Structural Alignment Using the Generalized Euclidean Distance Between Conformations

ALI R. MOHAZAB, STEVEN S. PLOTKIN

Department of Physics and Astronomy, University of British Columbia, Vancouver, BC, Canada

Received 17 January 2009; accepted 1 May 2009
Published online 6 July 2009 in Wiley InterScience (www.interscience.wiley.com).
DOI 10.1002/qua.22347

ABSTRACT: The usual Euclidean distance may be generalized to extended objects such as polymers or membranes. Here, this distance is used for the first time as a cost function to align structures. We examined the alignment of extended strands to idealized beta-hairpins of various sizes using several cost functions, including RMSD, MRSD, and the minimal distance. We find that using minimal distance as a cost function typically results in an aligned structure that is globally different than that given by an RMSD-based alignment. © 2009 Wiley Periodicals, Inc. Int J Quantum Chem 109: 3217–3228, 2009

Key words: protein folding; structural alignment; RMSD; MRSD; generalized distance; minimum distance; reaction coordinate; order parameter; optimization

Introduction

In a series of experiments starting in the late 1950s and culminating in a 1961 paper in the Proceedings of the National Academy of Sciences [1], Anfinsen and his colleagues showed that a protein such as bovine pancreatic ribonuclease would, under oxidizing conditions, undergo slow but spontaneous reshuffling of disulfide bonds from a state with initially random cross-linked pairs, to a state with correct disulfide pairing and full enzymatic activity. The spontaneous formation of correct disulfide pairs indicated that the amino acid sequence itself was guiding the process toward more thermodynamically favorable configurations, and the so-called thermodynamic hypothesis in protein folding was born.

This discovery underpinned the formalism that developed decades later to understand protein folding as a configurational diffusion process on an energy landscape that through molecular evolution had the overall topography of a rugged funnel [2–9]. The initial random crosslinkings and subsequent slow exchange of disulfide bonds observed by Anfinsen and coworkers argued against a mechanistic pathway picture, but there was nevertheless a lag phase before the energy landscape picture eventually took hold.

Correspondence to: S. S. Plotkin; e-mail: steve@phas.ubc.ca

Although important as a conceptual tool, real predictive power was brought to bear by quantifying the funnel notion to generate free energy surfaces as a function of a progress coordinate that measured the degree to which a protein was folded [10–12]. Soon thereafter, questions arose regarding what coordinate(s) best represented folding progress, or whether one could even find a simple geometric coordinate that would represent kinetically how folded a protein was. The kinetic proximity of a given configuration was quantified unambiguously as the probability a protein would fold first before unfolding, given that it was initially in that given configuration [13]. This idea had earlier analogues in the Brownian analysis of escape and recombination probabilities of an ionized electron [14].

Order Parameters in Protein Folding

The study of various order parameters that might best represent progress in the folding reaction has generated much interest [13, 15–29], with questions focusing on what parameter(s) or principle component-like motions might best correlate with splitting probability or probability of folding before unfolding.

On the other hand, analyses using intuitive geometric order parameters have been developed to understand folding and are now commonly used. These include the fraction of native contacts Q [21, 23, 30–34], which can be locally or globally defined, root mean square deviation (RMSD) between structures [35–38], structural overlap parameter χ [39–41], Debye-Waller factors [42, 43], or fraction of correct dihedral angles [34].

To find a simple geometrical order parameter that quantifies progress to the folded structure poses several challenges. These include an accurate account of the effects of polymer noncrossing [44], energetic and entropic heterogeneity in native driving forces [31, 45, 46], as well as non-native frustration and trapping [47–49]. Fortunately, it has been borne out experimentally that wild type proteins are sufficiently minimally frustrated that non-native interactions do not play a strong role in either folding rate or mechanism, and native structure-based models for folding rates and mechanisms have enjoyed considerable success [50–55].

In condensed matter systems, useful order parameters have historically had intuitive geometrical interpretations. Their definition did not require the knowledge of a particular Hamiltonian (although their temperature-dependence and time-evolution were affected by the energy function in the system). In chemical reactions, the distance between constituents in reactant and product has played a ubiquitous role in the construction of potential energy surfaces [56]. Moreover, from the point of view of stochastic escape and recombination, the distance perfectly correlates with the commitment probability for a freely diffusing particle between two absorbing boundaries.

