BIOINFORMATICS ORIGINAL PAPER Vol. 24 no. 24 2008, pages 2872–2879
doi:10.1093/bioinformatics/btn545

Structural bioinformatics

Integrated search and alignment of protein structures

Ahmet Sacan[1,2,*], I. Hakki Toroslu[1] and Hakan Ferhatosmanoglu[2,3]

[1]Department of Computer Engineering, Middle East Technical University, Ankara, Turkey, [2]Department of Computer Science and Engineering and [3]Department of Biomedical Informatics, The Ohio State University, Columbus, OH, USA

Received on July 8, 2008; revised on September 26, 2008; accepted on October 19, 2008

Advance Access publication October 22, 2008

Associate Editor: Anna Tramontano

ABSTRACT

Motivation: Identification and comparison of similar three-dimensional (3D) protein structures has become an even greater challenge in the face of the rapidly growing structure databases. Here, we introduce Vorometric, a new method that provides efficient search and alignment of a query protein against a database of protein structures. Voronoi contacts of the protein residues are enriched with the secondary structure information and a metric substitution matrix is developed to allow efficient indexing. The contact hits obtained from a distance-based indexing method are extended to obtain high-scoring segment pairs, which are then used to generate structural alignments.

Results: Vorometric is the first to address both search and alignment problems in the protein structure databases. The experimental results show that Vorometric is simultaneously effective in retrieving similar protein structures, producing high-quality structure alignments, and identifying cross-fold similarities. Vorometric outperforms current structure retrieval methods in search accuracy, while requiring comparable running times. Furthermore, the structural superpositions produced are shown to have better quality and coverage, when compared with those of the popular structure alignment tools.

Availability: Vorometric is available as a web service at http://bio.cse.ohio-state.edu/Vorometric

Contact: sacan@cse.ohio-state.edu

1 INTRODUCTION

A tremendous amount of sequence and structure data is being produced with the motivation of deriving biological insights through analysis of similarities, differences and interactions among biological macromolecules. Whereas the sequence comparison methods are generally sufficient for comparing proteins that share a high level of similarity, structure comparison becomes essential in discerning more distant evolutionary relationships. Moreover, the spatial organization of the protein residues provides stronger clues into the biochemical function of the proteins than can be derived from the sequence information alone.

Pairwise structure alignment is the basic step for comparing protein structures. Finding the optimal alignment has been proven to be NP-hard (Lathrop, 1994), and several heuristic approaches have been developed in DALI (Holm and Sander, 1993), CE (Shindyalov and Bourne, 1998) and MAMMOTH (Ortiz *et al*., 2002). The rapidly increasing size of the protein databases, however, has rendered exhaustive pairwise structure alignment infeasible.

To overcome the difficulties presented by the database size, several strategies that aim to quickly identify relevant protein structures have recently been proposed. These strategies can best be summarized in terms of the choice of protein structure representation and the indexing method utilized for fast searching. ProGreSS (Bhattacharya *et al*., 2004) maps windows of protein backbone to a feature vector space using the curvature and torsion angles and the amino acid type information, and performs spatial indexing in this feature space. Protdex2 (Aung and Tan, 2004) represents the protein as a set of feature vectors of the contact regions among secondary structure elements (SSE) and uses an inverted-file index for searching. Yakusa (Carpentier *et al*., 2005) describes the protein structure as a sequence of its backbone dihedral α angles and uses a method analogous to BLAST for searching blocks of this sequence. 3D-BLAST (Tung *et al*., 2007) clusters the κ and α angles to reduce the description to an alphabet and constructs a BLOSUM-like substitution matrix for this backbone angle alphabet, so that BLAST algorithm can be used without any modifications.

Currently available protein structure search methods provide database filtering, but defer a detailed structural alignment to further analysis by external alignment methods. More importantly, they focus on finding proteins that share similar overall topology or secondary structure composition, and are not sensitive to detect residue-level non-local interactions. Such non-local interactions are especially important in detecting functionally or evolutionarily significant similarities among proteins that span multiple structural folds (Brown *et al*., 2006; Friedberg and Godzik, 2005).

In this study, Vorometric is proposed as an integrated approach to both search and alignment tasks. We collect residue interactions from the protein structures using Voronoi tesselation and build a database of these *residue environments*. For a query protein, similar residue environments are retrieved from the database and extended to obtain high-scoring segment pairs (HSPs), which are then used for structural superposition. We have developed a sensitive metric substitution matrix for accurate comparison of both amino acid and secondary structure information of related residue environments. Whereas an exhaustive search of similar residue environments in the database is prohibitive, our metric matrix has made distance-based indexing possible so that similar environments can be retrieved very efficiently. To the best of our knowledge, Vorometric is the first study

*To whom correspondence should be addressed.

 © The Author 2008. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oxfordjournals.org

Downloaded from https://academic.oup.com/bioinformatics/article-abstract/24/24/2872/196843 by MIDDLE EAST TECHNICAL UNIVERSITY LIBRARY user on 23 July 2020

employing distance-based indexing to protein structure data. The main benefits of our approach can be summarized as follows:

- The correspondences obtained from search and extension of residue environments endorse integrated and accurate structural superpositions, so that further structural alignment by external programs is no longer necessary.
- Unlike other protein structure search methods that at best capture the inter-SSE contacts, Vorometric provides contact sensitivity at the residue level.
- The hit and extend methodology inherently detects local, flexible structure alignments, a feature not commonly available in pairwise structure alignment methods.

We demonstrate the advantages and limitations of Vorometric using both quantitative performance evaluation on large-scale datasets and on several detailed case studies. The experimental results show that Vorometric outperforms other structure search tools, and at the same time, yields high-quality structural alignments that are comparable to or better than those produced by other structure alignment tools.

2 METHODS

Vorometric exploits the fact that related protein structures share similar residue–residue interactions, in identification and alignment of related proteins. The residue interactions are captured using Voronoi tesselation and represented as a sequential string of residues. We incorporate both the amino acid type and secondary structure information into this representation. The residue contacts from all the proteins are then compiled in a database and metric indexing is used for fast similarity search in this database. For a query protein, the contacts that are similar to those formed by its residues are searched in the database, and hits are extended for structural alignment. In the next few sections, we describe each of these steps in detail.

