Structure Alignment via Delaunay Tetrahedralization

Jeffrey Roach,* Shantanu Sharma, Maryna Kapustina, and Charles W. Carter, Jr.
Department of Biochemistry and Biophysics, University of North Carolina, Chapel Hill, North Carolina

ABSTRACT **A novel protein structure alignment technique has been developed reducing much of the secondary and tertiary structure to a sequential representation greatly accelerating many structural computations, including alignment. Constructed from incidence relations in the Delaunay tetrahedralization, alignments of the sequential representation describe structural similarities that cannot be expressed with rigid-body superposition and complement existing techniques minimizing root-mean-squared distance through superposition. Restricting to the largest substructure superimposable by a single rigid-body transformation determines an alignment suitable for root-mean-squared distance comparisons and visualization. Restricted alignments of a test set of histones and histone-like proteins determined superpositions nearly identical to those produced by the established structure alignment routines of *DaliLite* and *ProSup*. Alignment of three, increasingly complex proteins: ferredoxin, cytidine deaminase, and carbamoyl phosphate synthetase, to themselves, demonstrated previously identified regions of self-similarity. All-against-all similarity index comparisons performed on a test set of 45 class I and class II aminoacyl-tRNA synthetases closely reproduced the results of established distance matrix methods while requiring 1/16 the time. Principal component analysis of pairwise tetrahedral decomposition similarity of 2300 molecular dynamics snapshots of tryptophanyl-tRNA synthetase revealed discrete microstates within the trajectory consistent with experimental results. The method produces results with sufficient efficiency for large-scale multiple structure alignment and is well suited to genomic and evolutionary investigations where no geometric model of similarity is known *a priori*. Proteins 2005;60:66–81.**

Key words: protein structure similarity; multiple structure alignment

INTRODUCTION

Protein structure comparison and alignment remains one of the most heavily investigated techniques in computational biology and bioinformatics.[1–22] Structural comparison can identify potential evolutionary relationships between more diverse classes of related proteins[23,24] than sequence comparison. Furthermore, sequence analysis cannot identify conformational changes in the same protein in different functional states.

The fundamental problem that must be addressed by any structure comparison method is how to define structural similarity precisely. Unlike sequence similarity,[25,26] biologically intuitive definitions of structural similarity tend to be highly context specific and do not translate readily into rigorous and universally applicable mathematical definitions. The lack of a universally accepted definition of structural similarity and the wide variety of applications to which each particular definition is suited has led to many distinct treatments of the question of measuring structural similarity. To assess each of these treatments individually has been the subject of numerous surveys.[27–30] Each method differs chiefly in the fundamental structural units under considerations: individual atoms, whole residues, secondary structure elements, and so forth; and the mathematical and statistical constructs used to define the similarity of the structure based on the similarity of its individual structural components. Broadly speaking, structural similarity measures can be divided into two classes: those that determine structure alignments,[1–16] and those that do not.[17–22] Typically, structural alignment is understood to mean a list of paired or *equivalenced* residues or atoms that can be superimposed on one another by a single proper, rigid-body transformation.

In general, methods that admit structure alignment are evaluated based on the number of equivalenced residues and the root-mean-squared distance (RMSD) between them. Comparison of these two criteria may not be as objective as it appears, considering many alignment methods are based on definitions of structural similarity intending to model more general similarities than RMSD of superposition by a single transformation. For example, *DaliLite*[10] and *FATCAT*[14] determine alignments that cannot be fully achieved with a single transformation. It is therefore important to distinguish between: (1) the proposed definition of similarity, (2) the alignment method used to optimize this similarity, and (3) how alignments based on different similarity definitions may be compared.

Rigid-body superposition is an appropriate measure of similarity for structures possessing a large substructure

Grant sponsor: National Science Foundation; Grant number: CCR-0086013; Grant sponsor: National Institutes for Health; Grant number: GM-48519.

*Correspondence to: Jeffrey Roach, Department of Biochemistry and Biophysics, University of North Carolina, Chapel Hill, NC 27599. E-mail: roachjm@email.unc.edu

Received 25 August 2004; Accepted 20 January 2005

Published online 26 April 2005 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.20479

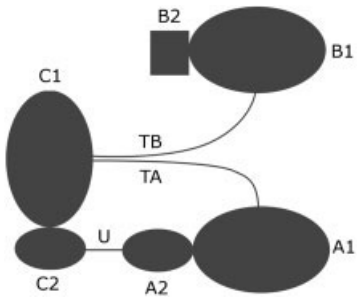


Fig. 1. A schematic diagram indicating the need to model multiple rigid-body transformations: clearly, object A is more similar to object C than is object B. Because the relative orientations of subobjects A_1 and A_2 differ from the relative orientation of subobjects C_1 and C_2, two transformations, T_A and U, are necessary to represent this similarity. If only a single rigid-body transformation is allowed, then objects A and B are equally similar to object C.

that can be superimposed nearly identically. Hypothetical examples of structures where RMSD of superposition will not capture an appropriate degree of similarity are easily envisioned. Consider, for example, three proteins A, B, and C. Here, each protein consists of a larger substructure A_1, B_1, C_1, and a smaller substructure A_2, B_2, and C_2, respectively. Suppose that two rigid-body transformations T_A and T_B superimpose A_1 on C_1 and B_1 on C_1 nearly identically. If a third rigid-body transformation U, not equal to T_A, can superimpose the smaller domain A_2 on C_2 nearly identically, yet no transformation can map B_2 to C_2 as well, then the RMSD of the largest superimposed substructures will consider protein A and B equally similar to protein C. Unless multiple transformations are allowed, the additional similarity of A_2 and C_2 will be lost. A schematic diagram describing this situation visually is given in Figure 1. Concrete examples of this sort, in particular homologous spectin repeats,[14] motivate the development of more flexible definitions of structure alignment. In general, introducing multiple transformations to achieve such flexibility significantly increases the necessary computation.

We introduce a very general definition of structural similarity based not on the exact atomic coordinates but on the geometric proximity relations encoded in the Delaunay tetrahedralization.[31] A succinct and comprehensive description of point set geometry, the Delaunay tetrahedralization has found a number of applications in structural biology: packing analysis,[32–34] fold recognition,[35] virtual mutagenesis,[36] and structure comparison[16,17,21,22] In this work each protein structure to be compared is represented as a one-dimensional string defined by the edges of the Delaunay tetrahedralization. Two one-dimensional representations are then compared by a dynamic programming scheme adapted from protein sequence analysis, thus reducing protein structural similarity to sequence similarity of the appropriate structure strings. One-dimensional structure string alignment determines an alignment of three-dimensional structure that is simultaneously more comprehensive and more flexible than can be achieved through superposition of a small number of components. Because actual atomic coordinates are not compared directly and rigid-body transformations are not modeled explicitly, the method compares structures more generally, albeit less precisely, than is measured by RMSD of the largest superimposable substructure. Furthermore, the efficiency of the Delaunay tetrahedralization as a network model of protein structure and the effectiveness of one-dimensional sequence alignment methods combine to make the method highly suitable for rapid, exploratory analysis of diverse protein structure classes.

Sequential Description of Tetrahedral Decomposition

Delaunay tetrahedralization decomposes the convex hull of a set of points, called *sites*, into a set of nonintersecting tetrahedra. By definition, four sites form a tetrahedron if their circumsphere contains no other site in the point set. The tetrahedral edges, therefore, encode much of the geometric proximity information in the point set and provide a more abstract, although less exact, description of its underlying geometry.

The coordinates of sites for the Delaunay tetrahedralization will necessarily affect its application. In structural studies, the sites are typically chosen to be atoms in the protein. In this case, the Delaunay tetrahedralization decomposes the interior space of the protein into a set of nonintersecting tetrahedra whose vertices correspond to atomic coordinates.

Although defining sites to be the atomic coordinates is most common, the definition of sites may be guided by specialized motives. Packing interactions, for example, may be expressed with more fidelity by the Delaunay tetrahedralization of sites corresponding to side-chain centroids.[36]

The sequential representation is derived from the Delaunay tetrahedralization of the set of Cα-atoms in the protein by describing each edge in a unique fashion. At a given residue, some tetrahedral edges connect to a previous residue (*backward-facing*) and some connect to a successive residue (*forward-facing*.) Beginning with the Cα-atom corresponding to the first residue and continuing to the last, each residue is encountered. If only backward-facing edges are recorded, then each edge of the tetrahedralization will be recorded exactly once.

To capture both the combinatorial and geometric information inherent in the tetrahedralization, each edge is recorded in the structure string as a relative residue difference and a length rank. Relative residue distance corresponds to the difference between residue numbers of the Cα-atoms joined by the edge. Edge length ranks, denoted A, B, C, correspond to short, intermediate range, and long interactions. The explicit definition of each class is a user-modified parameter of the representation. The default settings used here are: short lengths are defined to be less than 5 Å, intermediate lengths correspond to between 5 and 7.5 Å, and long edges fall between 7.5 and 10 Å. Edges of length greater than 10 Å are omitted. Thus, an edge of length 6.2 Å between residue 25 and residue 12 would be recorded as 13B. The delimiter | is used to

separate the edge sequence of the current residue from the edge sequence of the next residue.

A similar scheme can be used to describe forward-facing edges, that is, edges between a given residue and a residue that had not yet been encountered in the traversal. Forward-facing and backward-facing edges are not uniformly distributed throughout the protein. Forward-facing edges are more abundant than backward-facing edges near the N terminus of the protein. Conversely, near the C terminus backward-facing edges predominate. For structural similarity calculations, a convenient method of reducing bias toward one particular terminus is to represent the protein redundantly using both forward-facing and backward-facing edges. Note that if instead of traversing from the N-terminus to the C-terminus the tetrahedralization is traversed from the C-terminus to the N-terminus, that backward-facing edges and forward-facing edges are interchanged and the structure string is reversed.

The quality of structural similarity measures depends largely on how biologically important substructures are represented and compared. Particular substrings, *signatures*, in the sequential representation represent structural features of the protein. For example, due to packing interactions between protein Cα-atoms, the Delaunay tetrahedralization typically contains edges lying between two sequentially occurring Cα-atoms. These are short-range interactions and are denoted in the structure string as 1A. For complete proteins every edge sequence, both forward and backward-facing, contains 1A; however, occasionally coordinates for certain atoms are missing and the backbone is not continuous. In this case, 1A will be missing from the edge sequence.

Because the Delaunay tetrahedralization can be constructed on any point set, the structure string representation is not limited to single-chain structures. It is only necessary to distinguish between the chains. By introducing a chain order, for example, in a dimer, labeling one monomer's residues 1–500, and the other monomer's residues 501–1000, the structure string can be defined exactly as above, thus generalizing the sequential representation to account for multichained structures.

