Protein Similarity from Knot Theory and Geometric Convolution

Michael A. Erdmann

September 12, 2003

CMU-CS-03-181

School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

This research was supported in part by Carnegie Mellon University, the author, and the Pennsylvania Department of Health through the grant "Integrated Protein Informatics for Cancer Research".

The views and conclusions contained in this document are those of the author and should not be interpreted as representing the official policies, either expressed or implied, of the Pennsylvania Department of Health, or of any other government agency.

Keywords: Protein structure, homology, homotopy, writhing, knot theory, robot motion planning.

Abstract

Shape similarity is one of the most elusive and intriguing questions of nature and mathematics. Proteins provide a rich domain in which to test theories of shape similarity. Proteins can match at different scales and in different arrangements. Sometimes the detection of common local structure is sufficient to infer global alignment of two proteins; at other times it provides false information. Proteins with very low sequence identity may share large substructures, or perhaps just a central core. There are even examples of proteins with nearly identical primary sequence in which α-helices have become β-sheets.

Shape similarity can be formulated (i) in terms of global metrics, such as RMSD or Hausdorff distance, (ii) in terms of subgraph isomorphisms, such as the detection of shared substructures with similar relative locations, or (iii) purely topologically, in terms of the cohomology of structure-preserving transformations. Existing protein structure detection programs are built on the first two types of similarity. The third forms the foundations of knot theory.

The thesis of this paper is: Protein similarity detection leads naturally to an algorithm operating at the metric, relational, and homotopic scales. The paper introduces a definition of similarity based on atomic motions that preserve local backbone topology without incurring significant distance errors. Such motions are motivated by the physical requirements for rearranging subsequences of a protein. Similarity detection then seeks rigid body motions able to overlay pairs of substructures, each related by a substructure-preserving motion, without necessarily requiring global structure preservation. This definition is general enough to span a wide range of questions: One can ask for full rearrangement of one protein into another while preserving global topology, as in drug design; or one can ask for rearrangements of sets of smaller substructures, each of which preserves local but not global topology, as in protein evolution.

In the appendix, we exhibit an algorithm for answering the general question. That algorithm has the complexity of robot motion planning. In the text, we consider a more common case in which one seeks protein similarity by rearrangements of relatively short peptide segments. We exhibit an algorithm based on writhing numbers that runs in $O(n^2)$ time in practice. We define and use a new datastructure, called *geometric self-convolution*, within this algorithm.

Contributions: We believe that this is the first paper to consider carefully the need for combining metric and homotopic qualities in seeking protein similarity. We provide a parameterized definition of similarity that leads naturally to a metric in protein space. We exhibit algorithms for computing the metric and detecting similarity. We report results obtained with three pairs of proteins, each pair exhibiting different typical characteristics.

[This page intentionally left (quasi) blank.]

1 Introduction

Determining structural similarity between proteins is one of the most central and common problems within proteomics, yet there exist no simple universally accepted algorithms for solving this problem. Indeed, the most widely used existing 3D structural alignment tools (e.g., Dali [15], Vast [12], CE [35], and 3dSearch [37]) are likely to disagree in their specific atomic alignments and sometimes even in their top-scoring secondary structure alignments when presented with proteins that have low sequence similarity and low structural similarity.

As of late August 2003 the Protein Data Bank (PDB) [31, 5] contained in excess of 22,000 protein structures, up approximately 1,000 since early June. Many of these proteins are highly similar structures. There are only approximately 4,000 different folds represented in the PDB, roughly a ratio of 1:5 (fold:structure). Given a new protein, the probability is high that it is similar to an existing protein. Detecting such similarity quickly is essential for classifying a protein and understanding its biological function.

More importantly, as the growth in new structures outpaces the growth in new folds, it is likely that the role of structural similarity will need to become much more fine-grained than it is today. Biological discoveries will lie in unusual, possibly very sparse, structural similarities, rather than in rough fold-level classifications. For instance, in looking at the backbone CA atoms of a β-sheet, one can easily detect two roughly orthogonal families of curves in the sheet, one family parallel to the constituent β-strands, the other perpendicular to the strands. This observation raises at least the geometric question whether there are two proteins that place their β-sheets, viewed as two-dimensional sheets, in the same spatial locations, even though the underlying makeup of the one-dimensional strands is orthogonal. Such proteins would hint at some very interesting biochemical/genetic rearrangements. It is unclear whether any existing structural alignment tool today could detect such similarities automatically. It is easy to construct less exotic but equally interesting questions. As X-Ray and NMR methodologies enter high-throughput capability, the questions arise easily, yet many go unanswered.

Lacking is a good definition of "similarity", even for today's alignment tools. The Structural Classification of Proteins (SCOP) website [25] offers the following: "Proteins are defined as having a common fold if they have the same major secondary structures in the same arrangement and with the same topological connections." This sounds good, it is intuitive, and it is applied every day to classify proteins. But what really does it mean? When is a secondary structure "major", when is a collection of secondary structures in the "same arrangement", and which "topological connections" are really relevant?

This paper focuses on topological incidence and polygonal writhing as a gauge of geometric similarity. We take our inspiration from a recent fundamental paper [32] that classifies protein structures in terms of Gauss integrals, motivated by ongoing work on knot invariants [4]. In this paper we explore the connection further, leading naturally to a metric in protein space and a datastructure for representing geometric self-convolutions of polygonal curves. We comment on connections with methods from Robot Motion Planning. Finally, we implement an algorithm for detecting substructural similarities between proteins and report results on three pairs of proteins.

2 Structural Alignment (Related Work)

There are three major structural alignment tools in use today: Dali, Vast, and CE. All three are accessible off the PDB webpage. Since the appearance of these methods in the late 1990s, a host of other methods have appeared, which generally compare themselves to these three. One we have found useful is 3dSearch. We review these four methods here briefly.

Dali [15, 17, 16] aligns protein substructures using distance matrices. Distances are invariant to rigid body transformations, thereby avoiding the need for spatial alignment. Dali considers distances between alpha carbons; the distance matrices are indexed in residue order. Substructures that appear in similar relative spatial locations in the two proteins give rise to similar patterns between blocks of the distance matrices. Dali uses a clever Monte Carlo method to detect these patterns. It begins with small hexapeptides then repeatedly merges similarly related protein fragments into larger common substructures. One important aspect of Dali is an *elastic* similarity score; the significance of errors in distance alignments decreases with increasing distance. Consequently, substructures separated by larger distances can tolerate greater relative global motion, while residues nearer to each other must better preserve local shape. Dali is probably the gold standard for protein structure comparisons. Its main disadvantage is its relatively ad-hoc Monte Carlo structure and complexity.

CE [35, 36] searches for protein fragments in one protein that are locally similar to protein fragments in another protein. It then extends these local alignments by a sequential scan down the protein backbones. This scan is reminiscent of dynamic programming in sequence alignment, but CE actually employs a clever greedy algorithm. CE uses distances between alpha carbons and rigid body superposition to define similarity and to guide the extension scan. A limitation of CE is its requirement that matching substructures occur in sequential backbone order.

Vast [12, 13] and 3dSearch [37, 38] focus on elements of secondary structure to align proteins. Both methods begin with building blocks that are pairs of secondary structure elements, one pair in each protein. Vast matches pairs of secondary structural elements that have a similar type, relative orientation, and connectivity, then builds larger structures by considering substructure similarities that are statistically surprising. This probabilistic similarity function is both a strong advantage and a potential limitation of Vast; a class of "similar" structures is significant, but not necessarily easily circumscribed.

3dSearch first finds a pair of secondary structure vectors in one protein that best matches a pair of vectors in the other protein. This initial alignment seeds a loop whose basic step is a dynamic programming algorithm for aligning the two proteins' secondary structure vectors with each other. Atom-level alignment occurs subsequently. 3dSearch is limited by its initial best-pairing decision.

3 Protein Similarity from Knot Theory

3.1 Writhing and Knot Invariants

One of the difficulties with the previous alignment methods is the vagueness of their global similarity measures. Locally, these methods often measure similarity by the root-meansquare-deviation (RMSD) between aligned atoms, or some related variation. As Røgen and Fain [32] point out, RMSD of aligned atom coordinates is a wonderful measure of similarity for two shapes that are nearly identical. However, RMSD is a poor measure when the two shapes being compared differ significantly, particular when the two shapes contain some matching and some nonmatching subshapes. Existing alignment methods address this issue by seeding their routines with small matching subshapes, then repeatedly merging these into larger shapes. This process often succeeds well, but it is purely procedural. As a result, *automatic classification* of proteins remains brittle.

