# Protein Similarity from Knot Theory: Geometric Convolution and Line Weavings

Michael A. Erdmann

May 16, 2004

CMU-CS-04-138

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, isotopy, 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 $\alpha$-helices have become $\beta$-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 induced by 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 isotopic 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 two algorithms, one based on writhing numbers and one based on line weavings. The algorithms have time complexities ranging from $O(n^2)$ to $O(s^{11})$, depending on level of detail, where $n$ is the number of residues in the protein and $s$ is the number of secondary structure elements. In practice, the running times were nearly interactive. We define and use a new datastructure, called *geometric self-convolution*, within the writhing-based algorithm.


**Contributions:** We believe that this is the first paper to consider carefully the need for combining metric and isotopic qualities in seeking protein similarity. We provide a parameterized definition of similarity that leads naturally to a metric in protein space. The underlying topological approach leads further to a representation of proteins by line weavings. We exhibit algorithms for computing the metric and for detecting similarity. We report results obtained with a dozen pairs of proteins, exhibiting a range of typical features.

**This report supersedes and enhances Technical Report CMU-CS-03-181.**

# 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 [25], Vast [21], CE [52], and 3dSearch [54]) 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) [47, 10] 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 alpha carbons of a $\beta$-sheet, one can easily detect two orthogonal families of curves, one family parallel to the constituent $\beta$-strands, the other perpendicular to the strands. This suggests that nature may create the same two-dimensional $\beta$-sheet using orthogonal strand directions, hinting at interesting biochemical/genetic rearrangements. Indeed, proteins routinely create the same functional shapes using significantly different atom arrangements. Detecting such similarities is the goal of sequence order-independent comparison algorithms [38, 6]. As X-Ray and NMR methodologies enter high-throughput capability, even more exotic similarity searches will arise routinely, likely requiring additional methods of structure detection.

Lacking is a good definition of "similarity", even for today's alignment tools. The Structural Classification of Proteins (SCOP) website [37] 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 [48] that classifies protein structures in terms of Gauss integrals, motivated by ongoing work on knot invariants [9]. In this paper we explore the connection further, leading naturally to a metric in protein space and two datastructures for representing protein geometries:

(i) The first datastructure represents self-convolutions of polygonal curves. Applied to a protein, this datastructure delineates internal translations that may change the shape of the protein.

(ii) The second datastructure represents a protein by line weavings derived from its secondary structure elements. One may view this datastructure as topological essence extracted from the previous self-convolution datastructure.

We implement algorithms for detecting substructural similarities between proteins based on these two datastructures. We report results for twenty-four proteins. We comment on connections with methods from Robot Motion Planning.

**Paper Outline**

Section 2 reviews related work on structural alignment and discusses the role of metrics. Section 3 provides an intuitive introduction to topological similarity, structural isotopies, and line weavings. Section 4 reviews the basics of knot theory. Section 5 is the technical heart of the paper. That section defines isotopies, similarity, and a precise version of the structure problem, then proves computability of that problem in the Appendix. Section 6 provides the connection between the general structure problem and our approximation based on writhing numbers, describes the writhing-based algorithm, and reports results. Section 7 describes our approach based on line weavings, and reports results. Finally, Section 8 discusses future directions and Section 9 summarizes.

# 2 Related Work

## 2.1 Structural Alignment

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 [25, 27, 26] 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 [52, 53] 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.

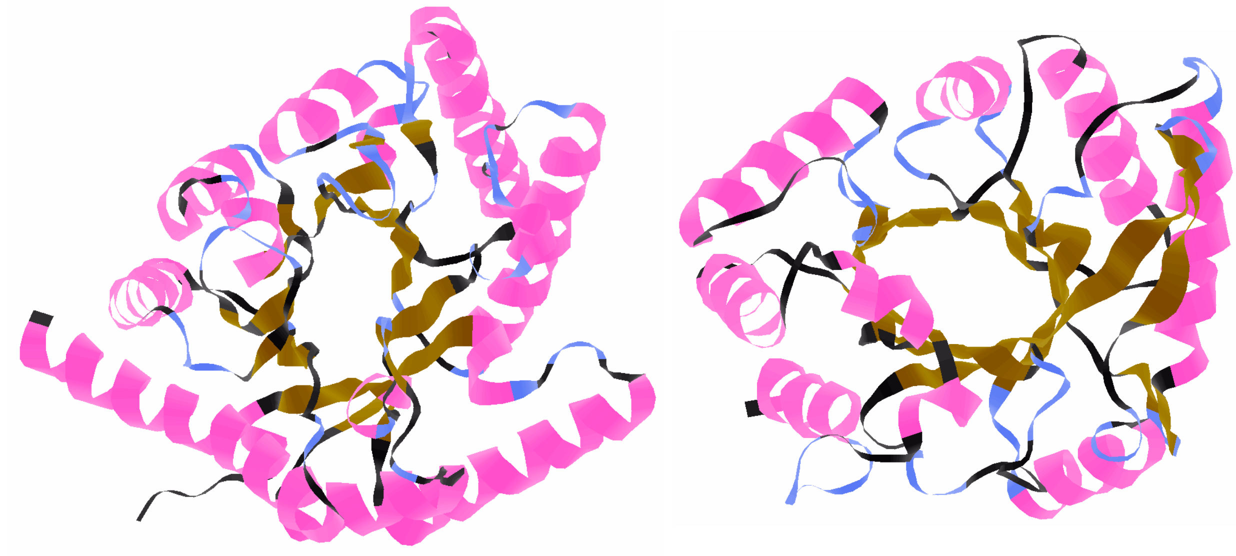
VAST [21, 22] and 3DSEARCH [54, 55] focus on elements of secondary structure to align proteins. Both methods begin with building blocks that are pairs of secondary structure elements, one pair in each protein. VAST matches pairs of secondary structural elements that have a similar type, relative orientation, 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 pairs of secondary structure vectors in one protein that match well with pairs of vectors in the other protein. These initial alignments repeatedly seed a dynamic programming algorithm for aligning all secondary structure vectors in one protein with those in the other protein. Atom-level alignment occurs subsequently. 3DSEARCH is potentially limited by its set of initial vector alignments.

## 2.2 Metrics

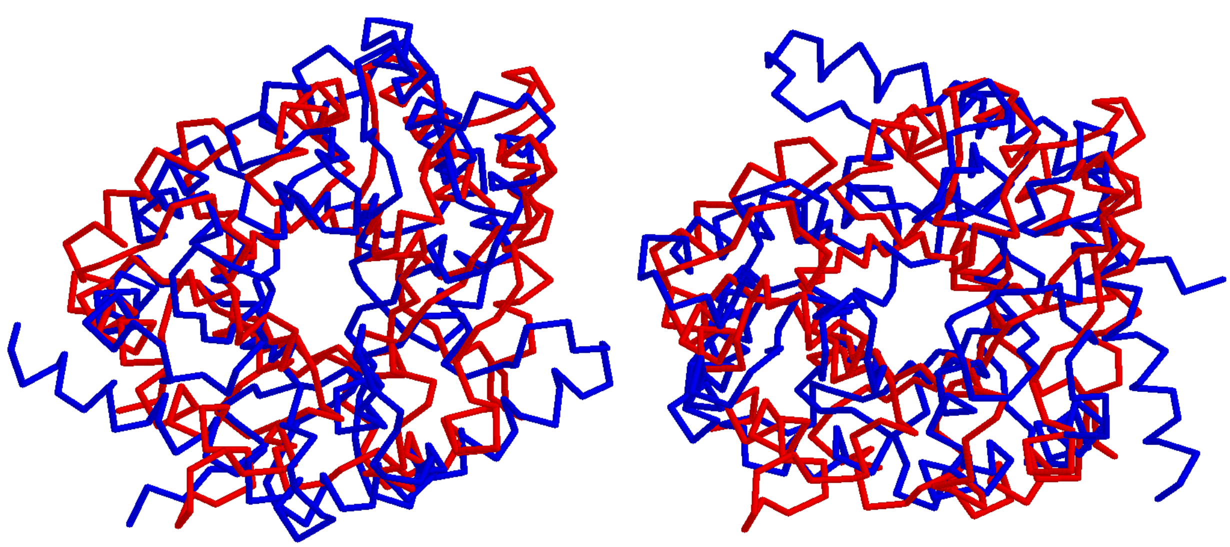
One of the difficulties with many alignment methods is the vagueness of their global similarity measures. Locally, these methods often measure similarity by the root-mean-square-deviation (RMSD) between aligned atoms, or some related variation. RMSD of aligned atom coordinates is a wonderful measure 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 [28]. More appropriate for proteins may be invariants derived from knot theory. Røgen and Fain [48] suggest a metric based on curve invariants. 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 [40, 39] (both SCOP and CATH are widely accepted protein classification databases, created by a combination of automatic and human judgments). The primary invariant in [48] is the *writhing number* of a curve; the others are built from this. Section 4.3 examines writhing numbers in detail.



1xis 1nar

Figure 1: Two TIM-barrels. On the left is Xylose Isomerase (PDB code: 1xis), minus its tail. On the right is Narbonin (PDB code: 1nar). Both proteins are displayed in `RASMOL`'s ribbon format [50].



Optimal Alignment Alternate Alignment

Figure 2: Two alignments of 1xis (blue) with 1nar (red). On the left is the optimal DALI alignment, on the right an alternate alignment formed by rotating 1xis approximately 1/3 turn about the TIM-barrel. Both proteins are displayed in `RASMOL`'s backbone format.

# 3 Goals and Intuition

## 3.1 Topology and Invariants

### Topology

The long-range goal of our research is to develop compact representations of protein geometry useful for structure comparison. The most fundamental representations of geometry are topological in nature, providing information about incidence, relative location, and allowable motions. For proteins, the elements of knot theory are likely to be useful, not because proteins are or are not knots, but because the geometry and motions of protein backbones may be modeled using techniques from knot theory.

Topology offers high-level descriptions of shape and motion. While precise folding paths of proteins depend intimately on the details of steric constraints, electrostatic potentials, and biochemical entropies, fundamental fold descriptions should not. One should be able to recognize the similarity in folds between two proteins based purely on topological considerations. By way of intuition, the weaving of the threads in my shirt is a characteristic of that shirt, independent of whether I am hunched over my terminal, standing straight, tugging on my shirt, or allowing it to hang loosely. The threads will move and turn, but their relative topological relationships will remain unchanged, so long as I do not tear the shirt. Similarly, two proteins may have very different three-dimensional coordinates yet be instances of the same fold. Not tearing the threads in my shirt is analogous to the assumption that a protein will not break the covalent bonds in its backbone as it moves.

**The key idea is that two proteins or subsegments of proteins are similar if there is a motion that transforms one into the other while avoiding backbone self-collisions.** The role of knot theory is to offer simple descriptors (called *invariants*) by which one can assess the similarity of two proteins rapidly.

### Invariants

Discovering useful invariants is at the core of modern knot theory. It is easy to find invariants that do not change as a curve deforms smoothly in space. It is much more difficult to find invariants that are sensitive enough to act as characteristics, meaning: (i) The invariant of a curve does not change with smooth deformations of that curve and (ii) the invariant can discriminate between two curves that are topologically dissimilar. (Two closed curves are topologically dissimilar if the curves cannot be deformed into each other except by tearing/cutting.) Research in modern knot theory entails discovering ever more sensitive knot invariants; finding a true characteristic is an open research question. We point to the nice introduction by Louis Kaufmann [29]. In the context of proteins, we also point to the work of Taylor [57] on defining fundamental arrangements of protein shapes and the work by Willett [23] on tertiary structure graphs.