Secondary Structure String Signatures

Each Cα-atom in an α-helical region occurs in close proximity to the Cα-atoms in the four previous residues and the Cα-atoms in the four successive residues. This observation leads to a particularly simple signature in its structure string. A residue takes part in an α-helix if the first four entries in the forward-facing and backward-facing edge sequences correspond to one through four. Recalling that the symbol | terminates the edge list for a particular residue, the string of backward-facing edges for residues 41–45 of seryl-tRNA synthetase (1SRY)[37] [Fig. 2(a)]:

. . . **1A 2B 3B 4B** | **1A 2B 3B 4B** | **1A 2B 3B 4B** | **1A 2B 3B 4B** | **1A 2B 3B 4B** . . .

and the string of forward-facing edges:

. . . **1A 2B 3B 4B** 42C 43C 46B | **1A 2B 3B 4B** | **1A 2B 3B 4B** 36C 40C | **1A 2B 3B 4B** 35C 36B 39B 40B | **1A 2B 3B 4B** . . .

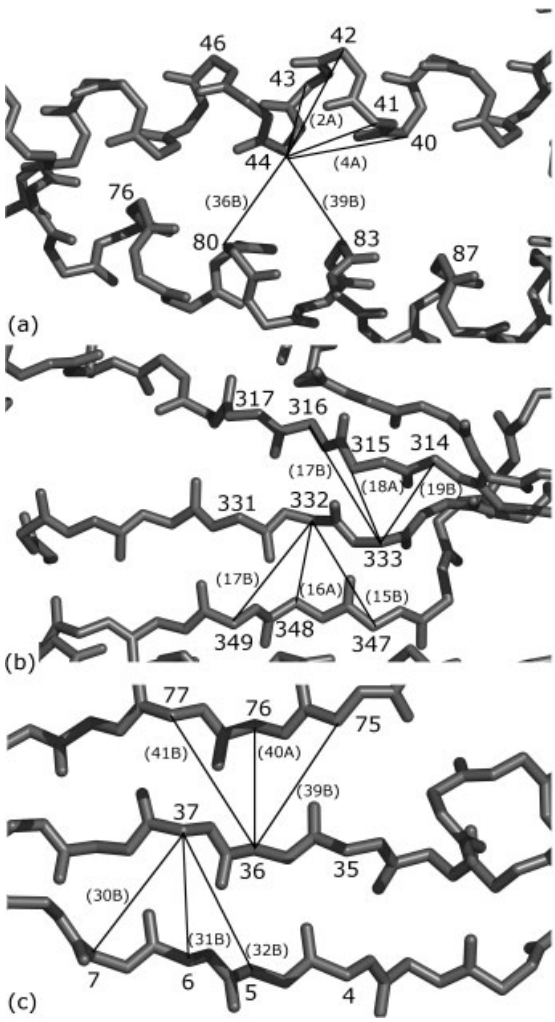


Fig. 2. (**a**) An α-helical region of seryl-tRNA synthetase (1SRY): residues 38 to 46 and 76 to 89. At Cα of residue 44, two of four forward-facing edges: 36B and 39B, and two of four backward-facing edges: 2B and 4B, marked. (**b**) An antiparallel β-sheet of seryl-tRNA synthetase (1SRY): residues 313 to 316, residues 331 to 334, and 347 to 349. At Cα of residue 332, three of six forward-facing edges: 15B, 16A, and 17B noted. Three of 10 backward-facing edges from Cα of residue 333: 17B, 18A, and 20B marked. (**c**) A parallel β-sheet of tryptophanyl-tRNA synthetase (1D2R): residues 3 to 7, 33 to 38, and 75 to 77. Three of six forward-facing edges from residue 36: 39B, 40A, and 41B, marked. Three of six backward-facing edges from residue 37: 29B, 30B, and 31B noted. All figures except Figure 3 produced by PyMol.[67]

determine an α-helical region indicated by the signature sequence in bold face. Note that additional edges in the forward-facing edge string: 42C, 43C, 46B, and so forth, describe the tertiary packing relations between the helical residues 41–45 and the helical residues 82–87.

Structure string signatures for β-sheets are described in a similar fashion. In a β-sheet, the Cα-atom for a given residue lies closer to Cα-atoms in a short sequence of residues occurring at least five residues upstream or downstream, than to Cα-atoms immediately surrounding it. Thus, structure strings for successive residues in

β-sheets do not include differences of 3 and 4 found in helical regions. Rather, they are generally marked by a sequential set of entries each of which is greater than four. For example [Fig. 2(b)] residues 331 to 335 of seryl-tRNA synthetase (1SRY) have a backward-facing edge string:

. . . 1A 2B **13B 14A 15B 16B** 32C 36C 211C | 1A 2B **14B 15B 16B 17B** 215C | 1A 2B **17B 18A 19B 20B** 27C 30C 31C 34C | 1A 2B **18C 19B 20B 21B** | 1A 2B **21B 22A 23B 24B** 29C . . .

and a forward-facing edge string:

. . . 1A 2B **17B 18B** | 1A 2B 13B **15B 16A 17B** | 1A 2B **12B 13B 14A 15B 16C** | 1A **9B 10B 11A 12B** | 1A 2B **8B 9B 10B 11B** 52C 55B . . .

determining a β-sheet. Note that the sequential backward-facing edges 13–16 originating from residue 331 increase to: 14–17 for residue 332, 17–20 for residue 333, 18–21 for residue 334, and 21–24 for residue 333. A similar, although decreasing, pattern is found in the forward-facing edges. An increasing pattern in the backward-facing edges and a decreasing pattern in the forward-facing edges indicate antiparallel β-strands. In this case, the backward-facing edges indicate the proximity of the 313 to 318 strand, whereas the forward-facing edges correspond to the 346–350 strand. As with the α-helical string, the remaining entries, backward-facing edges 32C, 36C, 211C, 215C, and forward-facing edges 52C, 55B, denote specific tertiary packing contacts with three different α-helices elsewhere in the structure.

Parallel β-sheets are generally indicated by patterns of four to five sequential values whose initial and terminal values remain largely fixed, in both backward-facing and forward-facing edges, as the residue number increases. For example [Fig. 2(c)], backward-facing edges for residues 36–40 in the Rossmann fold of tryptophanyl-tRNA synthetase (1D2R):[38]

. . . 1A 2B **30B 31A 32B 33C** | 1A 2B **29B 30B 31B 32B** | 1A 2B **30B 31A 32B 33C** | 1A 2B **29B 30B 31A 32B** | 1A 2B **30B 31B 32B 33B** | . . .

and forward-facing edges:

. . . 1A 2B **39B 40A 41B 42C** 100C 102C 271C | 1A 2B **25C 28C 33C 38B 39B 40A** | 1A 2B **39B 40B 41B 42B** 94C 269C | 1A 2B 19B 22C 23B **38C 39B 40A 41B** | 1A 2B 3B 4B **39B 40A** 89C 92C . . .

indicate the parallel β-sheet formed by strands: 3–7, 34–39, and 76–79.

String Alignment and Similarity Scoring

An interresidue similarity score that reflects as much as possible the biological structure encoded by the structure string is necessary for meaningful structural comparison. Consider, for example, two strands of a β-sheet connected through an arbitrarily long loop. For a particular residue on one of the strands, there will be several entries in its edge sequence corresponding to tetrahedral edges that connect it to residues on the other strand. As the length of the loop increases, the relative distances representing these edges in the structure sequence will increase because the number of residues between each strand in the β-sheet increases. However, increasing the length of the loop will not alter the edges that connect the residue to residues on the other strand. Therefore, although the actual values of the entries in the edge sequence change, their relative differences within the edge sequence remain the same.

The blocks of contiguous entries having relative difference one are of primary significance in defining secondary structure. These blocks provide the basis for structural alignment and similarity scoring. Thus, interresidue similarity is defined in terms of blocks of contiguous integer sequences in both forward- and backward-facing edge sequence. For example the forward-facing edge strings:

1A 2B 3A 4B 13C 87C 88B 89C 90B

is decomposed into:

{1A 2B 3A 4B} {13C} {87C 88B 89C 90B}.

Edge sequences are compared block by block. Edges in each block are paired off from smallest to largest until one block is exhausted. Each pair is assigned a similarity of 2 if the length ranks match and 1 if they do not. Therefore, similarity between two blocks corresponds to twice the size of the smaller block minus the number of edges whose length rank is mismatched.

Once the similarity between blocks is established, the similarity of two edge sequences is defined as the sum of block similarities. For example, for two forward-facing edge sequences,

1A 2B 3A 4B 13C 87C 88B 89C 90B 1A 2B 77C 78C 79C 301C 302C,

each is decomposed into blocks and compared. For the first block in each edge sequence:

{1A 2B 3A, 4B} {1A 2B} block similarity = 4

the smaller block contains two edges with distance rank A and B. The first two edges of the larger block have distance rank A and B, respectively; thus, the similarity score is 4. In the next pair of blocks,

{13C} {77C 78C 79C} block similarity = 2

the smaller block contains only one edge; because the first edge of the larger block also has distance rank C the similarity score is 2. In the final pair of blocks

{87C 88B 89C 90B} {301C 302C} block similarity = 3

the smaller block has two edges; however, the first two edges of the large block disagree on the distance rank of the second edge; therefore, the similarity score of the blocks is 3. The total edge sequence similarity is therefore 4 + 2 + 3 or 9. Finally, to reduce bias caused by the nonuniform distribution of edges over the length of the protein, the interresidue similarity is defined to be the sum of edge sequence similarities for the forward and backward-facing edges.

Given the sequential nature of this reduced protein representation, it is natural to define similarity in terms of sequence alignment with respect to interresidue similarity. By analogy to sequential similarity scores[25,26] let the similarity score of two protein structure sequences, and hence two protein structures, be the maximum sum of squared interresidue similarities, subject to a gap penalty, for all possible structure sequence alignments. Calculating this quantity is completely analogous to calculating sequence similarity, with the exception that the amino acid similarity matrix has been replaced by the residue similarity function defined above. In light of this observation, the Needleman-Wunsch algorithm,[25] can be amended easily to calculate similarity of structure strings and hence similarity of protein structures.

The Needleman-Wunsch algorithm has been successfully adopted from sequence similarity measures to structural similarity measures in cases where the individual structural components to be compared possess a natural ordering, or sequential structure as opposed to a Euclidean structure cf Vaisman et al.[21,22] Laiter et al.[4] have defined structural similarity in terms of the sequential alignment of pseudotorsional OCCO angles. Also, the TOPSCAN system,[19] extends the TOPS[18] description, to a sequential representation of "primary topology": secondary structure elements; and "secondary topology"': proximity, accessibility, and length of these structural components. The sequential representation of the tetrahedral decomposition encodes this information in a simple, unified expression.

Illustrative Examples: From Heuristic to More Abstract Similarities

To see how actual structures are represented by structure strings and to verify the effectiveness of producing a structure alignment from the string alignment consider the multiple structure alignment of a 27-residue segment flanking the HIGH signature in four aminoacyl-tRNA synthetases. Figure 3 displays a multiple structure alignment given by the tetrahedral decomposition alignment of four class I aminoacyl-tRNA synthetases. Of the three subclasses of class I: two methionyl-tRNA synthetase (1A8H),[39] shown in gold, and leucyl-tRNA synthetase (1H3N),[40] shown in red, belong to subclass Ia; glutaminyl-tRNA synthetase (1GTR),[41] shown in blue, belongs to subclass Ib; and tryptophanyl-tRNA synthetase (1M83),[42] shown in green, is from subclass Ic. Solid colors indicate regions of high similarity score, semitransparent segments have comparatively lower similarity scores. The highest similarity scores occur in the loop region and N terminal half of the α-helix. Dissimilar structures at the C-terminus of the α-A helix diverge substantially, with the two subclass Ia enzymes showing longer loops between α-A and β-1. Note also the significant distortion of the first loop in methionyl-tRNA synthetase that results from having a valine at position 10 in the tryptophanyl-tRNA synthetase numbering scheme instead of the consensus proline. Structural registration in methionyl-tRNA synthetase is restored by a proline immediately before the histidine in the HIGH signature. The backward-facing structure string indicates regions of tetrahedral similarity. The forward-facing edges, not shown, were included in the construction of the alignment. Table I contains the raw pairwise similarity scores for the entire catalytic domain of the four structures.