One possible alternative is to compare proteins using more general shape metrics, such as Hausdorff metrics [18]. More appropriate for proteins may be invariants derived from knot theory. Røgen and Fain [32] offer such a new perspective. Given a protein, they compute 30 different curve invariants, thereby mapping the protein to a point in $\Re^{30}$. They argue that this 30-dimensional measure satisfies the triangle inequality, and thus is a good method for grouping protein shapes into similarity classes at multiple levels of granularity. They demonstrate this claim empirically by classifying 20,937 protein domains into multiple levels, achieving 96% agreement with the CATH2.4 classification [27, 26] (both SCOP and CATH are widely accepted protein classification databases, created by a combination of automatic and human judgments).

The primary invariant in [32] is the *writhing number* of a curve; the others are built from this. The writhing number of a curve essentially measures self-linking of a curve. It is connected to the *linking number* of two curves by the following famous Călugăreanu-Fuller-White formula [9, 11, 44]:

$$Lk = Wr + Tw$$

The formula applies to a narrow closed orientable ribbon in three-dimensional space. Lk is the linking number of the two boundary curves of the ribbon, Wr is the writhing number of the central spine, and Tw is the *twisting number* of the two boundary curves. Lk is a purely topological number. **Purely topological means that it is invariant to any smooth deformation that avoids self-intersections.** The other two numbers are not topological invariants; they depend on the embedding of the ribbon. However, they are invariant to a large class of transformations, such as rigid body motions, even conformal (angle-preserving) mappings.

To gain intuition, suppose we orient the ribbon by orienting its spine. For proteins, the backbone plays the role of the spine, and it is naturally oriented by residue order. Now imagine projecting the ribbon onto a 2D plane orthogonal to a randomly chosen direction. The curves defining the ribbon will seem to cross each other at some locations in the plane

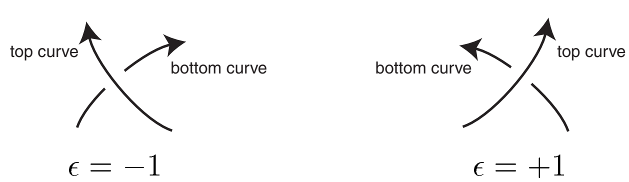


Figure 1: The two types of crossings and their crossing numbers.

of projection. At each crossing we keep track of which curve lies above the other, then assign the crossing a number, called ϵ, with value -1 or $+1$. The value depends on the orientation of the two curve segments at the crossing as well as on the over-under relationship. Specifically, imagine rotating the top curve locally so that its forward tangent at the crossing is parallel to the forward tangent of the bottom curve. Then ϵ is the sign of the smallest angle required. See Figure 1.

Keeping this geometric picture in mind, the linking number Lk counts the sum of signed crossings between the ribbon's two boundary curves, divided by two. This sum is independent of projection direction. The writhing number Wr counts the sum of signed self-crossings of the ribbon's spine, now averaged over *all* projection directions. Finally, the twist Tw is a torsion-dependent term that measures how much one boundary curves intertwines with the other. We will not have any need for it, and will not discuss it further. Instead, our focus will be on matching subsegments of proteins by comparing writhings.

The intuition that linking and writhing numbers count crossings is very important from the perspective of knot theory. From our perspective, we need an additional, related, piece of intuition. Consider two closed curves in space (see also Figure 2 and imagine that each of the edges is tangent to a curve). Place a finger on each curve and consider the unit direction vector pointing from one fingertip to the other. This is a point on the unit sphere. Sum up the signed area covered on the sphere for all possible finger placements, with sign given locally by the crossing number ϵ as a function of finger placements. With some effort one sees that the net area covered is the linking number Lk of the two curves. Amazingly, this number turns out always to be an integer. If the two curves are not linked, the net area covered will be zero, if the two curves are linked once (such as two magician's rings), the sphere will be covered once, and so forth. Intuitively, for proteins, **the extent to which the sphere is covered locally will provide us with a measure of the relative location and orientation of pairs of peptide segments.**

Linking: Formally, suppose c_1 and c_2 are two closed non-intersecting curves in three space, specifically disjoint embeddings of S^1 into $\Re^3$. Let G be the Gauss map applied to the difference between the curves, i.e., the function $G : S^1 \times S^1 \to S^2$ given by $G(s,t) = (c_2(t) - c_1(s))/||c_2(t) - c_1(s)||$. Then the linking number of the two curves can be written in terms of the Gauss integral:

$$Lk(c_1, c_2) \stackrel{\text{def}}{=} \frac{1}{4\pi}\int_{S^1\times S^1} G^*\omega = \frac{1}{4\pi}\int_{S^1}\int_{S^1} \frac{(c_1'(s)\times c_2'(t))\cdot(c_1(s)-c_2(t))}{||c_1(s)-c_2(t)||^3}\,ds\,dt. \qquad (1)$$

Here ω is the differential 2-form measuring area on S^2 and $G^*\omega$ is its pullback by G to $S^1 \times S^1$.

Writhing: It turns out that the writhing number of a curve has the same algebraic form; if $c : S^1 \to \Re^3$ is a closed curve in space, then its writhing number is simply $Wr(c) = Lk(c,c)$. In other words, "writhing is self-linking". Of course, in this case the function G is not well-defined on the diagonal (when $t = s$). *A priori* the integral $Lk(c,c)$ need not exist. Dealing with this issue leads to the twist Tw mentioned earlier [24]. The writhing number is almost never an integer.

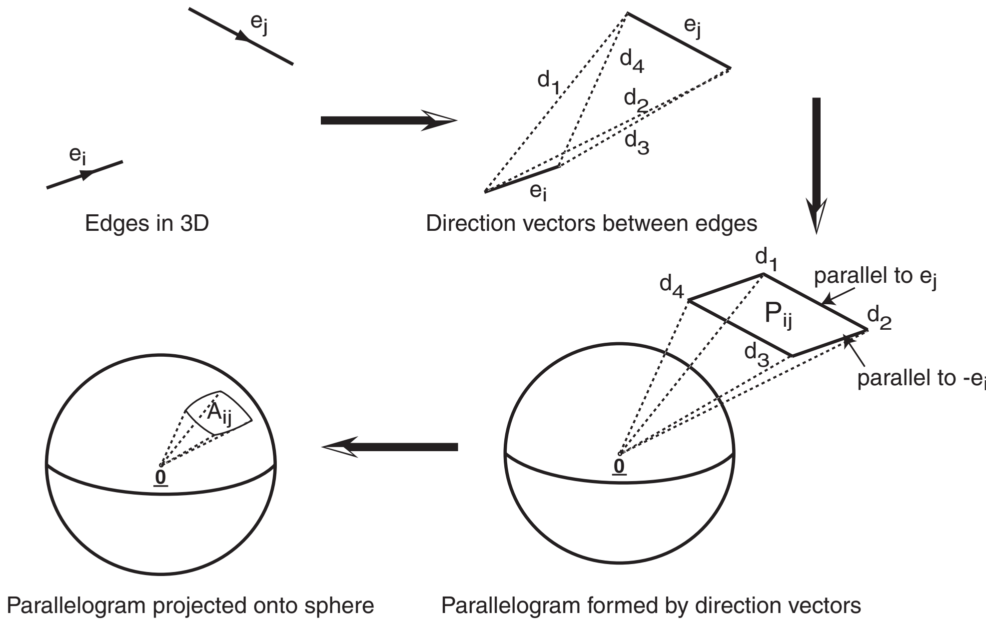


Figure 2: Two edges, e_i and e_j, generate a parallelogram P_{ij} of directions, with vertices d_1, d_2, d_3, d_4, whose ϵ-signed spherically projected area A_{ij} is their edge-edge writhing. (In this figure the edges are configured so that $\epsilon = +1$.)

Protein Fragments: The definitions continue to make sense for open curves, i.e., 3D embeddings of intervals rather than circles. In particular, we will find the component writhing numbers $Lk(c_1, c_2)$ of short protein backbone fragments, c_1 and c_2, to be useful shape indicators.

3.2 Writhing of Polygonal Curves

We will represent protein backbones as open polygonal curves[1], connecting sequential residues via their CA atoms (alpha carbons).[2] For a very nice exposition on writhing numbers of polygonal curves see [1]. That paper developed a clever $O(n^{1.6})$ algorithm and a sweepline algorithm for computing the writhing of a polygonal curve, then applied the second algorithm to various proteins. Considerable work has used knot theory to understand the supercoiling and knotting behaviors observed in DNA, another polygonal curve (see [33] for a sample). Also, see [19, 29] for some very interesting applications of robot motion planning to polygonal knot theory.

Polygonal curves simplify calculation of Equation (1). The integral becomes a finite sum:

$$Lk(c_1, c_2) = \sum_i \sum_j A_{ij}$$

where A_{ij} is the ϵ-signed area on the sphere covered by vectors pointing from edge e_i on the first curve to edge e_j on the second curve.