2.1 Representing the residue environments

Voronoi tesselation has been proposed as an effective method for extracting multi-body contacts from protein structures (Krishnamoorthy and Tropsha, 2003), and successfully utilized in packing analysis (Richards, 1974), protein folding (Gan *et al*., 2001), structure alignment and structural motif mining (Sacan *et al*., 2007). Ilyin *et al*. (2004) observed that structurally related proteins share common Voronoi contacts and used this observation to systematically match compatible tetrahedrons by shape, volume and backbone topology in order to obtain candidate seeds for structure alignment. Roach *et al*. (2005) and Birzele *et al*. (2007) use a different representation of the Voronoi contacts to obtain a sequential representation which allows direct use of dynamic programming. Roach *et al*. (2005) measure the compatibility of the contacts through discretization of the Voronoi edge lengths, whereas Birzele *et al*. (2007) use another level of dynamic programming to compare the residue contacts.

We acknowledge that a sequential representation of the residue environments is very effective for their comparison, and utilize a similar representation in this study. We use the location of C_α atoms to represent the amino acids of a protein structure as a set of points in 3D space. The region of space around each point closer to the enclosed point than any other point defines a Voronoi polyhedron (Fig. 1). Delaunay tessellation is obtained by connecting the points that share a Voronoi boundary. For each residue, we define the set of all of its Delaunay neighbors, ordered by their sequence number along the backbone, as its *environment* (also denoted as *contact string*). This definition of contacts encodes much of the geometric proximity information and provides an abstract description of the underlying geometry.

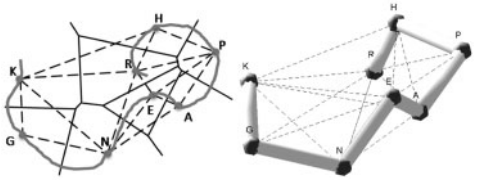


Fig. 1. Delaunay tessellation (dashed lines) of a set of points in 2D and 3D. The Voronoi diagram is shown for only 2D (solid lines). The 2D curve represents a projection of the 3D backbone segment from β2-microglobulin domain (3hla). The residue names are shown next to the C_α atoms.

The length of the contact strings, generated from a large structure dataset, has an average of 11.6 and a maximum of 23 residues.

We encode the amino acid type and the secondary structure assignment of the residues in each contact string. For instance, the contact string for the second, Histidine residue in Figure 1 (3D) is: $R_C H^*_C P_C A_C E_C K_C$, where the secondary structure state is given in subscript notation, and the central residue is marked with an asterisk '*'. We use DSSP (Kabsch and Sander, 1983) to obtain the secondary structure assignment for each residue and consider only the three primary states: α-helix (H), β-sheet (E), and turns (C).

2.2 Comparison of the contact strings

An SSE-enriched *distance matrix* (described below) is used to compare individual elements of two contact strings (e.g. comparing a helix forming Arginine, R_H, with a β-sheet forming asparagine, N_E). The optimal alignment that minimizes the edit distance between two SSE-enriched strings with respect to the distance matrix can be obtained using the classical dynamic programming method by Needleman and Wunsch (1970). Since the contact strings are relatively short, global alignment with linear gap penalty provides sufficient accuracy in comparing contact strings.

Note that, even though most of the related proteins share similar Voronoi contacts, slight differences in the backbone configurations, such as caused by the inherent noise in the crystallography experiment, or insertion and deletion of backbone segments, may induce significantly different Voronoi contacts. In order to increase the robustness of the comparison measure, we consider the contacts and the central residues separately when aligning two contact strings. The piecewise definition of the distance function is especially useful in recovering from the differences in Voronoi contacts of the residues flanking highly conserved structural cores. Furthermore, it ensures that the central residues, whose compatibility is being measured in the first place, are aligned together. Based on these considerations, the distance between two contact strings E and F is defined as follows:

$$D(E,F) = d(E^-,F^-) + \eta\, d(E^0,F^0) + d(E^+,F^+) \quad (1)$$

where d is the edit distance between two SSE-enriched sequences and η is a parameter used to adjust the importance of the similarity of the central residues (^{0}) compared with that of the contact residues preceding ($^-$) and following ($^+$) the central residues. A small value of η can stress the similarity of contacts better, whereas a large value would work better in recovering from differences in the contacts due to structural divergence. A very large value of η would disregard the similarity of the contacts, and would not be informative in identifying similar structures, generating many false positives that share the same central residue type. We determine a single value of η that is optimized for structure retrieval and alignment performance as described in Section 2.7. Note that the edit distance between the central residues is simply a lookup in the distance matrix and does not require dynamic programming.

2.3 Metric SSE-enriched distance matrix

Sellers (1974) has proved that if a metric distance matrix is used in the global alignment, then the resulting edit distance also forms a metric.

Downloaded from https://academic.oup.com/bioinformatics/article-abstract/24/24/2872/196843 by MIDDLE EAST TECHNICAL UNIVERSITY LIBRARY user on 23 July 2020

There has been a number of efforts to construct metric amino acid distance matrices (Zintzaras, 1999). On the other hand, a metric matrix that captures both the amino acid and SSE information is not available. Using an identity matrix is an obvious solution; however, the identity matrix is not sensitive to detect similarities between different types of amino acids.

We construct a 60×60 SSE-enriched distance matrix (M) using a weighted combination of a metric amino acid distance matrix (N) that we have previously derived from four-body Delaunay contact profiles of amino acids (Sacan and Toroslu, 2007) and a metric SSE exchange matrix (K) derived from an SSE similarity matrix (Wallqvist *et al*., 2000) using the inter-row distance method (Zintzaras, 1999). The elements of M are defined as follows:

$$M(\langle a,s\rangle,\langle b,t\rangle) = w_1 N(a,b) + w_2 K(s,t) \quad (2)$$

where a and b are types of amino acids, s and t are the SSE states and w_1,w_2 are positive weighing parameters to adjust the contributions of amino acid types and SSE states. In general, one would expect a smaller w_1 and a larger w_2 for proteins that are remote homologous, because the remote homologs usually maintain a well-conserved secondary structure despite their low sequence similarity. An analysis of the dependence of the w_1 and w_2 parameters on the sequence and structure similarity levels can be found in the Supplementary Material.