### A Range of Problems

This paper constitutes our first step in developing topological shape descriptors for proteins. There are several lines of attack, with different levels of topological emphasis. Section 5 defines protein structural similarity in terms of collision-free motions. The key result of that section is the construction of a metric in protein space and a proof of its computability (in Appendix A). Section 6 then offers a more practical approach, using writhing numbers as the basis of a structure comparison algorithm. Finally, Section 7 returns to the topological foundation, developing an approach for structure comparison based on line weavings of secondary structure elements.

It is instructive to realize that the existence of collision-free motions is a purely topological concept, while the definition of a metric is dependent on the precise coordinates of the proteins' atoms. Similarly, the set of crossing numbers associated with a line weaving is a topological concept, while the set of writhing numbers associated with a polygonal curve is dependent on embedding coordinates. We thus have the following list of problems:

- Knot Equivalence: Decide whether two closed curves are topologically equivalent, that is, whether one curve can be transformed into the other using a smooth collision-free motion.

- Polygonal Curve Similarity: Determine the smooth collision-free motion with least excursion that transforms one polygonal curve with $n$ vertices into another polygonal curve with $n$ vertices, preserving the existence and number of vertices during the motion. By the "excursion" of a curve we mean the maximum distance any vertex moves from its start or final position; Section 5 will define this notion precisely in terms of "$(E, \delta)$-isotopies".

- Weaving Equivalence: Decide whether two arrangements of infinite lines are isotopic to each other, that is, whether one arrangement of lines can be transformed into the other without causing any of the moving lines to intersect or become parallel.

- Embedding Similarity: Decide whether two polygonal curves with equal number of vertices are everywhere locally similar. By local similarity we mean that two edges in one curve have nearly the same relative separation and orientation as their corresponding edges in the other curve. A special case of this problem is the limit in which "nearly the same" means "exactly the same". That special case asks whether two curves are completely the same shape, merely transformed by a rigid body motion.

An approach for deciding Knot Equivalence exists, though with unknown complexity and uncertainty about its computability in the $\mu$-recursive sense [24]. A variant of this problem is Unknot, the problem to decide whether a closed curve is topologically equivalent to the unknotted loop. That problem is known to lie in NP and co-NP, but it is not known whether the problem is polynomial-time decidable [24, 3, 2]. If a closed curve is known to be the unknot then it can be flattened quickly [12, 11]. Observe that Knot Equivalence is a purely topological question.

Polygonal Curve Similarity is both a simplification and an elaboration of Knot Equivalence. Simplifying, the curves are now piecewise linear with an equal number of vertices and the transformation preserves the existence and number of vertices throughout the motion. Elaborating, the curves need not be closed and the problem asks for a motion that minimizes the greatest excursion any vertex needs to make in order to establish the similarity.

The main theoretical result of the current paper is that this problem is effectively computable. As an aside, the proof shows that a simplified version of KNOT EQUIVALENCE, in which the curves are polygonal and the number of vertices remains constant during motion, lies in PSPACE. We also point to [45, 14, 5] for related PSPACE-hardness and -completeness results. Observe that POLYGONAL CURVE SIMILARITY has a strong topological component, but the precise distance value computed is dependent on embedding coordinates.

WEAVING EQUIVALENCE is an open problem. It is not even known how many different isotopy classes exist for a given number of lines, when the number of lines is large. For small numbers of lines the problem is well understood, and the isotopy classes are characterized by simple invariants. Our weaving-based algorithm uses such small sets of lines as seeds to match up the secondary structure elements of two proteins. Observe that WEAVING EQUIVALENCE is a purely topological question.

EMBEDDING SIMILARITY is the general curve recognition problem. In this paper we address the problem using writhing numbers. Observe that a solution to this problem depends on the embedding coordinates of the two curves.

We view our algorithms for EMBEDDING SIMILARITY and WEAVING EQUIVALENCE as approximations to the general POLYGONAL CURVE SIMILARITY problem. These algorithms therefore provide a basis for detecting common protein substructures. The main practical contribution of the current paper is the use of writhings and weavings to generate protein structure alignments.

## 3.2 Structural Alignment Isotopies

We will illustrate our topological goals using the two proteins shown in Figure 1. On the left is the core of Xylose Isomerase (PDB code: 1xis), an enzyme that catalyzes the conversion of glucose into fructose. On the right is Narbonin (PDB code: 1nar), a plant seed protein with no known enzymatic function. Both proteins are TIM-barrels, and thus are structurally alignable in a variety of ways. Approximately 70% of the residues are structurally similar, even though the two proteins have only 7% sequence identity.

Figure 2 shows two alignments. On the left is the optimal DALI-alignment. On the right is an alternate alignment, in which Xylose Isomerase has been rotated approximately 120 degrees about the TIM-barrel. Alignments similar to these would likely appear in the top ten list produced by any comprehensive structural alignment program.

Structural alignment occurs at multiple scales, ranging from global superpositions to local residue alignments, possibly with a variety of scales in between, such as secondary structure superposition. Figure 2 displays its two alignments as rigid body superpositions. Such superpositions tell part of the story. From a biochemical perspective, structural alignment programs must also produce pairings at the residue level.

Geometrically, one may think of structural alignment as a sequence of motions that establishes similarity by transforming one protein shape into another. For residue alignments one must therefore exhibit motions that transform segments of one protein's backbone into corresponding segments of the other protein's backbone. In order to avoid geometrically and biochemically silly alignments we require these motions to avoid self-collisions. Such motions are called *isotopies*. Focusing on isotopies rather than arbitrary motions and alignments provides a basis for believing that the shapes are inherently similar, as opposed to coincidentally similar, from a topological perspective.

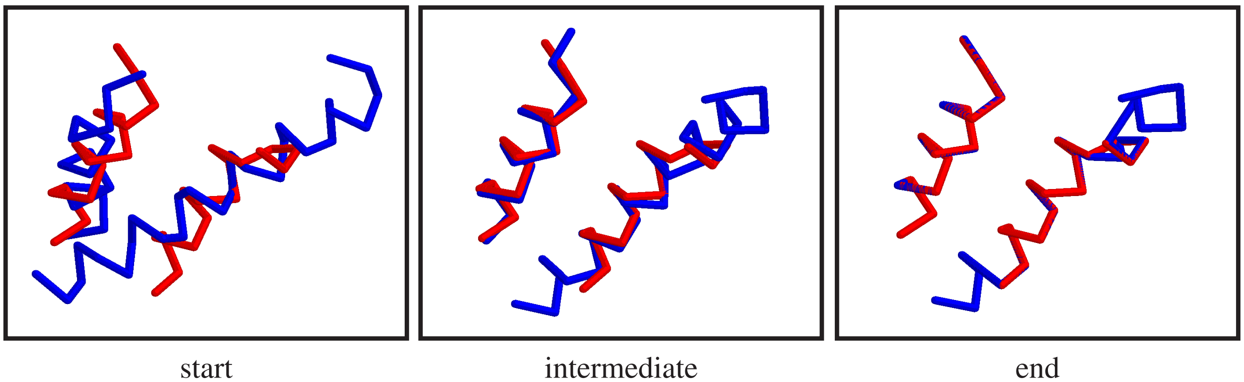



Figure 3: Isotopies of two pairs of helices, shown at three different snapshots in time. Two helices of 1xis (blue) morph into their counterparts in 1nar (red). The isotopy for the left pair of helices in each frame is essentially a quarter-turn rotation about the helical axis. The isotopy for the right pair of helices in each frame is essentially a rotation about an axis perpendicular to the helices, followed by a loop rearrangement near the top of the blue helix.

Figure 3 shows isotopies between two helices of Xylose Isomerase and the corresponding two helices of Narbonin. There is one isotopy for each pairing of helices; each isotopy morphs a backbone segment of Xylose Isomerase into a backbone segment of Narbonin. The start of the two isotopies is given by a high-level rigid body superposition of the two proteins, in this case the optimal DALI-alignment of Figure 2. The amount of motion required by each local isotopy provides a rough measure of how similar the two *pairs* of helices are to each other, as requested by the POLYGONAL CURVE SIMILARITY problem.

It is instructive to observe that the two local isotopies are very different from each other. To first approximation, one is a rotation about one of the helix axes, while the other is a rotation about a perpendicular axis. Neither isotopy determines the other. Instead, both isotopies are motions that occur subsequent to a global rigid body motion that roughly superimposes one pair of helices on the other pair. Determining such a global rigid body superposition is analogous to selecting a convenient origin in motion space, around which one can then compute finer-grained local motions to establish shape similarity between subsegments of the two proteins. In practice, the two scales influence each other. A rigid body superposition may suggest local isotopies. Conversely, a *collection* of local isotopies may suggest a global rigid body superposition. Section 5 will make this notion precise, by defining isotopies and shape similarity relative to rigid body motions.

### 3.3 Line Isotopies

In this subsection we illustrate the basic principles of our long-range goal, to develop topological characterizations of protein shape similarity. The exposition focuses on $\alpha$-helices but applies as well to $\beta$-strands.

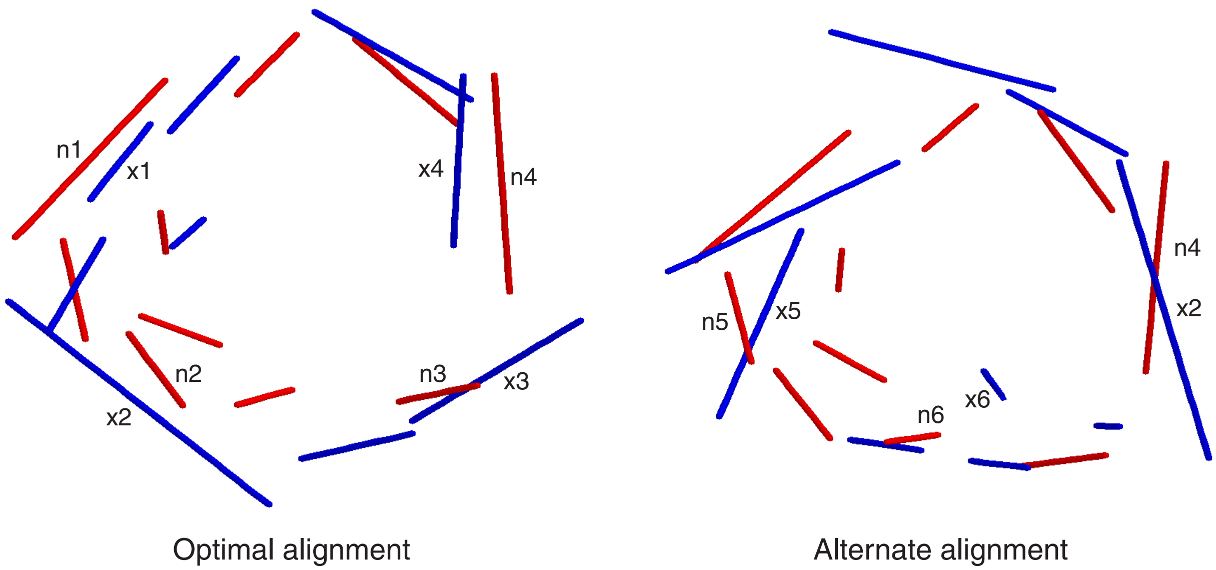



Figure 4: This figure again depicts the two alignments of Figure 2, now showing only the helix axes as line segments. 1xis is in blue, 1nar in red. Some of the line segments are labeled with identifiers (“`xi`” for helices in 1xis and “`ni`” for helices in 1nar). These generate the line weavings of Figures 5 and 6.