[1] "open" means that the start and endpoints are distinct; "polygonal" means that the curve is piecewise linear.

[2] In other contexts, e.g., NMR structure determination, amide protons (^{1}H^N) are more natural [8, 3].

Definition 1: We will refer to A_{ij} as the *edge-edge writhing* of the two edges e_i and e_j.

Computing A_{ij} is straightforward. Figure 2 illustrates the process. Algebraically, suppose the start and end points of the oriented edge e_i are p_1 and p_2, and suppose the start and end points of edge e_j are q_1 and q_2. Consider the four extremal cross directions between the two edges:

$$d_1 = q_1 - p_1, \qquad d_2 = q_2 - p_1, \qquad d_3 = q_2 - p_2, \qquad d_4 = q_1 - p_2.$$

For skew edges e_i and e_j, the four directions d_1, d_2, d_3, d_4 define the vertices of a parallelogram P_{ij} in three-dimensional space whose supporting plane does not intersect the origin. Projecting the parallelogram onto the unit sphere creates a spherical parallelogram. Its vertices are the unit direction vectors obtained from d_1, d_2, d_3, d_4, its edges are arcs of great circles connecting these vertices, and its absolute area multiplied by the crossing number of the two edges is the desired signed area A_{ij}. Computing the area of a spherical quadrilateral is also straightforward; one simply sums the interior angles of the quadrilateral and subtracts 2π. Observe that $A_{ij} = A_{ji}$.

3.3 Arrangements of Lines

Closely related to knot theory and potentially to protein structure similarity is the theory of line arrangements. For an introduction to this area see [42, 43]. Two arrangements of skew lines in three-dimensional space are said to be *topologically equivalent* if there is an isotopy that transforms one arrangement into the other. By an isotopy one means motions of the lines in which the lines remain skew. There is exactly 1 topological equivalence class consisting of 2 skew lines, 2 classes of 3-lines, 3 classes of 4-lines, 7 classes of 5-lines, 19 classes of 6-lines, and 74 classes of 7-lines. The classification of general collections of skew lines is an open research question. One approach is to transform line arrangements into elements of braid groups, construct the links induced by the braids, and apply methods from knot theory [28].

The potential application to protein structure comparison arises in three contexts. First, structural alignment programs often represent proteins by their secondary structure vectors [12, 37, 40]. Classifying such vector arrangements might provide simple invariants by which to label protein folds. Second, the peptide plane bond vectors (such as N-CA, N-H, and N-C(O)) fully determine a protein's shape. Again, a classification of the possible arrangements of these vectors might provide simple means for recognizing the shapes of unknown proteins. For instance, the orientations of these vectors relative to a global axis can be discerned using NMR [41, 2, 20, 10]. This may provide an efficient method for distinguishing proteins experimentally. Third, the techniques from line classifications may carry over to more general structures. The key idea is to consider the space of transformations that preserve certain topological properties, such as non-intersection, then to discover invariants that distinguish the connected components of this space of transformations.

4 Polygonal Curve Homotopies and the Structure Problem

As suggested by the SCOP definition, detecting protein similarity entails finding collections of paired substructures which are located roughly in the same relative locations in space.

Let us make this idea more precise. Recall that a *polygonal curve* is a piecewise linear embedding of the unit interval I into 3D space, $c : I \to \Re^3$. In particular, the curve is not self-intersecting. We can represent the curve as a sequence of *representative points* $\{p_1, \ldots, p_n\}$, namely the endpoints of the linear segments. In our case the points are the coordinates of a protein's alpha carbons. Any consecutive subsequence of a polygonal curve's representative points also defines a polygonal curve.

Definition 2: Suppose that $p = \{p_1, \ldots, p_n\}$ and $q = \{q_1, \ldots, q_m\}$ are two polygonal curves. Suppose that E is a Euclidean rigid body motion on $\Re^3$ (a rotation and translation). Let $\delta > 0$ be some positive number. We will say that curve p is (E, δ)*-homotopic* to curve q if the following two conditions are satisfied:

(i) $n = m$.

(ii) There is a polygonal-curve homotopy h mapping $E(p)$ to q such that no representative point moves further than δ from its initial or final location. More precisely, we require a continuous function $h : I \to (\Re^3)^n$, written as $h(t) = (h_1(t), \ldots, h_n(t))$, such that:

   (a) $h_i(0) = E(p_i)$, for all $i = 1, \ldots, n$.

   (b) $h_i(1) = q_i$, for all $i = 1, \ldots, n$.

   (c) The sequence $\{h_1(t), \ldots, h_n(t)\}$ is a polygonal curve for all t, meaning that the points $h_1(t), \ldots, h_n(t)$ define a curve that is not self-intersecting for all times $t \in I$.

   (d) $||E(p_i) - h_i(t)|| \leq \delta$ and $||q_i - h_i(t)|| \leq \delta$ for all $t \in I$ and $i = 1, \ldots, n$.

We will presently use this definition to compare subsegments of curves. **The motivating intuition is to regard two proteins as structurally similar if there is some rigid body transformation that places one on top of the other well enough that δ-perturbations of local coordinates permit atom alignment.** The homotopy requirement is phrased in a way that mimics the definition of isotopic lines we saw in Section 3.3, that is, via classes of motions that preserve structure. Thus, for instance, two helices might match if and only if one can be transformed into the other without backbone self-collisions. Observe that the transformation could be quite large, depending on δ, but at all times preserves the backbone topology. (We note in passing a generalization: it might be interesting to restrict the class of homotopies further by requiring that the polygonal curve $h(t)$ not intersect the *rest* of the protein at any time t.)

For large n and medium-sized δ, condition (ii) can be complicated to check. It basically entails solving a high-degree-of-freedom motion planning problem. Fortunately, for many short protein fragments and small δ, the condition is similar to enforcing low RMSDs of the final alignments. *The definition therefore addresses a wide tunable range of possible structural similarity questions.*

Definition 3: Define functions d_E and d on pairs of polygonal curves as follows:

$$d_E(p,q) = \inf\{\delta \mid p \text{ is } (E,\delta)\text{-homotopic to } q\} \qquad d(p,q) = \inf_E d_E(p,q)$$

Thus $d(p,q) = \infty$ iff p and q are not homotopic for any (E,δ), e.g., if the number of points differs.

Theorem 1: d is a metric and d is effectively computable. **Proof:** See Appendix B.1.

Monotonic Curve Homotopies Given a point p_i and a line ℓ in 3D space we can project the point orthogonally onto the line. We can do the same for all representative points of some polygonal curve. The curve is said to be *monotonic with respect to line* ℓ if the order of the projected points is the same as the order of the points in the curve. This order *orients* the line. Short protein segments, such as β-strands and α-helices, are often monotonic with respect to their best-approximating lines.

Lemma 1: Suppose $p = p_1, \ldots, p_n$ is a polygonal curve monotonic with respect to line ℓ. Let $\pi = \pi_1, \ldots, \pi_n$ be the polygonal curve obtained by projecting p onto ℓ. Then $d(p,\pi) \leq \max_i ||p_i - \pi_i||$.

Proof: Imagine drawing a line between p_i and π_i for each i. Define a homotopy that moves each p_i to π_i along these lines. The homotopy preserves the polygonal curve since the curve is monotonic.

Lemma 2: Suppose p and q are two polygonal curves with equal numbers of points, each monotonic with respect to some line. Let $\pi = \pi_1, \ldots, \pi_n$ and $\sigma = \sigma_1, \ldots, \sigma_n$ be the projections of the two curves onto their respective lines. Then $d(p,q) \leq d(p,\pi) + d(q,\sigma) + \inf_E \max_i ||\sigma_i - E(\pi_i)||$, where E is taken from the set of rigid body motions that align the two oriented lines.

See Appendix B.2 for a proof. The bound is often generous. The lemma tells us that **two monotonic curves whose line-projections are similar in 1D are also readily homotopic in 3D.**

For polygonal curves with equal numbers of points, d measures the spatial difficulty of transforming one curve into the other. It provides no such information for curves with different numbers of points. Instead, we now define structural similarity as the detection of local homotopies. We need one piece of additional notation. Suppose $p = p_1, \ldots, p_n$ is a polygonal curve; let us define p_i^k as the polygonal subcurve $p_{i-k}, \ldots, p_i, \ldots, p_{i+k}$ whenever $k+1 \leq i \leq n-k$. In other words, p_i^k is the curve segment centered at p_i, extending backwards and forwards by k points.

Definition 4: Suppose that $p = \{p_1, \ldots, p_n\}$ and $q = \{q_1, \ldots, q_m\}$ are two polygonal curves. Let $\delta > 0$ be a positive number, k a nonnegative integer, and $\mathcal{I}$ some set of index pairs $\{(i,j)\}$. We say that p is <u>*δ-structurally similar with k-strength alignment $\mathcal{I}$*</u> if there exists *some* rigid body transformation E such that $d_E(p_i^k, q_j^k) \leq \delta$ for *all* pairs $(i,j) \in \mathcal{I}$.

