Residue Contexts: Non-sequential Protein Structure Alignment Using Structural and Biochemical Features

Jay W. Kim[1] and Rahul Singh[2,*]

[1] Department of Biology
[2] Department of Computer Science,
San Francisco State University, 1600 Holloway Ave., San Francisco, CA 94132, USA
rsingh@cs.sfsu.edu

Abstract. The study of non-sequential alignments, with different connectivity of the aligned fragments in the proteins being compared can offer a more complete picture of the structural, evolutionary and functional relationship between two proteins, than what is possible purely with sequential alignments. The design of techniques for non-sequential protein structure alignment therefore, constitutes an important direction of research. This paper introduces a novel method for non-sequential protein structure alignment involving three principle technical facets: (1) determination of the seed alignments not just by matching features from a single residue or considering well defined regions in the structure such as α–helices and β-strands, but through rich and robust descriptors that can capture the structural similarities of the local 3D environment around arbitrary residues of interest. (2) Scoring alignments using both geometric criterion (RMSD) as well as the biochemical characteristics of the residues. (3) An iterative chaining process which alternates between refinement and non-sequential extension stages to build a final alignment. The efficacy of the approach is demonstrated using the RIPC reference set which includes 40 structural pairs that are problematic to align. The performance of the method was found to be comparable or better than established techniques across the experiments.

1 Introduction

Given two structures, the problem of determining their structural similarity involves determining the correspondence of homologous residues between them such that each pair of aligned residues fulfils equivalent functional and structural roles. The ability to reason about structure, in a comparative setting, is important in providing a mechanistic understanding of the structure-property relationships that constitute the process of “life”. Given that structure-level conservation is often much higher than sequence-level conservation, techniques for structure similarity can also provide clues to the unknown molecular function of a protein based on its structural similarity to one or more proteins of known function(s). Finally, structure similarity lies at the core of classifying protein structures and has been used in a variety of classification schemes such as SCOP, CATH and FSSP to name a few.

[*] Corresponding author. Research funded by NSF grant IIS-0644418.

M. Borodovsky et al. (Eds.): ISBRA 2010, LNBI 6053, pp. 77–88, 2010.

The problem of structure matching or alignment has been widely studied during the past three decades leading to an increasingly deeper understanding of the challenges. For closely related proteins, different methods generally output consistent alignments. However recent studies have revealed significant inconsistencies between alignment methods for distantly related proteins [1]. Such inconsistencies arise when two related proteins display considerable structural variability resulting from the evolutionary accumulation of mutations [2]. Determining non-sequential alignments, with different connectivity of the aligned fragments in the proteins being compared, constitute an intriguing problem in this context. One such example is a circularly permutated protein where the evolutionary divergence from an ancestor has resulted in a change in domain ordering. In such cases, an accurate alignment of a circularly permutated regions, region swaps and β-hairpin flips requires that a matching/alignment technique align individual residues or fragments while disregarding their natural sequence and order. A commonly encountered example is that of the Rossmann structure motif, which comprises of four α–helices and four β–strands and can be found with different SSE connectivity. It should be noted that proteins requiring non-sequential alignments comprise a non-trivial proportion of known protein structures (estimated to be between 17.4% and 35.2% of all alignments [3]). In proposing a solution to the non-sequential alignment problem, the method proposed in this paper seeks to focus on the following two important sub-problems:

- Design of algorithms for determining the initial (seed) alignments, based on which, the ultimate alignment is obtained. The goal is to determine the seed alignments, not just by matching features from a single residue or considering well-defined regions in the structure such as α–helices and β-strands, but through rich and robust descriptors that can capture the structural similarities of the local 3D environment around arbitrary residues of interest.
- Determination of the alignments, not just based on geometric criteria (such as RMSD), but also by involving biochemical characteristics of the residues.

To motivate the importance of the first sub-problem, a brief review of different non-sequential alignment techniques is necessary. These methods can be broadly classified into two groups based on how the initial (seed) correspondences between substructures of the two proteins are detected: residue-based seed matching (RSM) methods and secondary structure element-based (SSE) methods. Examples of RSM methods include STSA [4], and our method. In such methods, the initial correspondences are obtained by modelling and matching substructures in terms of their geometric properties (though in our method, we employ both geometric and biochemical characteristics). In SCALI [5], correspondence is established when the fragments being matched contain greater than five residues, do not have any gaps/insertions, do not have residues with backbone angles differing by greater than $90^{o,}$ and are not part of longer fragments considered earlier. In contrast to RSM methods, SSE-based methods ameliorate the complexity of finding initial correspondences by focussing on similar secondary structure elements (α-helices and β-strands). In GANGSTA [6], pair contacts and relative orientations between SSE are maximized using a genetic algorithm. Next, residue pair contacts between the best SSE-alignment are optimized. In SSM [7], correspondences are obtained by graph matching based on SSEs. In addition to sequential alignment, the method allows complete non-sequential alignment, where the connectivity is neglected, and a "soft" alignment, where the general order of SSEs is retained with the provision that any number of intervening unmatched/missing SSEs are allowed. Finally, TOPOFIT [3] constitutes a technique which does not clearly fall in either of the aforementioned two groups. In TOPOFIT, Delaunay triangulation of the points representing the proteins is used to construct tetrahedrons which are subsequently matched in terms of shape, volume, and backbone topology to find the seed correspondences. In summary, methods that use SSEs to find seed correspondences, ultimately treat the protein structure at a coarser level of granularity, than what is possible at the residue or atomic level. While this allows ameliorating the match complexity, it is possible to miss seed correspondences that do not fall in regions corresponding to well-defined SSEs. In contrast, such a risk is inherently lower in RSM methods which treat the structure at a finer granularity. However, this does require solving a more complex correspondence problem.