Distance as an Order Parameter

The distance is easy to define for a point particle, which we imagine to travel between two locations A at r_A and B at r_B. It is the variational minimum of the functional:

$$\int_{\mathbf{r}_A}^{\mathbf{r}_B} ds = \int_0^T dt\,\sqrt{\dot{\mathbf{r}}^2} \qquad (1)$$

where $\dot{\mathbf{r}} \equiv d\mathbf{r}/dt$, and the initial and final conditions, or equivalently boundary conditions, are $\mathbf{r}(0) = \mathbf{r}_A$ and $\mathbf{r}(T) = r_B$.

However, until recently [44, 57, 58], the distance had not been formulated for higher dimensional objects such as pairs of polymer configurations, despite close parallels in string theory [59].

In this article, after briefly reviewing two common reaction coordinates, Q and RMSD, and the two newer ones introduced and explored in [44, 57, 58], D and mean root squared distance (MRSD), we will further explore structural alignments based on D for idealized hairpins.

Some Problems With Commonly Used Reaction Coordinates

Many reaction coordinates have been used to describe the folding process, while still being flawed in principle. These characterizations have been largely successful because the majority of conformations during folding are well characterized by changes in these parameters: Proteins undergo some collapse concurrently with folding, lower their internal energy, and adopt structures geometrically similar to the native structure.

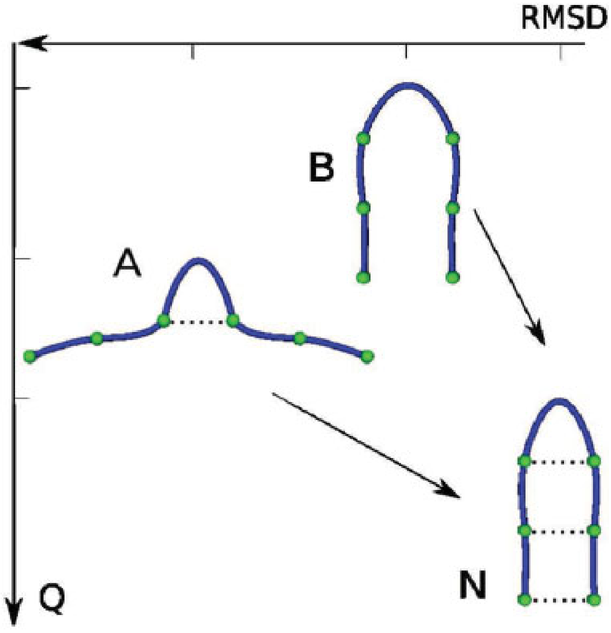


FIGURE 1. Order parameters do not always correlate with kinetic proximity. Structure *A* above is more native-like according to the fraction of native contacts, while structure *B* is more native-like according to RMSD, and is also closer kinetically to the native structure. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

Nevertheless, it is easy to point to simple examples of conformational transitions for which the adoption of native structure does not correlate with the change in commonly used order parameters. Although these conformational pairs may not be wholly representative of the total folding process, they point to situations where folding to a given structure would not be well-characterized by commonly used order parameters.

Figure 1 shows two structures *A* and *B* with different measures of structural similarity to a "native" hairpin fragment *N*. These structures have different measures of proximity depending on the coordinate used to characterize them. If we use the fraction of native contacts *Q* to describe native proximity[1], structure *A* has a *Q* of $Q_A = 1/3$ while $Q_B = 0$, so by this measure *A* is more native. If weuse the root mean square deviation RMSD[2], structure *B* is more native-like than *A*. Moreover, structure *B* would have a higher probability of folding before unfolding than *A*, that is, it has a larger value of p_{FOLD} [13], and so is closer kinetically to the native structure. The longer the hairpin, the more likely a slightly expanded structure is to fold, so the discrepancy between *Q* and RMSD for these pairs of structures becomes even larger.