A distance matrix (or function) f is metric if the following properties are satisfied for any three elements x, y and z:

(1) *Positivity:* $f(x,y) \geq 0$

(2) *Identity:* $f(x,y) = 0$ iff $x = y$

(3) *Symmetry:* $f(x,y) = f(y,x)$

(4) *Triangle Inequality:* $f(x,z) \leq f(x,y) + f(y,z)$

Now, we show that M, which is a weighted combination of the metric matrices N and K is also metric.

(1) The weights and matrices in Equation (2) are all positive, which makes M to be positive.

(2) If $M(\langle a,s\rangle,\langle b,t\rangle) = 0$, then $N(a,b) = 0$ and $K(s,t) = 0$ from Equation (2). Moreover, $a = b$, $s = t$ because N and K satisfy *identity*. Then, it follows that $\langle a,s\rangle = \langle b,t\rangle$. The reverse condition is also true using the same premises.

(3) $w_1 N(a,b) + w_2 K(s,t) = w_1 N(b,a) + w_2 K(t,s)$ because both N and K are symmetric, therefore M is also symmetric.

(4) $M(\langle a,s\rangle,\langle b,t\rangle) + M(\langle b,t\rangle,\langle c,u\rangle)$
$= w_1(N(a,b) + N(b,c)) + w_2(K(s,t) + K(t,u))$
$\geq w_1 N(a,c) + w_2 K(s,u) = M(\langle a,s\rangle,\langle c,u\rangle)$, therefore M also satisfies triangle inequality.

Note that, the distance function D defined for the contact strings is similar to M, in that it is also composed of a weighted combination of functions that are metric. According to the properties shown above, both M and D are metric.

2.4 Indexing and searching contact strings

Having a metric distance function D to compare contact strings allows us to utilize distance-based indexing for efficient retrieval. The main idea in distance-based indexing is to organize and partition the data into a hierarchical structure based on distances to representative elements of the partitions at each level. A partition whose representative entry is too dissimilar to a query can then be pruned using the triangle inequality, without having to examine the rest of the entries in that partition. This allows an efficient and focused search over the data for entries similar to the query. [Please refer to Taskin and Ozsoyoglu (2004) for a survey of distance-based indexing methods.] While any metric indexing method can be used to index and search the contact strings, we have implemented the Slim-tree method (Traina *et al*., 2000) which achieves sufficient time and memory performance for the large datasets used in this study.

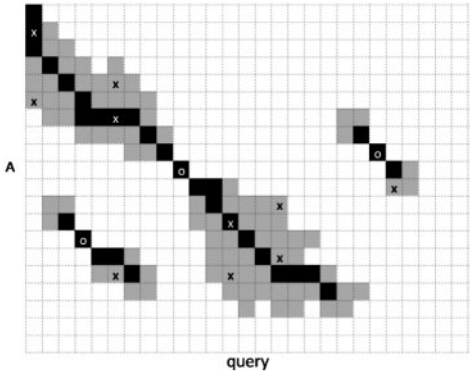


Fig. 2. Extension of the hits from a database protein A. The seeds being extended are marked with 'o', and those that are pruned are marked with 'x'. The gray area represents the cells that are explored by the dynamic programming and the black cells form the alignment paths of the HSPs.

We extract the contact strings from all of the protein structures in a dataset, and index them with respect to the distance function D. For a given query protein structure, we extract its contact strings, and search for similar entries in the database that are within the range δ from the query contact strings. The parameter δ provides a threshold on the similarity of the contact strings being searched. A loose threshold would capture the contact strings of all protein structures that are similar to the query but may also result in many false positive hits. Whereas, a tight threshold would seek only the proteins that share highly conserved structural cores with the query.

2.5 Generating HSPs

The pseudocode for generating HSPs from the contact string hits is outlined in Algorithm 1. The hits obtained for the individual residues of the query protein are first grouped based on which database proteins they belong to. These hits (also called *seeds*) correspond to a pair of residues, one from the query and one from the database protein, and are represented by the individual cells of the dynamic programming table, as illustrated in Figure 2. Please note that the substitution score of each residue pair is defined by the similarity of their contact strings, so the hit extension phase is, in fact, a 2nd level of dynamic programming.

Algorithm 1 Generate HSPs from contact string hits

```
Input: the contact string hits from the database
Output: HSPs: high scoring segment pairs
HSPs ← [];
foreach protein A that has contact string hits do
    H ← sort hits to A by their distance to query strings;
    foreach hit h ∈ H do
        if h is already explored in dynamic programming table then
            continue;
        hsp ← ExtendHit(h);
        if hsp can be merged into a previous hsp' ∈ HSPs then
            hsp' ← MergeHSP(hsp', hsp);
        else if Score(hsp) ≥ γ then
            add hsp to HSPs;
```

The extension heuristic employed for each seed is similar to that of BLAST sequence search tool (Altschul *et al*., 1997) in that we also construct gapped local alignments in both forward and backward directions and only consider the cells in the dynamic programming table whose score falls not more than a fraction of the best score yet found. However, we introduce several notable enhancements over the basic method that increase the efficiency while maintaining the same level of sensitivity.

Downloaded from https://academic.oup.com/bioinformatics/article-abstract/24/24/2872/196843 by MIDDLE EAST TECHNICAL UNIVERSITY LIBRARY user on 23 July 2020

The hits to a database protein *A* are sorted based on their distances to the query contact strings, such that the more similar hits, which are more likely to be part of the final HSPs, are explored first. Naturally, we would expect many seeds on the alignment path of an HSP; extending each of these seeds would be redundant because they would find the same HSP path. We therefore skip the seeds whose residue pairings have already been explored by the extension of the previous seeds (Fig. 2). This heuristic effectively eliminates about 42% of the seeds from consideration (based on randomized searches on the ASTRAL-25 database). Furthermore, to overcome the problem of generating many short HSPs, we merge a new HSP if its alignment path intersects with that of a previously generated HSP and if the merging produces a higher score than the individual HSP scores. This strategy results in merging 7% of the HSPs, which otherwise would have been generated as separate, shorter alignment segments.