#### Helix Line Weavings

Figure 4 displays line segments that model the helix axes of the two alignments shown earlier in Figure 2. Ideally, one would like to describe this arrangement of lines in a compact fashion that reveals commonalities and differences. One possibility is to look at small subsets of lines and decide whether they are topologically equivalent to each other as (oriented) line arrangements, meaning that there is an isotopy that transforms one arrangement into the other. By an isotopy of a line arrangement one means motions of the lines in which the lines remain skew.

For instance, we have labeled four pairs of the helix axes in the left panel of Figure 4. Imagine drawing infinite lines through these axes; Figure 5 shows the results in two panels, with blue lines for Xylose Isomerase and red lines for Narbonin. Each panel describes a *line weaving*. It is clear from the figure that the two weavings are topologically equivalent, meaning that we can move the lines of one color without collisions in such a way that they are completely identical to the lines of the other color. We have labeled the lines with their backbone orientations and the crossings with six *crossing numbers*, which happen to all be “+” in this case. Crossing numbers will be explained in more detail in Section 4.1. For now, what matters is that these six labels are identical for the blue and red weavings, indicating that the two arrangements of four lines are topologically equivalent.

In contrast, consider the three labeled pairs of helix axes in the right panel of Figure 4. The two associated line weavings are shown in Figure 6 (from a rotated perspective for better viewing). It is clear that the weavings are different. In fact, the alternate alignment of Xylose Isomerase with Narbonin is generally quite good, but there is one topologically troublesome helix-pairing as the two line weavings of Figure 4 indicate.

**Arrangements of Lines**

We have just described the rudiments of the theory of line arrangements, an area closely related to knot theory. For a more comprehensive introduction see [60, 61]. Research in this area seeks to classify the topological equivalence classes of line arrangements under isotopy. There is exactly 1 topological equivalence class consisting of 2 skew (unoriented) 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 [41, 42].

The potential application to protein structure comparison arises in three contexts. First, structural alignment programs often represent proteins by their secondary structure vectors [21, 54, 58]. Classifying such vector arrangements might provide simple invariants by which to label protein folds, as suggested by our previous examples. 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 [59, 4, 31, 19]. 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 induced equivalence classes.

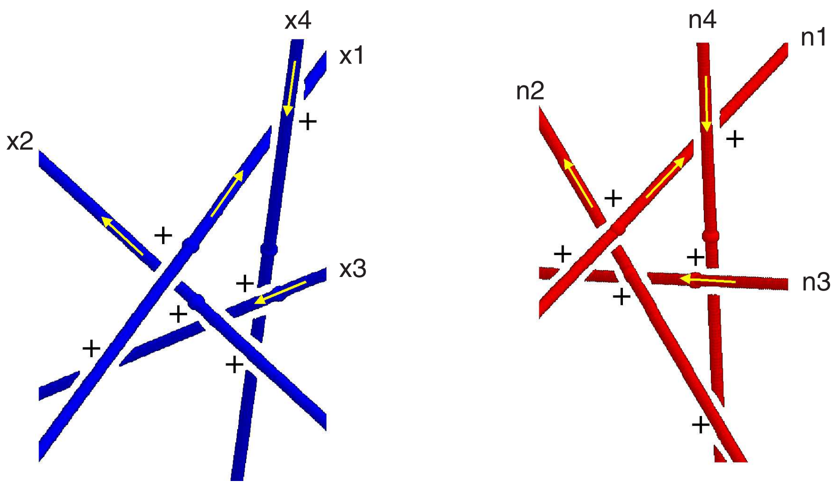



Figure 5: Line weavings generated from the four labeled pairs of edges shown in the optimal alignment of Figure 4. Each labeled helix edge generates a thick infinite line in the weaving. The yellow arrows indicate the backbone directions. The viewing perspective is the same as in Figure 4, looking square at the paper, only from further back so all crossings are visible.

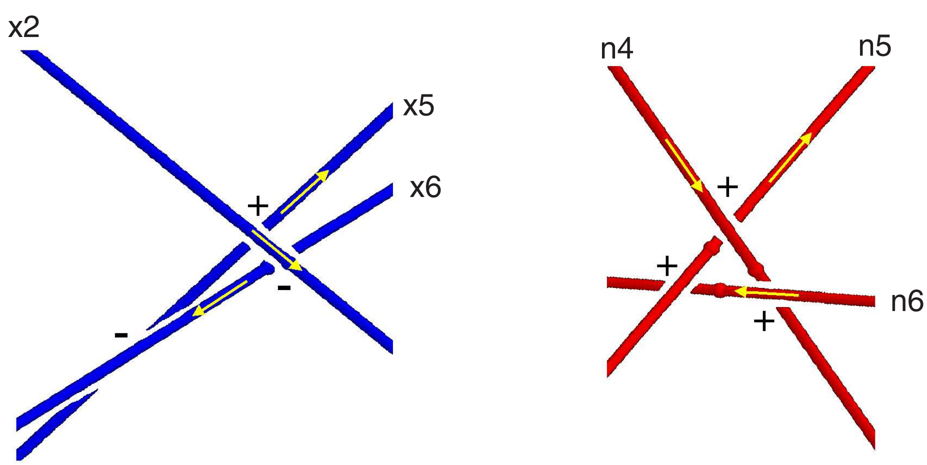



Figure 6: Line weavings generated from the three labeled pairs of edges shown in the alternate alignment of Figure 4. Again, each labeled helix edge generates a thick infinite line in the weaving. The viewing perspective is from the right side of the drawing depicted in Figure 4, looking tangential to the paper.

# 4 Elements of Knot Theory

## 4.1 Crossing and Linking Numbers

One of the fundamental invariants of knot theory is the *crossing number*. Imagine viewing two oriented curve segments in space. For some viewing directions these curve segments will seem to cross each other. One segment will be closer to the viewer than the other. Thus if one projects the curves into a plane perpendicular to the viewing direction, one curve will seem to cross over the other. This relationship defines a crossing number, written $\varepsilon$, with

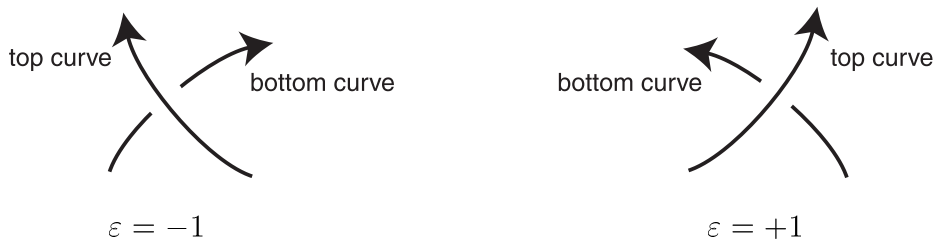



Figure 7: The two types of crossings and their crossing numbers.

value $-1$ or $+1$. Specifically, imagine rotating the top curve so that its forward tangent at the crossing is parallel to the forward tangent of the bottom curve. Then $\varepsilon$ is given by the sign of the smallest angle required. See Figure 7.

Observe that for two oriented, skew, infinite, straight lines in three-dimensional space the crossing number does not depend on the viewing direction. It is a purely topological property of the line directions and their relative locations in space. We saw these crossing numbers earlier, in the form of "+" and "−" labels in Figures 5 and 6.

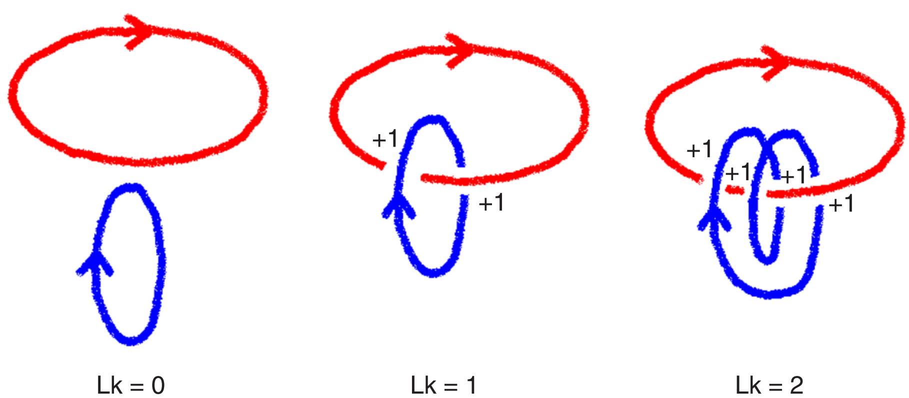



Figure 8: The linking of two curves is defined as the sum of the crossing numbers divided by two. This figure shows a pair of unlinked curves, a pair of singly linked curves, and a pair of doubly linked curves.

For two distinct closed oriented curves $c_1$ and $c_2$ in 3D space one can define the linking number $Lk(c_1, c_2)$ of the two curves as the sum of the crossing numbers divided by two. In turns out that this number does not depend on the viewing direction. Moreover, it is a topological invariant. This means that $Lk(c_1, c_2)$ is invariant to any smooth collision-free deformations of the two curves.

For simple curves the linking number provides a rough measure of how linked the curves are. See Figure 8 for examples. And in our previous discussion involving crossings of helix lines, we were essentially treating infinite lines as "half"-curves closed at infinity.

We caution that while crossing and linking numbers are topological invariants, they are not discriminating enough to be characteristics. For instance, the Whitehead link has linking number zero yet consists of two inseparable loops [29]. In the case of lines, there is an arrangement of six (unoriented) lines which is not isotopic to its mirror image, yet both the given arrangement and its mirror image have matrices of crossing numbers that lie in the same switching class [7]. This shows that crossing numbers are insufficient for classifying arbitrary arrangements of unoriented lines. According to folklore there is a similar example for oriented lines but we do not have a reference.

Also interesting in the case of oriented lines is an example in which a single line with specified crossing numbers relative to a set of fixed lines generates multiple isotopy classes [35]. Moreover, Chazelle et al. [16] conjecture that there are examples in which the orientation class of a single line may have $\Theta(n^2)$ isotopy classes, where $n$ is the number of fixed lines.

Fortunately, for small collections of oriented lines (e.g., 5 or fewer), crossing numbers fully characterize the isotopy classes. Consequently, if we see two weavings generated by a small set of oriented lines with permutation-equivalent crossing matrices, then we know there exists an isotopy that transforms one weaving into the other. Thus such weavings are good anchors by which to ground a search for global rigid body alignments. We will return to this topic in Section 7.

## 4.2 Gauss Integrals

It turns out that the linking number of two curves can be computed as a continuous integral. Formally, suppose $c_1$ and $c_2$ are two closed non-intersecting curves in 3D space, specifically disjoint embeddings of $S^1$ into $\Re^3$. Let $G$ be the Gauss map applied to the difference between the curves, that is, 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:

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

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

Amazingly, for two distinct closed curves, this integral is always an integer. To gain some intuition, consider two closed curves in space (see also Figure 9 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 on the curves, with sign given locally by the crossing number $\varepsilon$ of the two curve tangents. This is the value computed by the integral.

With some effort one sees that the net area covered is the linking number of the two curves as previously defined. In particular, if the two curves are not linked, as in the leftmost frame of Figure 8, then the net area covered on the sphere will be zero. If the two curves are linked once as in the middle frame, then the sphere will be fully 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.**

## 4.3 Writhing

### Writhing and Linking

The *writhing number* of a curve measures the curve's self-linking. Previously we defined the linking number for two curves. Linking and writhing are related by the following famous Călugăreanu-Fuller-White formula [18, 20, 62] defined for closed orientable ribbons in three-dimensional space:

$$Lk = Wr + Tw$$

Here $Lk$ 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 *twist* of the two boundary curves. While $Lk$ is a purely topological number, the other two numbers are not; 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. We note in passing that the writhing number and the twist are almost never integers.