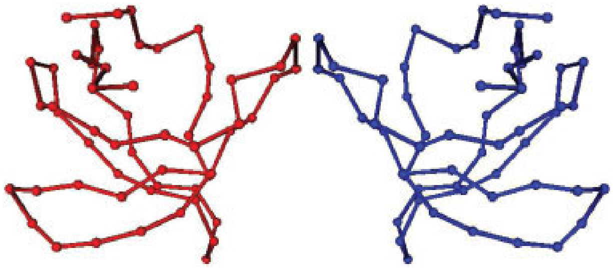


FIGURE 2. Native structure of SH3 (right) and its mirror image. Although dissimilar by RMSD, biologically nonfunctional, and disallowed by true dihedral potentials, this structure has a *Q* = 1, because native contacts remain intact after mirroring transformations. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

In contrast to RMSD, *Q* also does not distinguish between chiralities. Typically, the energy function forbids opposite chiralities; however, if the appropriate chirality is not enforced in the backbone dihedral potentials, mirror-image structures as in Figure 2 will be allowable, and are indistinguishable according to *Q* [58].

Although the RMSD is often characterized as a "distance" between structures, it is not equivalent or even proportional to the sum of the straight-line distances between the atoms or residues in the two structures (Fig. 3). This quantity is in fact given by the mean root squared distance (MRSD), defined for two structures *A* and *B* as:

$$\frac{1}{N}\sum_{n=1}^{N}|\mathbf{r}_{A_n}-\mathbf{r}_{B_n}| = \frac{1}{N}\sum_{n=1}^{N}\sqrt{(\mathbf{r}_{A_n}-\mathbf{r}_{B_n})^2} \qquad (2)$$

The RMSD between two structures is always greater than or equal to the MRSD between the same structures, with MRSD = RMSD in only the most trivial cases [58]. The RMSD is also less robust

[1] $Q_{AN} \equiv (\sum_{i<j}\Delta^A_{ij}\Delta^N_{ij})/(\sum_{i<j}\Delta^N_{ij})$ counts pairs of residues within some cut-off distance in both structure *A* and structure *N*. This result is then normalized by the number of contacts in the native structure.

[2] $RMSD \equiv \sqrt{N^{-1}\sum_{i=1}^{N}(\mathbf{r}_{A_i}-\mathbf{r}_{B_i})^2}$ is a least-squares measure of similarity between structures *A* and *B*. Typically, this quantity is minimized given two structures and so can be thought of as a "least squares fit." The sum may be over all atoms, or simply all residues in coarse-grained models.

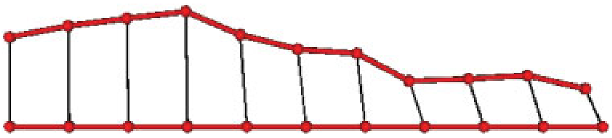


FIGURE 3. The MRSD is the average length of the black like segments between corresponding residues of the initial and final configuration. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

to large fluctuations of select residues in structural pairs [44].

MRSD has a simple intuitive physical meaning—the MRSD between two structures gives the average distance each residue in one structure would have to travel on a straight line to get to its counterpart in the other structure (Fig. 3).

Polymer Noncrossing in Protein Folding

The above interpretation of MRSD points to a shortcoming of both MRSD and RMSD, which is the importance of chain noncrossing constraints. Consider the two curves depicted in Figure 4, which differ by having opposite sense of underpass/overpass. When both curves are aligned by minimizing MRSD or RMSD, the respective values are almost zero. However, the physically relevant distance for one conformation to transform to the other is much larger, and must involve one arm of the backbone circumventing the other as it moves between conformations. The transformation that minimizes the distance has been shown previously to involve motions wherein one end of the polymer doubles back upon itself until it reaches the underpass/overpass, where it appropriately crosses under/over it, and then proceeds snake-like to extend itself to the final position [44, 57]. We will not deal further with the aspects of noncrossing in this article, other than some description around equation (12) below.