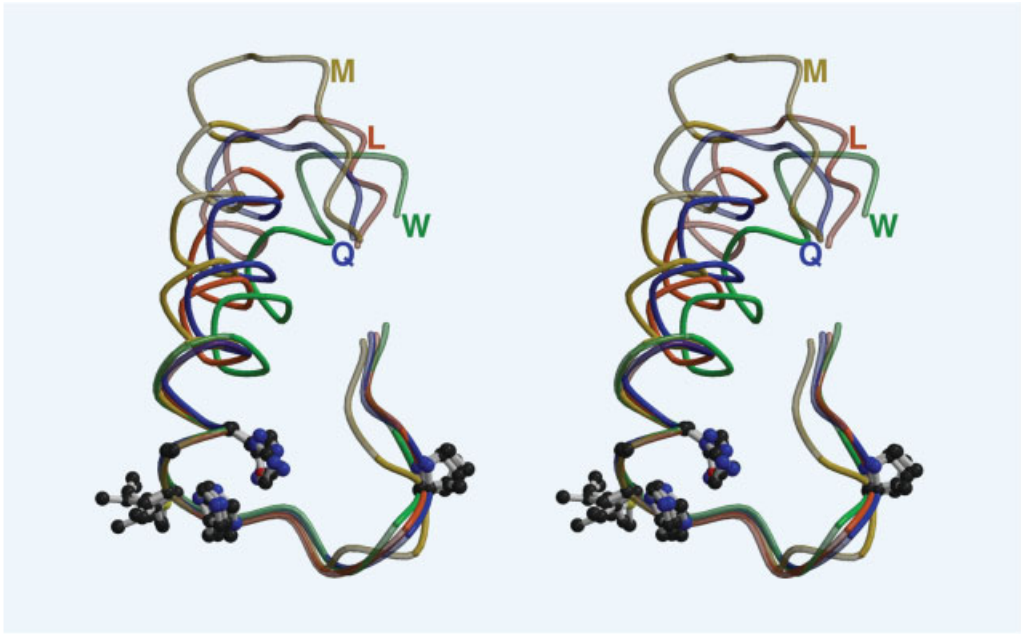
The tetrahedral decomposition alignment method next was applied to a test set of eight histones from the nucleosome core particle of chromatin from *Xenopus laevis* (1AOI chains A through H),[43] two recombinant histones from the hyper thermophyllic achaeon *Methanothermus fervidus* (1B67 chains A and B),[44] and two histone-like proteins from the hyper thermophyllic eubacterium *Thermotoga maritima* (1B8Z chains A and B).[45] Despite limited sequence identity, less than 20%, these histone and histone-like proteins exhibit a common histone fold.[46,47]

Each of the 12 proteins in the test set was aligned against histone H3 of the nucleosome core particle (1AOI chain A). The general alignments produced by the tetrahedral decomposition method identified more structurally equivalent residues than could be brought into alignment with a single rigid-body transformation. Restricting the general alignment to the largest set of pairs that could be equivalenced with a single transformation, a structure alignment was constructed by submitting these pairs to *lsqkab*[48–50] for rigid body superposition. For comparison, the alignment of the same proteins was performed with *DaliLite* and *ProSup*.

Figure 4 displays the alignment of the histone H2A (1AOI chain G) against H3 as determined by the three methods. The overall alignment [Fig. 4(a)] of H2A is essentially the same for all methods: the *ProSup* alignment of H2A shown in blue, the *DaliLite* alignment of H2A shown in green, the tetrahedral decomposition alignment shown in red, and the target, histone H3, shown in gray. Although all methods agree on the alignment of the helical regions, differences occur in between residues 61 and 65 and again between residues 74 and 81 of H3. In both cases the alignments of *DaliLite* and *ProSup* are nearly identical. The alignment given by the tetrahedral method falls closer to H3 on residues 74 to 81 [Fig. 4(b)] and strays from H3 on residues 61–65 [Fig. 4(c)]. Note that, because the alignment is given by a rigid body transformation, the deviation in these regions is not unrelated. In this case, letting the fit to residues 61–65 be less exact, allows a better fit to residues 74–81.

Figure 5(a) displays the alignment of the complete test set given by the tetrahedral method. Here each protein is aligned to the H3 histone of the nucleosome core particle (1AOI chain A). Note that the central α-helical region of the histone, the histone fold, is essentially preserved with significant deviation occurring only in the structure unique to each histone. In particular, both the archaeal recombinant protein rHMfA (1B67 chains A and B) and the eubacterial protein HU are composed completely of the histone fold and have only minimal N-terminal and C-terminal extensions.

RMSD and number of equivalenced residues for each method are shown in Table II. For the tetrahedral decomposition alignment, the number of structurally equivalent




```
GluRS 27 |2B 1A||2B 1A||2B 1A||4C 3C 2B 1A||   3C 1A||      6C 4C 2B 1A||   2B 1A||1A||2B 1A||    12C 10C 4B 3B 2B 36
LeuRS 43 |   1A||2B 1A||2B 1A||   3C 2B 1A||   3C 1A||   6C 5C 4C 2B 1A||   2B 1A||1A||2B 1A||12C 11C 10C 4B 3B 2B 52
MetRS 13 |   1A||2B 1A||2B 1A||4C 3B 2B 1A||3C 2B 1A||      6C 4C 2B 1A||3C 2B 1A||1A||2B 1A||13C 12C 10C 4B 3B 2B 22
TrpRS  9 |2B 1A||2B 1A||2B 1A||      2B 1A||   3B 1A||6C 5C 4C 3C 2B 1A||   2B 1A||1A||2B 1A||        10C 4B 3B 2B 18

GluRS 36 1A||5C 4B 3A 2B 1A||4B 3B 2B 1A||-----------||15B 14C 13C 4B 3B 2B 1A||    16B 14C 4B 3A 2B 1A||4B 3B 2B  41
LeuRS 52 1A||5B 4B 3A 2B 1A||4B 3B 2B 1A||-----------||    16B 15C 4B 3B 2B 1A||17B 16C 15C 4B 3A 2B 1A||4B 3B 2B  57
MetRS 22 1A||5B 4B 3A 2B 1A||4B 3B 2B 1A||-----------||    16B 15C 4B 3B 2B 1A||        17B 4B 3B 2B 1A||4B 3B 2B  27
TrpRS 18 1A||5B 4B 3A 2B 1A||4B 3B 2B 1A||4B 3B 2B 1A||16C 15C 14C 4B 3B 2B 1A||17C 16C 15C 9C 5B 4A 1A||5B 4B 2B  24

GluRS 41 1A||38C 37C 4B 3B 2B 1A||21B 20B 19B 4B 3B 2B 1A||21B 20C 4B 3B 2B 1A||4B 3B 2B 1A||42C 41C 38C 26C 24C   46
LeuRS 57 1A||        4B 3B 2B 1A||22B 21B 20B 4B 3B 2B 1A||        4B 3A 2B 1A||4B 3B 2B 1A||            27C 25B   62
MetRS 27 1A||        4B 3B 2B 1A||22B 21B 20B 4B 3B 2B 1A||        4B 3A 2B 1A||4B 3B 2B 1A||            27C 25C   32
TrpRS 24 1A||     5C 4C 3B 2B 1A||22C 21C 20C 4B 3B 2B 1A||        4B 3B 2B 1A||4C 3B 2B 1A||               25C   29

GluRS 46 4B 3B 2B 1A||-----------------------||27B 26B 25B 4B 3A 2B 1A||-----------||--------------------||         48
LeuRS 62 4B 3B 2B 1A||40C 28C 26B 4B 3A 2B 1A||    42C 41C 4B 3B 2B 1A||4B 3B 2B 1A||    31B 4B 3B 2B 1A||45C 44C  67
MetRS 32 4B 3A 2B 1A||28C 27C 26B 4B 3B 2B 1A||            4B 3B 2B 1A||4B 3B 2B 1A||33C 31B 4B 3B 2B 1A||          37
TrpRS 29 4B 3A 2B 1A||-----------------------||-----------------------||-----------||--------------------||         29

GluRS 48 4B 3B 2B 1A||4B 3B 2B 1A||45C 42C 41C 31C 30B 4B 3B 2B 1A||    32C  4B 3A 2B 1A||    33B 32A 31B 5A 4B    52
LeuRS 67 4B 3B 2B 1A||4B 3B 2B 1A||               36C 4B 3B 2B 1A||    38B 37C 3B 2B 1A||39B 38B 37B 36B 5B 4B    71
MetRS 37 4B 3B 2B 1A||4B 3A 2B 1A||               36C 4B 3A 2B 1A||        37B 3B 2B 1A||    38A 37B 36B 5B 4B    41
TrpRS 30 4B 3B 2B 1A||-----------||                  4C 3B 2B 1A||31B 30C  4C 3B 2B 1A||32B 31B 30B 29B 4B 3B    33

GluRS 52 2B 1A||-------------------------------------------------||-------------------------------------||-------  52
LeuRS 71 2B 1A||49C 48C 47B 46B 43B 42B 40B 39B 38A 37B 5B 2B 1A||50B 48B 39B 38B 37B 36B 10B 7B 6B 2B 1A||-------  73
MetRS 41 2B 1A||                     41C 39B 38A 37B 5B 2B 1A||            39B 38A 37B 10B 7B 6B 2B 1A||38A 11B  44
TrpRS 33 2B 1A||-------------------------------------------------||-------------------------------------||-------  33

GluRS 53 ------||                     34C 32A  6B  5B  2B 1A||    33B 32A 31A 30B 11B 10B  7B 6C 1A||-------    54
LeuRS 74 ------||51C 49B 48C 47C 45C 40C 39B 38A 37B 36C  2B 1A||52C 39B 38B 37B 36C 16C 15C 12B 2B 1A||-------    75
MetRS 44 2B 1A||                     39B 38A 37B 16B 15B 12B 1A||            40C 39B 38A 37B 17C 2B 1A||-------    46
TrpRS 34 ------||-------------------------------------------------||       33B 32B 31A 30B  4B  3C 2B 1A||32B 31A    35
```

Fig. 3. (**a**) Stereoview of multiple structure alignment by tetrahedral decomposition of the HIGH consensus sequence and flanking secondary structure from the first crossover connection of the Rossmann fold catalytic domains of TrpRS (green), GlnRS (blue), MetRS (gold), and LeuRS (red). Solid colors indicate regions of high similarity score, semitransparent segments have comparatively lower similarity scores. (**b**) One-dimensional structure string alignment for the structures in (**a**). Residues separated with |, white-space denotes edges missing from the tetrahedralization for a given residue, and dashes indicate alignment gaps. Edge structures identical in all four structures are highlighted. Figure produced by MOLSCRIPT[68] and raster3D.[69]

residues in both the general alignment and the alignment restricted to a single rigid-body transformation are reported. In all cases except for the alignment of HU proteins, the general alignment of the tetrahedral decomposition is nearly identical to the restricted alignment, indicating that a single rigid-body transformation is sufficient to model structural similarity. Furthermore, based on the probability model set forth in Levitt and Gerstein,[51] alignment by each method for all proteins except for those of HU had a p-value less than 0.00001. Therefore, although the additional flexibility gained by not directly using atomic coordinates may lead to a larger RMSD, the RMSD of the tetrahedral decomposition alignment do not become unacceptably large.

