence domains and thus recover domain similarity. In fact, FAST, CE and DALILite, and PRIDE2 only consider sequence domains for structural alignment. But, alternately, the contact vector representation accounts that, in general, the entire tertiary structure often is not reducible to a sum of its parts (i.e. to its sub-domains), and that proteins attain new folds and functions even though their sub-domains may be already known (50,51).

Also, because contact vectors represent all non-trivial contacts in the polypeptide chain, the contact metric becomes sensitive to large insertions and deletions between structures which overall still share a characteristic fold. PDB 8gchA represents a 237 residue long γ-chymotrypsin folded into a trypsin-like serine proteinase conformation (CATH 2.40.10.10). A much shorter (151 residues in length) viral analogue exists (PDB 2snv_), which differs from 8gchA by multiple deletions of α-helices, 3–10 helices and β-bridges, while preserving the common architecture (CATH 2.40.10.10). This difference in chain length and structure results in a large LCM distance, dl(8gchA,2snv_) = 0.33, which does not signalize similarity. Here, geometrical alignment has an advantage: FAST registers a low but significant similarity with RMSD = 2.68 Å and 54% of the residues aligned.

On the other hand, moderate insertions or deletions do not strongly perturb the contact metric because they result in small shifts in the contact vector profile. As long as the profiles considerably overlap after the shift, the contact metric still detects their similarity. For example, T4 lysozymes have been synthesized extensively with mutations, insertions and deletions. Specifically, a C54T, INS(A73-AAA), C97A mutant (PDB 209l_, and the only T4 lysozyme recorded in the PDB with a triple A73-AAA insert) has a nearest LCM neighbor (PDB 1qudA) which shares both C54T and C97A mutations but does not have insertions. Indeed, a comparison between both contact vectors (Supplementary Figure S5) shows a shift of a peak located at $k = 33$ (PDB 1qudA) to $k = 37$ (PDB 209l_).

It has been argued that entire fold comparison measures may fail to detect functional variations among structural homologs, for example among TIM barrels (7,52). The TIM barrel structural superfamily divides into 18-fold subfamilies (10), and to test LCM we considered FMN-dependent oxidoreductase and phosphate binding enzymes (FMOP), which all belong to the subfamily (CATH 3.20.20.70) with the highest functional diversity comprising 12 enzymatic classes (EC). In Supplementary Table S1 are listed 11 of the 12 representative proteins for each EC identifier (10) (we have left out the bifunctional enzyme 1pii_), as well as the nearest neighbor identified by LCM among 33 687 using the Contact Metric Server. In 10 out of 11 cases, the top match was statistically significant and had the same EC identifier as the original representative. Only the dehydrogenase 1ak5_ (EC 1.1.1.205; sixth entry in Table S1) had a closest match with a different function (1p0kA with EC 5.3.3.2). Given 91% correct functional assignments, this result suggests that even within a highly populated and redundant fold class like the TIM barrels, the length corrected contact metric is accurate enough to detect functional variations.

Another objection may be that the loss of information entailed by collapsing an $L \times L$ contact matrix into a contact vector with at most L non-zero components irretrievably

reduces the quality of structural comparisons and subsequent functional inference. But the data refute this view: the functional annotation sensitivity and specificity is nearly on par with FAST, the most accurate structure-based method of functional recognition tested here, and it outperforms other similarity measures that avoid structural alignment, e.g. SGM and PRIDE2.

One key finding to explain LCM performance is that small contact metric values correlate with small RMSD values showing that it captures absolute structural similarity and thus it can anticipate whether two structures will align with a small RMSD. This correlation is not perfect, however, even for smaller values. Thus the contact metric contains information about the topology of the protein chain—encoded in the inter-residue contacts—that is not part of geometrical alignment. This difference is indeed information, rather than noise, since LCM efficiently identifies structure-based functional similarities missed by FAST, and by other common sequence and structure-based similarity measures.

To further test our claim that the contact metric is sensitive to previously unrecognized fold similarities, one may turn to engineered proteins, designed specifically to explore novel regions of the fold space (18). These designed proteins may suggest new approaches to the prediction of tertiary structure (53), and may also lead to new functions (54–56). Engineering a novel fold, however, can be challenging. In small proteins, which is a well-populated region of natural folds (53), a 'new' design can easily possibly mimic an (naturally) occurring fold without this being readily apparent. We show below that in two cases of engineered proteins, Top7 and ANBP, the contact metric identifies meaningful similarities to known folds.