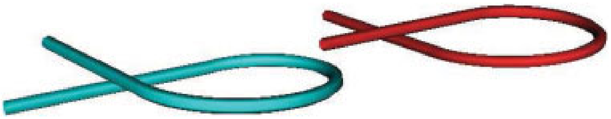


FIGURE 4. The MRSD and RMSD between the two curves are close to zero (the curves in this figure are displaced for better viewing but should be imagined to be superposed). However, because the curve cannot pass through itself, to undergo the transformation, one leg must undergo relatively large amplitude motions to travel from one conformation to another. Alternatively the loop must untwist and re-twist. This results in a nonzero distance between the conformations by accurate metrics, which can account for noncrossing. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

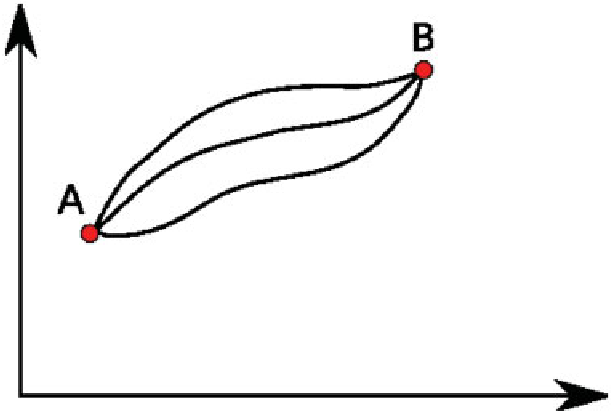


FIGURE 5. Distance between the two points *A* and *B* is the minimum length of the curve connecting the two points. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

The Generalized Distance *D*

The distance between two points can be cast as a variational problem, where the arclength of the curve between two points is minimized [Eq. (1), see Fig. 5]. The resultant Euler-Lagrange equations for the distance between two points are:

$$\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{\mathbf{r}}}\right) = 0 \qquad (3)$$

or

$$\dot{\mathbf{v}} = 0$$

which implies straight line motion, because this means that the direction of the velocity does not change.

As mentioned in the introduction, the notion of distance between two points can be generalized to two curves or higher-dimensional objects in general

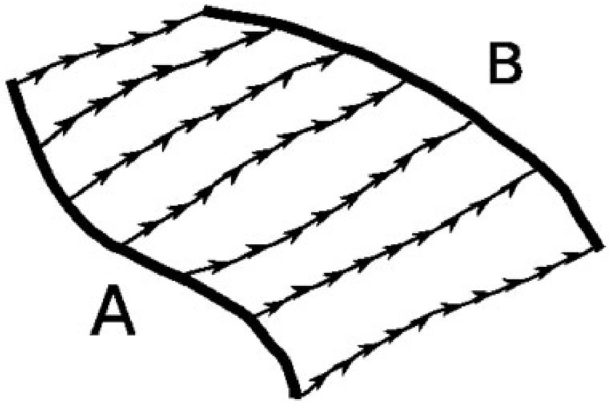


FIGURE 6. The distance D_{AB} is the accumulation of how much every part of the contour defining the space curve moves in the transformation between two conformations *A* and *B*.

[57]. As in the case of points, the distance between two curves can be thought as a variational problem, where one now minimizes the cumulative integrated arclength between the two space curves:

$$D[\dot{\mathbf{r}}] = \int_0^L ds \int_0^T dt \sqrt{\dot{\mathbf{r}}^2}. \qquad (4)$$

Here, $\mathbf{r} \equiv \mathbf{r}(s,t) = (x(s,t), y(s,t), z(s,t))$ and $\dot{\mathbf{r}} \equiv \partial\mathbf{r}/\partial t$. The independent variables in this formulation are: position along the contour of the polymer s and elapsed "time" during the transformation t.

Intuitively, the double integral in Eq. (4) measures how much every part of the polymer moves in going from one configuration to another (see Fig. 6 for a schematic).

The minimal distance problem Eq. (4) is not equivalent to a simple soap-film problem (see Fig. 7). It also has a lower symmetry than the relativistic world-sheet of a classical string [57], and so is inequivalent to that problem as well.

Minimizing Eq. (4) results in straight line motion of all points along the curve. This is because Eq. (4) models not an inextensible string but an effective "rubber band," which can expand and contract at no cost to facilitate the minimal-distance transformation. If the polymer cannot arbitrarily stretch and contract (a good approximation for real inextensible polymers), the trajectories of the constituent segments deviate from straight lines.