2.6 Structure superposition

The residue correspondences defined by the HSP alignments are used to obtain a structural superposition of the query and the related database proteins. Finding the optimal transformation that minimizes the root mean squared deviation (RMSD) between two structural alignments can be computed very fast: linear in the size of the proteins (Kabsch, 1978). Following the iterative optimization procedure commonly employed by the structure alignment tools, we derive a new set of correspondences from the superposition by finding the local alignment that minimizes the total distance of the aligned residues, and then repeat the iteration. A single iteration improves the structural alignment by readjusting the underlying residue correspondences to better reflect the spatial proximity of the residues in the structural superposition. The procedure is repeated until the transformation matrix no longer changes. Because the initial correspondences defined by the HSP alignments already optimize the structural compatibility of the aligned residues, the algorithm converges fast; in only a few iterations.

2.7 Parameter optimization

Parameters used in Vorometric are optimized on an independent training set using the Nelder–Mead simplex method (Lagarias *et al*., 1998). The objective function used for the optimization was the geometric mean of the precision and recall values of the results returned by Vorometric and the TM-score (Zhang and Skolnick, 2004) of the structural alignments between the queries and the resulting proteins. The geometric mean penalizes imbalance in the individual quality values more than the arithmetic mean, and is considered a better metric for combining different quality measures.

The training set used for optimization was taken from the representative ASTRAL v1.73 database with 25% sequence identity (Chandonia *et al*., 2004). We removed all the families that were used as queries in the evaluation of the Vorometric reported below, and kept remaining families that had at least 10 domains. From 13 such families, we randomly selected 10 members and assigned one of them to be the query and compiled the rest into a dataset. The training data are available from the supplementary web site.

3 EXPERIMENTAL RESULTS

Since Vorometric is proposed as a protein structure database search tool that at the same time produces high-quality structure alignments, we compare its performance with that of both pairwise structure alignment and database search tools. In the next few sections, we first demonstrate that the structural alignments produced by Vorometric are in fact comparable to or better than those of other pairwise structural alignment tools. We then show on large-scale experiments, that the structures in the database that are similar to a query protein are retrieved correctly, using the SCOP classifications (Murzin *et al*., 1995) as the gold standard.

Table 1. Comparison of alignment quality on 10 difficult pairs

Method	RMSD (Å)	%N (query)	Quality (TM)
CE	3.17	83.4	0.60
SSAP	4.37	**88.1**	0.59
DaliLite	2.82	80.0	0.61
Vorolign[a]	**2.28**	51.7	0.56
Vorometric	3.02	84.8	**0.65**

[a]Vorolign reports alignments for multiple substitution matrices; here we use the SM-THREADER matrix (Dosztanyi and Torda, 2001), which gives the best results. The best values in each column are shown in bold.

The experiments are performed on a Pentium 2.6 GHz personal computer with 2 GB of main memory. For each protein structure used in this study, the contact strings are extracted only once, requiring an average of 21 s per protein. In each experiment described below, a new database index is built for the contact strings of the protein structures used in the experiment. The insertion of a single protein structure into the database index takes 22 ms on the average.

3.1 Quality of the structural alignments

In order to demonstrate the quality of the structural alignments generated by Vorometric, we used the *10 difficult pairs* of protein structures that have previously been used to evaluate structural alignment methods (Fischer *et al*., 1996). A difficult pair is defined as a structurally similar pair that has a low sequence similarity and that had proven difficult to align with the available methods. For each pair, we use one of the proteins as query to search against the database composed solely of the other protein, and report the top-scoring HSP alignment. We compare the structural alignments produced by Vorometric with those by other popular structural alignment tools (Table 1). Note that this small and specific set of proteins is used only to demonstrate the alignment quality of Vorometric for very difficult cases. A larger and more general evaluation of the alignment quality can be found in the next section.

The comparison is made using the RMSD between the superimposed structures, the percentage of the query protein aligned (%N), and the TM-score (Zhang and Skolnick, 2004). The RMSD of a superposition is defined as:

$$\mathrm{RMSD} = \sqrt{\frac{\sum_i d_i}{N}}$$

where d_i is the Euclidean distance between the i-th corresponding residues of the aligned proteins and N is the length of alignment. Note that a tradeoff exists between the coverage of the alignment (*%N*) and the RMSD error. It is generally possible to produce short structural alignments with very low RMSD error (e.g. aligning only a single residue from each protein would trivially achieve zero error). And naturally, a higher RMSD error is incurred for longer alignments. In response to the difficulty of interpreting both RMSD and *%N* simultaneously, TM-score attempts to summarize the quality of the alignment in a single scoring function and has been shown to agree with the results of human expert visual assessments. TM-score is defined as:

$$TM\text{-}score = \frac{1}{L_{\mathrm{target}}} \sum_i^N \frac{1}{1+(\frac{d_i}{d_0(L_{\mathrm{target}})})^2}$$

Downloaded from https://academic.oup.com/bioinformatics/article-abstract/24/24/2872/196843 by MIDDLE EAST TECHNICAL UNIVERSITY LIBRARY user on 23 July 2020



Fig. 3. Structural alignment produced by Vorometric for 1ten (dark) and 3hhrb (light). Aligned regions are shown thicker.

where L_{target} is the length of the query protein, d_i is same as above, and $d_0(L_{\text{target}})$ is a normalizing factor so that the average TM-score is not dependent on the size of the protein. TM-score ranges from 0 to 1, with 1 being a perfect superposition.

For the 10 difficult pairs, CE and DaliLite give comparable coverage and RMSD values, whereas SSAP produces slightly longer alignments with significantly worse RMSD values. For instance, the alignment produced by SSAP for the 1ede–1crl pair had the worst RMSD (9.25 Å) among all the alignments. 1ede and 1crl belong to the α/β Hydrolases superfamily and are relatively large proteins (310 and 534 residues, respectively), having eight β-strands wrapped around by 11 α-helices. SSAP relies on aligning residues that share similar inter-residue distances; the high number of contacts formed by the residues at the core of these proteins makes their alignment difficult by SSAP.