2 Proposed Method

We approach the problem of non-sequential structure alignment, in context of the two aforementioned sub-problems, as a three-step process:

(1) *Determining the initial correspondence (seed determination)*: The initial correspondence provides a (possibly coarse) match between similar substructures in the two molecules and can be thought of as the first approximation of the alignment. To determine the initial match we propose a novel rotation invariant and geometrically rich local structure descriptor, which we call the *residue context*. The residue context is a quantized description of the *distribution* of atoms or residues in 3D space with respect to a given point on the protein structure. In this work, the residue context is determined at each C_α atom on the protein backbone. Thus, solving the initial correspondence problem reduces to finding for each C_α atom on one structure, the corresponding C_α atom on the other structure that has the most similar residue context. We formulate the problem of determining the similarity of two residue contexts in terms of the transportation problem, which is a special case of linear programming. This allows us to use the Earth Mover's Distance (EMD) [12] to efficiently address this question. A fundamental advantage of our matching formulation is that it naturally overcomes representation variations that occur due to quantization.

(2) *AFP Generation using Structural and Bio-Chemical information*: In this step, regions are identified using a geometric-fit criteria and analyzed based on the biochemical agreement of the aligned residues, to obtain aligned fragment pairs (AFP). Inspection through this "double lens" of geometric and physicochemical properties raises the likelihood of only procuring desirable AFPs, which serve as interchangeable building blocks for the construction of the final alignment.

(3) *AFP Chaining and Refinement*: In the final step, the AFPs undergo stages of assembly and restructuring as determined by a composite alignment score, to obtain longer and more accurate alignments. The chaining and refinement stages are iterative such that hard correspondences are not assigned until the alignment score does not improve further. The final alignment can be non-sequential or sequential and is driven solely by the structures being compared.

2.1 Definition of Residue Context

In the first step of the proposed method, similar substructures in the two molecules are determined by capturing the structural similarities of the local 3D environment around arbitrary residues of interest through their residue contexts. The design of this descriptor is motivated by research in computer vision on shape recognition [9]. The underlying insight utilizes results from stereopsis indicating that determining correspondences between shapes is easier with rich local descriptors (such as the one proposed in [9] as well as residue context) as opposed to features that are dependent on single shape primitives. Our research extends to 3D molecular structures, the basic idea of shape-context descriptors introduced in [9], namely that given a point on a shape, the distribution of other shape points around this point constitutes a compact, yet highly discriminative descriptor of the local shape geometry.

The notion of residue context, as proposed by us can be described as follows: given the protein backbone defined through the 3D coordinates of its constituent C_α-atoms, and a reference C_α-atom, consider the set of n-1 vectors originating from the reference C_α atom to all the other C_α-atoms of the backbone. These vectors describe the configuration of the entire backbone shape relative to the reference atom and can be thought of as to constitute its local shape context in 3D space. It may be noted that the set of n-1 vectors constitutes a rich description, since, as n (the number of C_α-atoms) increases, the representation of the backbone shape becomes exact. The distribution of the vectors centered at a reference C_α-atom can be succinctly represented using a 3D spherical histogram centered at the reference atom. Further, each of the vectors can be defined by three parameters in a spherical coordinate system: the radial distance $\vec{r}$ corresponding to the distance between the reference C_α-atom and another C_α-atom, the azimuthal (longitude) angle θ in the x-y plane from the x-axis with $0 \le \theta \le 2\pi$ and the polar (latitude) angle ϕ from the z-axis with $0 \le \phi \le \pi$. Following [9], we require the 3D spherical histogram centered on the reference C_α-atom to have the following two properties:

- The descriptor needs to be more sensitive to nearby residues than residues that are farther away. This property corresponds to the importance of proximity in defining intermolecular interactions. To ensure this property, the magnitude of r is logarithmically discretized and the longitude angle is uniformly discretized in the range [0, 2π].
- Bins equidistant from the center should cover the same surface area. This property ensures that the representation is isotropic in space. To support it, the latitude angle $\phi \in [-\pi/2, \pi/2]$, is discretized non-uniformly, such that each ϕ_i satisfies the relationship in Eq. (1), where the righthand side denotes the i^{th} fraction of the surface area of the upper hemisphere:

$$\int_{\theta=0}^{2\pi} \int_{\phi=0}^{\phi_i} r^2 \cos\phi \, d\phi \, d\theta = \frac{i}{N} 2\pi r \tag{1}$$

From Eq. (1), the required discretization of the latitude angle is $\phi_i = \arcsin(i/N)$.