It turns out that the writhing number of a curve has the same algebraic form as the linking number. If $c : S^1 \to \Re^3$ is a closed curve in space, then its writhing number is simply $Wr(c) = Lk(c, c)$. 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 $Tw$ [36]; it is a torsion-dependent term measuring 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.

**Protein Fragments:** The definitions continue to make sense for open curves, that is, 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.

### Writhing of Polygonal Curves

We will represent protein backbones as open polygonal curves[1], connecting sequential residues via their 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

[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 ($^1\mathrm{H}^N$) are more natural [17, 8].

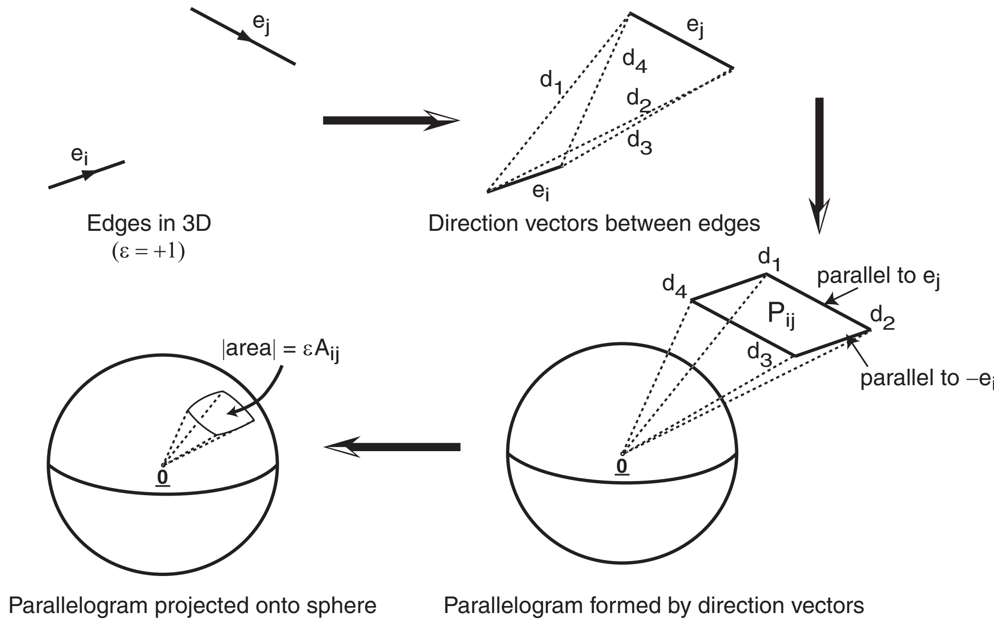



Figure 9: Edges $e_i$ and $e_j$ generate a parallelogram $P_{ij}$ of interedge directions, with vertices $d_1, d_2, d_3, d_4$. The absolute area of the parallelogram projected onto the unit sphere is $\varepsilon A_{ij}$, where $\varepsilon$ is the crossing number of the two edges (in the figure, $\varepsilon = +1$.) The *edge-edge writhing* is defined to be $A_{ij}/4\pi$.

proteins. Considerable work has used knot theory to understand the supercoiling and knotting behaviors observed in DNA, another polygonal curve (see [49] for a sample). Also, see [30, 43] 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:

$$Lk(c_1, c_2) = \frac{1}{4\pi} \sum_i \sum_j A_{ij}$$

where $A_{ij}$ is the $\varepsilon$-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.

**Definition 1** *We will refer to $A_{ij}/4\pi$ as the edge-edge writhing of the two edges $e_i$ and $e_j$.*

Computing $A_{ij}$ is straightforward. Figure 9 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 oriented edge $e_j$ are $q_1$ and $q_2$. Consider the four extremal cross directions between the two edges:

$$\begin{aligned} d_1 &= q_1 - p_1, & d_2 &= q_2 - p_1, \\ d_3 &= q_2 - p_2, & d_4 &= q_1 - p_2. \end{aligned}$$

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 $2\pi$. Observe that $A_{ij} = A_{ji}$.

# 5 Polygonal Curve Isotopies & 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 $\underline{(E, \delta)\text{-isotopic}}$ to curve $q$ if the following two conditions are satisfied:*

*(i) $n = m$.*

*(ii) There is a polygonal-curve isotopy $h$ mapping $E(p)$ to $q$ such that no representative point moves further than $\delta$ 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 all $i = 1, \ldots, n$.*

The $\delta$ appearing in this definition is the "excursion" to which we referred in the intuitive introduction of Section 3. 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 protein on top of the other well enough that $\delta$-perturbations of local coordinates permit atom alignment without backbone self-collisions.** The isotopy requirement mirrors formally the intuition of Sections 3.2 and 3.3: it measures similarity 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 $\delta$, but at all times preserves the backbone topology. (We note in passing a generalization: it might be interesting to restrict the class of isotopies 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 $\delta$, 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 $\delta$, 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:*

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

Thus $d(p,q) = \infty$ if and only if $p$ and $q$ are not isotopic for any $(E,\delta)$, e.g., if the number of representative points differs. Computing $d$ is the problem we called POLYGONAL CURVE SIMILARITY in the intuitive introduction of Section 3.

**Theorem 1** *$d$ is a metric and $d$ is effectively computable.*

PROOF. See Appendix A. □

**Monotonic Curve Isotopies** Given a point $p_i$ and a line $\ell$ in 3D space one can project the point orthogonally onto the line. One can do the same for all representative points $\{p_1,\ldots,p_n\}$ of some polygonal curve. The curve is said to be *monotonic with respect to line* $\ell$ 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 $\alpha$-helices and $\beta$-strands, 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 $\ell$. Let $\pi = \{\pi_1,\ldots,\pi_n\}$ be the polygonal curve obtained by projecting $p$ onto $\ell$. Then $d(p,\pi) \le \max_i ||p_i - \pi_i||$.*

PROOF. Imagine drawing a line between $p_i$ and $\pi_i$ for each $i$. Define a homotopy that moves each $p_i$ to $\pi_i$ along these lines. The homotopy preserves the polygonal curve (and thus is an isotopy) 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 $d(p,q) \le d(p,\pi) + d(q,\sigma) + \inf_E \max_i ||\sigma_i - E(\pi_i)||$, where $E$ is taken from the set of rigid body motions that align the two* oriented *lines.*

PROOF. See Appendix B. □

The bound in Lemma 2 is often generous. The lemma tells us that **two monotonic curves whose line-projections are similar in 1D are also readily isotopic 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 isotopies. 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>*$\delta$-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 isotopy 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 $\delta$ 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 the proof of Theorem 1. □

Although computable, the algorithm derived from our proof of Theorem 1 is horrendously exponential [13, 14, 32, 51]. One possibility is to use a motion planner specialized for knots, such as the untangling planner of [30]. Alternatively, for our purposes, Lemmas 1 and 2 suggest a simplification: In the next two sections we will examine one approach based on edge-edge writhings and a second approach based on line weavings, both of which attack the Structure Problem by aligning line projections of peptide segments.

# 6 Protein Similarity from Geometric Convolution

In this section we examine more closely the construction of Figure 9. Our observations will motivate us to define a self-convolution datastructure for detecting structural similarity in proteins.

## 6.1 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 [33, 34]. In particular, $Y \ominus X$ defines the set of translations of $X$ that cause collisions with $Y$.)*

**Lemma 3** *Assume $e_i$, $e_j$, and $P_{ij}$ are as defined at the end of Section 4.3. 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 $A_{ij}/4\pi$ of two oriented edges $e_i$ and $e_j$ is the absolute area of the convolution $e_j \ominus e_i$ projected onto the sphere $S^2$ times the crossing number $\varepsilon$ of the two edges, divided by $4\pi$.*

**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 edge-edge writhing $A_{ij}/4\pi$ 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 $A_{ij}/4\pi$ 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 $A_{ij}/4\pi$ 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. $P_{ij}\backslash\{0\}$ projects to one or two points on the sphere and the writhing $A_{ij}/4\pi$ 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.

## 6.2 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 oriented 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 C. □

**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 $\delta$, then there is a rigid body transformation $E$ such that $e_1$ and $e_2$ are $(E, \delta)$-isotopic to $f_1$ and $f_2$, respectively.*

PROOF. See Appendix D. □

When Corollary 4 applies we say that the polygons are *$\delta$-isotopic.*

**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 $p \ominus p$:*

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

A writhing polygon $P_{ij}$ delineates internal translations of a polygonal curve that cause self-collisions, namely of edge $e_i$ with edge $e_j$. The self-convolution $\otimes(p)$ therefore describes internal translations that may change the topological shape of 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 isotopy 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 features: 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.

## 6.3 Comparing Self-Convolutions

We now combine the isotopy 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 $\otimes^k(p)$.

**Convolution-based Matching 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 $\delta$-isotopic 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 $\delta$.

Figure 10 shows the magnitudes of the writhings $\{w_{ij}^k\}$ obtained from the self-convolution structures of 1xis and 1nar. These writhings were generated using polypeptide segments consisting of 11 residues, that is, with $k = 5$. The self-writhings of helices is evident in the bright red and orange bands along the diagonals of the matrices. The writhings of different $\beta$-strands appear as magenta off-diagonal peaks. The 8-fold symmetry of the TIM-barrel is clearly evident. Finally, the green speckle patterns indicate writhings of $\alpha$-helices with $\beta$-strands.

## 6.4 Analysis

The convolution-based 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 $\epsilon_P$ and $\epsilon_E$ are the resolutions of the polygon and rigid body hash tables, respectively.

In practice, $k$ and $\epsilon_E$ 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 $s$ can be $\Theta(n^2)$, it depends on $\epsilon_P$. Choosing this carefully, the ratio $s/\epsilon_P$ can become $O(n)$. In that case, the algorithm has $O(n^2)$ behavior, with $n$ the maximum protein length. The hash table resolutions constrain the observable distance $\delta$. The hash tables could be replaced by $k$-D trees, Voronoi diagrams, or other clustering methods [44], but we did not do so.

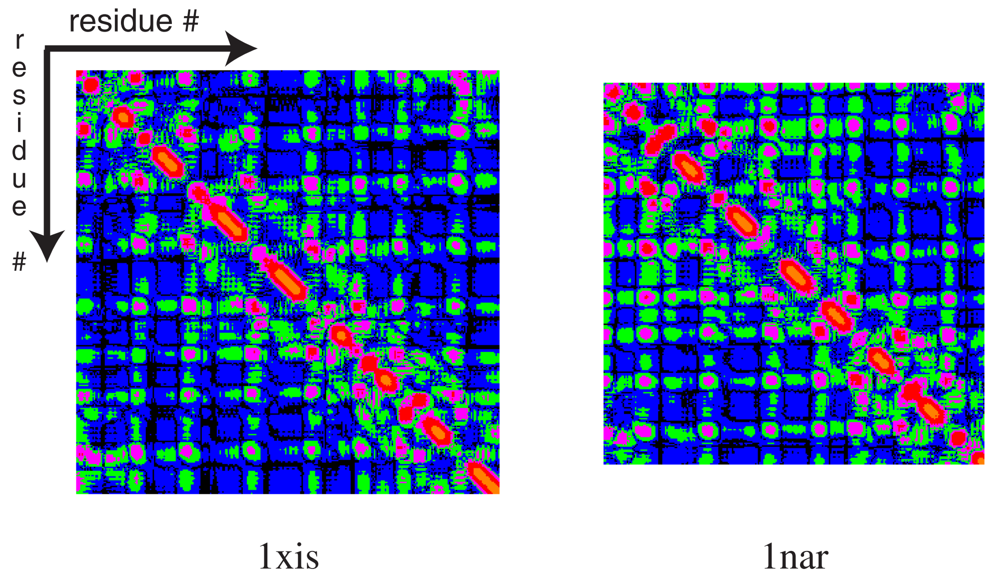



Figure 10: Writhing matrices for 1xis and 1nar. These matrices depict the writhing magnitudes computed by evaluating $|Lk(p_i^5, p_j^5)|$ for pairs of polypeptide strands, each consisting of 11 residues. These values are indexed by $i$ and $j$, that is, by the central residues of the two strands. Colors indicate writhing magnitudes as follows: Black: 0–0.01, Blue: 0.01–0.05, Green: 0.05–0.10, Magenta: 0.10–0.25, Red: 0.25–0.75, Orange: $\geq 0.75$.