The polymer is made inextensible by introducing the constraint

$$\sqrt{\left(\frac{\partial \mathbf{r}}{\partial s}\right)^2} \equiv \sqrt{\mathbf{r}'^2} = 1, \qquad (5)$$

whereupon the function to be minimized becomes

$$D = \int_0^L \int_0^T ds dt\, L(\dot{\mathbf{r}}, \mathbf{r}') \qquad (6)$$

with effective Lagrangian:

$$L(s,t) = \sqrt{\dot{\mathbf{r}}} - \lambda(\sqrt{\mathbf{r}'^2} - 1) \qquad (7)$$

and Lagrange multiplier $\lambda \equiv \lambda(s, t)$, a function of both s and t.

The new equations of motion obtained by extremizing the functional become:

$$\dot{\hat{\mathbf{v}}} = \lambda\boldsymbol{\kappa} + \lambda'\hat{\mathbf{t}} \qquad (8)$$

where $\hat{\mathbf{t}}$ is the unit tangent vector, and $\boldsymbol{\kappa}$ is the curvature vector [57].

Numerical solutions may be more readily obtained by discretizing the string as shown in Figure 8. This procedure is a particular example of the method of lines, used to obtain solutions of partial differential equations. After discretization, the functional to be minimized becomes

$$D[\mathbf{r}_i, \dot{\mathbf{r}}_i] = \int_0^T dt\, L(\mathbf{r}_i, \dot{\mathbf{r}}_i), \qquad (9)$$

where the effective Lagrangian L is:

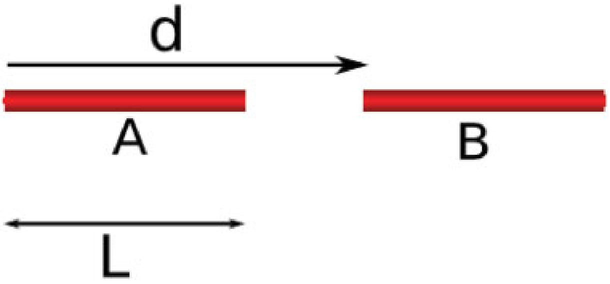


FIGURE 7. The line segment *A* is displaced by *d* along itself, to *B*. The soap film area A_{soap} between the two segments is 0. But the distance $D_{AB} = L\ d$. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

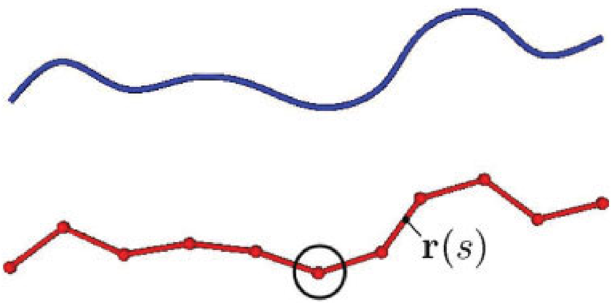


FIGURE 8. The lower curve is a discretized version of the upper one. After discretization, the PDE for the upper curve becomes a set of N coupled ODE's for the N residues in the lower chain (A sample residue is marked with a circle). [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

$$L(\mathbf{r}_i,\dot{\mathbf{r}}_i) = \sum_{i=1}^{N}\left(\sqrt{\dot{\mathbf{r}}_i^2} - \frac{\lambda_{i,i+1}}{2}((\mathbf{r}_{i+1} - \mathbf{r}_i)^2 - b^2)\right). \quad (10)$$

Here, b is the segment length which we set to unity. The distances we obtain are thus in units of b^2. The distance between space curves has the dimensions of area just as the distance between points has dimensions of length. Upon discretization, the PDE of the system becomes a set of N coupled ODE's, one for each residue:

$$\dot{\hat{\mathbf{v}}}_1 + \lambda_{12}\mathbf{r}_{2/1} = 0 \quad (11a)$$

$$\dot{\hat{\mathbf{v}}}_2 - \lambda_{12}\mathbf{r}_{2/1} + \lambda_{23}\mathbf{r}_{3/2} = 0 \quad (11b)$$

$$\vdots$$

$$\dot{\hat{\mathbf{v}}}_N + \lambda_{N-1,N}\mathbf{r}_{N/(N-1)} = 0. \quad (11c)$$

where e.g. $\mathbf{r}_{2/1} = \mathbf{r}_2 - \mathbf{r}_1$.