Given a reference C_α-atom and a spherical histogram centered on it, the residue context of this C_α-atom is constituted by the distribution, within the bins of the histogram, of the other n-1 C_α-atoms of the protein backbone, that fall within the radius r. Specifically, for a reference atom $C\alpha_j$, the histogram H_j of the relative coordinates of the remaining n-1 atoms is given as:

$$H_j(k) = |\{C\alpha_i \neq C\alpha_j : (C\alpha_i - C\alpha_j) \in bin(k)\}| \tag{2}$$

In Eq.(2), $C\alpha_j$ is the reference atom, $C\alpha_i$ indexes the set of neighboring atoms located within the radius r of the reference atom, and |.| denotes the number of the neighboring reference C_α-atoms that fall within the k^{th}-bin of the histogram. An example, illustrating this concept is shown in Figure 1. Finally, given a molecule M, consisting of m alpha-Carbon atoms $C\alpha_i$, $i=1,\ldots,m$, its residue context-based description, denoted as $R(M)$, consists of the set of m histograms H_i, with each centered on one of the alpha-Carbon atoms of M: $R(M) = \{H_1, H_2,\ldots, H_m\}$.

One important practical issue in defining residue contexts is that of the context scale (size), which is specified by the choice of the context radius r. A large r, which considers a more global environment around each residue, can be useful for simple alignments consisting of two proteins with high sequence and structural similarities. Conversely, a smaller r may be necessary for making difficult non-sequential alignments and aligning two proteins of low sequence and structural similarities. In such cases, a large residue context may be counter-productive since it would incorporate variances due to extensive insertions, deletions, repetitions, and conformational variability. In subsection 2.4 we further address the issue of automatic scale selection for alignment.

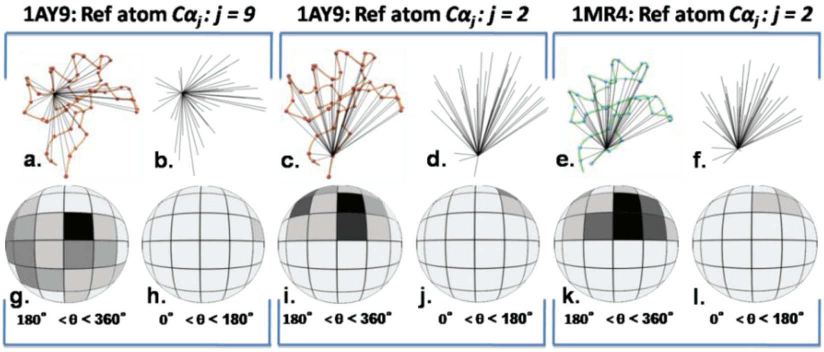

1AY9: Ref atom $C\alpha_j$: j = 9
1AY9: Ref atom $C\alpha_j$: j = 2
1MR4: Ref atom $C\alpha_j$: j = 2
a.
b.
c.
d.
e.
f.
g.
h.
i.
j.
k.
l.
180° <θ < 360°
0° <θ < 180°
180° <θ < 360°
0° <θ < 180°
180° <θ < 360°
0° <θ < 180°

Fig. 1. 3D backbone representations of 1AYJ **(a,c)** and its homolog 1MR4 **(e)** where positions of C_α-atoms are shown as red or blue dots. 3D vectors originating from reference residue j=9 and j=2 of 1AYJ to all other C_α-atoms are shown in **(b)** and **(d)**,respectively. The corresponding "front" (180° <θ< 360°) and "back" (0° <θ< 180°) views of the residue contexts at these positions are shown in figures **(g-h)**, and figures **(i-j)**. One may note that the residue contexts are clearly distinct for these positions. In **(f)** the 3D vectors originating from reference residue j=2 of 1MR4 are shown. Since 1MR4 is a homolog of 1AYJ, we expect residue contexts at similarly located C_α-atoms in 1AYJ and 1MR4 to be similar. The "front" and "back" views of the residue contexts at j=2 of 1MR4 are shown in figures (k) and (l). The reader may note the similarity of the residue contexts (at reference residue j=2) for 1AYJ and 1MR4 and be comparing the "front" and "back" views from figures **(i)** and **(j)** with the corresponding views in figures **(k)** and **(l)**. In all the figures, darker bins are more heavily populated.