## 6.5 Results from Self-Convolution

We implemented the algorithm in (an old 8-bit) `Lisp` on a 1GHz Windows PC. Running times for proteins with 300 residues were typically a minute or two, half of that garbage collection. (We chose that particular implementation simply because the author had written an extensive geometric and numerical library over the years in it, permitting easy interactive prototyping of ideas. We expect that a production-quality implementation in `C++` would likely be 10–100 times faster.) Here are three interesting pairs of proteins:

**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 [56]. 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 11.

**3adk vs. 1gky:** Adenylate Kinase (PDB code: 3adk) and Guanylate Kinase (PDB code: 1gky) are two transferases catalyzing two ATP-dependent phosphorylations. 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 Figure 12. 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.

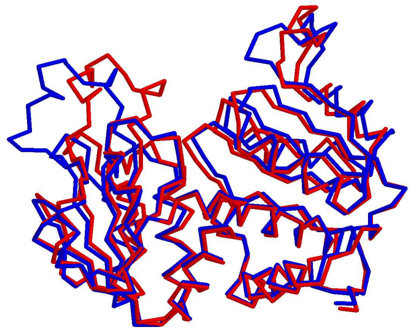


Figure 11: Alignment of 5at1_A (blue) and 8ATC_A (red) found by our convolution-based algorithm. The backbones match nearly perfectly, except where they should not, namely two loops that undergo significant conformational change (these appear near the top left and the top right in the figure).

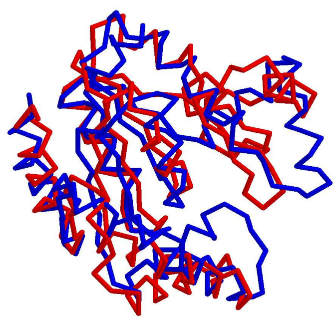


Figure 12: Alignment of 3adk (blue) and 1gky (red). The proteins have mere 19% sequence identity and include both matching and nonmatching secondary structures. Roughly 80% of the two proteins should align. One can see this in the figure, with the left parts matching well and some of the right clearly not.

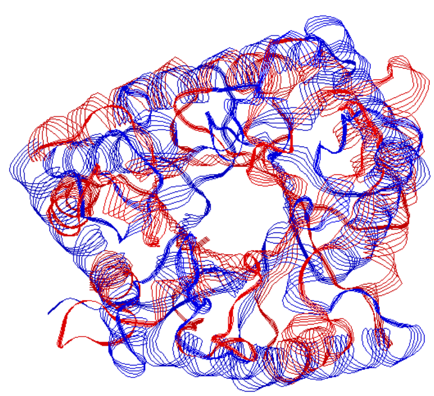


Figure 13: Maximal alignment of 1xis (blue) and 1nar (red), closely matching the optimal Dali-alignment. The proteins have 7% sequence identity.

**1xis vs. 1nar:** These are the two TIM-barrels we used extensively to illustrate the ideas of Section 3. We considered the 321 residues of 1xis without its tail versus the 289 residues of 1nar. The two protein chains have 7% sequence identity. As we mentioned earlier, there are several possible alignments, related by rotation around the central barrel (see again Figure 2). This pair of proteins is interesting because even in optimal alignment there are significant angular differences between aligned helices. Such comparisons originally motivated our isotopy definitions. Our code finds an alignment with RMSD 3.3Å, differing by 14° and 1.5Å from the optimal DALI-alignment. See Figure 13.

# 7 Alignments from Weavings

## 7.1 Comparing Crossing Numbers

Section 4 discussed line weavings of helix axes as a topological gauge of similarity. We have implemented that idea using line weavings derived from a protein's secondary structure elements, namely its $\alpha$-helices and $\beta$-strands. An $\alpha$-helix generates an oriented line representing the helix axis, while a $\beta$-strand generates an oriented line that best approximates the strand.

In order to deal with geometric singularities we model crossing numbers using three values, namely $-1$, 0, and $+1$. We assign the value 0 whenever two lines are nearly coplanar. Our code considers pairs, triples, and quadruples of lines, depending on the number of secondary structure elements available. Pairs of lines generate a single crossing number, triples generate three crossing numbers, and quadruples generate six crossing numbers. The code hashes sets of lines based on the number of their positive and negative crossing numbers, allowing 0 to act as a wild card. For instance, the three sets of crossing numbers $\{+1,+1,+1,+1,+1,-1\}$, $\{0,+1,+1,+1,+1,-1\}$, and $\{+1,+1,+1,+1,+1,0\}$ are all comparable. All three sets of crossing numbers might represent essentially the same quadruple of lines, except that two lines are (nearly) coplanar in two of the quadruples. The code looks for topologically similar weavings between proteins first by checking that their crossing numbers hash to the same bucket, then by checking whether their crossing matrices are related by a permutation matrix. Throughout, crossing number 0 acts as a wild card.

Each pairing of topologically similar line weavings between proteins generates a rigid map that aligns one quadruple (or triple or pair) of lines in one protein with a quadruple (or triple or pair) of lines in the other protein as well as is possible using a rigid map. Our code discards rigid maps that do not properly align their generating lines within some tolerance when viewed as points in the space of lines. Given such a core alignment of generating secondary structure elements, the code then extends the alignment to other secondary structure elements by looking for nearby neighbors.

In summary, the basic algorithm is:

**Weaving-based Matching Algorithm**

Given two proteins, detect structural similarity as follows:

1. Compute approximating lines for secondary structure elements in the two proteins.
2. Generate weavings of such lines in each protein (primarily quadruples).
3. Match topologically similar weavings.
4. Extend the matchings to other secondary structures in the two proteins.

In short, instead of hashing on geometric writhing numbers as in Section 6, we now hash on a topological invariant of the line weavings. Our results for the three pairs of proteins mentioned earlier are very similar using the two approaches. Table 1 lists some more; details in Section 7.2.

We note in passing that Step 4 can be performed in many ways. We use a bipartite graph matching algorithm, in which the underlying cost function is an $L_2$ measure, described further in Section 7.2.3. In addition, at various locations the code uses a variety of measures to prune or extend alignments. We omit the details.

The complexity of the weaving-based approach is potentially high — there are $O(s^4)$ quadruple-line weavings in a protein, where $s$ is the number of secondary structure elements in the protein, leading potentially to $O(s^8)$ comparisons between proteins. Extending an alignment of one pairing of weavings to all the secondary structure elements in the two proteins may require $O(s^2)$ effort to compute similarity and $O(s^3)$ to run an optimizing bipartite graph matcher. This suggests an overall complexity of $O(s^{11})$ for a straightforward implementation. In practice, we did not encounter exorbitant runtimes. In fact, with some exceptions, we generally found that our weaving-based matcher executed much faster than our writhing-based matcher. For many examples, the code ran in seconds to minutes, despite being implemented in an old 8-bit `Lisp`, though for proteins with large numbers of secondary structures the code sometimes ran for 20–60 minutes. Again, a production-quality implementation in `C++` would likely be 10-100 times faster.

One reason we observed reasonable runtimes is that we restricted the focus of our weaving-based matcher in the following three ways:

(a) When generating quadruples (or triples or pairs) of lines, the code requires the underlying secondary structure elements to lie spatially within some distance cutoff of each other. The precise distance is an input parameter to our code. We consistently used 30Å, which is about three-quarters the diameter of a typical protein domain.

(b) When generating quadruples and their associated rigid maps, the code first considers quadruples of $\alpha$-helices, turning to quadruples of $\beta$-strands only if there are insufficiently many helices, and then turning to mixed quadruples of helices and strands, if necessary.

(c) When extending alignments from quadruples to all the secondary structure elements in the two proteins, the code only matches secondary structure elements of the same type ($\alpha$ to $\alpha$ and $\beta$ to $\beta$). (Of course, it would be easy to remove that restriction.)

Restriction (a) in particular is good at limiting the number of generating quadruples. Since secondary structure elements are physical, the number of such elements that can be packed into a volume of 30Å is bounded by some constant. Thus the algorithm effectively only considers $O(s)$ weavings in each protein, leading to an overall complexity of $O(s^4)$–$O(s^5)$, depending on how Step 4 is implemented. We note in passing that geometric hashing could reduce this complexity even further.

## 7.2 Results from Line Weavings

Table 1 shows the alignments obtained by our weaving-based matching algorithm. As explained in the previous subsection, the code first matches weavings of a small number of lines in one protein with topologically equivalent weavings in the other protein. Each such pairing of line weavings seeds a routine that computes alignments between larger sets of secondary structure elements in the two proteins. The remainder of this section explains Table 1 further.

### 7.2.1 Alignment Rankings and Backbone Sequentiality

"Rank" in the table refers to an ordering given by the similarity measure $p_{12}$. Section 7.2.4 describes this measure further.

The table depicts two alignments of 1wsy_A with 2rus_A, namely those ranked #1 and #15. These proteins are TIM-barrels, exhibiting considerable rotational symmetries. The helices and strands are analogous to teeth in a gear, with consequent symmetry. The nominally correct alignment, as determined by CE, happens to rank #15. Interestingly, it is the first alignment in the ranking that preserves backbone sequentiality. If one asks the code to favor backbone order-preserving alignments, then the nominally correct alignment appears as the overall winner.

The comparison of 3adk with 1gky also has an ambiguity in its ranking. The #1 ranked alignment differs slightly from the nominally correct alignment. Again, this alignment also does not completely preserve backbone sequentiality. The first alignment that does preserve backbone sequentiality is indeed the nominally correct alignment, which happens to be ranked #2.

In all other cases, the first ranked alignment is also the nominally correct one, as measured by Dali, CE, and/or 3dSearch.

### 7.2.2 Crossing Consistency

An alignment between a set of $n$ secondary structures in one protein and a set of $n$ secondary structure in a second protein generates an associated crossing matrix in each protein. Each protein's crossing matrix contains the crossing numbers associated with the infinite lines that represent the aligned secondary structures. Each matrix is an $n{\times}n$ symmetric matrix with zeros on the diagonal.