In English, this definition requires one curve to move rigidly over the other curve such that two paired collections of subcurves are nearly identical to each other, as measured by subsequent homotopy deformations. For $k = 0$, this definition is similar to aligning pointsets. For large k, the definition amounts to detecting overall curve similarity. In between, the definition captures the notion of structural alignment with rearrangements. In particular, the order of indices in the index set $\mathcal{I}$ need not be sequential. This leads to the following:

Structure Problem: For given curves p and q, for δ positive and k a nonnegative integer, compute all index sets $\mathcal{I}$ and their associated rigid body transformations E satisfying Definition 4.

Theorem 2: The Structure Problem is effectively computable.

Proof: Follows from Theorem 1.

Although computable, the algorithm derived from our proof of Theorem 1 is horrendously exponential [6, 7, 21, 34]. One possibility is to use a motion planner specialized for knots, such as the untangling planner of [19]. Alternatively, for our purposes, Lemmas 1 and 2 suggest a simplification: In the next two sections we will examine an approach based on edge-edge writhings that attacks the Structure Problem by aligning line projections of peptide segments.

5 Understanding Protein Similarity by Geometric Convolution

In this section we examine more closely the construction of Figure 2. Our observations will motivate us to define a self-convolution datastructure for detecting structural similarity in proteins.

5.1 Writhing and Convolution

Definition 5: Suppose X and Y are two sets of points in R^3. Then the *geometric convolution of Y with X* is the set of points $Y \ominus X = \{y - x \mid x \in X \text{ and } y \in Y\}$. (Sometimes this is defined by saying that the geometric convolution of Y with X is the *Minkowski sum* of Y and $-X$. There are again strong connections to robot motion planning [22, 23]).

Lemma 3: Assume e_i, e_j, and P_{ij} are as defined at the end of Section 3.2. Then $P_{ij} = e_j \ominus e_i$.

Proof: Definitional: P_{ij} is the set of all vectors pointing from a point on e_i to a point on e_j.

Corollary 1: The edge-edge writhing A_{ij} of two edges e_i and e_j is the area of the convolution $e_j \ominus e_i$ projected onto the sphere S^2 times the crossing number ϵ of the edges' supporting lines.

Corollary 2: Suppose edges e_i and e_j are given. The following four possibilities exist:

(a) The edges are skew. In this case P_{ij} is a 2D polygon whose plane of support does not include the origin. The writhing A_{ij} is therefore well-defined and nonzero.

(b) The edges are coplanar but not parallel. In this case P_{ij} is again a 2D polygon, but now its plane of support does include the origin. The polygon P_{ij} may or may not touch the origin. $P_{ij}\backslash\{0\}$ projects to a great-circle arc on the sphere, and the writhing A_{ij} is therefore zero.

(c) The edges are parallel but not colinear. In this case the polygon P_{ij} degenerates to colinear line segments lying on a line that does not pass through the origin. The writhing A_{ij} is zero.

(d) The edges are colinear. In this case the polygon P_{ij} degenerates to colinear line segments lying on a line that passes through the origin. The polygon P_{ij} may or may not touch the origin. It projects to one or two points on the sphere and the writhing A_{ij} is again zero.

Corollary 3: The edges e_i and e_j intersect if and only if polygon P_{ij} touches the origin.

Corollary 3 tells us that we can count edge incidence by counting polygons touching the origin. Suitably generalized, that hints at a method for determining structural similarity.

5.2 Self-Convolution

Earlier we observed that many successful structural alignment programs compare arrangements of pairs of lines. We now extend that idea to writhing polygons. In reading Lemma 4 imagine that we are comparing *a pair* of peptide segments in one protein with *another pair* in another protein.

Lemma 4: Consider four edges: e_1, e_2, f_1, f_2. There is a rigid body transformation E mapping the edges (e_1, e_2) to the edges (f_1, f_2) if and only if there is a rotation R about the origin such that $R(e_2 \ominus e_1) = f_2 \ominus f_1$ while preserving vertex correspondence.
Proof: See Appendix B.3.

Corollary 4: If R is a rotation such that the maximum distance between corresponding vertices of the two polygons $R(e_2 \ominus e_1)$ and $f_2 \ominus f_1$ is δ, then there is a rigid body transformation E such that e_1 and e_2 are (E, δ)-homotopic to f_1 and f_2, respectively.
Proof: See Appendix B.4.

When Corollary 4 applies we say that the polygons are *δ-homotopic.*

Definition 6: If p is a polygonal curve, we define the *geometric self-convolution of* p, written $\otimes(p)$, to be the generating polygons of $p \ominus p$:

$$\otimes(p) = \{P_{ij} \mid P_{ij} = e_j \ominus e_i, \text{with } e_i \text{ and } e_j \text{ edges in the curve } p\}.$$

Given two curves p and q, we will seek structural similarity by comparing the curves' self-convolutions. Lemma 4 suggests that we mod out by rotations and translations, and focus instead on comparing the *configurations* of the polygons $\{P_{ij}\}$. Corollary 4 relates configuration similarity to homotopy distance. A writhing polygon has six configuration parameters: the two edge lengths, the angle between the edges, the distance from the origin, and two orientation parameters describing the polygon normal. We have found it useful to cluster using two characteristics: edge-edge writhing and distance from the origin. Writhing provides a mixed measure of all six degrees of configuration freedom; retaining distance mitigates the roughly inverse-square effect of distance on writhing. Similarity is easily checked, using for instance a best-aligning rotation in Corollary 4.

6 Structure Detection

We now combine the homotopy and self-convolution ideas to implement an algorithm for detecting common protein structure. There is one additional wrinkle, needed to deal with the segment length parameter k in Definition 4. When constructing the self-convolution $\otimes(p)$, we replace the polygon P_{ij} with a polygon formed from the best-line projections of the peptide segments p_i^k and p_j^k, as motivated by Lemmas 1 and 2. Denote this polygon by P_{ij}^k. For the writhing number we use the true writhing of the two peptide segments, that is, $w_{ij}^k(p) = Lk(p_i^k, p_j^k)$. Let $d_{ij}(p) = ||p_i - p_j||$. Denote the resulting combinatorial structure consisting of all $\{(P_{ij}^k(p), w_{ij}^k(p), d_{ij}(p))\}$ by the symbol $\otimes^k(p)$.

Algorithm: Given polygonal curves p and q, distance $\delta > 0$, and integer $k \geq 1$, detect structural similarity as follows:

1. Compute $\otimes^k(p)$ and $\otimes^k(q)$.
2. Hash the polygons $\{P_{ij}^k(p)\}$ and $\{P_{ij}^k(q)\}$ based on w_{ij}^k and d_{ij}, ignoring near zeros.
3. For each nonempty (or sufficiently full) hash bucket B_{wd} of polygons do the following:
   - For each pair of δ-homotopic polygons $P \in \otimes^k(p)$ and $Q \in \otimes^k(q)$ in B_{wd}, compute the rigid map E implied by Corollary 4. Hash the rigid map with its generating polygons.

The generating polygons and rigid maps associated with a hash bucket in Step 3 • offer an approximate solution $(\mathcal{I}, E)$ to the Structure Problem. The entire hash table describes all nontrivial alignments at the given hash table resolutions. We ignore polygons with near zero writhing or distance to avoid degeneracies. The solutions are approximate in the sense that the polygons P_{ij}^k are based on best-approximating edges and the maps E are clustered, potentially dilating δ.

6.1 Analysis

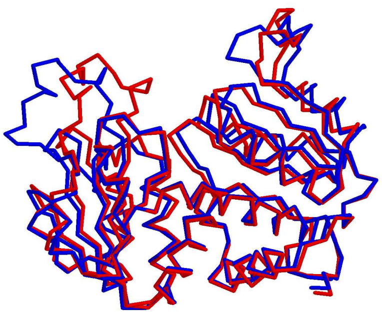
The Algorithm runs in time $O(k^2n^2 + k^2m^2 + s^2/\epsilon_P^2 + 1/\epsilon_E^6)$ and space $O(n^2 + m^2 + 1/\epsilon_B^2 + 1/\epsilon_E^6)$ where n and m are the number of points in p and q, k is the half-length of a peptide segment, s is the maximum number of pairwise similar polygons appearing in a polygon hash bucket, and ϵ_P and ϵ_E are the resolutions of the polygon and rigid body hash tables, respectively.

