Protein Structure Comparison and Alignment Using Residue Contexts

Tobias Sayre[1] and Rahul Singh[2]
[1]Department of Biology, [2]Department of Computer Science
San Francisco State University, San Francisco, CA 94132
[1]tobias@sfsu.edu, [2]rsingh@cs.sfsu.edu

Abstract

We introduce a method for comparing protein structures using the notion of residue contexts based on protein C_α-atom backbones. The residue context is derived from the set of vectors from a given C_α-atom to each other C_α-atom in the molecule. A three-dimensional histogram is generated from these vectors, containing a relative distribution of the other C_α-atoms for each C_α-atom on the backbone for a protein. Histograms are compared using the χ^2 test, resulting in the cost for matching any two given C_α-atoms in a pair of protein molecules. An optimal alignment is made using the Smith-Waterman algorithm, and a score is calculated based on the length of the alignment and the RMSD, yielding a best alignment that can be displayed in an interactive user interface. Resulting alignments are compared with alignments generated by CTSS, DALI, and CE, yielding different aligned protein regions.

1. Introduction

Comparison of protein structures is a fundamental problem in drug discovery and structural molecular biology. A drug molecule that has unwanted interaction with a target protein molecule may work well with a similar molecule identified by protein structure comparison. Conversely, a drug targeting one protein may have unintended consequences for proteins with similar recognition sites. Better understanding of the relationship between protein structure and function could also lead to deeper understanding of the molecular basis of diseases.

Protein structure is more highly conserved than protein sequence, and structural similarity is often associated with functional similarity or common phylogeny [2], making protein comparison crucial to protein structure prediction, classification of proteins into families and folds, finding relevant motifs, and mapping phylogenetic trees. The utility of protein structural comparison has led to attempts to classify known protein structures into a map of the protein universe [3]. To this end, structural databases have been compiled including Structural Classification of Proteins (SCOP) [4], Families of Structurally Similar Proteins (FSSP) [5], Molecular Modeling DataBase (MMDB) [6], and Class, Architecture, Topology and Homologous superfamily (CATH) [7]. Compounding the problem is the rapidly increasing number of known protein structures, exemplified by the count of protein structures listed in the Protein Databank (PDB), in which there are currently more than forty-two thousand protein structures.

Structural alignment is an NP-hard problem [8], leading most research into comparison algorithms to take advantage of heuristics for reducing the complexity of a protein molecule, typically reducing the molecule to the set of positions of the C_α-atoms of the protein's amino acid residues. Many widely used techniques [9, 10] use this approach, treating global structure as the set of inter-atomic distances. Other methods [11] represent protein structure as the set of secondary structure elements (SSEs), either α-helices or β-strands. Another set of methods uses localized features for structure representation, using techniques such as geometric hashing [12, 13] or residue spherical shell neighborhood [14] for defining a local feature.

An approach that relies solely on protein structure geometry as its descriptor will fail to fully capture the underlying biochemical properties that enable molecular interactions. In this paper, we present a method that uses the notion of the *residue context* of a protein structure that can be used to capture both the geometry and biochemical properties of a protein molecule. To capture geometry, the backbone of C_α-atoms is used, and to capture biochemical information, appropriate properties can be associated with each residue, yielding a rich molecular representation. Given the description of a protein in terms of its residue contexts, we also propose an efficient algorithm that finds meaningful correspondences between protein structures.

Residue context is defined as the set of vectors from a given C_α-atom to each other C_α-atom in the

978-0-7695-3096-3/08 $25.00 © 2008 IEEE
DOI 10.1109/WAINA.2008.22

molecule, or for a higher resolution, every other non-H atom in the molecule. This residue context of an atom can be succinctly described by a log-polar histogram that contains a relative distribution of other atoms on a protein backbone. In this manner, residue context inherently captures significance of the localized 3D environment of residues. Histograms are compared using the chi-squared test, resulting in the cost for matching any two given C_α-atom in a pair of protein molecules. Finally, an optimal alignment is made using the Smith-Waterman algorithm [15], and the aligned substructures are superimposed for visualization and calculation of RMSD.