Top7 was synthesized *de novo* (PDB 1qys chain A) (57) after a computational search identified an amino acid sequence predicted to fold to a new tertiary structure. It was then found to overlap geometrically with the computationally predicted, novel protein fold. Specifically, Top7 had no significant matches after a structural search of the PDB with DALI, and verification through the TOPS server (58) also indicated novelty in fold topology.

A contact metric search across the PDB for structures near Top7 revealed four significant matches at almost equal distance ($P = 0.003$). A remarkable biophysical aspect of Top7 is its thermal stability which is significantly higher than in most proteins of comparable size (57). One the four contact metric matches is a small DNA-binding protein Sso10b2 from the hyperthermophilic archaeon *Sulfolobus solfataricus* (PDB 1udv chain A) (59), that thrives at 87°C and acidic pH (60). Top7 and Ssso10b2 both share a 2-layer sandwich architecture characterized by a β1−α1−β2−α2−β3−β4 order of secondary structure elements, as shown in Figure 5A. The main difference between Top7 and Sso10b2 is the N-terminus β-sheet domain: the former shows two anti-parallel β-strands, whereas 1udvA has only one short β-strand accompanied by an unstructured N-terminal coil. Otherwise, both chains have comparable fold topology. Thus, LCM analysis suggests that Top7 is a designed protein sequence that achieves a variant of the natural fold of 1udvA (CATH 3.30.110.20). Structural alignment using FAST indicates rather poor similarity with 3.9 Å RMSD for 67% of the residues aligned.

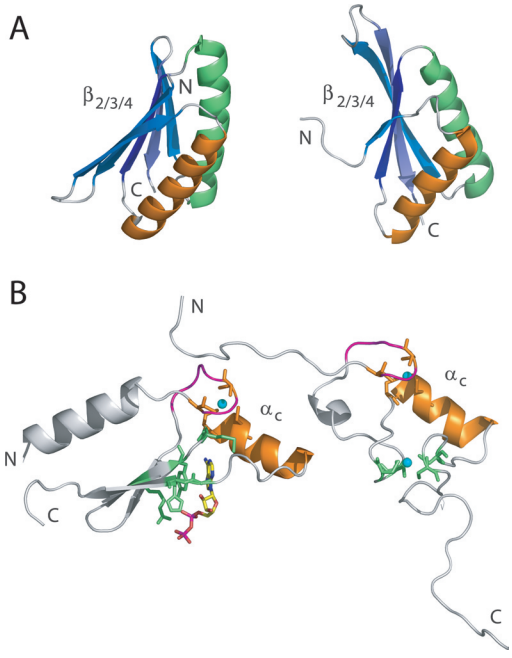

Figure 5. (**A**) Structural similarity between Top7 (PDB code 1qysA; left) and DNA-binding protein Sso10b2 (PDB code 1udvA) from the thermophyllic archaeon *S.solfataricus*. Both chains share a 2-layer sandwich architecture and the same fold topology. (**B**) Structural similarity between ANBP and RING finger domain of Not4. The treble clef finger in ANBP (1uw1A, left) is also found in 1e4uA (right), located at the N-terminal of the RING finger and consists of a loop (formed by an N-terminal hairpin and a knuckle; residues 14–20) and a central α-helix (α_c, residues 38–48) which is enclosed by two other loops surrounding a second zinc ion. The relative orientation and geometry of the first loop, the central α-helix and the zinc ion, i.e. the entire treble clef finger domain, are remarkably congruent in both structures: geometrical alignment between 1uw1A and 1e4uA by means of the four characteristic cysteine residues located in the clef finger motif C-X2-CX20/21-C-X2-C yields an RMSD of 0.3 Å and a similar orientation of the cysteine side-chains, and by aligning the C_α atoms of the entire zinc finger domain taking the residues 23–29, 46–56 from 1uw1A and residues 14–20, 38–48 from 1e4uA gives an RMSD of 1.2 Å.

A second example is ANBP. This artificial protein was created through *in vitro* evolution that selected for ATP/ADP binding activity; remarkably, zinc binding arose as well (61,62). Others (63) have noted that a stretch of 18 residues from the ANBP zinc binding region (1uw1A, residues 21–27 and 44–56) is a structural analogue to the zinc binding treble clef fingers motif of Pyk2–associated protein β ARF-GAP (PDB 1dcqA).