In practice, k and ϵ_E are constants. We took $k = 5$ and $\epsilon_E = 0.1$. $1/\epsilon_E^6$ is the size of the hash table for Euclidean transformations. We represented each transformation as a 4D quaternion and a 3D translation, projected the quaternion into 3D, then hashed the resulting 6 numbers. Although s could be $\Theta(n^2)$, it depends on ϵ_P. Choosing this carefully, the ratio s/ϵ_P becomes $O(n)$. Thus, the algorithm has $O(n^2)$ behavior, with n the maximum protein length. The hash tables could be replaced by k-D trees, Voronoi diagrams, or other clustering methods [30], but we did not do so.

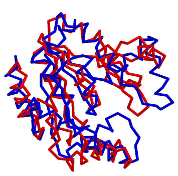
6.2 Results

We implemented the algorithm in Lisp on a 1GHz PC running Windows. Running times for proteins with 300 residues were typically 1 minute or less, half of that garbage collection. Here are three interesting pairs of proteins (see Appendix A for larger figures):

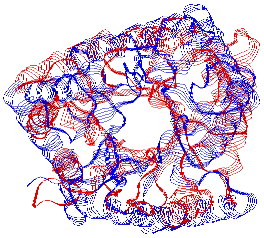
5at1_A vs. 8atc_A: These are two different conformations of the catalytic chain A in Aspartate Carbamoyltransferase (ATC), a famous allosteric protein involved in the synthesis of pyrimidine nucleotides [39]. Chain A has two domains, that rotate with respect to each other as part of the process. Two loops change conformation drastically. Our algorithm detects both the similarities and the differences. The rigid map with the greatest number of aligned segments lies within 2° in rotation and 0.6Å in translation of the correct alignment. Our subsequent atom-alignment code assigns 289 of the 310 residues with RMSD 1.0Å; the remaining residues constitute the two non-alignable loops. See figure to left (large view in Figure 3).



3adk vs. 1gky: These two proteins have mere 19% sequence identity, are different lengths (194 vs. 186 residues), and include both matching and nonmatching secondary structures. Our code finds the alignment shown in the figure to the left (large view in Figure 4). The rigid map lies within 5° and 0.5Å of the CE-alignment. Our subsequent atom-alignment assigns 165 atoms with RMSD 2.9Å, closely matching CE.



1xis vs. 1nar: These are two TIM barrels, with 7% sequence identity. We considered the 321 residues of 1xis without its tail versus the 289 residues of 1nar. There are several possible alignments, related by rotation around the central barrel (Figure 5 shows two). This pair of proteins is interesting because even in optimal alignment there are significant angular differences between aligned helices. Such comparisons motivated our homotopy definitions. Our code finds an alignment with RMSD 3.3Å, differing by 14° and 1.5Å from the optimal Dali-alignment. See figure to left (large view in Figure 6).



7 Summary

This paper introduced the notion of homotopy deformations into structural alignment. The paper explored the relationship between writhing and self-convolution. Self-convolution is a compact way of registering edge-edge interactions, and extends naturally to interactions of curve segments. Writhing and separation are useful shape descriptors for clustering pairs of curve segments. The paper presented an algorithm for matching substructures by clustering similar segment pairs, then clustering among the induced rigid maps.

Acknowledgment

I am very grateful to Dr. Gordon S. Rule in the Department of Biological Sciences for countless wonderful conversations and discussions regarding protein structure and biochemistry over the past five years.

References

[1] P.K. Agarwal, H. Edelsbrunner, and Y. Wang. Computing the writhing number of a polygonal knot. *Proceedings Thirteenth Symposium on Discrete Algorithms (SODA)*, 13:791–799, 2002.

[2] H.M. Al-Hashimi, H. Valafar, M. Terrell, E.R. Zartler, M.K. Eidsness, and J.H. Prestegard. Variation of molecular alignment as a means of resolving orientational ambiguities in protein structures from dipolar couplings. *Journal of Magnetic Resonance*, 143:402–406, 2000.

[3] C. Bailey-Kellogg, A. Widge, J.J. Kelley, M.J. Berardi, J.H. Bushweller, and B.R. Donald. The NOESY Jigsaw: Automated protein secondary structure and main-chain assignment from sparse, unassigned NMR data. *Journal of Computational Biology*, 7(3–4):537–558, 2000.

[4] Dror Bar-Natan. On the Vassiliev knot invariants. *Topology*, 34:423–472, 1995.

[5] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, and P.E. Bourne. The Protein Data Bank. *Nucleic Acids Research*, 28:235–242, 2000.

[6] John Canny. *The Complexity of Robot Motion Planning.* MIT, Cambridge, Massachusetts, 1988.

[7] John Canny. Some algebraic and geometric computations in PSPACE. *Proceedings of the Twentieth Annual ACM Symposium on Theory of Computing (STOC)*, 20:460–467, 1988.

[8] G.M. Crippen and T.F. Havel. *Distance Geometry and Molecular Conformation.* Research Studies Press, Taunton, England, 1988.

[9] G.C. Călugăreanu. Sur les classes d'isotopie des noeuds tridimensionnels et leurs invariants. *Czechoslovak Mathematics Journal*, 11:588–625, 1961.

[10] M.A. Erdmann and G.S. Rule. Rapid protein structure detection and assignment using residual dipolar couplings. Technical Report CMU-CS-02-195, Carnegie Mellon University, December 2002.

[11] F.B. Fuller. The writhing number of a space curve. *Proceedings of the National Academy of Sciences USA*, 68:815–819, 1971.

[12] J.-F. Gibrat, T. Madej, and S.H. Bryant. *Vector Alignment Search Tool*. National Center for Biotechnology Information (NCBI), `http://www.ncbi.nlm.nih.gov/Structure/VAST/vast.html`.

[13] J.-F. Gibrat, T. Madej, and S.H. Bryant. Surprising similarities in structure comparison. *Current Opinion in Structural Biology*, 6:377–385, 1996.

[14] J. Hass, J.C. Lagarias, and N. Pippenger. The computational complexity of knot and link problems. Technical Report math.GT/9807016, arXiv.org e-Print archive, July 1998.

[15] L. Holm, A. de Daruvar, C. Sander, and C. Dodge. *DALI*. European Bioinformatics Institute (EMBL-EBI), `http://www2.ebi.ac.uk/dali`.

[16] L. Holm and C. Sander. Protein structure comparison by alignment of distance matrices. *Journal of Molecular Biology*, 233:123–138, 1993.

[17] L. Holm and C. Sander. Mapping the protein universe. *Science*, 273:595–602, 1996.

[18] D.P. Huttenlocher and K. Kedem. Distance metrics for comparing shapes in the plane. In B.R. Donald, D. Kapur, and J.L Mundy, editors, *Symbolic and Numerical Computation for Artificial Intelligence*, pages 201–219. Academic Press Limited, London, England, 1992.

[19] A.M. Ladd and L.E. Kavraki. Motion planning for knot untangling. *Proceedings Fifth International Workshop on the Algorithmic Foundations of Robotics*, Nice, France, December 15–17, 2002.

[20] C.J. Langmead, A.K. Yan, L. Wang, R. Lilien, and B. R. Donald. A polynomial time nuclear vector replacement algorithm for automated NMR resonance assignments. *Proceedings Seventh International Conference on Computational Molecular Biology (RECOMB)*, 7, 2003.

[21] Jean-Claude Latombe. *Robot Motion Planning*. Kluwer Academic Publishers, Boston, Massachusetts, 1991.

[22] T. Lozano-Pérez. Automatic planning of manipulator transfer movements. *IEEE Transactions on Systems, Man, and Cybernetics*, 11(10):681–698, 1981.

[23] T. Lozano-Pérez and M. Wesley. An algorithm for planning collision-free paths among polyhedral obstacles. *Communications of the ACM*, 22(10):560–570, 1979.

[24] Daniel Moskovich. Framing and the self-linking integral. Technical Report math.QA/0211223, arXiv.org e-Print archive, November 2002.

[25] A.G. Murzin, L. Lo Conte, A. Andreeva, D. Howorth, B.G. Ailey, S.E. Brenner, T.J.P. Hubbard, and C. Chothia. *Introduction to Structural Classification of Proteins.* SCOP, `http://scop.mrc-lmb.cam.ac.uk/scop/intro.html`.

[26] C.A. Orengo, A.D. Michie, S. Jones, D.T. Jones, M.B. Swindells, and J.M. Thornton. CATH — A hierarchic classification of protein domain structures. *Structure*, 5(8):1093–1108, 1997.

[27] F.M.G. Pearl, I. Sillitoe, M. Dibley, J. Thornton, and C.A. Orengo. *Protein Structure Classification.* CATH, `http://www.biochem.ucl.ac.uk/bsm/cath/`.

[28] Rudi Penne. The Alexander Polynomial of a configuration of skew lines in 3-space. *Pacific Journal of Mathematics*, 186(2):315–348, 1998.