This method includes a number of important features that contribute to its potential for use in protein structure comparison. First, the computability is efficient, because the extremely rich descriptor of all of the interatomic features is reduced to a histogram encoding feature distributions. This compact representation allows for rapid calculation of alignments. Residue context also captures the greater influence of local atoms on prediction of binding sites where usual methods calculate unweighted interatomic distances as the principal feature of their descriptors. The incorporation of data from amino acid side chains both geometric and biochemical also deepens the value of residue context for a given C_α-atom. All of these features of residue context correspond to our understanding of the biological reality of a protein that does not depend so much on the structure of the backbone as the qualities across the surface of the molecule.

The structural comparison proposed here is rotationally and translationally invariant and the aligned features are local to each protein, increasing the chances of finding a biologically relevant match. The alignment is robust, i.e. resilient to small perturbations of C_α-atom positions, and the descriptors are compact, taking a very rich set of inter-atomic distances and capturing that information in a much smaller histogram. Finally, the method takes into account biochemical features of amino-acid residues, ensuring that consequential alignments are made.

2. Prior Work

Prior work in protein structure alignment has produced many methods that have attempted to obtain a relevant alignment of two protein structures. Early work in this area included the DALI method [9], which simply considers the interatomic distances between all C_α-atoms along the protein backbone. Other approaches have been made using combinatorial extension of aligned fragment pairs [10]. More recently, published methods include approximate structural alignment achieved in polynomial time [16], a method that defines protein structural alignment as a mixed integer programming (MIP) problem, and using a mean field annealing technique [17]. Another technique uses the TM-score, which is up to 20-fold faster than some popular methods [18]. Towards the goal of finding biologically relevant alignments, other recent work has attempted to find structurally similar proteins that are diverse in chain-topology [19]. Several methods have also treated protein molecules as sets of secondary structure elements [20], which can apply other information such as curvature and torsion of the C_α-atom backbone represented as a spline [1], or probabilistic methods [21]. Finally, there have also been attempts to organize known protein structures into a space such that similar proteins are grouped. One study [22] using the Monte Carlo algorithm found several *fold attractors*, grouping proteins largely based on secondary structure element composition. Another method [23] clusters proteins on a basis of inter-C_α distances. Most of these types of universal methods allow visualization of proteins on 3D axes, including a method using SSE triplets [11], using DALI [24, 25], and gene ontology functional classification [3].

In prior research, the method technically closest in spirit to residue context is CTSS [1], which identifies local similarity in curvature and torsion of the protein backbone represented as a spline. The general idea of these methods is to identify locally similar features, and to extend the alignment of residues until the longest possible alignment is found within a given superimposition metric (usually RMSD). The notion of residue contexts differs subtly but significantly from CTSS. The geometric features in CTSS are highly localized in that they do not provide a complete representation of the local environment of the C_α-atoms. Furthermore, the spacing of C_α-atoms leads to numerical difficulties in accurately computing values of curvature and torsion that require computation of derivatives. Residue context does not suffer from these drawbacks and as results in the experimental section show, often provides alignments that have lower RMSD than CTSS.

3. Method

The idea of residue contexts builds on research directed at shape matching in computer vision [26]. In residue context, a protein molecule is considered as the set of its C_α-atoms. While this information does not tell us everything about the molecule, it gives a reasonable representation of the structure of the protein. A more detailed depiction of the molecule can be obtained by using all of the non-hydrogen atoms in the side-chains. The goal of this approach is to find an alignment between substructures of the C_α-atoms backbones of two molecules such that there is an atom p_i on the first molecule that corresponds to an atom q_j on the second molecule.

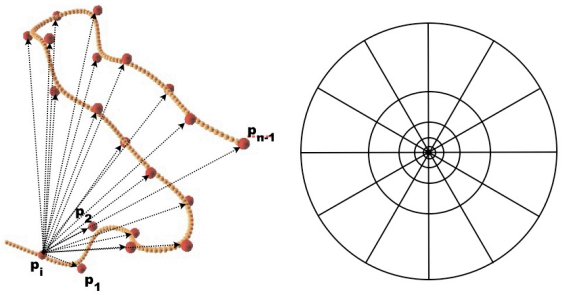
3.1 Defining residue context

Residue context is defined for each atom as the set of relative positions of every other atom in the molecule (Figure 1). This set of $n - 1$ vectors describes the arrangement of the full C_α-atom backbone relative to the reference atom. Using such a rich descriptor for every atom ensures that the biological significance of protein structure is captured.