Likewise, the most significant nearest neighbor of ANBP by the contact metric also contains a treble clef finger protein. The match ($P = 0.001$) is a 78 residue long N-terminal RING finger domain of human Not4 (1e4uA, Really Interesting New Gene), which is part of a complex regulating RNA polymerase transcription (64). Figure 4B shows the geometrical alignment between 1uw1A and 1e4uA treble clef finger at the N-terminal of the RING finger. Specifically, ANBP binding of ADP involves the residues Arg41 and Tyr43 (bind to ADP phosphate groups), Gly63 and His64 (connect to the ADP ribose unit), and Met45 and Gly63 (link to the adenine moiety of ADP) (61). These residues divide into two motifs, R–X–Y–X–M, which is placed within

a five residue section right before the N-terminus of the central α-helix (α_c), and a G–H motif located at the end of an eight residue stretch adjacent to the C-terminus of the central α-helix (Figure 5B). Similarly, the cysteine residues in the RING finger specific for binding to the second zinc ion also group in two motifs located in structural domains very similar to those in ANBP: the C–X–C motif is found within a seven residue section right before the N-terminus of the central α-helix, while the C–X–X–C motif again reaches out to eight residues located after the carboxyl end of α_c. Thus, the common treble clef finger fold exhibits domain modularity at the ADP/zinc binding site of while it preserves its overall structural order, and the RING finger domain of Not4 found with the contact metric is another structural analog to the zinc ion binding region of ANBP.

As controls to both examples, DALI, VAST, PRIDE2 and CE searches for structures similar to Top7 and ANBP did not retrieve significant matches, and, in particular, the Sso10b2 match to Top7 and the RING finger domain of Not4 to ANBP could not be found.

In conclusion, our study shows that the contact metric enhances and complements the current repertoire of protein similarity measures. But since every method tested here exhibited some level of complementarity, our results also indicate that no single approach can capture all biologically relevant structural and functional relationships between protein molecules (35), and that a proper combination of several complementary approaches may lead to further improvement (19).

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors thank Zoltan Gaspari for providing the PRIDE2 program code. Financial support was provided by a training fellowship from the Gulf Coast Consortia through the W. M. Keck Center for Computational and Structural Biology (A.M.L). This work was also supported by grants from the National Science Foundation (DBI-0547695), National Institutes of Health (R01 GM066099) and March of Dimes (1-FY06-371) to O.L. Funding to pay the Open Access publication charges for this article was provided by the NSF.

Conflict of interest statement. None declared.