For each alignment, one can compare the crossing numbers in the two crossing matrices generated by that alignment. The entries "`Bad/Sig:Tot`" in Table 1 do just that. "`Tot`" counts the number of crossings, that is the number of entries in the upper triangle of the

| Proteins | | Seq Sim | $\lvert SSE_1 \rvert$ | $\lvert SSE_2 \rvert$ | Rank | Alignment | | Crossing Consistency | | Deviation from correct rigid | | CA RMSD |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Prot1 | Prot2 | (%) | | | | \|SSE\| | $L2$ | Bad/Sig:Tot | $\rho_{12}$ | Å | deg | (Å) |
| 5at1_A | 8atc_A | 100 | 22 | 22 | 1 | 22 | 0.9 | 4 / 180 : 231 | 0.9988 | 0.2 | 1.5 | 1.4 |
| 3adk | 1gky | 18.8 | 15 | 14 | 1 | 11 | 2.1 | 2 / 37 : 55 | 0.7253 | 0.6 | 9.5 | 2.8 |
| " | " | " | " | " | 2 | 11 | 2.5 | 3 / 32 : 55 | 0.6915 | 0.1 | 11.7 | 2.9 |
| 1xis | 1nar | 7.1 | 20 | 18 | 1 | 13 | 2.5 | 2 / 59 : 78 | 0.6456 | 0.3 | 7.2 | 3.4 |
| 1fpk_A | 1fpk_B | 100 | 19 | 21 | 1 | 19 | 0.9 | 1 / 139 : 171 | 0.9993 | 0 | 1.6 | 0.6 |
| 1a6m | 1lhs | 64.2 | 7 | 7 | 1 | 7 | 0.8 | 0 / 18 : 21 | 0.9995 | 0 | 0.7 | 0.8 |
| 1pbg_A | 1gow_A | 26.8 | 27 | 31 | 1 | 24 | 1.5 | 5 / 231 : 276 | 0.8855 | 0.4 | 2.1 | 2.6 |
| 1ki7_A | 1qhi_A | 100 | 16 | 19 | 1 | 16 | 0.9 | 1 / 91 : 120 | 0.9994 | 0 | 1.2 | 1.5 |
| 1hyq_A | 1cp2_A | 21.5 | 19 | 15 | 1 | 14 | 1.6 | 2 / 64 : 91 | 0.7350 | 0.4 | 6.8 | 2.3 |
| 1atn_A | 3hsc | 13.6 | 28 | 25 | 1 | 20 | 2.1 | 8 / 145 : 190 | 0.7064 | 1.0 | 4.1 | 3.0 |
| 1d9c_A | 2rig | 40.0 | 7 | 6 | 1 | 6 | 1.4 | 1 / 10 : 15 | 0.7686 | 0.1 | 4.5 | 2.0 |
| 1wsy_A | 2rus_A | 11.1 | 19 | 29 | 1 | 17 | 1.9 | 0 / 111 : 136 | 0.8870 | 0.3 | 88.2 | 3.0 |
| " | " | " | " | " | 15 | 16 | 2.1 | 1 / 94 : 120 | 0.8347 | 0.3 | 7.6 | 3.0 |
| 1mjc | 1a62 | 24.6 | 5 | 9 | 1 | 5 | 1.9 | 0 / 9 : 10 | 0.9886 | 2.3 | 17.1 | 3.0 |

Table 1: Alignment of proteins from weaving topologies.

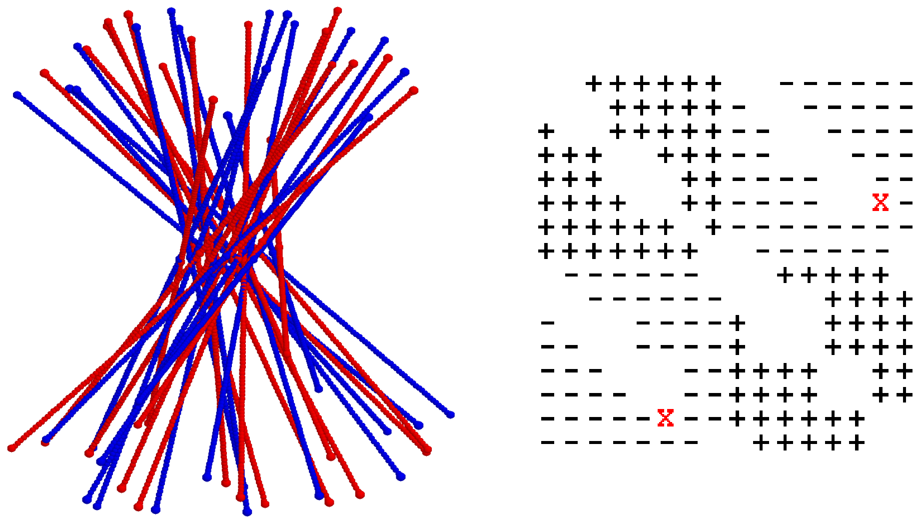
Each row represents an alignment of two protein chains. The alignments were seeded using line weavings as explained in the text.

The left set of columns lists the protein chain names (Prot1 and Prot2), their sequence similarity as a percentage, and the number of secondary structure elements (SSEs) eligible for alignment in each chain. The code only considers $\alpha$-helices with at least five residues and $\beta$-strands with at least three residues.

The middle set of columns depicts the results of an alignment: the rank of the alignment, the number of lines matched between the two proteins (|SSE|), a measure of the deviation between paired lines ($L2$), a measure of the line crossing consistency (Bad/Sig:Tot), and a cumulative similarity measure ($\rho_{12}$).
$L2$ measures a deviation, so small values are preferred; 0 is the smallest possible value.
$\rho_{12}$ measures similarity, so large values are preferred; 1 is the largest possible value.
The overall "Rank" is based on $\rho_{12}$.

The right set of columns assesses the accuracy of the results obtained. The first two columns show the deviation, in terms of distance offset and angular rotation, between the rigid map inferred directly from the line alignments and the optimal rigid map obtained from DALI, CE, or 3DSEARCH. The last column shows the RMSD between aligned CA atoms (alpha carbons), as computed by our atom alignment code (this alignment code starts with a rigid map computed from the line alignments, then tries to align both proteins, not just the secondary structures, using an iterative bipartite-graph closest-point routine).

crossing matrix; it has value $n(n-1)/2$, where $n$ is the the number of secondary structures in each protein that have been aligned. Some of these entries will be 0, indicating (nearly) coplanar lines. "`Sig`" counts the number of corresponding entries that are nonzero in both crossing matrices. "`Bad`" counts the number of these entries that are inconsistent, meaning that two secondary structure lines have crossing number "+1" in one protein while their aligned counterparts have crossing number "−1" in the other protein.




**(a)** **(b)**

Figure 14: Panel (a): Line weavings for the optimal alignment of 1wsy_A (blue) and 2rus_A (red). Panel (b): Overlay of the crossing matrices for the two weavings. An entry is blank if one or both of the crossing numbers is zero, it is the sign of the crossing number if the crossing numbers agree, and it is a red `X` if they disagree.

The weavings used to seed an alignment always have fully consistent crossing matrices. However, one would not necessarily expect the crossing matrices corresponding to an overall alignment induced by that seed to be consistent. After all, $\alpha$-helices and $\beta$-strands are actually finite-length polypeptide segments, not infinite lines. Thus a motion of a helix or strand could preserve the overall topology of a protein but change the crossing numbers associated with the protein's representation by infinite lines. It thus comes as a pleasant discovery that crossing matrices generally are indeed fairly consistent globally for good structural alignments.

By way of example, Panel (a) of Figure 14 depicts the line weavings for the correct alignment of 1wsy_A with 2rus_A. Panel (b) shows the overlay of the crossing matrices for the two weavings. It is interesting to observe both the roughly hyperbolic shape formed by the line weavings as well as the block diagonal structure of the crossing matrix. The first 8 rows and columns in the matrix represent lines of $\alpha$-helices; the last 8 rows and columns represent lines of $\beta$-strands. Internal to each of these two sets of lines, the crossings are primarily positive. Crossings across sets, that is, between an $\alpha$-line and a $\beta$-line, are primarily negative. The reason for this is the symmetry of the TIM-barrel and the fact that the backbones of $\alpha$-helices and $\beta$-strands are oriented oppositely relative to the barrel axis, as inspection of the proteins shows.

### 7.2.3 The $L2$ Measure

In ranking and extending alignments, the code considers various error measures, including the length of the alignment and an $L2$ *measure of line embeddings*, which we now explain. While weavings are constructed from infinite lines, the $L2$ measure is based on finite line segments that represent the protein's secondary structure elements. A finite line segment is a straight-line embedding of the unit interval $[0, 1]$ into 3D space. Given two oriented line segments $h : [0, 1] \to \Re^3$ and $k : [0, 1] \to \Re^3$, a standard *least-squares* metric for measuring their similarity is:

$$L_2(h, k) = \left( \int_0^1 ||h(t) - k(t)||^2 \, dt \right)^{1/2}.$$

Given a collection of line segments $\{h_1, \ldots, h_n\}$ in one protein, paired with a corresponding collection of line segments $\{k_1, \ldots, k_n\}$ in a second protein, one can measure the goodness of the alignment as follows:

$$L2 = \sqrt{\frac{\sum_{i=1}^{n} (L_2(h_i, k_i))^2}{n}}.$$

The value "$L2$" thus obtained appears in Table 1. It is an analogue for oriented line-alignments of the RMSD measure often used for atom-alignments.

### 7.2.4 Similarity and Rank

In Table 1, the value $\rho_{12}$ provides yet another measure of how well one protein (Prot1) may be aligned with a second protein (Prot2). The "Rank" column of Table 1 refers to a ranking by $\rho_{12}$ value. The value lies in the range $[0, 1]$, with 1 optimal. It combines three different measurements, namely the number of aligned secondary structure elements, the $L2$ measure, and the crossing consistency, as follows:

$$\rho_{12} = \left( s_1^4 + s_2^4 + s_3^4 \right)^{-1/4}$$

Here $s_1$ is the ratio of secondary structure elements in Prot1 to the number of elements appearing in the optimal alignment, $s_2$ is $L2/(4\text{Å})$, and $s_3$ is $10{*}$`Bad/Sig`. When combining multiple measures, small exponents reduce the significance of any one deviation, while large exponents increase the significance; we use exponent 4 to amplify any deviations above 1 in the values $\{s_1, s_2, s_3\}$. Thus the divisor 4Å in $s_2$ simply asserts that deviations below 4Å are not terribly significant; similarly the multiplier 10 in $s_3$ asserts that crossing errors exceeding 10% are significant. We picked these numbers without any tuning, based simply on intuition developed in observing protein alignments. Likely other values would be equally good or better.

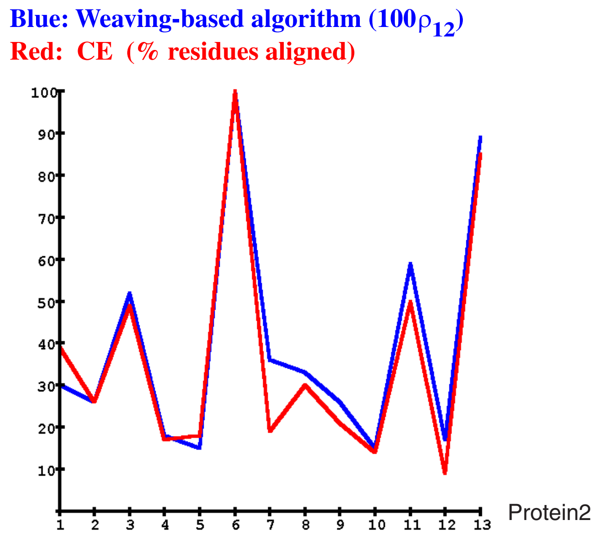
The precise value of $\rho_{12}$ is not significant; we caution against reading too much into its absolute value. Instead, it is a rough qualitative dimensionless number for assessing how well

| Weaving-based 100 $\rho_{12}$ | | Protein 2 | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Protein 1 | \|SSE\| | 5at1_A | 3adk | 1xis | 1fpk_A | 1a6m | 1pbg_A | 1ki7_A | 1hyq_A | 1atn_A | 1d9c_A | 1wsy_A | 1mjc | Table 1 |
| 5at1_A | 22 | 100 | 32 | 27 | 32 | 18 | 36 | 32 | 39 | 24 | 18 | 49 | 18 | 100 |
| 3adk | 15 | 47 | 100 | 27 | 46 | 27 | 32 | 75 | 66 | 33 | 27 | 44 | 26 | 73 |
| 1xis | 20 | 30 | 20 | 100 | 30 | 20 | 70 | 34 | 50 | 35 | 20 | 69 | 22 | 65 |
| 1fpk_A | 19 | 36 | 32 | 31 | 100 | 26 | 31 | 42 | 32 | 29 | 21 | 32 | 26 | 100 |
| 1a6m | 7 | 57 | 57 | 57 | 71 | 100 | 57 | 57 | 57 | 57 | 59 | 57 | 0 | 100 |
| 1pbg_A | 27 | 30 | 26 | 52 | 18 | 15 | 100 | 36 | 33 | 26 | 15 | 59 | 17 | 89 |
| 1ki7_A | 16 | 44 | 68 | 42 | 31 | 25 | 52 | 100 | 56 | 43 | 25 | 37 | 24 | 100 |
| 1hyq_A | 19 | 44 | 53 | 52 | 32 | 21 | 47 | 47 | 100 | 38 | 21 | 47 | 23 | 74 |
| 1atn_A | 28 | 23 | 25 | 14 | 23 | 14 | 18 | 25 | 28 | 100 | 17 | 31 | 18 | 71 |
| 1d9c_A | 7 | 57 | 57 | 57 | 57 | 57 | 57 | 57 | 57 | 57 | 100 | 57 | 0 | 77 |
| 1wsy_A | 19 | 52 | 36 | 73 | 33 | 25 | 83 | 32 | 47 | 42 | 21 | 100 | 23 | 83 |
| 1mjc | 5 | 78 | 30 | 69 | 97 | 0 | 79 | 39 | 73 | 79 | 0 | 39 | 100 | 99 |