2.3 Efficient Matching of Residue Contexts

The problem of comparing residue contexts can directly be interpreted as that of matching two histograms. Several measures have been proposed to address this problem and they can be broadly classified into two categories. Most fall into the first category of bin-by-bin dissimilarity measures. This includes χ^2 statistics (used in [9] to compare 2D shape contexts), histogram intersection, L_p distances, Kullback-Leibler divergence, Jeffry divergence, and Jensen-Shannon divergence. A fundamental assumption underlying these techniques is that the domain of the histograms can be aligned. However, in practice, this assumption can be violated due to noise, sub-optimal quantization (binning), different number of bins, or the inherent nature of the data. The second category of measures is called cross-bin measures. Cross-bin measures utilize the ground distance between representative features in different bins to compare both aligned and non-aligned bins. The earth-mover's distance (EMD) [8] is an example of such a measure and is used by us.

Given two residue contexts, defined in terms of their respective histograms P and Q, one of them can be interpreted as a mass distribution spread on the underlying space and the other as a collection of holes in that same space. If a unit of work corresponds to transporting a unit of mass by a unit of ground distance, then the matching problem can be defined as determining the least amount of work required to fill the holes. This precisely corresponds to the EMD between the two distributions. Following [12], we formalize our problem as follows: Let the first histogram be represented by a set of tuples P = {<$\boldsymbol{p}_1$, w_{p1}>, <$\boldsymbol{p}_2$, w_{p2}>,…,<$\boldsymbol{p}_m$, w_{pm}>}, where the i^{th} bin is represented by the tuple <$\boldsymbol{p}_i$, w_i> with $\boldsymbol{p}_i$ denoting an appropriately chosen bin representative (such as its mean, centroid, or medoid) and w_{pi} the weight of the i^{th} bin, given by the fraction of residues from the context that fall into this bin. Similarly, let Q = {<$\boldsymbol{q}_1$, w_{q1}>, <$\boldsymbol{q}_2$, w_{q2}>, …,<$\boldsymbol{q}_m$, w_{qm}>} be the tuple set representing the second histogram and d_{ij} denote the ground distance between bins $\boldsymbol{p}_i$ and $\boldsymbol{q}_j$ (we use the Euclidean distance as the ground distance). Matching the residue contexts by computing the EMD requires solving the following minimization problem, where f_{ij} denotes the flow between $\boldsymbol{p}_i$ and $\boldsymbol{q}_j$:

$$\underset{f_{ij}}{\arg\min} \sum_{i=1}^{m}\sum_{j=1}^{n} d_{ij} f_{ij} \tag{3}$$

The minimization is subject to the constraints (4) – (7) below, where the constraint (4) ensures that the mass is moved in only one direction, constraint (5) and (6) ensure that the mass sent by bins in P and the mass received by bins in Q is limited to their weights, and constraint (7) requires that the maximum possible amount of mass is moved.

$$f_{ij} \geq 0, i = 1,2,...,m; j = 1,2,...,n \tag{4}$$

$$\sum_{j=1}^{n} f_{ij} \leq w_{p_i}, i = 1,2,...,m \tag{5}$$

$$\sum_{i=1}^{m} f_{ij} \leq w_{q_j}, j = 1,2,...,n \tag{6}$$

$$\sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij} = \min\left(\sum_{i=1}^{m} w_{p_i}, \sum_{j=1}^{n} w_{q_j}\right) \tag{7}$$

Given the optimal flows f_{ij} obtained from solving the transportation problem as described above, the EMD between the two residue contexts is defined as:

$$EMD(P,Q) = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n} d_{ij} f_{ij}}{\sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij}} \quad (8)$$

In Eq. (8), the numerator denotes the resulting work and the denominator describes the total flow.

2.4 Determining the Scale of the Residue Context

The size of the environment around each residue of a protein chain is determined by the radius *r* which is logarithmically discretized. For arbitrary alignments it is not possible to determine, *a priori*, the value of *r* for the optimal context size. We use a data driven procedure where the optimal value of the radius is defined as the one which best captures the similarity between two sub-chains across all possible values of the radius. This is done by computing *n* cost matrices C_n, which correspondingly store the costs associated with residue matches using contexts of varying radii r_n. The optimal cost matrix corresponds to the most similar contexts given by the lowest matching costs (Eq. 9) across the radii.

$$C_{ij}^{opt} = \min\left(\frac{C_1}{a_1+b_1}, \frac{C_2}{a_2+b_2}, \ldots, \frac{C_i}{a_i+b_i}\right) \quad (9) \qquad M_{ij} = low + \frac{-C_{ij}^{opt} + \min}{\max - \min}(high - low) \quad (10)$$

In Eq. (9), a_n and $b_n$are the respective number of residues from protein chain *A* and *B* with context radii r_n. Next, the entries in the optimal cost matrix are normalized to lie in the interval [*low*, *high*], with *low* set to -100 and high set to 100 using Eq. (10). Matching two histograms whose bins are identically populated yields a score of 100. An example illustrating the intuition underlying the notion of residue context-based description and matching is described in Fig. 1.