Vorometric produces better alignments than any other method as measured by the TM-score. The coverage of the alignments by Vorometric are as long as those of SSAP's, while Vorometric at the same time achieves the best average RMSD, when compared with CE and SSAP.

Vorolign (Birzele *et al*., 2007), which is a pairwise alignment method also based on Voronoi contacts, generates the shortest alignments (about 30% smaller than Vorometric) and consequently, achieves better RMSD. However, the average alignment quality evaluated by TM-score is poorer than the other methods. Furthermore, Vorolign fails to generate an alignment for the 1ten–3hhrb pair. Both 1ten and 3hhrb are in the Fibronectin type III family; 1ten is composed of only one domain of the immunoglobulin-like β-sandwich fold, whereas 3hhrb contains two such domains, one of which aligns well with 1ten (Fig. 3). We attribute Vorolign's failure to its sensitivity to differences in residue contacts introduced by the additional domain in 3hhrb.

3.2 Database search for similar proteins

Large-scale comparison of different structure alignment or search methods is in itself a serious undertaking which is neither straightforward, nor completely fair, because each such method uses different databases and accuracy measures [see Kolodny *et al*. (2005) for a comprehensive evaluation]. Furthermore, some methods are made available only as a web service, which makes large-scale experimentation with newly crafted datasets impossible, if not prohibitive. For these reasons, we use the same dataset used by Aung and Tan (2004) and Tung *et al*. (2007), and compare our results with those reported by them.

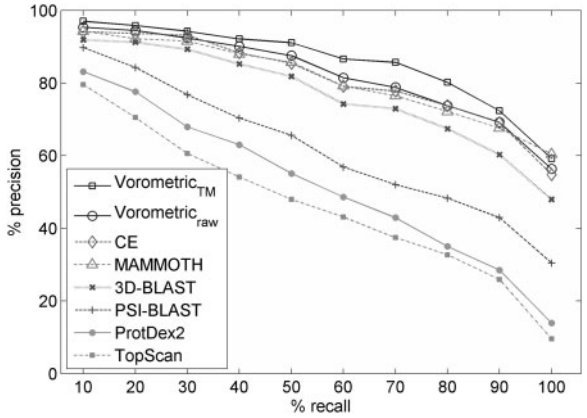


Fig. 4. Average precision–recall curves for 108 queries on the database of 34 055 proteins.

Table 2. Average precision and running times on the database of 34 055 proteins

	Avg. precision (%)	Time per query	Superposition
Vorometric-TM	**82.9**	51 s	**yes**
Vorometric-raw	79.7	44 s	no
CE	80.9	14 h	**yes**
MAMMOTH	80.8	1.6 h	**yes**
3D-BLAST	76.2	14 s	no
PSI-BLAST	61.8	**8 s**	no

Average precision is calculated as the mean of precision values for different recall levels. The time results for Vorometric are based on returning top 100 hits, performed on a Pentium 2.6 GHz personal computer. Vorometric-raw does not include the time spent for optimization of structural superposition, whereas Vorometric-TM does. The times for CE, MAMMOTH, 3D-BLAST and PSI-BLAST are approximate values interpolated from Tung *et al*. (2007) using the running times of CE as basis of comparison. The best values in each column are shown in bold.

The dataset consists of 34 055 proteins which cover about 90% of the ASTRAL database. From medium-size families, 108 queries are selected and have <40% sequence homology to each other. The precision of the results for different recall levels is shown in Figure 4 and summarized in Table 2. Even when the hits returned by Vorometric are ranked according to their raw HSP alignment scores (Vorometric-raw), the accuracy is better than other search methods and is comparable to that of detailed pairwise structure alignment methods CE and MAMMOTH, which indicates that the contact string representation and comparison used by Vorometric accurately captures the structural compatibility of the residues. When the results are ranked by their superposition TM-scores, Vorometric-TM achieves higher accuracy than any other method; giving slightly worse accuracy than MAMMOTH only above the 95% recall level.

Please note that CE (Shindyalov and Bourne, 1998) and MAMMOTH (Ortiz *et al*., 2002) are pairwise structure *alignment* methods, and for each query, they exhaustively scan the entire database. On the other hand, 3D-BLAST, ProtDex2 and TopScan (Martin, 2000) are structure *search* methods which are proposed as fast filters for similar structures, and do not produce structural superpositions. PSI-BLAST (Altschul *et al*., 1997) is a sequence profile search method, which interestingly performs better and faster than ProtDex2 and TopScan. Vorometric achieves the best precision

Downloaded from https://academic.oup.com/bioinformatics/article-abstract/24/24/2872/196843 by MIDDLE EAST TECHNICAL UNIVERSITY LIBRARY user on 23 July 2020

Table 3. Classification of ASTRAL v1.65-v1.67 difference set

	Family	Superfam	Fold	TM	%N	RMSD
Vorometric-TM	**90.7**	**94.9**	97.6	0.74	**87.2**	2.43
Vorometric-raw	85.9	91.2	97.0	—	—	—
Vorolign	86.4	92.4	**97.7**	0.74	76.3	**1.9**
CE	84.6	91.9	94.1	**0.77**	78.2	1.95
SSEA	60.8	68.9	75.6	—	—	—
BLAST	48.9	52.5	52.8	—	—	—

Vorolign and CE scan only the top 250 proteins returned by SSEA. The classification accuracy and the structural alignment metrics are based on top-hit assignments and alignments. The best values in each column are shown in bold.

while running in a comparable time scale as the other database search methods, and additionally produces detailed structural superpositions for the returned hits.

3.3 Protein classification

Another task that is closely related to the structure similarity search is to identify the structural or functional class of a protein via comparison with already annotated set of protein structures. In order to evaluate the classification performance, we used the dataset previously used by Birzele *et al*. (2007), where the difference set between SCOP v1.67 and v1.65 are queried against the ASTRAL-25 v1.65 containing 4358 proteins. The classification performance is measured as the percentage of the 979 query proteins correctly classified when compared with their actual classifications in SCOP v1.67.

Although more elaborate voting schemes are possible for this task, the most commonly employed strategy is to assume that the query has the same class as the top-1 hit returned from a database search. In order to provide a fair comparison, we also use the top-1 hit for assignment. Vorometric-TM achieves the best classification accuracy in family and superfamily levels (Table 3), and only slightly worse accuracy than Vorolign at the fold level.