Table 2: Weaving-based similarities for cross comparisons of 12 proteins with each other. The table depicts $100\rho_{12}$, producing values in the range $[0, 100]$, with 100 optimal. For otherwise good alignments, $\rho_{12}$ is roughly the fraction of Protein1's secondary structure elements that have been aligned. For reference, the column labeled "Table 1" refers to the nominally-correct comparison of Protein1 with its counterpart in Table 1.

| CE % aligned | | Protein 2 | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Protein 1 | size | 5at1_A | 3adk | 1xis | 1fpk_A | 1a6m | 1pbg_A | 1ki7_A | 1hyq_A | 1atn_A | 1d9c_A | 1wsy_A | 1mjc | Table 1 |
| 5at1_A | 310 | 100 | 29 | 33 | 28 | 26 | 59 | 23 | 32 | 26 | 21 | 31 | 15 | 95 |
| 3adk | 195 | 46 | 100 | 57 | 37 | 31 | 64 | 49 | 57 | 37 | 31 | 45 | 21 | 81 |
| 1xis | 387 | 26 | 29 | 100 | 23 | 17 | 60 | 23 | 40 | 25 | 19 | 55 | 10 | 55 |
| 1fpk_A | 335 | 26 | 21 | 26 | 100 | 19 | 24 | 26 | 27 | 19 | 21 | 24 | 17 | 98 |
| 1a6m | 151 | 53 | 40 | 44 | 42 | 100 | 56 | 48 | 48 | 58 | 50 | 42 | 21 | 100 |
| 1pbg_A | 468 | 39 | 26 | 49 | 17 | 18 | 100 | 19 | 30 | 21 | 14 | 50 | 9 | 85 |
| 1ki7_A | 331 | 22 | 29 | 27 | 27 | 22 | 27 | 100 | 29 | 27 | 22 | 19 | 12 | 99 |
| 1hyq_A | 233 | 43 | 48 | 66 | 39 | 31 | 61 | 41 | 100 | 38 | 31 | 42 | 21 | 97 |
| 1atn_A | 373 | 21 | 19 | 26 | 17 | 24 | 26 | 24 | 24 | 100 | 19 | 19 | 11 | 77 |
| 1d9c_A | 121 | 53 | 50 | 60 | 60 | 63 | 53 | 60 | 60 | 60 | 100 | 66 | 33 | N/A |
| 2wsy_A | 268 | 36 | 33 | 79 | 30 | 24 | 87 | 24 | 37 | 27 | 30 | 100 | 15 | 85 |
| 1mjc | 69 | 70 | 58 | 58 | 81 | 46 | 58 | 58 | 70 | 58 | 58 | 58 | 100 | 91 |

Table 3: CE alignments. The table depicts the percentage of Protein1's residues aligned by CE with each of the other proteins. "size" is the number of residues considered by CE in Protein1.




**Comparison of 1pbg_A with the proteins of Tables 2 & 3.**

Figure 15: Overlay of $\rho_{12}$ and residue percentages for alignments of 1pbg_A with the proteins listed in Tables 2 and 3. Blue shows the weaving-based similarity, while red shows the percentage of residues aligned by CE. The data is taken from row #6 in each of Tables 2 and 3. The axis "Protein2" uses integers; these refer to the columns in which each protein appears in the tables.

the lines of Prot1 may be placed onto the lines of Prot2. The number is purposefully not symmetric. For instance, a small protein domain might appear as a subdomain of another protein. This would mean that $\rho_{12}$ might be near 1, while the opposite comparison $\rho_{21}$ might be considerably less than 1. For good alignments in which the $L2$ and `Bad/Sig` values are low, $\rho_{12}$ effectively measures the fraction of secondary structure elements in Prot1 that have been aligned.

Table 2 depicts the similarity values for all possible comparisons between 12 of the 24 proteins from Table 1, using the version of the weaving-based alignment algorithm that favors preserving backbone order. The values of $100\rho_{12}$ in a single row give a rough relative comparison of how well one protein matches all the others. For some proteins a clear baseline is apparent, indicating an alignment of four secondary structure elements, the minimum possible using quadruples. For instance, the value 57 appears frequently in the row for 1a6m, indicating an alignment of 4 of the 7 possible secondary structure elements.

We ran the same comparisons using CE, obtaining qualitatively similar results. Table 3 depicts the results, showing for each protein the percentage of its residues aligned with the other proteins. This data roughly mirrors the data of Table 2. For instance, Figure 15 graphs the alignment data for 1pbg_A, a large protein for which the two methods agree quite well. As a reminder, $\rho_{12}$ values are based only on secondary structure alignments, whereas the CE-derived percentages are based on all residues.

### 7.2.5 Protein Descriptions

In selecting proteins we examined SCOP, focusing primarily on classes "a+b" and "a/b", plus a few others. We chose proteins with a range of sequence similarities. Two of the protein pairs, (1a6m, 1lhs) and (1d9c_A, 2rus), are all-helical. One protein (1mjc) is all-sheet. The others contain a mix of $\alpha$-helices and $\beta$-sheets. Here is a brief description of all the proteins appearing in Tables 1 and 2:

**(5at1_A, 8atc_A)** The taut and relaxed conformations, respectively, of the catalytic chain A in Aspartate Carbamoyltransferase, a protein involved in the synthesis of pyrimidine nucleotides. Chain A has 310 residues, forming two domains, each consisting of a $\beta$-sheet and several $\alpha$-helices. The two domains are joined at a hinge.

**(3adk, 1gky)** These are transferases catalyzing two ATP-dependent phosphorylations. 3adk consists of 194 residues, forming one $\beta$-sheet and several $\alpha$-helices. 1gky consists of 187 residues, forming two $\beta$-sheets and somewhat fewer $\alpha$-helices than 3adk.

**(1xis, 1nar)** These are two TIM-barrels. Xylose Isomerase (1xis, 387 residues, now including its tail) catalyzes the conversion of glucose into fructose. Narbonin (1nar, 290 residues) is a plant seed protein with no known enzymatic function.

**(1fpk_A, 1fpk_B)** These are chains A and B of the dimer Fructose-1,6-Bisphosphatase, a hydrolase involved in gluconeogenesis in the liver. Each chain consists of 335 residues, forming three $\beta$-sheets and several $\alpha$-helices.

**(1a6m, 1lhs)** 1a6m (151 residues) is myoglobin from the sperm whale, while 1lhs (153 residues) is myoglobin from the loggerhead sea turtle. These proteins consist purely of $\alpha$-helices.

**(1pbg_A, 1gow_A)** These are TIM-barrel hydrolases. Chain A of 6-Phospho-Beta-D-Galactosidase (1pbg) consists of 468 residues, forming three $\beta$-sheets and many $\alpha$-helices. Chain A of Beta Glycosidase (1gow) consists of 489 residues, forming five $\beta$-sheets and many $\alpha$-helices.

**(1ki7_A, 1qhi_A)** These are two different complexes of thymidine kinase from the herpes simplex virus. Thymidine Kinase is a phosphotransferase. Chain A consists of 374 residues (329 of which are represented in the PDB file) forming one $\beta$-sheet and numerous $\alpha$-helices.

**(1hyq_A, 1cp2_A)** 1hyq is a bacterial cell-division regulator (minD). Chain A consists of 263 residues (233 of which are represented in the PDB file), forming one large $\beta$-sheet surrounded by several $\alpha$-helices. 1cp2 is a nitrogenase iron protein. Chain A consists of 269 residues, again forming a large $\beta$-sheet surrounded by $\alpha$-helices.

**(1atn_A, 3hsc)** 1atn is Actin from rabbit, while 3hsc is Heat Shock Cognate 70 from cow. Chain A of 1atn consists of 372 residues, forming five $\beta$-sheets and several $\alpha$-helices. 3hsc consists of 384 residues, forming five $\beta$-sheets and several $\alpha$-helices. The two proteins share a common ATPase domain [26].

**(1d9c_A, 2rig)** 1d9c is Interferon-Gamma from cow, while 2rig is Interferon-Gamma from rabbit. Chain A of 1d9c consists of 121 residues, while 2rig consists of 119 residues, in both cases forming an all-helical protein.

**(1wsy_A, 2rus_A)** Chain A of Tryptophan Synthase (1wsy) consists of 265 residues, forming a TIM-barrel in the Tryptophan family. Chain A of (2rus) consists of 457 residues, forming a TIM-barrel in the RuBisCo family; there are several additional $\beta$-strands and there is additional domain structure outside the common barrel motif.

**(1mjc, 1a62)** 1mjc is Cold Shock from E. coli with 70 residues. It is an all $\beta$ protein, with a six-strand barrel fold. 1a62 is the RNA binding domain of E. coli rho factor (often called Rho130). It consists of two subdomains, an amino terminal helical region and a $\beta$-barrel carboxy terminal domain. The $\beta$-barrel domain binds either ssDNA or RNA and is structurally homologous to the oligonucleotide-saccharide binding domain. Thus, 1mjc should appear homologous to the $\beta$ domain of 1a62.

# 8 Future Work

**Topology:** We suspect that considerable additional information may be gained by focusing more closely on the pure topology of proteins. For instance, a purely topological hashing scheme would look more closely at the crossing matrix. In the case of quadruples, the following two numbers fully characterize isotopy classes of four oriented pairwise-skew lines: (i) The sum of all the triple linking numbers in the weaving, and (ii) the cardinality of the positive (or negative) crossing numbers. (A *triple linking number* is the product of the three crossing numbers defined by some triple of lines. Invariant (i) is the sum of all such products for all possible triples of lines in a given weaving. See [60].) As mentioned, we currently hash on invariant (ii), expanded to accommodate coplanar lines, then compare permutations of crossing matrices to select topologically equivalent weavings. We subsequently discard line alignments that do not make physical sense. Our approach currently therefore is not purely topological, but takes rough account of the line embeddings. Future research should explore further in both directions: the more topological direction as well as the geometrically more specific direction. As we have indicated, for large numbers of lines, many of the topological problems are wide open.