Despite the fact that the HU proteins, shown in brown, appear similar to the other histones in Figure 5(a), only

TABLE I.

Protein	MetRS	LeuRS	GlnRS	TrpRS
MetRS	23354	8170	5283	5330
LeuRS	8170	32300	6169	5528
GlnRS	5283	6169	22826	5147
TrpRS	5330	5528	5147	23248

Raw pairwise similarity scores for the alignment of the catalytic domains of: methionyl-tRNA synthetase (MetRS), leucyl-tRNA synthetase (LeuRS), glutaminyl-tRNA synthetase (GlnRS), and tryptophanyl-tRNA synthetase (TrpRS).

TABLE II.

Protein	PDB	*ProSup*	*DaliLite*	*TetraDA*
H4	1AOIB	1.86 (70)	2.20 (74)	1.374 (61/58)
H2A	1AOIC	1.92 (68)	2.10 (72)	2.392 (71/65)
H2B	1AOID	1.42 (66)	1.70 (68)	0.835 (58/58)
H3	1AOIE	0.38 (98)	0.40 (98)	0.306 (89/89)
H4	1AOIF	2.10 (71)	3.20 (81)	1.206 (63/60)
H2A	1AOIG	1.94 (68)	2.10 (72)	2.311 (73/67)
H2B	1AOIH	1.47 (66)	1.90 (69)	1.714 (62/62)
rHMFA	1B67A	1.49 (65)	1.70 (68)	1.412 (44/38)
rHMFA	1B67B	1.47 (64)	1.50 (64)	1.541 (59/57)
HU	1B8ZA	2.88 (37)	5.50 (48)	2.850 (50/22)
HU	1B8ZB	2.88 (37)	2.30 (41)	2.850 (49/22)

Structure alignment of histones and histone-like proteins against *X. laevis* histone H3. Root-mean-squared distance (RMSD) and number of equivalenced residue reported for *ProSup, DaliLite,* and tetrahedral decomposition alignment. Number of equivalenced residues for tetrahedral decomposition methods given for general alignment and alignment restricted to a single rigid-body transformation.

the alignment of 1B8ZB by *DaliLite* was significant at reasonable levels. Even in this case p is equal to 0.002, which, although small, is still several orders of magnitude larger than the p-values for the other histones. The poor fit of the HU proteins against the other histones is due to the differing orientation of the central α-helix in relation to the N terminal α-helices. A single transformation that closely fits the initial helices of HU proteins to the histone H3 will necessarily increase the distance between the proteins at the top of the central helix and lead to a large RMSD. Further evidence that single rigid-body transformation RMSD is unsuitable as a similarity measure for the comparison of *X. laevis* histones and eubacterial HU proteins is given by the general alignment of the tetrahedral decomposition that contains more than twice as many structurally equivalent residues as the restricted alignment. In this case the general alignment is defining more residues to be structurally equivalent than can be superimposed with a single transformation.

Figure 5(b) displays the restricted tetrahedral decomposition alignment of the *X. laevis* H3 histone, shown in gray, against the eubacterial HU protein, shown in brown. This alignment superimposes the two initial α-helices, residues 67–77 of H3 with residues 3–13 of HU, both shown in red, and residues 86–97 of H3 with residues 17–18 of HU, both shown in green. The general alignment also equivalences the top of the central helix, residues 106–120 of H3 with residues 30–40 of HU, both shown in blue. It is unclear from Figure 5(b) that these structures are indeed similar, because a second rigid-body transformation is needed to superimpose them. Figure 5(c) displays the alignment equivalencing the top of the central helix. Note that this transformation eliminates the appearance of similarity in the initial α-helices. The general alignment also aligns a five-residue helical region near the C-terminus of each protein, shown in orange in Figure 5(b) and 5(c). Aligning these regions requires a third, distinct, rigid-body transformation.

To investigate the differences between the general and restricted alignments on larger proteins, cysteinyl-tRNA synthetase of *Escherichia coli* (1LI5 chain A)[52] was aligned against glutamyl-tRNA synthetase of *Thermus thermophilius* (1J09 chain A)[53] by the tetrahedral composition method. Sequence identity for these proteins is 17%. For comparison, the same alignment was performed by *DaliLite* and *ProSup*.

Figure 6(a) shows the restricted alignment of glutamyl-tRNA synthetase given by the tetrahedral decomposition method, displayed in red, compared with the alignments determined by *DaliLite*, in green, and *ProSup*, in blue. Figure 6(b) displays a more detailed view of the alignments near the highly conserved Rossmann fold parallel β-sheet of cysteinyl-tRNA synthetase. The unrefined restricted alignment equivalenced 190 residues at 3.3 Å RMSD compared to *ProSup* equivalencing 213 residues at 2.4 Å RMSD and *DaliLite* equivalencing 241 residues at 2.8 Å RMSD. It is clear from the illustration that the three alignments indicate largely the same similarity, and that optimizing the restricted alignment with respect to RMSD would reduce the variation in the number of equivalenced residues and RMSD between the three methods.

Although the restricted alignment identifies similarity in the Rossmann fold domain, the general alignment by *TetraDA* also indentifies approximately 80 additional residues in the anticodon binding domains. These 80 residues correspond to seven structurally equivalent components that cannot be superimposed with the same transformation that superimposes the large domain, recall Figure 1.

In Figure 6(c) one particular component, the loop and helix shown in green, corresponding to residues 331–345 of cysteinyl-tRNA synthetase and residues 339–356 of glutamyl-tRNA synthetase, are superimposed. In addition to this component, Figure 6(c) shows three other components equivalenced by the general alignment: cysteinyl-tRNA synthetase residues 313–327 with glutamyl-tRNA synthetase residues 323–338, shown in the blue, cysteinyl-tRNA synthetase residues 346–355 with glutamyl-tRNA synthetase residues 359–370, shown in red, and cysteinyl-tRNA synthetase residues 356–372 with glutamyl-tRNA synthetase 385–401, shown in orange. Note that the orange helices, which superimpose poorly by RMSD standards, are nonetheless equivalenced by a more flexible similarity measure. An additional 25 residues are equivalenced by the method in the general alignment that are not shown in this figure. Similarity scores relating the anticodon binding domains are weaker than those relating the

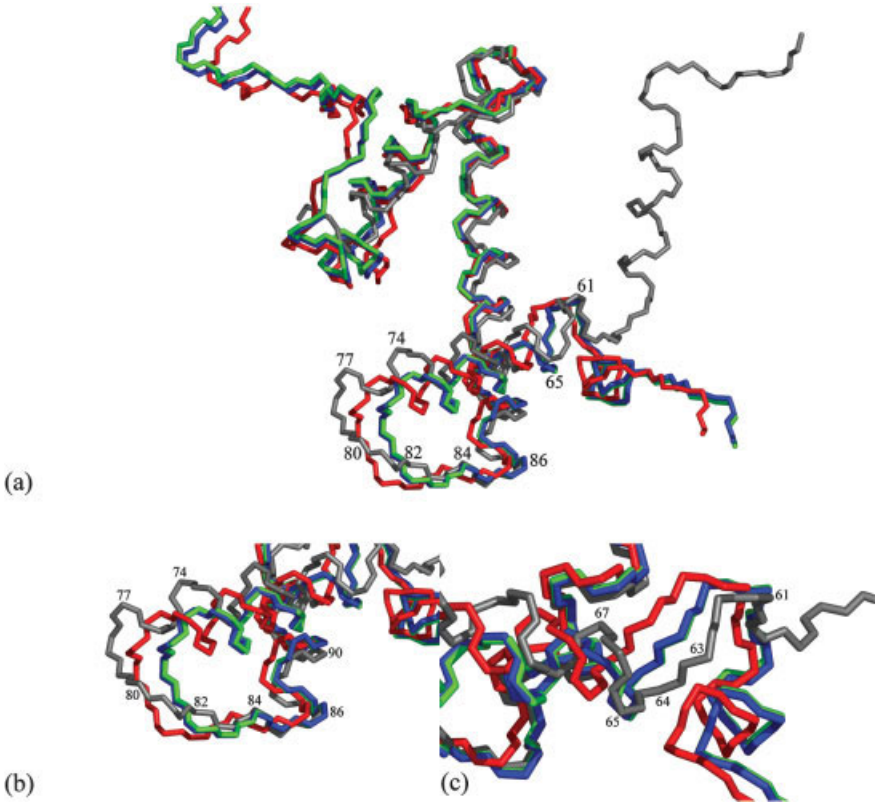


Fig. 4. (**a**) Alignment of *X. laevis* histones (1AOI) H2A (chain G) against H3 (chain A) by three different methods: *DaliLite* shown in green, *ProSup* shown in blue, and the tetrahedral decomposition method shown in red. Target, histone H3, shown in dark gray. All three methods show reasonable alignment, with *DaliLite* and *ProSup* having nearly identical configuration. Residue numbers for H3 noted. (**b**) Detailed view of residues 74 to 90. Note that in this region the tetrahedral decomposition method, shown in dark gray, aligns more closely with the target. (**c**) Detailed view of residues 61 to 67. In contrast (**b**), the tetrahedral decomposition shows the largest deviation from the target of all three methods.