2.5 AFP Generation and Scoring

Given two structures to be aligned as input, we define an aligned fragment pair (AFP) as a correspondence of residues between fragments from each structure. Our definition of an AFP differs from the original definition given by Shindyalov and Bourne [10] in that we allow single-residue gaps in either fragments' sequence to accommodate for single-residue insertions/deletions encountered in aligning structures displaying low sequence similarity. Larger gaps are naturally accommodated by the mechanics of our AFP chaining algorithm described later. Given the entire set of residues $\{p\}$ from molecule A and $\{q\}$ from molecule B, for the identification of AFPs, we first locate all triplets of residue pairs $\tau_{ij} = \{(p_{i-1},q_{j-1}),(p_i,q_j),(p_{i+1},q_{j+1})\}$ which occur continuously along a diagonal of the similarity matrix *M* such that $M_{i-1,j-1}$, $M_{i,j}$, and $M_{i+1,j+1}$ all exceed an AFP initiation threshold value *t*. The set of residues $\{q\}$ is transformed such that the three pairs of residues defined by τ are optimally superimposed and thedistances between the residues of $\{p\}$ and $\{q\}$ are stored in a matrix D_{ij}. Next, each triplet is extended in both the N-terminal and C-terminal directions based on the following two conditions, as long as the EMD score $M_{i+1,j+1}$ stays below an extension threshold *e*and the aligned distance of the extended correspondence does not exceed the aligned distance of the C-terminal correspondence (prior to extension) by 3Å. The two conditions are: ***cond1:*** $M_{i+2,j+1} \geq M_{i+1,j+1}$ && $M_{i+3,j+2} \geq M_{i+2,j+2}$ && $D_{i+2,j+1} < D_{i+1,j+1}$; ***cond2***: $M_{i+1,j+2} \geq M_{i+1,j+1}$ && $M_{i+2,j+3} \geq M_{i+2,j+2}$ && $D_{i+1,j+2} < D_{i+1,j+1}$. If only ***cond1*** holds, the correspondence (p_{i+2},q_{j+1}) is added to the AFP. Similarly, if only ***cond2*** holds, (p_{i+1},q_{j+2}) is added. Finally, if both conditions hold, then the correspondence with the lowest RMSD is added. The resulting set of AFPs $F = \{f\}$ is filtered to ensure that an AFP contains a minimum of 4 residue correspondences. Further, AFPs that are completely contained within a larger AFP are discarded. Note that although each $f \in F$ is unique in its entirety, AFPs are allowed to extend freely with partial overlap to avoid introducing bias based on the initial triplet locations. At this point, any two or more overlapping AFPs represent a collection of residue pair correspondences whose final alignment path has not yet been determined. This uncertainty is resolved during the subsequent AFP chaining step in section 2.6.

The AFPs are next ranked using an AFP alignment score *AS* (Eq. 11) which is the weighted sum of two component scores; the structural score *SS* is defined simply as the sum of the similarity scores along the length *l* of the AFP (Eq. (12)).

$$AS = w_{SS} + w_{BS} \qquad (11) \qquad\qquad SS = \sum_{i,j=1}^{l} M_{ij} \qquad (12)$$

The second component of the alignment score is the biochemical score *BS*. The biochemical score captures the likelihood of the evolutionary occurrence of each pair-wise amino acid substitution as suggested by the residue correspondences defined by the AFP. Its use is motivated by the assumption that among structurally and functionally conserved proteins, the frequency of amino acid substitutions at a given site is correlated with the physicochemical similarities between exchanged amino acids. Thus by rewarding biochemical agreement between aligned residues, we seek to select a pool of AFPs that contain conserved functional alignments between two proteins. To compute the biochemical score *BS*, the Blosum62 matrix is used to estimate the likelihood of occurrence of each possible pair-wise amino acid substitution. Depending on the likelihood of substitution, each pair of aligned residues in an AFP earns a predefined numerical score towards the total biochemical score for the AFP as follows: the Blosum62 matrix values are first normalized over the interval [*low*, *high*] with *low* set to -100 and *high* set to 100. The normalized Blosum62 matrix B_{ij} is computed using the following equation and values less than zero are reset to zero:

$$B_{ij} = low + \frac{-Blosum_{ij}^{opt} + \min}{\max - \min}(high - low) \qquad (13)$$

In Eq. (13), *max* and *min* denote the highest and lowest values found in the Blosum62 matrix. In our current investigations, the SS and BS components are given equal weight, that is, $w_{SS} = w_{BS} = 1$. However, these weights can be changed, if needed, to emphasize either of the components.