[29] J. Phillips, A. Ladd, and L.E. Kavraki. Simulated knot tying. *Proceedings IEEE International Conference on Robotics and Automation*, pages 841–846, 2002.

[30] F.P. Preparata and M.I. Shamos. *Computational Geometry — An Introduction.* Springer Verlag, New York, 1985.

[31] Research Collaboration for Structural Bioinformatics (RCSB), `http://www.rcsb.org`. *Protein Data Bank (PDB).*

[32] P. Røgen and B. Fain. Automatic classification of protein structure by using Gauss integrals. *Proceedings of the National Academy of Sciences, USA*, 100(1):119–124, 2003.

[33] V. Rossetto and A.C. Maggs. Writhing geometry of open DNA. *Journal of Chemical Physics*, 118:9864–9874, 2003.

[34] J.T. Schwartz and M. Sharir. On the Piano Movers' Problem: II. General techniques for computing topological properties of real algebraic manifolds. *Advances in Applied Mathematics*, 4:298–351, 1983.

[35] I.N. Shindyalov and P.E. Bourne. *Databases and Tools for 3-D Protein Structure Comparison and Alignment.* National Partnership for Advanced Computational Infrastructure (NPACI) and National Biomedical Computation Resource (NBCR), `http://cl.sdsc.edu/ce.html`.

[36] I.N. Shindyalov and P.E. Bourne. Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Engineering*, 11(9):739–747, 1998.

[37] A.P. Singh and D.L. Brutlag. *3dSearch — Secondary Structure Superposition*. Stanford Bioinformatics Group, `http://gene.stanford.edu/3dSearch`.

[38] A.P. Singh and D.L. Brutlag. Hierarchical protein structure superposition using both secondary structure and atomic representations. *Proceedings International Conference on Intelligent Systems for Molecular Biology*, 5:284–293, 1997.

[39] Lubert Stryer. *Biochemistry*. W.H. Freeman, New York, fourth edition, 1995.

[40] W.R. Taylor. Protein structure comparison using bipartite graph matching and its application to protein structure classification. *Molecular & Cellular Proteomics*, 1(4):334–339, 2002.

[41] N. Tjandra and A. Bax. Direct measurement of distances and angles in biomolecules by NMR in a dilute liquid crystalline medium. *Science*, 278:1111–4, 1997.

[42] J. Viro and O. Viro. Configuration of skew lines. Technical Report (this is an easy introduction based on the journal article [43]), Uppsala University, Sweden, 2000.

[43] O.Ya. Viro and Yu.V. Drobotukhina. Configuration of skew lines. *Leningrad Mathematics Journal*, 1(4):1027–1050, 1990.

[44] J. White. Self-linking and the Gauss integral in higher dimensions. *American Journal of Mathematics*, 91:693–728, 1969.

Appendix A: Large Alignment Figures

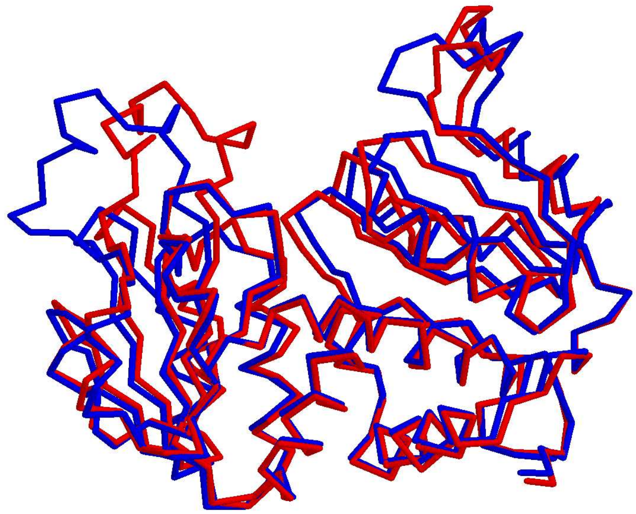


Figure 3: Alignment of 5at1_A (blue) and 8ATC_A (red) found by our writhing-convolution-based Algorithm. The backbones match nearly perfectly, except where they shouldn't, namely two loops that undergo significant conformational change. These loops appear near the top left and top right in the figure.

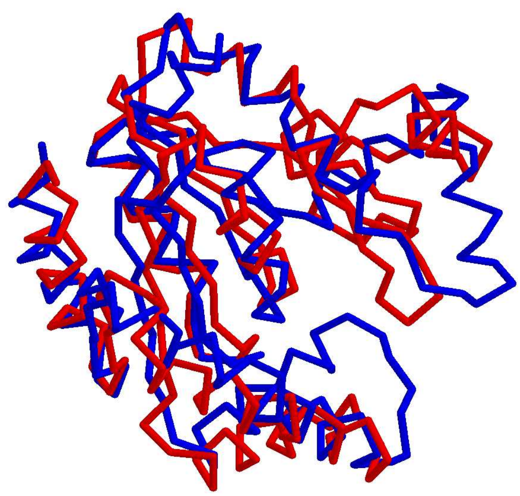


Figure 4: Alignment of 3ADK (blue) and 1GKY (red). The proteins have mere 19% sequence identity and include both matching and nonmatching secondary structures. Roughly 80% of the two proteins should align. One can see this in the figure, with the left parts matching well and some of the right clearly not.

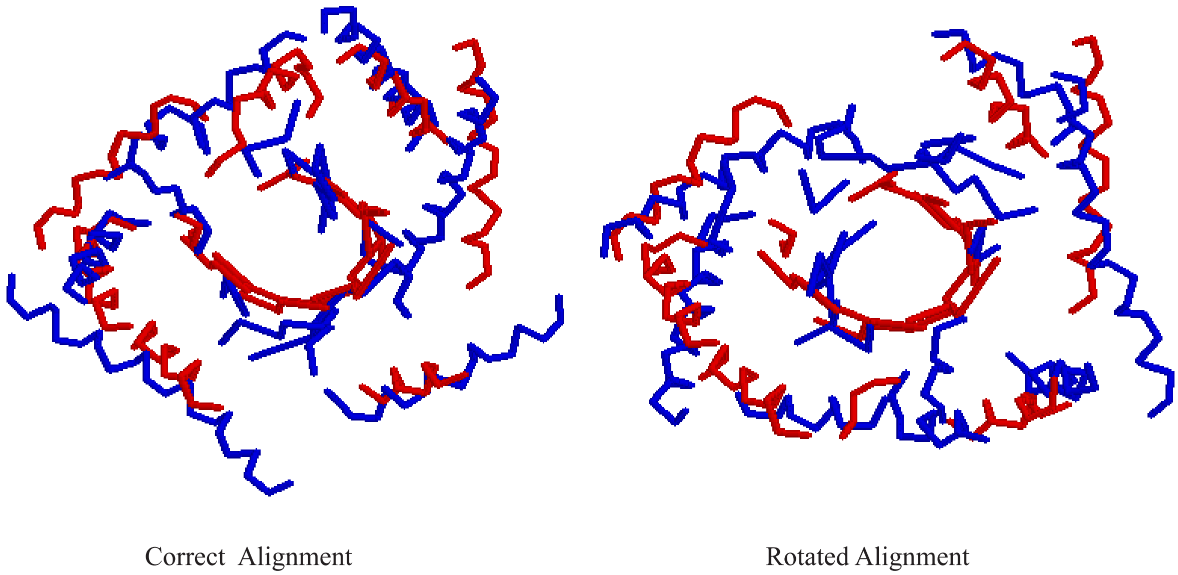


Figure 5: Two possible structural alignments of 1xis (blue) with 1nar (red). The left one is the best-scoring choice in Dali. (Only some of the strands and helices are depicted.) Note the large angular differences between some aligned helices. The two proteins have 7% sequence identity.

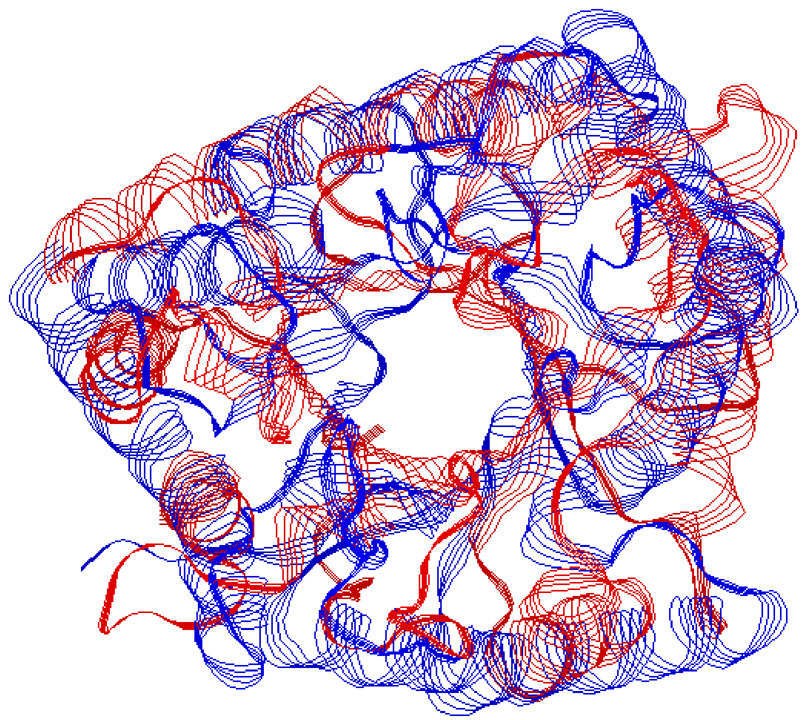


Figure 6: Maximal alignment of 1xis (blue) and 1nar (red) obtained by the Algorithm reported in this paper, closely matching the "correct alignment" in Figure 5.