Note that the average structural divergence between the queries and their top hits are less for this dataset than that of the 10 difficult pairs discussed above, which results in less pronounced differences in the alignment qualities. The average sequence identity of the 10 difficult pairs is 10%, whereas that between the queries used in this classification study and their best matches is 20%. While this may suggest that Vorometric performs better than Vorolign and CE when the level of homology is lower, we did not observe a consistent correlation between the comparative alignment quality between these methods and the level of homology of the proteins being aligned. Nevertheless, Vorometric-TM produces longer alignments, while maintaining similar TM-score alignment quality.

Vorometric-raw, which uses the raw HSP alignment scores and does not generate structural superimpositions has similar classification accuracy as Vorolign and CE. SSEA (Fontana *et al*., 2005) uses alignment of secondary structure elements to search the database, whereas BLAST is based on local alignment of primary sequences. The classification by these two database search methods are significantly worse than other methods. Please note that due to time constraints, Birzele *et al*. (2007) use SSEA to prefilter the database and use only the top 250 proteins to perform detailed pairwise structure alignment by Vorolign and CE. On the contrary, the integrated approach we employ in Vorometric relieves the dependence on prefiltering the database with a coarse-level retrieval method.

A number of the misclassifications by Vorometric were due to low quality of the query entries. One of the extreme cases is 1oau:I; 85% of whose residues were not located in the X-ray experiment. A more subtle example is the 1r1g:A short-chain of scorpion neurotoxin, whose few missing residues cause the structural alignment with 1aho:A domain, a long-chain scorpion toxin of the same superfamily (TM-score 0.62, 28% sequence identity), better than that with the correct family member 1jlz:A (TM-score 0.28, 48% identity).

A large fraction of the other misclassifications was due to the cross-fold similarities, especially in highly conserved domains such, as the immunoglobulin-like β-sandwich, zinc-finger and OB-fold. It must also be noted that the SCOP classifications are based not only on structural similarity, but also on functional and sequence similarity considerations, and even on the dimerization state of the proteins (e.g. the distinction between c.3.1.1 and c.3.1.5 families). As such, even though Vorometric places the structurally most similar protein as the top hit, it can be evaluated as a misclassification. For instance, 1urf:A of the a.2.6.1 family is structurally aligned better with 1lrz:A1 of the a.2.7.4 family, instead of 1cxz:B of the a.2.6.1 family. In most such cases, the correct family member were among the top few hits, and a simple analysis of the sequence homology was able to identify it correctly, indicating that a more elaborate strategy considering the top-k hits can be developed for highly accurate and fully automated classification of proteins.

3.4 Cross-fold similarities

We remark that there is no obvious or unambiguous way of clustering the proteins into discrete groups, and a significant number of overlaps will inevitably exist between proteins that are treated as unrelated based on hierarchical classification schemes (Kolodny *et al*., 2006). While the ability to replicate these classifications demonstrate the performance of the structure search and alignment methods and is useful in functional annotation, we believe that the ability to identify the cross-fold similarities is also as important in identifying more distant evolutionary and functional relationships that may help to understand the biochemical mechanisms of the particular biological function. While a systematic and exhaustive analysis of such cases is beyond the scope of this study, here we present a few examples to demonstrate that Vorometric is able to identify such relationships. The examples are also available on the Supplementary web site.

When Vorometric is used to query the first Ferrodoxin domain of the small subunit of FDH (1h0h:B,d.58.1.5), the second high scoring hit is the immunoglobulin-binding domain of protein L (1hz6:A,d.15.7.1) (Fig. 5a). The similarity between these two proteins has previously been used to put forth a mechanism of structural drift during evolution (Krishna and Grishin, 2005). Other significant cross-fold similarities were found between Beta-D-xylosidase (d1uhva1,b.71.1.2) and Chondroitin ABC lyase I (d1hn0a3,b.24.1.1), and between sucrose phosphorylase (d1r7aa1, b.71.1.1) and acidic mitochondrial matrix protein p32 (1p32:A,d.25.1.1).

We have also observed that many of the toxins share a significant structural similarity with proteins whose normal function is critical for the organism. The similarity between the assassin bug toxin AD01 (1lmr:A,g.3.6.3) and human insulin-like growth

Downloaded from https://academic.oup.com/bioinformatics/article-abstract/24/24/2872/196843 by MIDDLE EAST TECHNICAL UNIVERSITY LIBRARY user on 23 July 2020

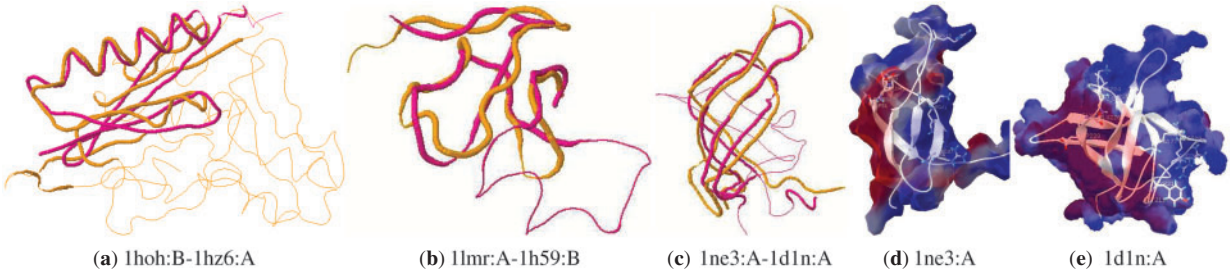


Fig. 5. Examples of cross-fold similarities.

factor-binding protein-5 (1h59:B,g.3.9.1) shown in Figure 5b; between the Chinese scorpion neurotoxin (1r1g:A,g.3.7.2) and human transcription initiation factor TAF(II)18 (1bh9:A,a.22.1.3) and between short-chain scorpion Cobatoxin 1 (1pjva:A,g.3.7.2) and human methylation-dependent transcriptional repressor MBD1/PCM1 (1owt:A,d.230.3.1) are only some of such instances. We believe that a detailed analysis of these similarities may provide insight into the biochemical mechanism of the toxins, and that of the respective proteins they mimic.