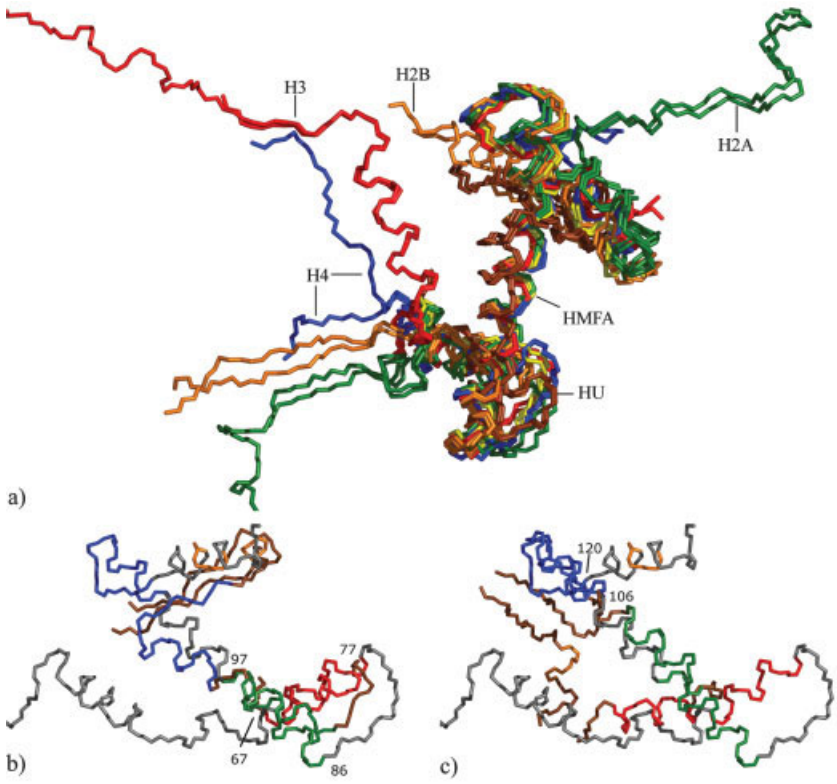


Fig. 5. (**a**) Tetrahedral decomposition structure alignment all *X. laevis* nucleosome core particle histones (1AOI): two copies H3 (chains A and E) shown in red, two copies H4 (chains B and F) shown in blue, two copies H2A (chains C and G) shown in green, and two copies H2B (chains D and H) shown in orange. Two recombinant histones from *M. fervidus* (1B67 chains A and B) shown in yellow and two histone-like proteins from *T. maritima* (1B8Z chains A and B) shown in brown. (b) Superposition of N-terminal helices HU, shown in brown, on to N-terminal helices of H3, shown in gray. (**c**) Superposition of the central helix of HU on the central helix of H3. In both (**a**) and (**c**) structurally equivalent components under the general alignment are shown in blue, green, red, and orange.

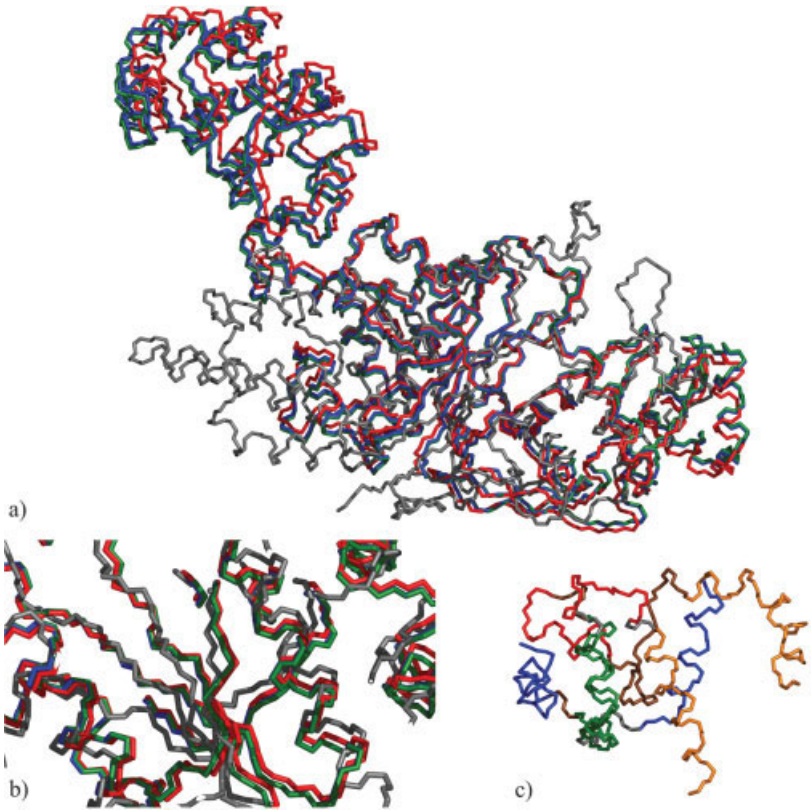


Fig. 6. (**a**) Alignment of glutamyl-tRNA synthetase (1J09) to target cysteinyl-tRNA synthetase, shown in gray, by three methods: *DaliLite* shown in green, *ProSup* shown in blue, and restricted tetrahedral decomposition alignment shown in red. (**b**) Detailed view of conserved regions. All three alignments essentially agree. (**c**) General alignment of anticodon binding domains. Cysteinyl-tRNA synthetase shown in gray and glutamyl-tRNA synthetase shown in brown. Structurally equivalent components under the general alignment shown in blue, green, red, and orange.

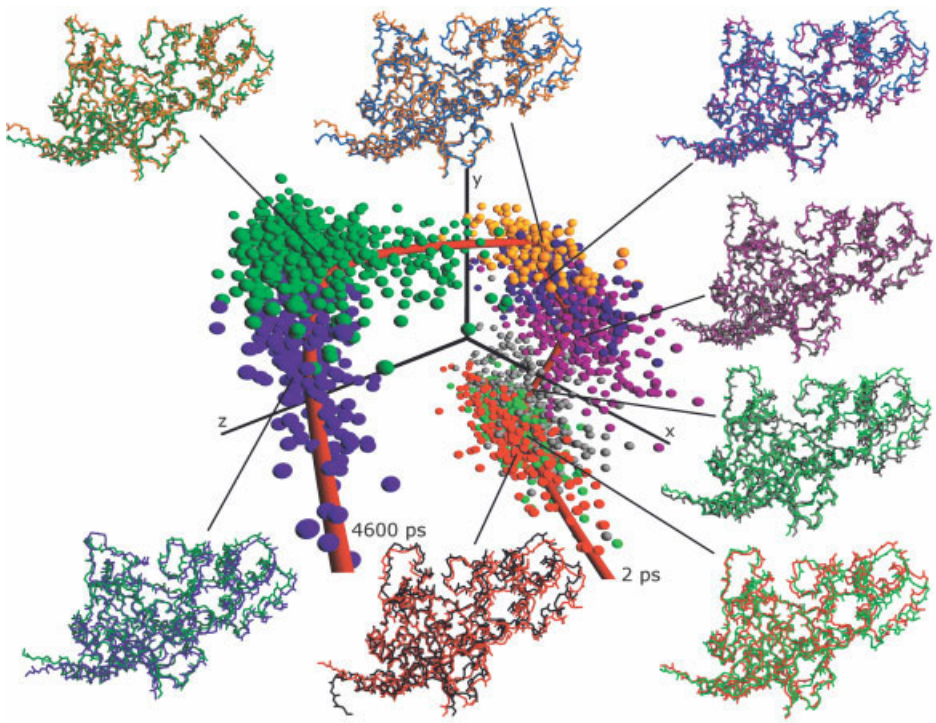


Fig. 10. Generalized geometric configuration of 2300 2-ps snapshots of molecular dynamics trajectory of the enzyme tryptophanyl-tRNA synthetase. Trajectory indicated by heavy brown curve. Distinct structural clusters indicated by color. Insets display cluster representative against previous representative along trajectory. Initial structure, shown in black, is crystal structure 1M83. Note that clusters were calculated in the first 14 principal components; thus, apparent lack of separation between clusters is due to projection to three dimensions.

catalytic domains of the two proteins, consistent with a more abstract similarity.

Because no one rigid-body transformation is sufficient, similarity measures based on superposition RMSD cannot account for these similarities in the anticodon binding domain. Each of these seven components requires its own rigid-body transformation to be superposed. In several cases, for example the components shown in orange in Figure 6(c), rigid-body transformation will lead to large RMSD. From the perspective of superposition, similarities in the anticodon binding domain appear weak and difficult to evaluate. The general alignment, however, is able to assess simultaneously the nearly identical Rossmann fold domains as well as the more abstract potential homologies present in the anticodon binding domains.

Identifying Internal Symmetry

All structural alignment methods aim to determine maximally similar alignments of two proteins. These alignments, however, may not be the only biologically interesting alignments. Often, suboptimal alignments produce evidence supporting certain evolutionary hypotheses. Consider, for example, the bacterial eight-iron ferredoxin. A small metalloprotein of only 54 residues, ferredoxin constructs a cavity for its two Fe_4S_4 clusters by forming two nearly identical halves: the first half corresponding to residues 1–25, and the second half corresponding to residues 25–50.[54] The ramification of this repetition for structure alignment is that ferredoxin can be aligned against itself in a nontrivial way and a suboptimal, but biologically relevant, alignment will equivalence residues 1 and 25, residues 2 and 26, and so forth. Sequence identity of residues 1–25 aligned with residues 26–50 is approximately 46%. The structural similarity score derived from the tetrahedral decomposition alignment and mean residue number shift between equivalenced residues for each possible self-alignment is plotted in Figure 7(a). The largest peak, naturally, corresponds to the trivial self-alignment. However, the self-alignment of residues 1–25 on to residues 26–50 corresponds to a suboptimal peak, with a mean separation of approximately 25.

To test the ability of the method to isolate other structurally important suboptimal alignments, the tetrahedral decomposition was used to identify internal self-similarity by aligning two larger proteins to themselves: cytidine deaminase[55] (1CTU) and the large subunit of carbamoyl phosphate synthetase[56] (1KEE). Although the simple alignment corresponds to the largest similarity score, the local optima of the objective function correspond to structurally important alignments reflecting internal twofold symmetry.

The similarity score and mean residue number shift between equivalenced residues of a set of possible cytidine deaminase self-alignments is displayed in Figure 7(b). The largest peak, again, corresponds to the trivial self-alignment. However, a suboptimal peak occurs with a mean separation between equivalenced residues of approximately 135. The alignment that corresponds to this suboptimal peak reveals a region of the protein approximately 85 residues in length beginning with residue 53, which is repeated in a nearly structurally identical fashion approximately 135 residues downstream beginning at residue 189. Figure 7(b) shows the alignment of this region against its repeat: residues 53–142 against residues 189–285. This approximate structural duplication accounts for slightly less than half of the protein suggesting that cytidine deaminase evolved by developing a nearly identical copy of itself and then specializing each half to a particular purpose. The similarity score of this suboptimum is approximately 5% of the similarity score of the trivial self-alignment. The sequence identity of the repeated region is approximately 9%.

A similar structural duplication scheme is suggested in the large subunit of carbamoyl phosphate synthetase. Carbamoyl phosphate synthetase (1KEE), being a tetramer of heterodimers, consists of four large subunits (chains A, C, E, and G) and four small subunits (chains B, D, F, H), possessing three twofold noncrystallographic symmetries mapping each pair of large and small subunits to each other pair of large and small subunits. Within each large subunit, a structural repeat of approximately 385 residues occurring first at residues 8–400 and repeated at residues 560–936 accounts for approximately 770 of 1037 residues [Fig. 7(c)]. The alignment superimposing these regions corresponds to a suboptimal peak with mean residue difference about 530 and similarity score approximately 5% of the similarity score given by the trivial self-alignment. Again, this does not take into account the cyclic alignment possible by mapping the second copy of the region back on to the first. The sequence identity of the structural repeat is approximately 36%.