This method can be used either to obtain the global best alignment between two protein structures, or a local alignment between a subset of amino acid residues. Many biological problems require global alignment, such as phylogenetic mapping and other evolutionary biology questions. However, for most applications of protein structure alignment in biomecidine and pharmacology, the interest is in finding similar active and binding sites within proteins. These types of approaches require methods that determine the best local alignment between two molecules. A histogram is constructed for each C_α-atom, containing the distribution of positions of other C_α-atoms in the molecule based on the set of vectors from p_i to each other C_α-atom p_k:

$$h_i(k) = |\{q \neq p_i : (q - p_i) \in bin(k)\}| \qquad (1)$$

In Eq(1), |.| denotes the size of the set. This yields a three-dimensional histogram for each atom, with binning in log-polar space to give higher sensitivity to closer atoms than distant atoms. Giving higher precision to atoms closest to p_i preserves the regional aspects of protein geometry that are important in identifying relevant binding and active sites. Rotational invariance is maintained by using a relative frame, based on the vector from p_i to $p_{(i+1)}$, in the N-terminal direction.



(a) (b)

Figure 1. (a) Residue context for an atom p_i is defined as the set of vectors to each other atom p_k. (b) The vectors are binned in three-dimensional log-polar space, i.e. a 3D version of the figure, giving higher sensitivity to atoms closest to p_i.

3.2 Pairwise Comparison and Normalization

In constructing a structural alignment between two protein molecules, we define a cost of matching two C_α-atoms p_i and q_j, where $C_{ij} = C(p_i, q_j)$. To calculate the cost, the χ^2-test is used to make an all-by-all comparison of each histogram from each molecule. The cost of matching two histograms is equal to the χ^2 distance between the histograms, yielding a two-dimensional matrix of χ^2 distances between each C_α-atom on each molecule:

$$C_{ij} \equiv C(p_i, q_j) = \frac{1}{2}\sum_{k=1}^{K}\frac{[h_i(k) - h_j(k)]^2}{h_i(k) + h_j(k)} \qquad (2)$$

The cost may also include an additional value representing the biochemical difference between two amino acid residues, based on hydrophobicity, charge, size, or function as a hydrogen-bond donor or acceptor. The strength in adding this type of matching cost is that each protein molecule has a descriptor based on a combination of geometric and biochemical information.

The cost matrix C_{ij} is normalized using the interval [*low*, *high*] for use in the dynamic programming local alignment part of the method, with *low* set to a default of *-10.0*, and *high* set to *20.0*. These values comprise a range similar to the PAM matrix (31). The final matrix M is then computed using the following equation:

$$M_{ij} = low + \frac{-C_{ij} + (256\sqrt{2} + c)}{256\sqrt{2} + 2c} \times (high - low) \qquad (3)$$

3.3 Alignment and Superimposition

Given a normalized cost matrix M containing the difference in residue context between each C_α-atom in each protein molecule, an alignment of protein backbones is made using the Smith-Waterman dynamic programming algorithm [15]. An affine gap cost model is used, where opening a gap and extending a gap have different costs, the defaults being *14* and *10* respectively. The best local alignment may have gaps in places where the protein backbone turns at sharp angles, potentially giving the alignment a high RMSD. Therefore, an additional step of superimposing the two aligned regions is applied. Given a set of corresponding C_α-atoms, the optimal rotation and translation are computed using a fast non-iterative least-squares solution from [1] that uses the singular value decomposition (SVD) with some adjustments to ensure a accurate rotation matrix. The final score equals the length of alignment divided by RMSD.