Appendix B: Proofs

Appendix B.1 [Theorem 1, p. 8]

Theorem 1: d is a metric and d is effectively computable.

Proof: Throughout, let $p = p_1, \ldots, p_n$ and $q = q_1, \ldots q_m$ be two polygonal curves.

Part 1 (metricity): We will show that d is a metric on the quotient space of polygonal curves moded out by $SE(3)$, the special Euclidean group of 3D rigid body motions. Observe that $d(p, q) \geq 0$.

(i) Reflexivity: Clearly $d(p, p) = 0$ for all p. Now suppose $d(p, q) = 0$. Then $n = m$ and, since d is defined as an inf of an inf, for every $\delta > 0$ there is some Euclidean motion E_δ such that p is (E_δ, δ)-homotopic to q, implying that $||E_\delta(p_i) - q_i|| \leq \delta$, for all $i = 1, \ldots, n$. Intuitively, it is now clear that p and q must have the same shape, but let's walk through a detailed argument: The set $\{E_\delta\}$ sits inside a compact subspace of $SE(3)$, so has a limit point $E \in SE(3)$. This means that for any positive ϵ, we can pick a δ_ϵ such that $0 < \delta_\epsilon < \epsilon/2$ and $||E(p_i) - E_\delta(p_i)|| < \epsilon/2$ for all $i = 1, \ldots, n$ and all δ with $0 < \delta < \delta_\epsilon$. Therefore $||E(p_i) - q_i|| \leq ||E(p_i) - E_\delta(p_i)|| + ||E_\delta(p_i) - q_i|| < \epsilon/2 + \delta < \epsilon$. It follows that $E(p_i) = q_i$, for all $i = 1, \ldots, n$. In short, if $d(p, q) = 0$ then p and q are the same curve, differing by at most a Euclidean rigid body transformation.

(ii) Symmetry: We need to show that $d(p, q) = d(q, p)$. We will do so by showing that if p is (E, δ)-homotopic to q then q is (E^{-1}, δ)-homotopic to p. Suppose h is a homotopy satisfying Definition 2 (see p. 7) establishing that p is (E, δ)-homotopic to q. Define $g : I \to (\Re^3)^n$ by the rules $g(t) = (g_1(t), \ldots, g_n(t))$ with $g_i(t) = E^{-1}(h_i(1 - t))$ for all $t \in I$ and $i = 1, \ldots, n$. We claim that g is a homotopy establishing that q is (E^{-1}, δ)-homotopic to p. Let us verify the properties of Definition 2 explicitly. Observe that g is continuous since E^{-1} and h are. Then:

(a) $g_i(0) = E^{-1}(h_i(1)) = E^{-1}(q_i)$, for all $i = 1, \ldots, n$.

(b) $g_i(1) = E^{-1}(h_i(0)) = E^{-1}(E(p_i)) = p_i$ for all $i = 1, \ldots, n$.

(c) The sequence $\{g_1(t), \ldots, g_n(t)\}$ is the sequence $\{E^{-1}(h_1(1 - t)), \ldots, E^{-1}(h_n(1 - t))\}$ which is just the sequence $\{h_1(1 - t), \ldots, h_n(1 - t)\}$ rigidly moved in space by E^{-1}. Hence $\{g_1(t), \ldots, g_n(t)\}$ is a polygonal curve for all $t \in I$.

(d) $||E^{-1}(q_i) - g_i(t)|| = ||E^{-1}(q_i) - E^{-1}(h_i(1 - t))|| = ||q_i - h_i(1 - t)|| \leq \delta$ and $||p_i - g_i(t)|| = ||p_i - E^{-1}(h_i(1 - t))|| = ||E(p_i) - h_i(1 - t)|| \leq \delta$, for all $t \in I$ and $i = 1, \ldots, n$.

(iii) Triangle Inequality: Suppose $r = r_1, \ldots, r_k$ is another polygonal curve. We need to show that $d(p, q) \leq d(p, r) + d(r, q)$. First we observe that if k differs from either n or m, then the inequality is trivially true, so we can assume without loss of generality that $m = n = k$. Now let ϵ be an arbitrary positive number. Then there are Euclidean motions $E, F \in SE(3)$

and homotopies e, f establishing that p is $(E, d(p,r)+\epsilon/2)$-homotopic to r and that r is $(F, d(r,q)+\epsilon/2)$-homotopic to q. Now let $H = F \circ E$ and define $h : I \to (\Re^3)^n$ by the rules $h(t) = F(e(2t))$ for $t \in [0, \frac{1}{2}]$ and $h(t) = f(2t-1)$ for $t \in [\frac{1}{2}, 1]$. h is continuous since e, f, and F are continuous and since $h(\frac{1}{2}) \equiv F(e(1)) = F(r) = f(0) \equiv h(\frac{1}{2})$.

Let's look at the conditions (ii) of Definition 2 with data H and h:

(a) $h_i(0) = F(e_i(0)) = F(E(p_i)) = H(p_i)$.

(b) $h_i(1) = f_i(1) = q_i$.

(c) $h(t)$ is either $F(e(2t))$ or $f(2t-1)$, both of which define polygonal curves.

(d) First, suppose that $t \in [0, \frac{1}{2}]$. In that case:

- $||H(p_i) - h_i(t)|| = ||H(p_i) - F(e_i(2t))|| = ||E(p_i) - e_i(2t)|| \leq d(p,r) + \epsilon/2$.
- $||q_i - h_i(t)|| \leq ||q_i - F(r_i)|| + ||r_i - e_i(2t)|| \leq d(r,q) + d(p,r) + \epsilon$.

Similarly, if $t \in [\frac{1}{2}, 1]$:

- $||H(p_i) - h_i(t)|| \leq ||E(p_i) - r_i|| + ||F(r_i) - f_i(2t-1)|| \leq d(p,r) + d(r,q) + \epsilon$.
- $||q_i - h_i(t)|| = ||q_i - f_i(2t-1)|| \leq d(r,q) + \epsilon/2$.

Thus h establishes that p is (H, δ)-homotopic to q with $\delta = d(p,r) + d(r,q) + \epsilon$. Since ϵ is arbitrary this shows that $d(p,q) \leq d(p,r) + d(r,q)$.

Part 2 (computability): The relevant decision question is:

If p and q are polygonal curves and s is a rational number, is $d(p,q) < s$?

This question can be formulated as a sentence in the first order theory of the reals, hence is decidable. Here is a sketch of the proof:

We can assume again that the two curves each have n points. Suppose that $E \in SE(3)$ and $\delta \geq 0$ are given. We then have a robot motion planning for n point robots moving in three dimensions. The start configuration for robot #i is $E(p_i)$, the goal configuration is q_i. Robot #i is constrained to move within the intersection of two balls of radius δ, one centered at its start, the other at its goal. Moreover, edges drawn between two different pairs of successively indexed points may not move so as to intersect. More precisely, let $h_i(t)$ be the location of robot #i at time t, and let $e_i(t)$ be the line segment $[h_i(t), h_{i+1}(t)]$, for $i = 1, \ldots, n-1$. Then for all times t, we require that $e_i(t) \bigcap e_j(t) = \emptyset$ if $1 \leq i \leq j-2 \leq n-3$ and $e_i(t) \bigcap e_{i+1}(t) = \{h_{i+1}(t)\}$ if $1 \leq i \leq n-2$. This problem is effectively decidable as a question within the first order theory of the reals, in fact it lies within PSPACE. See, for instance, [6, 7, 21, 14]. We thus have a procedure for deciding whether p is (E, δ)-homotopic to q. Let $P(p, q, E, \delta)$ be the corresponding predicate, then formulate the following sentence:

$$\exists\delta,\ \exists E\ :\ (0 \leq \delta)\ \wedge\ (\delta < s)\ \wedge\ P(p, q, E, \delta)$$

For suitable parameterization of $SE(3)$, this sentence is a rewording of the original decision problem as a problem within the first order theory of the reals and thus is decidable. If we want, we can further quantify over s to isolate the value $d(p,q)$ as accurately as we desire.