Perhaps the most striking similarity we have discovered is that between the ribosomal protein S28e (1ne3:A,b.40.4.5) and the translation initiation factor IF2/eIF5b (1d1n:A,b.43.3.1). The SCOP classification of these proteins differ at the fold level; similarly, the CATH classification (Orengo *et al*., 1997) differ at the topology level (with CATH classifications 2.40.50.140 and 2.40.30.10 for 1ne3:A and 1d1n:A, respectively). While there is a significant body of work comparing IF2 with the translation elongation factors EF-tu and EF-G, it has not been compared with S28. It has been established that the initiation and elongation factors bind aminoacyl-tRNA, carry it to the ribosome, and detach from the ribosome after a conformational change caused by GTP hydrolysis, leaving the aminoacyl-tRNA attached at the A-site (Andersen *et al*., 2003; Meunier *et al*., 2000). On the other hand, the function of the ribosomal protein S28 is not characterized, although it has been conjectured to bind RNA, based on the analogy of its surface to the OB-fold proteins (Wu *et al*., 2003). The Vorometric for S28e finds IF2 as a significant hit; their structure alignment (Fig. 5c–e) shows a conserved [RK]EGD motif which provides a negatively charged site on both proteins. A comparison of their surface electrostatic potentials reveals a large, positively charged, arginine-rich face on both proteins. The structure and surface similarities suggest that the ribosomal protein S28 may be responsible for taking over the aminoacyl-tRNA from the IF2 as it is being detached from the ribosome, and support the codon–anticodon binding as the peptidyl-tRNA is translocated from A-site to the P-site.

4 DISCUSSION

The pairwise structural alignment methods hitherto proposed rely on coarse-level filtering methods to scan the database of protein structures for candidates that are worthy of alignment. We have introduced Vorometric, a fast protein structure database search and alignment tool that uses the same sensitive representation of residue interactions for both identifying similar proteins and generating high-quality structural alignments. The heuristic that structurally similar proteins share similar residue interactions is exploited through a metric comparison of these interactions which has allowed efficient distance-based indexing and retrieval of related proteins.

The additional accuracy achieved by Vorometric does not incur significant time and memory requirements. The whole index structure for the large dataset of 34 055 proteins is constructed incrementally in 12 min, requiring less than 600 MB, and is kept in the main memory for fast access. Note that the index structure allows efficient insertions and deletions, which means that one does not need to rebuild the whole database index structure for the addition of a new protein or the removal of an obsolete protein. The querying speed achieved by the distance-based indexing method is complemented by the hit-extension strategy, which allows fast exploration of the search space by effectively pruning redundant or unpromising hits. The search of a query protein against a large database takes less than a minute, including detailed superposition of the retrieved proteins.

Evaluation of Vorometric on large-scale datasets shows that it provides the accuracy of pairwise structural alignment tools and the speed of database search tools. Vorometric performs better than other methods on the database search and classification tasks and produces longer, high-quality structure alignments, relieving the dependence on separate structural alignment tools. Finally, Vorometric successfully identifies cross-fold similarities between proteins so that distant evolutionary and functional relationships can be discerned.

The representation of residue environments used in Vorometric for indexing and alignment provides fast and effective structure retrieval and alignment but does come with its own drawbacks. Specifically, homologous proteins whose sequences are circularly permuted (Jung and Lee, 2001) or inverted (Ausiello *et al*., 2007) and non-topological structural motifs (Alesker *et al*., 1996) are not captured by Vorometric. This is a common problem of most structure search or alignment methods and is mainly due to the limitation of the dynamic programming used for seed extension or alignment. In Vorometric, additionally, the linear representation of the residues in contact strings limits the method to a sequence-order-dependent comparison of different environments. We are currently exploring different ways of overcoming this limitation; extracting multiple representations for a single environment, including both forward and reverse sequence order, provides promising results for handling inverted sequences but still fails to handle topology-independent similarities.

Note that Vorometric is a specific implementation of a more general and extensible framework. Particularly, different substitution matrices or distance functions can be developed and used,

Downloaded from https://academic.oup.com/bioinformatics/article-abstract/24/24/2872/196843 by MIDDLE EAST TECHNICAL UNIVERSITY LIBRARY user on 23 July 2020

provided that they satisfy the metric properties or that they allow other efficient indexing strategies. The substitution matrix by Sacan and Toroslu (2007) was the matrix of choice in this study due to its accuracy in alignment tasks, and its structure search and alignment performance in our preliminary studies (data not shown); the algorithm would easily accommodate other metric matrices. The extension phase of Vorometric can also incorporate other candidate evaluation methods (such as structural compatibility) to further filter the seed contact strings. Finally, even though we have used a single set of optimized parameters in Vorometric, it is possible to obtain and use different parameter sets for various homology levels or for different secondary structure or topological classifications.

Funding: US National Science Foundation (Grants IIS-0546713 and DBI-0750891); and Turkish Scientific and Research Council (TUBITAK) (Grant 107E173).

Conflict of Interest: none declared.

REFERENCES

Alesker,V. *et al*. (1996) Detection of non-topological motifs in protein structures. *Protein Eng.*, **9**, 1103–1119.

Altschul,S. *et al*. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.

Andersen,G.R. *et al*. (2003) Elongation factors in protein biosynthesis. *Trends Biochem. Sci.*, **28**, 434–441.

Aung,Z. and Tan,K. (2004) Rapid 3D protein structure database searching using information retrieval techniques. *Bioinformatics*, **20**, 1045–1052.

Ausiello,G. *et al*. (2007) Local comparison of protein structures highlights cases of convergent evolution in analogous functional sites. *BMC Bioinformatics*, **8** (Suppl. 1), S24.

Bhattacharya,A. *et al*. (2004) Progress: simultaneous searching of protein databases by sequence and structure. *Pac. Symp. Biocomput.*, **9**, 264–275.

Birzele,F. *et al*. (2007) Vorolign: fast structural alignment using Voronoi contacts. *Bioinformatics*, **23**, e205–e211.