Multidimensional Similarity on Sets of Structures

The ability to identify biologically relevant similarities within a larger set of structures and classify these structures according to mutual similarity is of paramount importance in structural bioinformatics. Success in classification requires not simply to determine a binary property, similarity or dissimilarity; but in fact, the ability to assess all of the aggregate distance or similarity relations occurring between objects, differentiate degrees of similarity or dissimilarity, and weight the contributions of various similarity relations appropriately. A set of structures required to test the method in this respect should, necessarily, contain structures that are both similar and dissimilar to varying degrees.

The aminoacyl-tRNA synthetases provide a natural example of such a test set. In terms of sequential analysis, individual aminoacyl-tRNA synthetases have minimal sequence identity with one another. Structurally, however, these enzymes roughly evenly divided into two structurally distinct classes.[57,58] These two classes belong to entirely different fold classes. Class I synthetases all have Rossmann dinucleotide binding folds, and therefore, in terms of the SCOP[59] hierarchy, belong to the α/β class. Class II synthetases, in contrast, are based on a large, antiparallel β-sheet, and consequently fall into the $\alpha+\beta$ SCOP class.

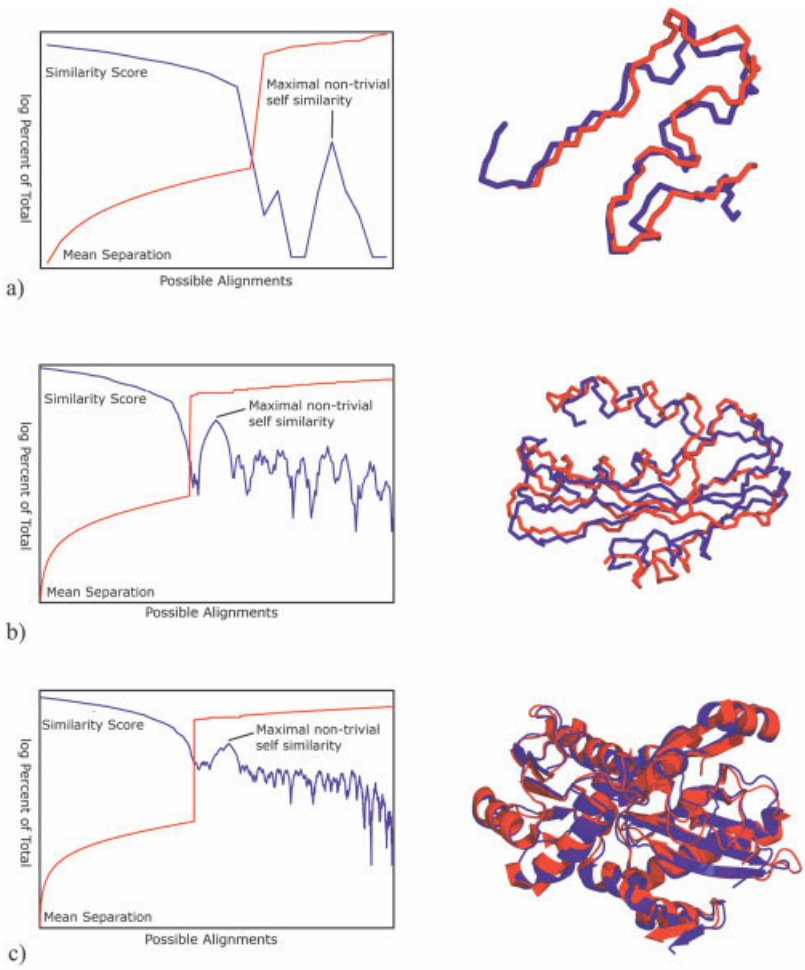


Fig. 7. Chart of similarity score and mean residue number difference between equivalenced pairs for a set of possible alignments. Scores and mean separation are reported in terms of the logarithm of the percent of maximum value attained. Self-alignment corresponding to the maximum non-trivial peak displayed at right. (**a**) Self similarity of ferredoxin (1DUR) determined by tetrahedral decomposition structure alignment. Residues 1 to 25 shown in red and residues 26 to 55 shown in blue. (**b**) Self-similarity of cytidine deaminase (1CTU). Residues 53 to 82 shown in red and residues 189 to 285 shown in blue. (**c**) Self-alignment of carbamoyl phosphate synthetase (1KEE) corresponding to maximal nontrivial alignment. Residues 8 to 400 appear in red and residues 560 to 936 appear in blue.

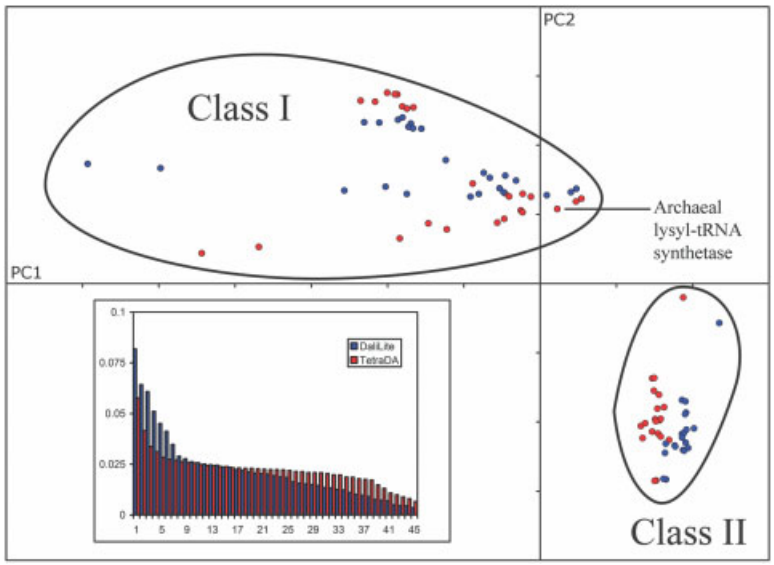


Fig. 8. Geometric description of aminoacyl-tRNA synthetases provided by principal component analysis, first two principal components plotted. Red points denote abstract coordinates determined by the tetrahedral decomposition method and blue points denote abstract coordinates determined by *DaliLite*. Inset charts the amount of variance described by each principal component for the tetrahedral decomposition method, red, and *DaliLite*, blue.

TABLE III.

4TS1A	1M83A	1LI5A	1BBUA	1ASZA
1JILA	1MB2A	1ILEA	1KMMA	1B8AA
1H3FA	1D2RA	1QU2A	1QE0A	1C0AA
1N3LA	1MAWA	1IVSA	1H3VB	1G51A
1J1UA	1GTRA	1H3NA	1ATIA	1PYSA
1I6LA	1J09A	1F7UA	1QF6A	1PYSB
1I6MA	1IRXA	1IQ0A	1EVLA	1HC7A
1I6KA	1A8HA	1SRYA	1FYFA	1NJ6A
1MAUA	1F4LA	1E1OA	1EOVA	12ASA

PDB code and chain identity for test set of class I and class II amino acyl-tRNA synthetases.

Each class can be further decomposed into three subclasses with respect to more subtle distinctions occurring in the catalytic domain of each enzyme. Within each subclass, enzymes of different species that catalyze the synthesis of the same amino acid tend to be more similar than enzymes from the same species catalyzing the synthesis of different amino acids. Finally, the most subtle differences correspond to conformational differences between enzymes from the same species, catalyzing the same amino acid. The set of animoacyl-tRNA synthetases, therefore, provides an exacting test of the flexibility, sensitivity, and discriminating power of the tetrahedral decomposition alignment method.

A test set of 45 aminoacyl-tRNA synthetases was selected (Table III) to include structures similar and dissimilar at all levels. In cases where a single chain corresponds to multiple domains, the entire chain, and thus all domains, was used. This test set was then submitted to all-against-all similarity comparisons using the tetrahedral decomposition similarity measure and, for comparison, *DaliLite*.

To better understand the structural similarities between members of a given set of proteins, it is convenient to define geometric coordinates for each object such that distance between coordinates correspond to the distances and similarities given by the similarity measure. This process is called *ordination* or *scaling*. Principal component analysis[60] defines a point in space corresponding to each object in the test set, providing the requisite geometry.

In general, the coordinate space describing the set of objects will be of high dimension: equal to the size of the set. However, the principal components analysis will decompose the variance in terms of decreasing variance along its axis. Therefore, gross properties of the similarity relations within the set are described using only the initial principal components, with higher order principal components reserved for smaller variations.

Similarity scores between two proteins were normalized to a similarity index to account for differing protein length. The similarity index for a given method was calculated from the raw, unscaled similarity scores and organized in a 45×45 matrix S:

$$S_{i,j} = \frac{s(p_i,p_j)}{\sqrt{s(p_i,p_i)}\ \sqrt{s(p_j,p_j)}}$$

Here, $s(p_i,p_j)$ denotes the similarity score of protein p_i and protein p_j. Note that the matrix S is analogous to Table I: the essential difference being that the matrix S contains indices derived from the raw similarity scores given in Table I. The raw scores have the property that larger structures have higher similarity scores when compared to themselves. Similarity indices reduce the length dependence of the raw scores.

Because similarity relations within the set will remain unchanged if the entire set is submitted to any particular rigid-body transformation, the geometric configurations produced by principal component analysis represents only one particular spatial orientation. To compare two different similarity measures on the same set, it is necessary to find a single rigid-body motion that will minimize the difference between geometric coordinates for the same object given by two different similarity measures. This technique is the so-called orthogonal Procrustean analysis[60] and will determine an orientation suitable to compare the similarity results given by tetrahedral decomposition similarity with the results of *DaliLite*. It is of independent interest that the method used to align objects described by two different similarity measures is mathematically identical to the Kabsch algorithm[48] used to align equivalenced pairs of residues in two proteins.

The use of principal component analysis deepens the comparative scope of the tetrahedral decomposition method, demonstrating its ability to discriminate between structural differences at all levels. The biplot of the first two principal components (Fig. 8) shows the geometric organization of the test set determined by the tetrahedral decomposition method, red points, and *DaliLite*, blue points. The overall organization of the biplot is nearly identical for both the tetrahedral decomposition method and *DaliLite*, demonstrating the essential equivalence of *TetraDA* similarity scores to those produced by *DaliLite*. Both methods define a clear distinction between class I and class II amino acyl-tRNA synthetases, and consequently between the SCOP classifications α/β and $\alpha + \beta$. There are no "false positives," that is, none of the proteins in class I are mistakenly classified as class II. In particular, both methods correctly place the archaeal lysyl-tRNA synthetase in class I as a close relative of glutaminyl-tRNA synthetase and glutamyl-tRNA synthetase, as opposed to class II, where other lysyl-tRNA synthetases group together with aspartyl-tRNA synthetase and asparaginyl-tRNA synthetase.

It is important to note that the notion of "false positive" is poorly defined in the sense that it assumes an unspecified discrete topology. That is, objects are defined to be either near one another or they are defined to be infinitely distant. The term disallows relative degrees of similarity. Given a similarity measure describing a continuum of possible degrees of similarity, it is straightforward to see that no single threshold can be universally appropriate. For example, in terms of the aminoacyl-tRNA synthetases, the appropriate definition of "different structures," be it belonging to different classes, belonging to different subclasses, or belonging to different conformational configura-

tions, depends entirely upon the context. Certainly no single definition will be appropriate for all applications. Principal component analysis allows us to rank various structural similarities in terms of variance within the set of structures.