Appendix B.2 [Lemma 2, p. 8]

Lemma 2: Suppose p and q are two polygonal curves with equal numbers of points, each monotonic with respect to some line. Let $\pi = \pi_1, \ldots, \pi_n$ and $\sigma = \sigma_1, \ldots, \sigma_n$ be the projections of the two curves onto their respective lines. Then $d(p, q) \leq d(p, \pi) + d(q, \sigma) + \inf_E \max_i ||\sigma_i - E(\pi_i)||$, where E is taken from the set of rigid body motions that align the two oriented lines.

Proof Sketch: Suppose $E \in SE(3)$ aligns the two oriented lines. We decompose the homotopy from $E(p)$ to q into three motions: The first moves $E(p)$ to $E(\pi)$, the second moves $E(\pi)$ to σ, and the third moves σ to q. The first and third motions are simple linear interpolations between start and end configurations. In fact, as in Lemma 1, all points move orthogonally to the lines of projection, so the first and third motions are curve-preserving. The second motion can also be taken as a curve-preserving linear interpolation, as we will see shortly. Thus the distance between any point and its start or end configuration is bounded by the sum of the lengths of the three linear motions. This gives the desired bound.

We need to verify that a linear interpolation moving $E(\pi)$ to σ is curve-preserving. This is easy: simply project into 2D, with time along the x-axis and the location of the points along the y-axis. The spacetime curves of the points are straight lines; the lines do not cross since the original curves were monotonic and since E is orientation-preserving.

Appendix B.3 [Lemma 4, p. 10]

Lemma 4: Consider four edges: e_1, e_2, f_1, f_2. There is a rigid body transformation E mapping the edges (e_1, e_2) to the edges (f_1, f_2) if and only if there is a rotation R about the origin such that $R(e_2 \ominus e_1) = f_2 \ominus f_1$ while preserving vertex correspondence.

Proof: Write $E(x) = Rx + t$ where R is a rotation matrix and t is a translation vector. We will prove the lemma for this choice of R. First, write the four edges in terms of their endpoints:

$$e_1 = [p_{11}, p_{12}] \qquad e_2 = [p_{21}, p_{22}] \qquad f_1 = [q_{11}, q_{12}] \qquad f_2 = [q_{21}, q_{22}]$$

And now write each convolution in terms of its four corners:

$$e_2 \ominus e_1 = \{p_{21} - p_{11},\ p_{22} - p_{11},\ p_{22} - p_{12},\ p_{21} - p_{12}\}$$

$$f_2 \ominus f_1 = \{q_{21} - q_{11},\ q_{22} - q_{11},\ q_{22} - q_{12},\ q_{21} - q_{12}\}$$

Consider the following two statements:

(1) E maps (e_1, e_2) to (f_1, f_2). (This means that $E(p_{ij}) = q_{ij}$, for $i, j = 1, 2$.)

(2) $R(e_2 \ominus e_1) = f_2 \ominus f_1$ while preserving vertex correspondence. (This means that $R(p_{2i} - p_{1j}) = q_{2i} - q_{1j}$, for $i, j = 1, 2$.)

We need to show that (1) is true if and only if (2) is true:

Suppose (1) is true. Then $q_{2i} - q_{1j} = E(p_{2i}) - E(p_{1j}) = R(p_{2i} - p_{1j})$, so (2) is true as well.

Suppose (2) is true. We need to construct a rigid motion E establishing (1). Let R be the rotational part of it and define the translation vector as $t = q_{11} - R\,p_{11}$. Now observe:

$$\begin{aligned}
E(p_{11}) &= R\,p_{11} + q_{11} - R\,p_{11} &&= q_{11} \\
E(p_{21}) &= R\,p_{21} + q_{11} - R\,p_{11} &&= R(p_{21} - p_{11}) + q_{11} = (q_{21} - q_{11}) + q_{11} = q_{21} \\
E(p_{22}) &= R\,p_{22} + q_{11} - R\,p_{11} &&= R(p_{22} - p_{11}) + q_{11} = (q_{22} - q_{11}) + q_{11} = q_{22}
\end{aligned}$$

Finally, observe that
$$\begin{aligned}
E(p_{12}) &= R\,p_{12} + q_{11} - R\,p_{11} \\
&= R(p_{21} - p_{11}) - R(p_{21} - p_{12}) + q_{11} \\
&= (q_{21} - q_{11}) - (q_{21} - q_{12}) + q_{11} \\
&= q_{12}
\end{aligned}$$

So (1) is true as well.

Appendix B.4 [Corollary 4, p. 10]

Corollary 4: If R is a rotation such that the maximum distance between corresponding vertices of the two polygons $R(e_2 \ominus e_1)$ and $f_2 \ominus f_1$ is δ, then there is a rigid body transformation E such that e_1 and e_2 are (E, δ)-homotopic to f_1 and f_2, respectively.

Proof: We will use similar notation, algebra, and techniques as in the proof of Lemma 4.

First, define $E(x) = Rx + t$ with $t = \frac{1}{2}(q_{11} + q_{12} - Rp_{11} - Rp_{12})$.

Then write $R(p_{2i} - p_{1j}) = q_{2i} - q_{1j} + \Delta_{ij}$, with Δ_{ij} a vector, for $i, j = 1, 2$.

By assumption

$$\max_{i,j} ||\Delta_{ij}|| = \delta.$$

We see that:

$$\begin{aligned}
E(p_{11}) &= Rp_{11} + \tfrac{1}{2}(q_{11} + q_{12} - Rp_{11} - Rp_{12}) \\
&= \tfrac{1}{2}(R(p_{21} - p_{12}) - R(p_{21} - p_{11}) + q_{11} + q_{12}) \\
&= \tfrac{1}{2}(q_{21} - q_{12} + \Delta_{12} - q_{21} + q_{11} - \Delta_{11} + q_{11} + q_{12}) \\
&= q_{11} + \tfrac{1}{2}(\Delta_{12} - \Delta_{11}) \\
E(p_{21}) &= Rp_{21} + \tfrac{1}{2}(q_{11} + q_{12} - Rp_{11} - Rp_{12}) \\
&= \tfrac{1}{2}(R(p_{21} - p_{11}) + R(p_{21} - p_{12}) + q_{11} + q_{12}) \\
&= \tfrac{1}{2}(q_{21} - q_{11} + \Delta_{11} + q_{21} - q_{12} + \Delta_{12} + q_{11} + q_{12}) \\
&= q_{21} + \tfrac{1}{2}(\Delta_{11} + \Delta_{12}) \\
E(p_{22}) &= Rp_{22} + \tfrac{1}{2}(q_{11} + q_{12} - Rp_{11} - Rp_{12}) \\
&= \tfrac{1}{2}(R(p_{22} - p_{11}) + R(p_{22} - p_{12}) + q_{11} + q_{12}) \\
&= \tfrac{1}{2}(q_{22} - q_{11} + \Delta_{21} + q_{22} - q_{12} + \Delta_{22} + q_{11} + q_{12}) \\
&= q_{22} + \tfrac{1}{2}(\Delta_{21} + \Delta_{22}) \\
E(p_{12}) &= Rp_{12} + \tfrac{1}{2}(q_{11} + q_{12} - Rp_{11} - Rp_{12}) \\
&= \tfrac{1}{2}(R(p_{21} - p_{11}) - R(p_{21} - p_{12}) + q_{11} + q_{12}) \\
&= \tfrac{1}{2}(q_{21} - q_{11} + \Delta_{11} - q_{21} + q_{12} - \Delta_{12} + q_{11} + q_{12}) \\
&= q_{12} + \tfrac{1}{2}(\Delta_{11} - \Delta_{12})
\end{aligned}$$

It follows that:

$$\begin{aligned}
||E(p_{11}) - q_{11}|| &\leq \tfrac{1}{2}(||\Delta_{12}|| + ||\Delta_{11}||) \leq \delta \\
||E(p_{21}) - q_{21}|| &\leq \tfrac{1}{2}(||\Delta_{11}|| + ||\Delta_{12}||) \leq \delta \\
||E(p_{22}) - q_{22}|| &\leq \tfrac{1}{2}(||\Delta_{21}|| + ||\Delta_{22}||) \leq \delta \\
||E(p_{12}) - q_{12}|| &\leq \tfrac{1}{2}(||\Delta_{11}|| + ||\Delta_{12}||) \leq \delta
\end{aligned}$$

It is always possible to construct a homotopy between two directed edges that morphs one into the other. The previous inequalities show that we can in fact construct (E, δ) homotopies between e_i and f_i, for $i = 1, 2$.