Brown,S. *et al*. (2006) A gold standard set of mechanistically diverse enzyme superfamilies. *Genome Biol.*, **7**, R8.

Carpentier,M. *et al*. (2005) Yakusa: a fast structural database scanning method. *Proteins*, **61**, 137–151.

Chandonia,J. *et al*. (2004) The ASTRAL compendium in 2004. *Nucleic Acids Res.*, **32**, 189–192.

Dosztanyi,Z. and Torda,A. (2001) Amino acid similarity matrices based on force fields. *Bioinformatics*, **17**, 686–699.

Fischer,D. *et al*. (1996) Assessing the performance of fold recognition methods by means of a comprehensive benchmark. *Pac. Symp. Biocomput.*, World Scientific Publishing, Hawaii, pp. 300–318.

Fontana,P. *et al*. (2005) The SSEA server for protein secondary structure alignment. *Bioinformatics*, **21**, 393–395.

Friedberg,I. and Godzik,A. (2005) Connecting the protein structure universe by using sparse recurring fragments. *Structure*, **13**, 1213–1224.

Gan,H. *et al*. (2001) Lattice protein folding with two and four-body statistical potentials. *Proteins*, **43**, 161–174.

Holm,L. and Sander,C. (1993) Protein structure comparison by alignment of distance matrices. *J. Mol. Biol.*, **233**, 123–138.

Ilyin,V.A. *et al*. (2004) Structural alignment of proteins by a novel TOPOFIT method, as a superimposition of common volumes at a topomax point. *Protein Sci.*, **13**, 1865–1874.

Jung,J. and Lee,B. (2001) Circularly permuted proteins in the protein structure database. *Protein Sci.*, **10**, 1881–1886.

Kabsch,W. (1978) A discussion of the solution for the best rotation to relate two sets of vectors. *Acta Crystallogr.*, **A34**, 827–828.

Kabsch,W. and Sander,C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577–2637.

Kolodny,R. *et al*. (2005) Comprehensive evaluation of protein structure alignment methods: scoring by geometric measures. *J. Mol. Biol.*, **346**, 1173–1188.

Kolodny,R. *et al*. (2006) Protein structure comparison: implications for the nature of fold space, and structure and function prediction. *Curr. Opin. Struct. Biol.*, **16**, 393–398.

Krishna,S.S. and Grishin,N.V. (2005) Structural drift: a possible path to protein fold change. *Bioinformatics*, **21**, 1308–1310.

Krishnamoorthy,B. and Tropsha,A. (2003) Development of a four-body statistical pseudo-potential to discriminate native from non-native protein conformations. *Bioinformatics*, **18**, 1540–1548.

Lagarias,J.C. *et al*. (1998) Convergence properties of the nelder-mead simplex method in low dimensions. *SIAM J. Optim.*, **9**, 112–147.

Lathrop,R. (1994) The protein threading problem with sequence amino acid interaction preferences is NP-complete. *Protein Eng.*, **7**, 1059–1068.

Martin,A. (2000) The ups and downs of protein topology: rapid comparison of protein structure. *Protein Eng.*, **13**, 829–837.

Meunier,S. *et al*. (2000) Structure of the fMet-tRNAfMet-binding domain of B.stearothermophilus initiation factor IF2. *EMBO J.*, **19**, 1918–1926.

Murzin,A.G. *et al*. (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.*, **247**, 536–540.

Needleman,S.B. and Wunsch,C.D. (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol Biol.*, **48**, 443.

Orengo,C. *et al*. (1997) CATH: a hierarchic classification of protein domain structures. *Structure*, **5**, 1093–1108.

Ortiz,A. *et al*. (2002) Mammoth (matching molecular models obtained from theory): an automated method for model comparison. *Protein Sci.*, **11**, 2606–2621.

Richards,F. (1974) The interpretation of protein structures: total volume, group volume distributions and packing density. *J. Mol. Biol.*, **82**, 1–14.

Roach,J. *et al*. (2005) Structure alignment via Delaunay tetrahedralization. *Proteins*, **60**, 66–81.

Sacan,A. and Toroslu,I.H. (2007) Amino acid substitution matrices based on 4-body Delaunay contact profiles. In *IEEE Trans. of the 7th International Symposium on Bioinformatics and Bioengineering (IEEE-BIBE2007)*. Boston, MA, USA, pp. 796–802.

Sacan,A. *et al*. (2007) LFM-pro: a tool for detecting significant local structural sites in proteins. *Bioinformatics*, **23**, 709–716.

Sellers,P. (1974) On the theory and computation of evolutionary distances. *J. Appl. Math.*, **26**, 787–793.

Shindyalov,I. and Bourne,P. (1998) Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Eng*, **11**, 739–747.

Taskin,M. and Ozsoyoglu,Z.M. (2004) Improvements in distance-based indexing. *In IEEE Proceedings of the 16th International Conference on Science and Statistic. Database Management*. Washington, DC, USA, pp. 161–170.

Traina,J.C. *et al*. (2000) Slim-trees: High performance metric trees minimizing overlap between nodes. In *Proceedings of the 7th International Conference on Extending Database Technology*. Springer, Konstanz, Germany, pp. 51–65.

Tung,C.-H. *et al*. (2007) Kappa-alpha plot derived structural alphabet and BLOSUM-like substitution matrix for rapid search of protein structure database. *Genome Biol.*, **8**, R31.1–R31.16.

Wallqvist,A. *et al*. (2000) Iterative sequence/secondary structure search for protein homologs. *Bioinformatics*, **16**, 988–1002.

Wu,B. *et al*. (2003) Solution structure of ribosomal protein S28E from Methanobacterium thermoautotrophicum. *Protein Sci.*, **12**, 2831–2837.

Zhang,Y. and Skolnick,J. (2004) Scoring function for automated assessment of protein structure template quality. *Proteins*, **57**, 702–710.

Zintzaras,E. (1999) A comparison of amino acid distance measures using procrustes analysis. *Comput. Biol. Med.*, **29**, 283–288.

Downloaded from https://academic.oup.com/bioinformatics/article-abstract/24/24/2872/196843 by MIDDLE EAST TECHNICAL UNIVERSITY LIBRARY user on 23 July 2020