The magnitudes of the principal component singular values were transformed to correspond to a percent of total variance along each axis. A plot of these normalized scores is inset in Figure 8. In both methods, the tetrahedral decomposition alignment shown in red, and *DaliLite* shown in blue, the first two to three principal components carry significantly more variance than the others. Notably, however, the spectrum of the tetrahedral decomposition method distributes the remaining variance more evenly over the high order principal components, consistent with its sensitivity to more abstract similarities.

Higher order pairs of principal components reveal a series of increasingly subtle structural distinctions within each class. The subclassification[57,58] within class I aminoacyl-tRNA synthetases is shown in the biplot of principal components 3 and 4 [Fig. 9(a)]. The primary distinction between subclasses occurs on the third principal component axis. Conformational differences induced by ligand binding prevent all the structures of subclass A from forming a compact cluster. The subclass A ligand free structures form a fairly tight cluster together with the isoleucyl-tRNA synthetase complexed with mupirocin;[61] however, the methionyl-tRNA synthetase complexed with methionine,[62] the leucyl-tRNA synthetase complexed with the leucyl-adenylate,[40] and the valyl-tRNA synthetase complexed with the valyl-adenylate[63] are outliers. Even more detailed cluster analysis of the tryptophanyl-tRNA synthetase enzymes correctly distinguishes "open" and two different types of "closed" conformations [Fig. 9(a) inset].

The subclassification of class II aminoacyl-tRNA synthetases is captured by the fifth principal component axis. The biplot of the second and fifth principal components [Fig. 9(b)] shows clusters for subclasses A, B, and C of class II as well as their distinction from class I. Differences in the anticodon binding domain introduce separation within the subclass IIb cluster between asparaginyl-tRNA synthetase and the lysyl-tRNA and aspartyl-tRNA synthetases. The subclass IIc cluster contains α- and β-subunits of phenylalanyl-tRNA synthetase. Additional domains not present in the other class II aminoacyl-tRNA synthetases cause the β-subunit of phenylalanyl-tRNA synthetase to separate somewhat from the class. The fifth principal component [displayed in Fig. 9(b)] necessarily carries less variance than the third and fourth [shown in Fig. 9(a)]. Thus, the available class II structures display less conformational variance than those from class I.

Cluster Analysis of Molecular Dynamics Trajectories

The efficiency of the tetrahedral decomposition offers substantial advantages in evaluative similarity measures on large sets of structures requiring all-by-all pairwise comparison. Tetrahedral decomposition alignment becomes particularly effective in the analysis of large-scale all-by-all comparisons. Molecular dynamics simulations of the enzyme tryptophanyl-tRNA synthetase produced 2300 structures corresponding to 2-ps snapshots over a particular trajectory exhibiting a functionally significant conformational change. Similarity indices for each pair of snapshots were calculated, and the resulting similarity matrix was submitted to principle component analysis. Approximately 80% of the variance within the set could be described by 14 principal components. Hierarchical clustering based on these principal components revealed eight clusters corresponding to maximally distinct microstates along the trajectory. Figure 10 displays a three-dimensional projection of the geometric configuration described by the tetrahedral decomposition similarity scores. Distinct clusters are indicated by each color. Note that clusters calculated in 14 dimensions need not appear separated in projection. For each cluster, the structure closest to the cluster centroid was chosen as a representative and illustrated around the circumference of the figure.

Implementation and Availability

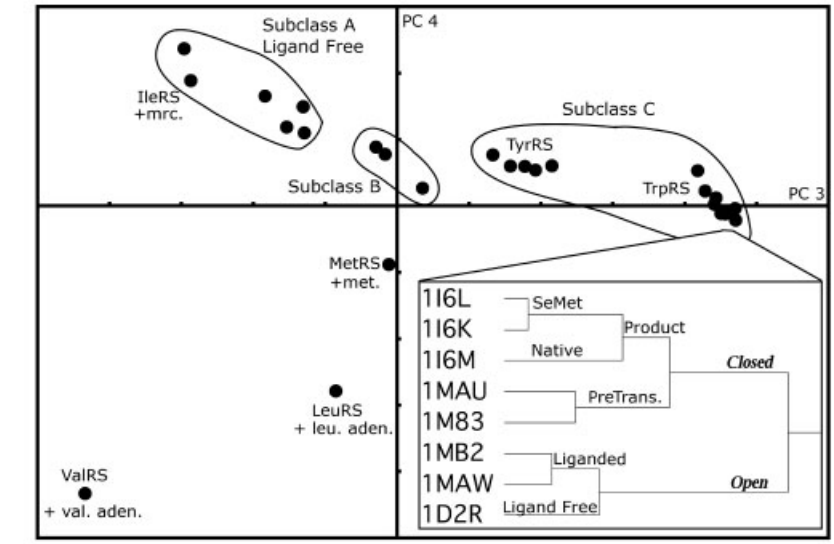
Sequential representations and alignments methods have been implemented in C++ as the program *TetraDA*. Delaunay tetrahedral decompositions are constructed by linking to the *qhull*[64] library. The program has been tested on Linux and Mac OS X platforms. Effort has been made to make *TetraDA* as portable as possible, and it should compile with any C++ compiler. Source code is available at this time from the corresponding author by request.

The all-by-all comparison of aminoacyl-tRNA synthetases performed on Sharp PC-UM10M laptop required 1 h, 27 min using the tetrahedral decomposition method compared to 24 h, 20 min for *DaliLite*. Singular value decompositions necessary in this analysis were calculated by LAPACK.[65]

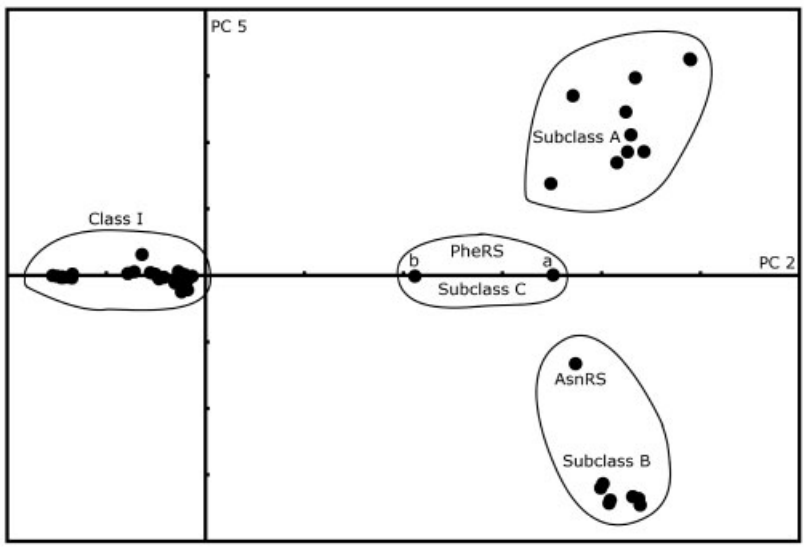
Calculation of pairwise similarity for snapshots along the molecular dynamics trajectory was performed on an IBM BladeServer Linux Cluster comprising 61 1.6-GHz nodes running RedHat Linux 7.2. Because similarity comparison between pairs of structures are independent, the process was highly suited to parallel processing. Depending on processor availability, the entire calculation required approximately 5 h. Principal component analysis of the similarity matrix and hierarchical clustering was performed using the statistical package JMP.[66]

CONCLUSION

Alignments of the sequential representation of the tetrahedral decomposition express both nearly exact local similarities and similarities that cannot be represented with rigid-body transformations in a single alignment. Using the actual atomic coordinates only to define the initial tetrahedral decomposition, similarity based on the adjacency relations within the structure determines analogous substructure by a more flexible model than rigid-body superposition. For the histone test set, single rigid-body superposition derived from the tetrahedral decomposition was consistent with, although nonetheless distinct from,



a)



b)

Fig. 9. Principal components of higher order correspond to more subtle differences between the protein structures. (**a**) Principle components three and four determine subclassification of class I aminoacyl-tRNA synthetases. Subclass A is divided between liganded and ligand free conformations. Inset, a cluster analysis of the tryptophanyl-tRNA synthetases showing distinction between "open" and "closed" conformations. (**b**) Principle components two and five determine subclassification of class II aminoacyl-tRNA synthetases. Asparaginyl-tRNA synthetase and both α- and β-subunits of phenylalanyl-tRNA synthetase are indicated.

those produced by established methods such as *DaliLite* and *ProSup*. Tetrahedral decomposition alignment also determines extensive regions of self-similarity within a single protein. Applied to ferredoxin, cytidine deaminase, and carbamoyl phosphate synthetase, the method readily reproduced previously identified regions of approximate internal twofold symmetry.

The method is ideal when no precise geometric model of structural similarity is apparent, for example: describing phylogenetic relationships between more distant evolutionary ancestors than is possible with sequence analysis, identifying structural ramifications of protein sequence variation in multiple sequence alignment, and investigations of conformational changes within the same protein.

Combining a concise network model of the protein with techniques from bioinformatic sequence analysis, structural similarity based on the sequential representation of the tetrahedral decomposition allows rapid classification of large numbers of structures necessary for multiple structure alignment and clustering of molecular dynamics trajectories. Similarity indices calculated on a test set of the Class I and II aminoacyl-tRNA synthetases produced results in close agreement with *DaliLite* in 1/16 the time. The efficiency of calculating pairwise similarity from the tetrahedral decomposition alignment is sufficient for large-scale clustering and multiple structure alignment. All-by-all pairwise tetrahedral decomposition similarity of 2300 molecular dynamics snapshots of tryptophanyl-tRNA snapshots produced eight clusters of snapshots with one representative from each cluster corresponding to a unique, discrete microstate within the trajectory consistent with experimental results. Reducing 2300 snapshots along a trajectory to a small number of maximally distinct representatives greatly facilitates the analysis of molecular dynamics trajectories.

No single definition of structural similarity may be appropriate for all contexts. The tetrahedral decomposition alignment method complements established atomic coordinate-based methods by addressing similarities of the geometry underlying protein structure. Reducing proximity relations within the protein to properties of a particular string of symbols, the sequential representation allows structural information to be exploited much more efficiently and effectively in structure alignment, providing general similarity measures with surprisingly little loss of information.

ACKNOWLEDGMENTS

We would like to thank Liisa Holm for providing a copy of *DaliLite* for comparison. We would also like to thank Phil Carl, Nikolay Dokholyan, Marshall Edgell, and Jan Hermans for helpful comments on the manuscript.