**Sheets:** This paper approximated both $\alpha$-helices and $\beta$-strands using lines, then developed an algorithm for matching lines and line weavings. While successful, we suspect that such an approach only captures part of the structure contained in $\beta$-sheets. Such sheets have both a two-dimensional surface structure and a component one-dimensional line structure. In other contexts, such as our work on detecting protein similarities from sparse NMR data [19], we have discovered and used very natural two-dimensional polytope structures for representing $\beta$-sheets, based on hydrogen-bonding. We suspect that higher-dimensional generalization of line weavings to surface foliations may provide additional useful topological information. Such generalizations may prove particularly useful in contexts where $\beta$-strands are only poorly approximated by straight lines, due to twisting and bending.

**General Loops:** Our topological approach currently is limited in a practical sense to secondary structure elements. Section 5 outlined the theoretical foundation of a general approach able to deal with arbitrary polypeptide segments, not just those defining $\alpha$-helices and $\beta$-strands. Future work needs to extend the current results in that direction. The driving goal should be to derive compact topological descriptors characteristic of protein shapes and to circumscribe the hypervolume of potential topological shapes actually inhabited by proteins.

# 9 Summary

This paper introduced the notion of isotopy deformations into structural alignment. The paper explored the relationship between writhing and self-convolution. Self-convolution compactly describes 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. The paper also explored line weavings as a means for characterizing protein structures by arrangements of $\alpha$-helices and $\beta$-strands. The paper presented an algorithm for matching proteins based on line weaving topology. Future work should extend these knot theoretic ideas to include surface representations and general loops, then classify protein shapes topologically.

# 10 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 six years.

# References

[1] Agarwal, P., Edelsbrunner, H., and Wang, Y. Computing the writhing number of a polygonal knot. *Proc SODA 13* (2002), 791–799.

[2] Agol, I. Research blog 3/18/03. See `http://www.math.uic.edu/~agol/blog/030318.pdf` and `http://www.math.uic.edu/~agol/coNP/coNP01.html`, Mar. 2003.

[3] Agol, I., Hass, J., and Thurston, W. 3-manifold knot genus is NP-complete. *Proceedings of the Thirty-Fourth Annual ACM Symposium on Theory of Computing, Montreal, Quebec, May 19–22* (2002), 761–766.

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

[5] Alt, H., Knauer, C., Rote, G., and Whitesides, S. On the complexity of the linkage reconfiguration problem. Tech. Rep. B 03-02, Freie Universität Berlin, Jan. 2003.

[6] Bachar, O., Fischer, D., Nussinov, R., and Wolfson, H. A computer vision based technique for 3-D sequence-independent structural comparison of proteins. *Protein Engineering 6(3)* (1993), 279–288.

[7] Bacher, R., and Garber, D. Spindle configurations of skew lines. Tech. Rep. math.GT/0205245, arXiv.org, May 2002.

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

[9] Bar-Natan, D. On the Vassiliev knot invariants. *Topology 34* (1995), 423–472.

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

[11] Birman, J., Feb. 2004. Private communication.

[12] Birman, J., and Menasco, W. Studying links via closed braids. V: The unlink. *Transactions of the American Mathematical Society 329(2)* (1992), 585–606.

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

[14] Canny, J. Some algebraic and geometric computations in PSPACE. *Proceedings ACM Symposium Theory of Computing (STOC) 20* (1988), 460–467.

[15] Canny, J. Computing roadmaps of general semi-algebraic sets. *Computer Journal 36(5)* (1993), 504–514.

[16] Chazelle, B., Edelsbrunner, H., Guibas, L., Sharir, M., and Stolfi, J. Lines in space: Combinatorics and algorithms. *Algorithmica 15* (1996), 428–447.

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

[18] Călugăreanu, G. Sur les classes d'isotopie des noeuds tridimensionnels et leurs invariants. *Czech Mathematics Journal 11* (1961), 588–625.

[19] Erdmann, M., and Rule, G. Rapid protein structure detection and assignment using residual dipolar couplings. Tech. Rep. CMU-CS-02-195, CMU, `http://reports-archive.adm.cs.cmu.edu/anon/2002/abstracts/02-195.html`, 2002.

[20] Fuller, F. The writhing number of a space curve. *Proc Natl Acad Sci USA 68* (1971), 815–819.

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

[22] Gibrat, J.-F., Madej, T., and Bryant, S. Surprising similarities in structure comparison. *Current Op in Struct Biology 6* (1996), 377–385.

[23] Grindley, H., Artymiuk, P., Rice, D., and Willett, P. Identification of tertiary structure resemblance in proteins using a maximal common subgraph isomorphism algorithm. *Journal of Molecular Biology 229* (1993), 707–721.

[24] Hass, J., Lagarias, J., and Pippenger, N. The computational complexity of knot and link problems. Tech. Rep. math.GT/9807016, arXiv.org, July 1998.

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

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

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

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

[29] Kaufmann, L. *Knots*. `http://www2.math.uic.edu/~kauffman/Tots/Knots.htm`.

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

[31] Langmead, C., Yan, A., Wang, L., Lilien, R., and Donald, B. R. A polynomial time nuclear vector replacement algorithm for automated NMR resonance assignments. *Proc Seventh RECOMB* (2003).

[32] Latombe, J.-C. *Robot Motion Planning.* Kluwer Academic Publishers, Boston, Massachusetts, 1991.

[33] Lozano-Pérez, T. Automatic planning of manipulator transfer movements. *IEEE Trans Sys Man Cyber 11(10)* (1981), 681–698.

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

[35] McKenna, M., and O'Rourke, J. Arrangements of lines in 3-space: A data structure with applications. *Proceedings of the Fourth ACM Symposium on Computational Geometry, Urbana-Champaign, Illinois, June 6–8* (1988), 371–380.

[36] Moskovich, D. Framing and the self-linking integral. Tech. Rep. math.QA/0211223, arXiv.org, Nov. 2002.

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

[38] Nussinov, R., and Wolfson, H. Efficient detection of three-dimensional structural motifs in biological macromolecules by computer vision techniques. *Proc Natl Acad Sci USA 88(23)* (1991), 10495–9.

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

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

[41] Penne, R. Moves on pseudoline diagrams. *European Journal of Combinatorics 17* (1996), 569–593.

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

[43] Phillips, J., Ladd, A., and Kavraki, L. Simulated knot tying. *Proc IEEE Intl Conf Robotics and Automation* (2002), 841–846.

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

[45] Reif, J. The complexity of the Mover's Problem and generalizations. *Proceedings ACM Symposium Foundations of Computer Science (FOCS) 20* (1979).

[46] Renegar, J. On the computational complexity and geometry of the first-order theory of the reals, I–III. *SIAM Journal of Symbolic Computation 13(3)* (1992), 255–352.

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

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

[49] Rossetto, V., and Maggs, A. Writhing geometry of open DNA. *Journal of Chemical Physics 118* (2003), 9864–9874.

[50] Sayle, R. *RasMol.* `http://www.umass.edu/microbio/rasmol/index2.htm`. A protein visualization tool.

[51] Schwartz, J., and Sharir, M. On the Piano Movers' Problem: II. General techniques for computing topological properties of real algebraic manifolds. *Advances in Applied Mathematics 4* (1983), 298–351.

[52] Shindyalov, I., and Bourne, P. *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`.

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

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

[55] Singh, A., and Brutlag, D. Hierarchical protein structure superposition using both secondary structure and atomic representations. *Proc Intel Sys Mol Bio 5* (1997), 284–293.

[56] Stryer, L. *Biochemistry*, fourth ed. W.H. Freeman, New York, 1995.

[57] Taylor, W. A 'periodic table' for protein structures. *Nature 416* (11 Apr 2002), 657–660.

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

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

[60] Viro, J., and Viro, O. Configuration of skew lines. Tech. Rep. (this is an easy introduction based on the journal article [61]), Uppsala University, Sweden, 2000.

[61] Viro, O., and Drobotukhina, Y. Configuration of skew lines. *Leningrad Mathematics Journal 1(4)* (1990), 1027–1050.

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

# Appendix A Proof of Theorem 1

**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 $E_\delta$ such that $p$ is $(E_\delta, \delta)$-isotopic 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 us walk through a detailed argument: The net $\{E_\delta\}$ sits inside a compact subspace of $SE(3)$, so it contains a subnet $\mathcal{E}$ that converges to some $E \in SE(3)$. This means that for any positive $\epsilon$, we can pick a $\delta_\epsilon$ 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 $\delta$ with $0 < \delta < \delta_\epsilon$ and $E_\delta \in \mathcal{E}$. 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 $(E, \delta)$-isotopic to $q$ then $q$ is $(E^{-1}, \delta)$-isotopic to $p$. Suppose $h$ is an isotopy satisfying Definition 2 establishing that $p$ is $(E, \delta)$-isotopic 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 an isotopy establishing that $q$ is $(E^{-1}, \delta)$-isotopic 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) For all $t \in I$ and all $i = 1, \ldots, n$:

- $||E^{-1}(q_i) - g_i(t)|| = ||E^{-1}(q_i) - E^{-1}(h_i(1-t))||$
  $= ||q_i - h_i(1-t)|| \leq \delta.$
- $||p_i - g_i(t)|| = ||p_i - E^{-1}(h_i(1-t))||$
  $= ||E(p_i) - h_i(1-t)|| \leq \delta.$

(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 $m = n = k$. We then need to show how isotopies $p \to r$ and $r \to q$ imply an isotopy $p \to q$. Let $\epsilon$

be an arbitrary positive number, and pick Euclidean motions $E, F \in SE(3)$ and isotopies $e$, $f$ establishing that $p$ is $(E, d(p,r)+\epsilon/2)$-isotopic to $r$ and that $r$ is $(F, d(r,q)+\epsilon/2)$-isotopic 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 us 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 $(H, \delta)$-isotopic to $q$ with $\delta = d(p,r) + d(r,q) + \epsilon$. Since $\epsilon$ 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 $\delta$, one centered at its start, the other at its goal. Robots may not collide. Moreover, edges drawn between two different pairs of successively indexed robots 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, [13, 14, 15, 51, 32, 46, 24]. We thus have a procedure for deciding whether $p$ is $(E, \delta)$-isotopic 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 iteratively isolate the value $d(p, q)$ as accurately as we desire. □

# Appendix B  Proof of Lemma 2

**Lemma 2:** Suppose $p$ and $q$ are two polygonal curves with equal numbers of points, each monotonic with respect to some line. Let $\pi = \{\pi_1, \dots, \pi_n\}$ and $\sigma = \{\sigma_1, \dots, \sigma_n\}$ 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. Since $d$ is a metric, the triangle inequality says that

$$d(p, q) \leq d(p, \pi) + d(\pi, \sigma) + d(q, \sigma).$$

Now let $E$ be a rigid body motion that aligns the oriented line containing $\pi$ with the oriented line containing $\sigma$. Define a linear interpolation that moves $E(\pi)$ to $\sigma$, by moving $E(\pi_i)$ to $\sigma_i$ uniformly in time for each $i = 1, \dots, n$. We claim that this linear interpolation is an isotopy. As a consequence, $d(\pi, \sigma) \leq \max_i ||\sigma_i - E(\pi_i)||$. Taking infima over all such orientation-preserving $E$ establishes the lemma.

To see that the linear interpolation of $E(\pi)$ to $\sigma$ is an isotopy 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 C  Proof of Lemma 4

**Lemma 4:** Consider four oriented 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 D  Proof of Corollary 4

**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 $\delta$, then there is a rigid body transformation $E$ such that $e_1$ and $e_2$ are $(E, \delta)$-isotopic to $f_1$ and $f_2$, respectively.

PROOF. We will use similar notation 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 $\Delta_{ij}$ 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 an isotopy between two directed edges (viewed as polygonal curves) that morphs one edge into the other. The previous inequalities show that we can in fact construct $(E, \delta)$-isotopies between $e_i$ and $f_i$, for $i = 1, 2$. $\square$