REFERENCES

1. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res*., **28**, 235–242.
2. Berman,H.M. and Westbrook,J.D. (2004) The impact of structural genomics on the protein data bank. *Am. J. Pharmacogenomics*, **4**, 247–252.
3. Chandonia,J.M. and Brenner,S.E. (2006) The impact of structural genomics: expectations and outcomes. *Science*, **311**, 347–351.
4. Pruitt,K.D., Tatusova,T. and Maglott,D.R. (2005) NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res*., **33**, D501–504.
5. Todd,A.E., Orengo,C.A. and Thornton,J.M. (2001) Evolution of function in protein superfamilies, from a structural perspective. *J. Mol. Biol*., **307**, 1113–1143.
6. Russell,R.B. (1998) Detection of protein three-dimensional side-chain patterns: new examples of convergent evolution. *J. Mol. Biol*., **279**, 1211–1227.
7. Stark,A., Shkumatov,A. and Russell,R.B. (2004) Finding functional sites in structural genomics proteins. *Structure*, **12**, 1405–1412.
8. Kristensen,D.M., Chen,B.Y., Fofanov,V.Y., Ward,R.M., Lisewski,A.M., Kimmel,M., Kavraki,L.E. and Lichtarge,O. (2006) Recurrent use of evolutionary importance for functional annotation of proteins based on local structural similarity. *Protein Sci*., **15**, 1530–1536.
9. Martin,J.L. (1995) Thioredoxin—a fold for all reasons. *Structure*, **3**, 245–250.
10. Nagano,N., Orengo,C.A. and Thornton,J.M. (2002) One fold with many functions: the evolutionary relationships between TIM barrel families based on their sequences, structures and functions. *J. Mol. Biol*., **321**, 741–765.
11. Babbitt,P.C. (2003) Definitions of enzyme function for the structural genomics era. *Curr. Opin. Chem. Biol*., **7**, 230–237.
12. Najmanovich,R.J., Torrance,J.W. and Thornton,J.M. (2005) Prediction of protein function from structure: insights from methods for the detection of local structural similarities. *Biotechniques*, **38**, 847–849851.
13. Orengo,C.A., Todd,A.E. and Thornton,J.M. (1999) From protein structure to function. *Curr. Opin. Struct. Biol*., **9**, 374–382.
14. Thornton,J.M., Todd,A.E., Milburn,D., Borkakoti,N. and Orengo,C.A. (2000) From structure to function: approaches and limitations. *Nature Struct. Biol*., **7**, 991–994.
15. Dobson,P.D., Cai,Y.D., Stapley,B.J. and Doig,A.J. (2004) Prediction of protein function in the absence of significant sequence similarity. *Curr. Med. Chem*., **11**, 2135–2142.
16. Koehl,P. (2001) Protein structure similarities. *Curr. Opin. Struct. Biol*., **11**, 348–353.
17. Sierk,M.L. and Kleywegt,G.J. (2004) Deja vu all over again: finding and analyzing protein structure similarities. *Structure*, **12**, 2103–2111.
18. Kolodny,R., Petrey,D. and Honig,B. (2006) Protein structure comparison: implications for the nature of 'fold space', and structure and function prediction. *Curr. Opin. Struct. Biol*., **16**, 393–398.
19. Watson,J.D., Laskowski,R.A. and Thornton,J.M. (2005) Predicting protein function from sequence and structural data. *Curr. Opin. Struct. Biol*., **15**, 275–284.
20. Holm,L. and Sander,C. (1994) Searching protein structure databases has come of age. *Proteins*, **19**, 165–173.
21. Shindyalov,I.N. and Bourne,P.E. (1998) Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Eng*., **11**, 739–747.
22. Jones,T.A. and Kleywegt,G.J. (1999) CASP3 comparative modeling evaluation. *Proteins*, **37**, 30–46.
23. Singh,A.P. and Brutlag,D.L. (1997) Hierarchical protein structure superposition using both secondary structure and atomic representations. *Proc. Int. Conf. Intell. Syst. Mol. Biol*., **5**, 284–293.
24. Kawabata,T. and Nishikawa,K. (2000) Protein structure comparison using the markov transition model of evolution. *Proteins*, **41**, 108–122.
25. Carugo,O. and Pongor,S. (2002) Protein fold similarity estimated by a probabilistic approach based on C(alpha)-C(alpha) distance comparison. *J. Mol. Biol*., **315**, 887–898.
26. Krissinel,E. and Henrick,K. (2004) Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. *Acta. Crystallogr. D. Biol. Crystallogr*., **60**, 2256–2268.
27. Lu,G. (2000) TOP: a new method for protein structure comparisons and similarity searches. *J. Appl. Crystallogr*., **33**, 177–183.
28. Martin,A.C. (2000) The ups and downs of protein topology; rapid comparison of protein structure. *Protein Eng*., **13**, 829–37.
29. Gibrat,J.F., Madej,T. and Bryant,S.H. (1996) Surprising similarities in structure comparison. *Curr. Opin. Struct. Biol*., **6**, 377–385.
30. Madej,T., Gibrat,J.F. and Bryant,S.H. (1995) Threading a database of protein cores. *Proteins*, **23**, 356–369.