The best scoring alignment is then displayed as the superimposition of the subset of residues from each molecule in the alignment. Superimposition also lends itself to easy visualization of the alignment, which can then be used to make an intuitive check of

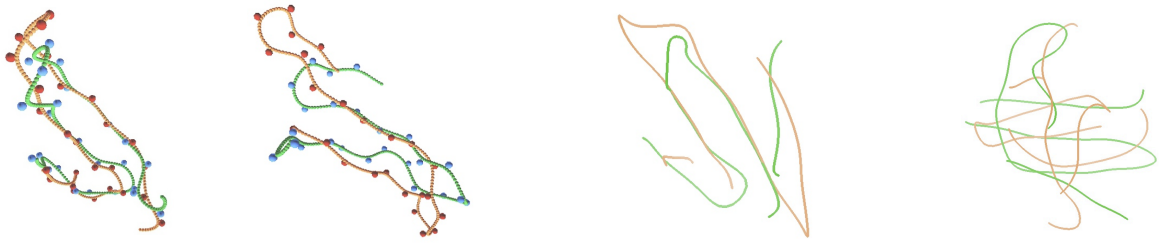
Table 1. Comparison of alignments of several proteins made by different algorithms.

PDB IDs	Protein Lengths	% Sequence Identity	Residue Context RMSD	Residue Context Alignment Length	CTSS RMSD	CTSS Alignment Length	DALI RMSD	DALI Alignment Length	CE_{medium} RMSD	CE_{medium} Alignment Length
2AQM 2C9V	154 153	29	1.9	43	2.3	40	2.2	140	2.0	139
1BYI 2IGD	224 61	5	1.5	25	1.5	21	3.6	41	5.8	48
1SBY 1ZK4	254 251	23	2.6	43	3.6	186	2.8	222	2.4	215
1PJX 2IIM	314 62	6	2.7	26	7.3	24	4.4	33	4.9	48
2FBA 2G58	492 121	5	2.7	25	3.0	26	4.0	65	6.0	80
1K5N 1L9L	276 74	5	2.3	21	6.0	30	4.9	39	5.5	48
1H97 1KQP	147 271	8	2.5	21	16.9	142	6.4	63	6.3	80
1R0R 2G58	274 121	5	2.6	20	2.7	23	4.6	61	7.4	72
1GU2 2FBA	124 492	5	3.0	22	1.2	21	*	*	5.7	72
2AVM 2H5C	99 198	11	2.9	21	18.9	20	3.3	45	3.5	64
1OK0 2IGD	74 61	8	3.0	22	*	*	*	*	3.9	40
1YFQ 2H5C	342 198	6	3.1	20	5.1	22	3.4	35	5.6	64
1H97 1PSR	147 100	4	3.1	20	9.6	91	3.0	53	5.8	85
1C9O 1YS1	66 320	9	3.1	20	12.6	46	*	*	6.4	48

** No alignment found by software.*

(a) (b) (c) (d)

Figure 2. Superimposed alignments made using residue context method compared with RMSD and alignment length from CTSS. (a) Superimposed alignment between 1BYI and 2IGD, with RMSD 1.5 Å and length of 25. CTSS returns an alignment with RMSD 1.5 Å and length of 21. (b) Superimposed alignment between 1SBY and 1ZK4, with RMSD 2.6 Å and length of 43. CTSS returns an alignment with RMSD 3.6 Å and length of 186. (c) Superimposed alignment between 1K5N and 1L9L, with RMSD 2.3 Å and length of 21. CTSS returns an alignment with RMSD 6.0 Å and length of 30. (d) User interface with superimposed alignment between 2AQM and 2C9V, with RMSD 1.9 Å and length of 43. CTSS returns an alignment with RMSD 2.3 Å and length of 40. Superimposed residues with the highest scoring alignment for 2AQM and 2C9V are shown on the left, with the matrix of χ^2-distances on the right. Darker patches represent corresponding C_α-atoms with lower χ^2-distances. Highest scoring alignments are shown in yellow, with the best alignment shown in red.



(a) (b) (c) (d)

Figure 3. Superimposed alignments made using residue context method compared with superimposed alignments made with CTSS, DALI, and CE. (a) Superimposed residue context alignment between 1PJX and 2IIM, with RMSD 2.7 Å and length of 26. (b) Superimposed CTSS alignment between 1PJX and 2IIM, with RMSD 7.3 Å and length of 24. (c) Superimposed DALI alignment between 1PJX and 2IIM, with RMSD 4.4 Å and length of 33. (d) Superimposed CE alignment between 1PJX and 2IIM, with RMSD 4.9 Å and length of 48.

the alignment. An interactive user interface from [1] was implemented to allow detailed examination of the superimposed alignment (Figure 2).