REFERENCES

1. Taylor WR, Orengo CA. Protein structure alignment. J Mol Biol 1989;208:1–22.
2. Holm L, Sander C. Protein structure comparison by alignment of distance matrices. J Mol Biol 1993;233:123–138.
3. Subbiah S, Laurents DV, Levitt M. Structural similarity of DNA-binding domains of bacteriophage repressors and the globin core. Curr Biol 1993;3:141–148.
4. Laiter S, Hoffman DL, Singh RK, Vaisman II, Tropsha A. Pseudotorsional OCCO backbone angle as a single descriptor of protein secondary structure. Protein Sci 1995;4:1633–1643.
5. Feng ZK, Sippl MJ. Optimum superimposition of protein structures: ambiguities and implications. Fold Des 1996;1:123–132.
6. Gilbrat JF, Madel T, Spouge JL, Bryant SH. The vast protein structure comparison method. Biophys. J 1997;72:MP298.
7. Shindyalov IN, Bourne PE. Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. Protein Eng 1998;11:739–747.
8. Taylor WR. Protein structure comparison using iterated double dynamic programming. Protein Sci 1999;8:654–665.
9. Schneider T. Objective comparison of protein structures: error-scaled difference distance matrices. Acta Crystallogr 2000;D56: 714–721.
10. Holm L, Park J. DaliLite workbench for protein structure comparison. Bioinform 2000;6:566–567.
11. Lackner P, Koppensteiner WA, Sippl MJ, Domingues FS. ProSup: a refined tool for protein structure alignment. Protein Eng 2000;13: 745–752.
12. Schneider T. A genetic algorithm for the identification of conformationally invariant regions in protein molecules. Acta Crystallogr 2002;D58:195–208.
13. Ye Y, Jaroszewski L, Li W, Godzik A. A segment alignment approach to protein comparison. Bioinform 2003;19:742–749.
14. Ye Y, Godzik A. Flexible structure alignment by chaining aligned fragment pairs allowing twists. Bioinform 2003;19:ii246–ii255.
15. Shapiro J, Brutlag D. FoldMiner: structural motif discovery using an improved superposition algorithm. Protein Sci 2004;13:278–294.
16. Ilyin VA, Abyzov A, Leslin C. Structural alignment of proteins by a novel TOPOFIT method, as a superimposition of common volumes at a topomax point. Protein Sci 2004;13:1865–1874.
17. Wako H, Yamato T. Novel method to detect a motif of local structure in different protein conformations. Protein Eng 1998;11: 981–990.
18. Gilbert D, Westhead D, Nagano N, Thornton J. Motif-based searching in TOPS protein topology databases. Bioinformatics 1999;15:317–326.
19. Martin ACR. The ups and downs of protein topology; rapid comparison of protein structure. Protein Eng 2000;13:829–837.
20. Rogen P, Fain B. Automatic classification of protein structures using Gauss integrals. Proc Natl Acad Sci USA 2003;100:119–124.
21. Bostick D, Vaisman II. A new topological method to measure protein structure similarity. Biochem Biophys Res Commun 2003; 302:320–325.
22. Bostick D, Shen M, Vaisman II. A simple topological representation of protein structure: implications for new, fast, and robust structural classification. Proteins 2004;56:487–501.
23. Sunyaev SR, Bogopolsky GA, Oleynikova NV, Vlasov PK, Finkelstein AV, Roytberg MA. From analysis of protein structural alignments toward a novel approach to align protein sequences. Proteins 2004;54:569–582.
24. Dokholyan NV, Shakhnovich B, Shakhnovich EI. Expanding protein universe and its origin from the biological Big Bang. Proc Natl Acad Sci USA 1999;99:14132–14136.
25. Needleman S, Wunsch C. A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol 1970;48:443–453.
26. Smith TF, Waterman MS. Identification of common molecular subsequences. J Mol Biol 1981;147:195–197.
27. Koehl P. Protein structure similarities. Curr Opin Struct Biol 2001;11:348–353.
28. Novotny M, Madsen D, Kleywegt GJ. Evaluation of protein fold comparison servers. Proteins 2004;54:260–270.
29. Orengo C. Classification of protein folds. Curr Opin Struct Biol 1994;4:429–440.
30. Gibret J-F, Madej T, Bryant SH. Surprising similarities in structure comparison. Curr Opin Struct Biol 1996;6:377–385.
31. Preparata F, Shamos M. Computational geometry. New York: Springer-Verlag; 1985.
32. Finney JL. Random packing and the structure of simple liquids I. The geometry of random close packing. Proc R Soc 1970;319:479–493.
33. Richards FM. The interpretation of protein structures: total volume, group volume distributions, and packing density. J Mol Biol 1974;82:1–14.
34. Tropsha A, Carter CW Jr, Cammer S, Vaisman II. Simplicial neighborhood analysis of protein packing (SNAPP) a computational geometry approach to studying proteins. Methods Enzymol 2003;374:509–544.
35. Zheng W, Cho SJ, Vaisman II, Tropsha A. A new approach to protein fold recognition based on Delaunay tessellation of protein structure. In: Altman RB et al., Eds. Pacific Symposium on Biocomputing '97. Singapore: World Scientific; 1997. p 487–496.
36. Carter CW Jr, LeFebvre BC, Cammer SA, Tropsha A, Edgell MH. Four-body potentials reveal protein-specific correlations to stability changes caused by hydrophobic core mutations. J Mol Biol 2001;311:625–638.
37. Fujinaga M, Berthet-Colominas C, Yaremchuk AD, Tukalo MA, Cusack S. Refined crystal structure of the seryl-tRNA synthetase from *Thermus thermophilus* at 2.5A resolution. J Mol Biol 1993; 234:222–233.
38. Ilyin VA, Temple B, Li G-P, Vachette YP, Carter CW Jr. 2.9A

Crystal structure of ligand-free tryptophanyl-tRNA synthetase: domain movements fragment the adenine nucleotide binding site. Protein Sci 2000;9:218–231.
39. Sugiura I, Nureki O, Ugaji Y, Kuwabara S, Shimada A, Tateno M, Lober B, Giege R, Moras D, Yokoyama S, Konno M. The 2.0 A crystal structure of *Thermus thermophilus* methionyl-tRNA synthetase reveals two RNA-binding modules present in other class I enzymes. Struct Fold Des 2000;8:197–208.
40. Cusack S, Yaremchuk A, Tukalo M. The 2A structure of leucyl-tRNA synthetase and its complex with a leucyl-adenylate analogue. EMBO J 2000;19:2351–2361.
41. Rould MA, Perona JJ, Steitz TA. Structural basis of anticodon loop recognition by glutaminyl-tRNA synthetase. Nature 1991;352:213–218.
42. Retailleau P, Huang X, Yin Y, Hu M, Weinreb V, Vachette P, Vonrhein C, Bricogne G, Roversi P, Ilyin V, Carter CW Jr. Interconversion of ATP binding and conformational free energies by tryptophanyl-tRNA synthetase: structures of ATP bound to open and closed, pre-transition-state conformations. J Mol Biol 2002;325:39–63.
43. Luger K, Mader AW, Richmond RK, Sargent DF, Richmond TJ. Crystal structure of the nucleosome core particle at 2.8 A resolution. Nature 1997;389:251–260.
44. Decanniere K, Babu AM, Sandman K, Reeve JN, Heinemann U. Crystal structures of recombinant histones Hmfa and Hmfb from the hyperthermophilic Archaeon *Methanothermus fervidus*. J Mol Biol 2000;303:35–47.
45. Christodoulou E, Vorgias CE. Cloning, overproduction, purification and crystallization of the DNA binding protein Hu from the hyperthermophilic Eubacterium *Thermotoga maritima*. Acta Crystallogr 1998;D54:1043–1045.
46. Arents G, Moudrianakis EN. Topography of the histone octamer surface: repeating structural motifs used in the docking of nucleosomal DNA. Proc Natl Acad Sci USA 1993;90:10489–10493.
47. Ramakrishnan V. Histone structure and the organization of the nucleosome. Annu Rev Biophys Biomol Struct 1997;26:83–112.
48. Kabsch W. A solution for the best rotation to relate two sets of vectors. Acta Crystallogr 1976;A32:922–923.
49. Kabsch W. A discussion of the solution for the best rotation to relate to sets of vectors. Acta Crystallogr 1978;A34:827–827.
50. Collaborative Computational Project, Number 4. The CCP4 suite: programs for protein crystallography. Acta Crystallogr 1994;D50:760–763.
51. Levitt M, Gerstein M. A unified statistical framework for sequence comparison and structure comparison. Proc Natl Acad Sci USA 1998;95:5913–5920.
52. Newberry KJ, Kohn J, Hou Y-M, Perone JJ. Crystallization and preliminary diffraction analysis of escherichia coli cysteinyl-tRNA synthetase. Acta Crystallogr 1999;D55:1046–1047.
53. Sekine S, Nureki O, Dubois DY, Bernier S, Chenevert R, Lapointe J, Vassylyev DG, Yokoyama S. ATP binding by glutamyl-tRNA synthetase is switched to the productive mode by tRNA binding. EMBO J 2003;22:676–688.
54. Rossmann MG, Argos P. Exploring structural homology of proteins. J Mol Biol 1976;105:75–95.
55. Betts L, Xiang S, Short SA, Wolfenden R, Carter CW Jr. Cytidine deaminase. The 2.3 A crystal structure of an enzyme: transition-state analog complex. J Mol Biol 1994;235:635–656.
56. Thoden JB, Holden HM, Wessenberg G, Raushel FM, Rayment I. Structure of carbamoyl phosphate synthetase: a journey of 96A from substrate to product. Biochemistry 1997;36:6305–6316.
57. Cusack S. Eleven down and nine to go. Nat Struct Biol 1995;2:824–831.
58. Moras D. Structural and functional relationships between aminoacyl-tRNA synthetases. Trends Biol Sci 1992;17:159–164.
59. Murzin AG, Brenner SE, Hubbard T, Chothia C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 1995;247:536–540.
60. Krzanowski WJ, Marriott FHC. Multivariate analysis: Part 1—Distributions, ordination and inference. London: Edward Arnold; 1994.
61. Silvian LF, Wang J, Steitz AT. Insights into editing from an Ile-tRNA synthetase structure with $tRNA^{Ile}$ and Mupirocin. Science 1999;285:1074–1077.
62. Serre L, Verdon G, Choinowski T, Hervouet N, Risler JL, Zelwer C. How methionyl-tRNA synthetase creates its amino acid recognition pocket upon L-methionine binding. J Mol Biol 2001;306:863–876.
63. Fukai S, Nureki O, Sekine S-I, Shimada A, Vassylyev DG, Yokoyama S. Mechanisms of modular interactions for $tRNA^{Val}$ recognition by valyl-tRNA synthetase. RNA 2003;9:100–111.
64. Barber CB, Dobkin DP, Huhdanpaa HT. The Quickhull algorithm for convex hulls. ACM Trans Math Software 1996;22:469–483.
65. Anderson E, Bai Z, Bischof C, Blackford S, Demmel J, Dongarra J, Du Croz J, Greenbaum A, Hammarling S, McKenney A, Sorensen D. LAPACK users' guide. Philadelphia: Society for Industrial and Applied Mathematics; 1999.
66. SAS Institute Inc. Statistics and graphics guide. Cary: SAS Institute Inc.; 2002.
67. Delano WL. The PyMol users' manual. San Carlos: Delano Scientific; 2002.
68. Kraulis PJ. MOLSCRIPT: a program to produce both detailed and schematic plots of protein structures. J Appl Crystallogr 1991;24:946–950.
69. Merritt EA, Bacon DJ. Raster3D: photorealistic molecular graphics. Methods Enzymol 1997;277:505–524.