2.6 Chaining and Refinement Using Structural and Biochemical Scores

Given two molecules *A* and *B*, the chaining process starts with the seed alignments captured by the AFPs. A crucial challenge in extending the seed alignments, is that of avoiding spurious alignments. Specifically, during protein alignment numerous short AFPs of length 4~6 residues can be encountered which can be superimposed at low RMSD values. However, such alignments are often misleading. For example, two β-strands of length 4~6 or two segments consisting of 1~2 turns of an α-helix are often structurally similar in any two protein. Thus, a strategy that simply minimizes RMSD can lead to incorrect alignments. We therefore chain the AFPs by using the alignment score *AS* defined earlier. It may be noted that this score consists of the EMD score (capturing topological local shape similarity) and the biochemical scores (reflecting biochemical similarity). The initial chain is constructed as follows: we begin with the set of k AFPs denoted as $\{f^k\}$. This set is sorted by the *AS* score and the highest scoring AFP is denoted as f^1. The initial alignment is $c^{init}=f^1$. Next, an optimal (in the least square sense) Euclidean transformation T is calculated to align the subset of residues $\{q\}$ from molecule B with the corresponding residues $\{p\}$ from molecule A, where the correspondences are given by c^{init}. This optimal transformation T is applied to molecule B to give B^* and the RMSD between subset $\{p\}$ and transformed subset $\{q^*\}$ from B^* is stored as the chain RMSD. Further, a chain alignment score *CS* is determined as follows:

$$CS = \frac{w_{SS} + w_{BS}}{RMSD} \tag{14}$$

Subsequently, the remaining APFs in $\{f^k\}$ are treated as follows. Any AFPs that overlap with the residues contained in the current chain are discarded. Thus each residue p_i contained in the chain must have a unique corresponding residue q_j and vice versa. A non-overlapping AFP is added to the chain only if its addition increases the alignment score *CS*. After all AFPs have been considered, the resulting chain is stored in the set of chains *C*. The initial chaining process is repeated, substituting the next highest scoring AFP denoted as f^2 for the initial seed alignment, and iterated for each of the top 50 AFPs. The highest scoring chain from *C* is passed to the refinement step.

For the refinement of a chain, we first reduce $\{f^k\}$ to only include AFPs which overlap the immediate vicinity of a residue correspondence stored in the chain. Given a correspondence (p_i,q_j), its immediate vicinity is defined on an i x j alignment matrix as the area enclosed by: $(p_{i-\delta},q_{j-\delta})$,$(p_{i-\delta},q_{j+\delta})$,$(p_{i+\delta},q_{j-\delta})$, and $(p_{i+\delta},q_{j+\delta})$. In all our experiments, δ is set to 3. As in the initial chaining step, each AFP in the reduced set is considered and included only if *CS* increases after its addition while giving priority to the new correspondences in case of overlap. Thus any redundant residue and its corresponding partner are removed from the current chain before the new correspondence given by the AFP is added. The chaining and refinement steps are iterated using the refined chain as input for each chaining step. The process is stopped when the chain alignment score *CS* converges or changes between successive iterations become smaller than a predefined threshold. Before the final alignment is output, the N and C-terminal correspondences of chained AFPs are briefly extended as described in section 2.5.

3 Experimental Investigations and Results

The RIPC set comprises 40 structural pairs that are problematic to align [1]. Each pair in this set is characterized by repetitions, extensive insertions and deletions, circular permutations, and/or conformational variability. Human-curated reference alignments based on conservation of sequence and function are provided for 23 out of 40 protein pairs. Agreement to the reference alignments is measured for each pair by the fraction of correctly aligned residues, f_{CAR}, numerically defined as:

$$f_{CAR} = \text{\# of correctly aligned residues / \# of reference pairs .} \tag{15}$$

We compared our performance against 3 non-sequential alignment methods (GANGSTA, TOPOFIT, STSA), and 4 sequential alignment methods (DALI [11], CE [8], MATT [12], FATCAT [13]) (Fig. 2). MATT and FATCAT are also flexible aligners that allow twists and translations to the protein backbone to accommodate for conformational variability.