3.4 Complexity

Residue context histogram construction requires computation of the vector of each C_α-atom with respect to each other C_α-atom, so the complexity for this step is $O(n^2)$. An important note is that residue context computation is offline. The matrix C_{ij} containing an all-by-all comparison of the residue contexts represented by χ^2-values requires comparison of each bin in each histogram to each other bin in each other histogram, so the complexity for constructing this matrix is $O(kmn)$, where k is the number of bins, and m and n are the number of residues in each molecule. The metric-properties of the χ^2 distance (specifically, the triangle inequality), may also be used during comparisons, to exclude comparing molecules that are significantly different. The Smith-Waterman algorithm also requires the calculation of alignment score for an all-by-all comparison of two protein molecules, so an additional complexity of $O(mn)$ is added. Finally the least-squares superimposition algorithm used to compute the optimal superimposition is done in $O(n)$ time. The complexity then adds to:

$$O(n^2) + O(kmn) + O(mn) + O(n) \qquad (4)$$

4. Experiments

For our experiments, we selected a number of protein pairs that illustrate the ability of the residue context method to represent protein molecules in such a way that lends itself to meaningful structural alignments. (Table 2). Protein pairs were chosen with sequence identity ≤30%. These types of alignments are inherently more meaningful than those with higher primary sequence identify, because they treat the structure of the molecules without reliance on the fact that proteins with near-identical primary sequences will always have similar structures. A total of 241 protein structures were compared in an all-by-all alignment, with a subset of 14 alignments of interest for which we show results, compared to alignment results using CTSS [1], DALI [9], and CE_{medium} [10]. The two metrics we use to compare alignment results are RMSD and alignment length. We have chosen the *medium* parameter for CE, corresponding to the default similarity threshold heuristic [10]. Several alignments based on residue context are shown in Figure 2.

Next, we compare an alignment made using Residue Context with the alignments made using CTSS, DALI and CE (Figure 3). In aligning 1PJX and 2IIM, with sequence identity of 6%, Residue Context produces a significantly lower RMSD of 2.7 Å, while CTSS, DALI, and CE produce RMSD values of 7.3 Å, 4.4 Å, and 4.9 Å, respectively. Both Residue Context and CTSS produce alignments with no gaps between aligned residues, while DALI and CE produce alignments with several gaps each. All of the protein comparison algorithms identify β-strand motifs, but each algorithm selects different locations on each of the protein molecules to align.

5. Conclusions

We have presented a novel method for protein structure comparison based on the notion of Residue Context. The principal contribution of our proposed method is the extraction of relevant features of protein molecules that have been shown to contribute to formation of protein binding and active sites. Because each atom is considered within the neighborhood of surrounding atoms, this method diverges from typical protein structural comparison and can capture the impact of both local and distant neighborhoods of each

residue. This representation of residue context for a protein supports a highly efficient matching strategy that takes advantage of the compactness of the extracted features.

The experimental and comparative results using proteins of low sequence identity highlight the ability of residue context alignments to find correspondences between similar regions of molecules. Residue context alignments tend to be shorter than those obtained using CTSS, DALI, or CE, which may aid in future efforts towards applying this method to finding further known binding site alignments. Another experimental approach is implied by the "fine-tuned" alignment found by residue context: A query protein could be used to find a rough overall alignment using algorithms like CTSS, DALI, or CE. After locating proteins with similar overall structure, the residue context could be used to identify substructures of those proteins that are actually likely to exhibit similar binding and functional activity.

While residue context as defined here provides a rich descriptor for protein structure, there are many possible additional protein features that could be included in the residue context. Future efforts will be directed at using different measures of similarity and extend the use of the algorithm to structural features such as residue side chains.

6. References

[1] T. W. Can, Y.F., "CTSS: a robust and efficient method for protein structure alignment based on local geometrical and biological feature," in *Proc. of the 2003 IEEE Bioinformatics Conference*, 2003, pp. 169-179.

[2] C. Chothia and A. Lesk, "The relation between the divergence of sequence and structure in proteins," *EMBO J.,* vol. 5, pp. 823-826, 1986.

[3] M. Vendruscolo and C. Dobson, "A glimpse at the organization of the protein universe," *Proc. Natl. Acad. Sci. U.S.A.,* vol. 102, pp. 5641-5642, 2005.