31. Ye,Y. and Godzik,A. (2004) FATCAT: a web server for flexible structure comparison and structure similarity searching. *Nucleic Acids Res.*, **32**, W582–585.
32. Comin,M., Guerra,C. and Zanotti,G. (2004) PROuST: a comparison method of three-dimensional structures of proteins using indexing techniques. *J. Comput. Biol.*, **11**, 1061–1072.
33. Zhu,J. and Weng,Z. (2005) FAST: a novel protein structure alignment algorithm. *Proteins*, **58**, 618–627.
34. Kolodny,R., Koehl,P. and Levitt,M. (2005) Comprehensive evaluation of protein structure alignment methods: scoring by geometric measures. *J. Mol. Biol.*, **346**, 1173–1188.
35. Godzik,A. (1996) The structural alignment between two proteins: is there a unique answer? *Protein Sci.*, **5**, 1325–1338.
36. Lathrop,R.H. (1994) The protein threading problem with sequence amino acid interaction preferences is NP-complete. *Protein Eng.*, **7**, 1059–1068.
37. Holm,L. and Sander,C. (1993) Protein structure comparison by alignment of distance matrices. *J. Mol. Biol.*, **233**, 123–138.
38. Rogen,P. and Fain,B. (2003) Automatic classification of protein structure by using Gauss integrals. *Proc. Natl Acad. Sci. USA*, **100**, 119–124.
39. Gaspari,Z., Vlahovicek,K. and Pongor,S. (2005) Efficient recognition of folds in protein 3D structures by the improved PRIDE algorithm. *Bioinformatics*, **21**, 3322–3323.
40. Novotny,M., Madsen,D. and Kleywegt,G.J. (2004) Evaluation of protein fold comparison servers. *Proteins*, **54**, 260–270.
41. Orengo,C.A., Bray,J.E., Buchan,D.W., Harrison,A., Lee,D., Pearl,F.M., Sillitoe,I., Todd,A.E. and Thornton,J.M. (2002) The CATH protein family database: a resource for structural and functional annotation of genomes. *Proteomics*, **2**, 11–21.
42. Fischer,D., Elofsson,A., Rice,D. and Eisenberg,D. (1996) Assessing the performance of fold recognition methods. *Proc. Pacific. Symp. Biocomput.*, **96**, 300–318.
43. Harris,M.A., Clark,J., Ireland,A., Lomax,J., Ashburner,M., Foulger,R., Eilbeck,K., Lewis,S., Marshall,B., Mungall,C. *et al.* (2004) The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.*, **32**, D258–261.
44. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
45. Vendruscolo,M., Kussell,E. and Domany,E. (1997) Recovery of protein structure from contact maps. *Fold Des.*, **2**, 295–306.
46. Hobohm,U. and Sander,C. (1994) Enlarged representative set of protein structures. *Protein Sci.*, **3**, 522–524.
47. Liebman,M.N., Venanzi,C.A. and Weinstein,H. (1985) Structural analysis of carboxypeptidase A and its complexes with inhibitors as a basis for modeling enzyme recognition and specificity. *Biopolymers*, **24**, 1721–1758.
48. Porto,M., Bastolla,U., Roman,H.E. and Vendruscolo,M. (2004) Reconstruction of protein structures from a vectorial representation. *Phys. Rev. Lett.*, **92**, 218101.
49. Kinjo,A.R. and Nishikawa,K. (2005) Recoverable one-dimensional encoding of three-dimensional protein structures. *Bioinformatics*, **21**, 2167–2170.
50. Kinch,L.N. and Grishin,N.V. (2002) Evolution of protein structures and functions. *Curr. Opin. Struct. Biol.*, **12**, 400–408.
51. Soding,J. and Lupas,A.N. (2003) More than the sum of their parts: on the evolution of proteins from peptides. *Bioessays*, **25**, 837–846.
52. Meng,E.C., Polacco,B.J. and Babbitt,P.C. (2004) Superfamily active site templates. *Proteins*, **55**, 962–976.
53. Zhang,Y. and Skolnick,J. (2005) The protein structure prediction problem could be solved using the current PDB library. *Proc. Natl Acad. Sci. USA*, **102**, 1029–1034.
54. Korkegian,A., Black,M.E., Baker,D. and Stoddard,B.L. (2005) Computational thermostabilization of an enzyme. *Science*, **308**, 857–860.
55. Looger,L.L., Dwyer,M.A., Smith,J.J. and Hellinga,H.W. (2003) Computational design of receptor and sensor proteins with novel functions. *Nature*, **423**, 185–190.
56. Dwyer,M.A., Looger,L.L. and Hellinga,H.W. (2004) Computational design of a biologically active enzyme. *Science*, **304**, 1967–1971.
57. Kuhlman,B., Dantas,G., Ireton,G.C., Varani,G., Stoddard,B.L. and Baker,D. (2003) Design of a novel globular protein fold with atomic-level accuracy. *Science*, **302**, 1364–1368.
58. Gilbert,D., Westhead,D., Nagano,N. and Thornton,J. (1999) Motif-based searching in TOPS protein topology databases. *Bioinformatics*, **15**, 317–326.
59. Chou,C.C., Lin,T.W., Chen,C.Y. and Wang,A.H. (2003) Crystal structure of the hyperthermophilic archaeal DNA-binding protein Sso10b2 at a resolution of 1.85 Angstroms. *J. Bacteriol.*, **185**, 4066–4073.
60. Shehi,E., Granata,V., Del Vecchio,P., Barone,G., Fusi,P., Tortora,P. and Graziano,G. (2003) Thermal stability and DNA binding activity of a variant form of the Sso7d protein from the archeon *Sulfolobus solfataricus* truncated at leucine 54. *Biochemistry*, **42**, 8362–8368.
61. Lo Surdo,P., Walsh,M.A. and Sollazzo,M. (2004) A novel ADP- and zinc-binding fold from function-directed *in vitro* evolution. *Nature Struct. Mol. Biol.*, **11**, 382–383.
62. Keefe,A.D. and Szostak,J.W. (2001) Functional proteins from a random-sequence library. *Nature*, **410**, 715–8.
63. Krishna,S.S. and Grishin,N.V. (2004) Structurally analogous proteins do exist!. *Structure*, **12**, 1125–1717.
64. Hanzawa,H., de Ruwe,M.J., Albert,T.K., van Der Vliet,P.C., Timmers,H.T. and Boelens,R. (2001) The structure of the C4C4 ring finger of human NOT4 reveals features distinct from those of C3HC4 RING fingers. *J. Biol. Chem.*, **276**, 10185–11290.