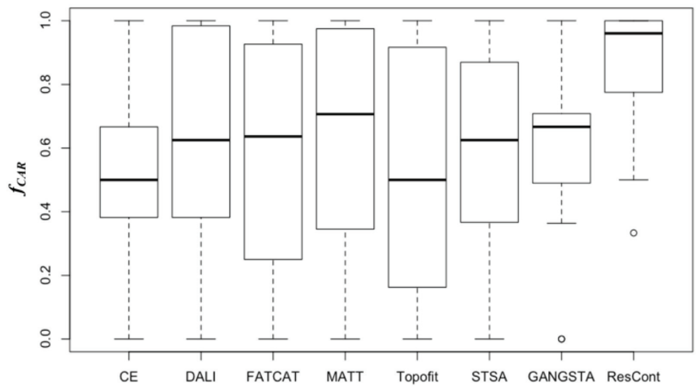


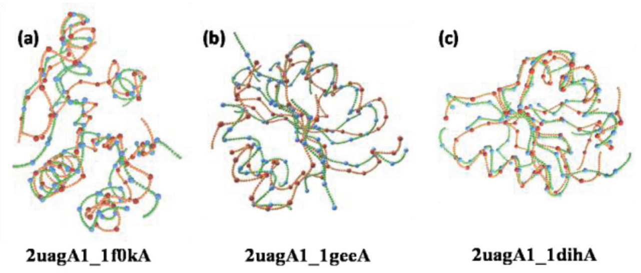
Fig. 2. Comparison of various methods' performances on the RIPC reference set. Box and whisker plot properties are as follows: bottom whisker – min sample, lower box boundary – 1st quartile, bolded line – median, upper box boundary – 3rd quartile, top whisker – max sample. The statistical median for each method was calculated from 23 samples (alignments) where the f_{CAR} for each alignment served as a measure of agreement with the RIPC reference.

Among the methods tested, Residue Context showed the highest agreement with the reference set (median = 96%) and Matt was second highest (median = 71%). Residue Context was the only method which correctly aligned at least one reference pair for each of the 23 alignments. The lowest f_{CAR} obtained using our method was for the alignment of an E6AP-UbcH7 complex (d1d5fa_) to a HECT domain E3 ligase (d1nd7a_) for which 4 of 6 reference pairs were missed. The alignment requires accounting for considerable conformational variability to correctly align all reference pairs. DALI and FATCAT managed to correctly align all 6 reference pairs. The alignment of an L-2-haloacid dehalogenase (d1qq5a_) and *E. Coli* CheY (d3chy__) confounded most methods. This alignment involves a circular permutation, and also extensive insertions are present in d1qq5a_ with respect to d3chy__. We found that non-sequential methods as well as sequential methods produced inconsistent alignments. Only Residue Context was able to align all 3 reference residues correctly. On the other hand, Topofit and STSA missed all 3 reference residues (Table 1). When only considering non-sequential alignments, Residue Context had the highest median and mean among 4 methods at 94% and 89%, respectively. Topofit had the 2nd highest median, but also displayed the greatest inconsistency between alignments as evidenced by the disparity between its Q3 (94%) and Q1 (25%) values. Residue Context was the most consistent (Q3 = 100%; Q1 = 81%) method across the dataset.

Table 1. Statistical comparison of performances of three non-sequential methods on non-sequentially related pairs from the RIPC set.The mean, median, 1st quartile (Q1), and 3rd quartile (Q3) were calculated using f_{CAR} values from 10 non-sequential alignments of the RIPC set. The highest scores (including ties) for each alignment and statistical category are bolded.

	Residue Context					**STSA**				
Aligned Pair	**Length**	**RMSD**	**Aligned**	**Total**	f_{CAR}	**Length**	**RMSD**	**Aligned**	**Total**	f_{CAR}
d1nkl__-d1qdma1	73	2.61	54	72	0.75	74	2.26	72	72	**1.00**
d1nls__-d2bqpa_	221	1.44	6	6	**1.00**	212	1.50	2	6	0.33
d1qasa2-d1rsy__	113	1.87	72	75	**0.96**	111	1.94	67	75	0.89
d1b5ta_-d1k87a2	227	3.80	5	8	**0.63**	188	2.84	5	8	**0.63**
d1jwyb_-d1puja_	148	3.34	11	12	**0.92**	116	2.34	9	12	0.75
d1jwyb_-d1u0la2	124	3.69	11	11	**1.00**	99	2.68	8	11	0.73
d1nw5a_-d2adma_	166	3.96	13	13	**1.00**	120	2.57	11	13	0.85
d1gsa_1-d2hgsa1	83	3.29	4	5	**0.80**	229	2.59	2	5	0.40
d1qq5a_-d3chy__	107	3.48	3	3	**1.00**	92	2.73	0	3	0.00
d1kiaa_-d1nw5a_	162	3.83	10	12	0.83	90	3.37	0	12	0.00
	Mean	**Q1**	**median**	**Q3**		**mean**	**Q1**	**median**	**Q3**	
	0.89	**0.81**	**0.94**	**1.00**		0.56	0.35	0.68	0.82	
	GANGSTA					**Topofit**				
Aligned Pair	**Length**	**RMSD**	**Aligned**	**Total**	f_{CAR}	**Length**	**RMSD**	**Aligned**	**Total**	f_{CAR}
d1nkl__-d1qdma1	74	2.41	72	72	**1.00**	56	1.65	28	72	0.39
d1nls__-d2bqpa_	222	3.23	4	6	0.67	212	1.01	6	6	**1.00**
d1qasa2-d1rsy__	115	2.95	44	75	0.59	105	1.16	71	75	0.95
d1b5ta_-d1k87a2	181	3.34	5	8	**0.63**	134	1.85	1	8	0.13
d1jwyb_-d1puja_	137	2.83	9	12	1.75	108	1.65	11	12	**0.92**
d1jwyb_-d1u0la2	111	2.60	11	11	**1.00**	99	1.58	11	11	**1.00**
d1nw5a_-d2adma_	146	2.98	13	13	**1.00**	93	1.61	11	13	0.85
d1gsa_1-d2hgsa1	75	2.38	2	5	0.40	66	1.44	1	5	0.20
d1qq5a_-d3chy__	101	3.36	2	3	0.67	63	1.65	0	3	0.00
d1kiaa_-d1nw5a_	150	2.94	8	12	0.67	132	1.73	11	12	**0.92**
	Mean	**Q1**	**median**	**Q3**		**mean**	**Q1**	**median**	**Q3**	
	0.74	0.64	0.67	0.94		0.64	0.25	0.89	0.94	