[4] Murzin et al., "SCOP: A structural classification of proteins database for the investigation of sequences and structures," *J. Mol. Biol.,* vol. 247, pp. 536-540, 1995.

[5] L. Holm and C. Sander, "Touring protein fold space with Dali/FSSP," *Nucleic Acids Res.,* vol. 26, pp. 316-319, 1998.

[6] H. e. a. Ohkawa, "MMDB: an ASN.1 specification for macromolecular " *ISMB,* vol. 3, pp. 259-267, 1995.

[7] Orengo, C. et al., "CATH: a hierarchic classification of protein domain structures," *Structure,* vol. 5, pp. 1093-1108, 1997.

[8] A. Godzik, "The structural alignment between two proteins: Is there a unique answer?," *Protein Science,* vol. 5, pp. 1325-1338, 1996.

[9] L. Holm and C. Sander, "DALI: a network tool for protein structure comparison," *Trends Biochem. Sci.,* vol. 20, pp. 478-480, 1995.

[10] I. Shindyalov and P. Bourne, "Protein structure alignment by incremental combinatorial extension (CE) of the optimal path," *Protein Engineering,* vol. 11, pp. 739-747, 1998.

[11] M. M. Young, A. G. Skillman, and I. D. Kuntz, "A rapid method for exploring the protein structure universe," *Proteins: Structure, Function, and Genetics,* vol. 34, pp. 317-332, 1999.

[12] R. Nussinov and H. J. Wolfson, "Efficient detection of three-dimensional structural motifs in biological macromolecules by computer vision techniques," *Biophysics,* vol. 88, pp. 10495-10499, 1991.

[13] X. Pennec and N. Ayache, "A geometric algorithm to find small but highly similar 3D substructures in proteins," *Bioinformatics,* vol. 14, pp. 516-522, 1998.

[14] N. Leibowitz, Z. Y. Fligelman, R. Nussinov, and H. J. Wolfson, "Multiple Structural Alignment and Core Detection by Geometric Hashing," *n Proc. of the 7th International Conference on Intelligent Systems for Molecular Biology (ISMB),* pp. 169-177, 1999.

[15] T. Smith and W. M., "Identification of Common Molecular Subsequences," *J. Mol. Biol.,* vol. 147, pp. 195-197, 1981.

[16] Kolodny R and L. N., "Approximate protein structural alignment in polynomial time.," *Proc. Natl. Acad. Sci. U.S.A.,* vol. 101, pp. 12201-12206, 2004.

[17] L. Chen, T. Zhou, and Y. Tang, "Protein structure alignment by deterministic annealing," *Bioinformatics,* vol. 21, pp. 51-62, 2005.

[18] Y. Zhang, & Skolnick, J., "TM-align: a protein structure alignment algorithm based on the TM-score," *Nucleic Acids Research,* vol. 33, pp. 2302-2309, 2005.

[19] Dundas J, et al., "Topology independent protein structural alignment," *BMC Bioinformatics,* vol. 8, p. 388, 2007.

[20] E. Krissinel and H. K., "Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions," *Acta Cryst,* vol. 60, pp. 2256-2268, 2004.

[21] E. S. Shih and M. J. Hwang, "Protein structure comparison by probability-based matching of secondary structure elements," *Bioinformatics,* vol. 19, pp. 735-741, 2003.

[22] L. Holm and C. Sander, "Mapping the Protein Universe," *Science,* vol. 273, pp. 595-602, 1996.

[23] R. Sowdhamini, S. Rufino, and T. Blundell, "A database of globular protein structural domains: clustering of representative family members into similar folds," *Folding & Design,* vol. 1, pp. 209-220, 1996.

[24] J. Hou, S. Jun, C. Zhang, and S. Kim, "Global mapping of the protein structure space and application in structure-based inference of protein function," *Proc. Natl. Acad. Sci. U.S.A.,* vol. 102, pp. 3651-3656, 2005.

[25] J. Hou, G. Sims, C. Zhang, and S. Kim, "A global representation of the protein fold space," *Proc. Natl. Acad. Sci. U.S.A.,* vol. 100, pp. 2386-2390, 2003.

[26] S. Belongie, J. Malik, and J. Puzicha, "Shape Matching and Object Recognition Using Shape Contexts," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 24, pp. 509-522, 2002.