In the next experiment we compared the performances of three non-sequential methods on the alignment of structures involving the Rossmann fold (data from [3]). GANGSTA generally produced the longest alignments while Residue Context produced alignments of comparable lengths at significantly lower RMSDs (Table 2). Topofit generated significantly shorter alignments at lower RMSDs. Examples of 3D structural representations of alignments generated by Residue Context are shown in Figure 3.



2uagA1_1f0kA 2uagA1_1geeA 2uagA1_1dihA

Fig. 3. 3D structural representations of 3 non-sequential alignments involving the Rossmann fold. (a-c) obtained using the Residue Context method

Table 2. Comparison of three methods on alignments involving the Rossmann fold

	Residue Context			GANGSTA			Topofit		
Structures	**Length**	**RMSD**	**Length/RMSD**	**Length**	**RMSD**	**Length/RMSD**	**Length**	**RMSD**	**Length/RMSD**
2uagA1_1f0kA	83	2.63	31.6	85	3.52	24.1	41	1.34	30.6
2uagA1_1geeA	85	2.93	29.0	89	3.28	27.1	60	1.56	38.5
2uagA1_1dih_1	83	2.46	33.7	82	3.07	26.7	63	1.60	39.4

4 Conclusions

This paper considers the problem of non-sequential protein structure alignment. We have presented a novel approach that involves determining initial (seed) correspondences using a rich descriptor that can capture structural similarities of the local 3D environment around arbitrary residues of interest. Based on these seed correspondences the alignments are constructed using geometric and biochemical characteristics of the involved residues. Experiments indicate that, in terms of alignment quality, the proposed method either exceeds or is comparable with leading methods at the state of the art.

References

1. Mayr, G., Domingues, F.S., Lackner, P.: Comparative Analysis of Protein Structure Alignments. BMC Structural Biology 7(50), 564–577 (2007)
2. Grishin, N.: Fold change in evolution of protein structures. J. Struct. Biol. 134, 167–185 (2001)
3. Ilyin, V.A., Abyzov, A., Leslin, C.M.: Structural alignment of proteins by a novel TOPOFIT method, as a superimposition of common volumes at atopomax point. Protein Sci. 13(7), 1865–1874 (2004)
4. Salem, S., Zaki, M.J., Bystroff, C.: Iterative Non-Sequential Protein Structural Alignment. Journal of Bioinformatics and Computational Biology 7(3), 571–596 (2009)
5. Yuan, X., Bystroff, C.: Non-sequential structure-based alignments reveal topology-independent core packing arrangements in proteins. Bioinformatics 21(7), 1010–1019 (2005)
6. Kolbeck, B., May, P., Schmidt-Goenner, T., Steinke, T., Knapp, E.W.: Connectivity independent protein-structure alignment: a hierarchical approach. BMC Bioinformatics 7, 510 (2006)
7. Krissinel, E., Henrick, K.: Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. ActaCrystallogr D BiolCrystallogr 60(1), 2256–2268 (2004)
8. Rubner, Y., Tomasi, C., Guibas, L.J.: A Metric for Distributions with Applications to Image Databases. In: Proceedings of the 1998 IEEE Conf. on Computer Vision, pp. 59–66 (1998)
9. Belongie, S., Malik, J., Puzicha, J.: Shape Matching and Object Recognition Using Shape Contexts. IEEE Trans.on Pattern Analysis and Machine Intelligence 24, 509–522 (2002)
10. Shindyalov, I.N., Bourne, P.E.: Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. Protein Engineering 11(9), 739–747 (1998)
11. Holm, L., Sander, C.: Protein structure comparison by alignment of distance matrices. J. Mol. Biol. 233(1), 123–138 (1993)
12. Menke, M., Berger, B., Cowen, L.: Matt: local flexibility aids protein multiple structure alignment. PLoSComput Biol. 4(1), e10 (2008)
13. Ye, Y., Godzik, A.: Flexible structure alignment by chaining aligned fragment pairs allowing twists. Bioinformatics 19(2), 246–255 (2003)