The solutions of the first and last (Nth) residues or beads consist of either straight-line motion of the bead, pure rotation of the link terminating on the bead, or a stationary solution where the residue remains at rest. Moreover, Weierstrass-Erdmann corner conditions or transversality conditions demand smooth curves for solutions by disallowing discontinuities or cusps in the trajectories [58].

Given two conformations that serve as boundary conditions on the equations of motion 11(a–c), several solutions yielding slightly (nonextensively) different D's can be constructed. It can be shown that they are all local minima [58]. In Figure 9, two solutions are shown. Figure 9(A) depicts the global minimum transformation, and Figure 9(B) a subminimal "excited-state" transformation. The solutions both involve either rotations of the constituent links or straight line motion of the constituent beads. In Figure 9(A), rotation occurs away from the straight-line conformation and results in a distance $D = 45.793$, while in 9B rotation occurs from the curved conformation and results in $D = 46.278$.

The fact that a real polymer cannot cross itself can be incorporated into the problem of finding the minimal distance [44]. Noncrossing is manifested as an inequality constraint [60–63], which appears in Eq. (10) as a Lagrange parameter for each residue i, multiplying the excluded volume constraint. To describe this, let the unit vector from the kth to the (k + 1)th bead be $\hat{\mathbf{e}}_k \equiv (\mathbf{r}_{k+1} - \mathbf{r}_k)/b$, then the vector to position $r(s)$ at contour length s on the chain (e.g., see Fig. 8) is

$$\mathbf{r}(s) = b\sum_{i=0}^{k-1}\hat{\mathbf{e}}_i + (s - kb)\hat{\mathbf{e}}_k$$

$$= \mathbf{r}_k + (s - kb)\hat{\mathbf{e}}_k.$$

To constrain the motion of the beads so that the chain cannot cross itself, we add the term

$$\lambda_i\left(\int_0^s ds(|\mathbf{r}(s) - \mathbf{r}_{\mathrm{i}}| + \epsilon_{\mathrm{i}}^2)\right) \quad (12)$$

to the summand of Eq. (10). Note that by discretizing the problem to find the motion of residues,

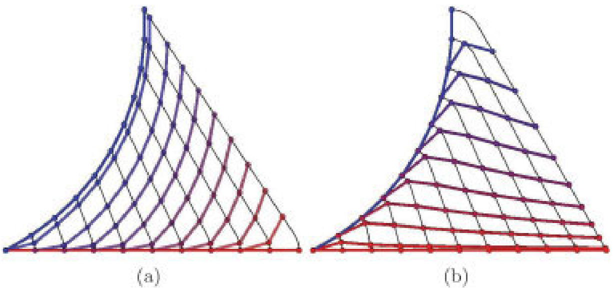


FIGURE 9. Minimal and subminimal transformations between a straight line and a quarter circle (see text for description). For the left transformation $D = 45.793$ and for the right one $D = 46.278$, in units of the link-length squared. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

there must be an asymmetry in the way that the chain is treated: in a continuum treatment the term in the integrand of (12) would be $|\mathbf{r}(s) - \mathbf{r}(s')|$. The quantity ϵ^2 in (12) is an "excess parameter" which is zero unless a residue is directly constrained (touching some part on the rest of the chain). If $\epsilon_i^2 = 0$ the problem of finding minimal distance is a "free" problem for residue i, and the equations of motion 11(a–c) are unchanged. However, the corner conditions mentioned above induce an implicit "knowledge" of the sterically avoided boundary, so that the motion of the residues are altered to travel most directly to the steric surface constituting the constraint or obstacle. At this point, the residue is constrained to be on the surface of the obstacle, and the trajectory is defined accordingly. Subsequently, the residue leaves the constraining surface, and the problem becomes a free problem once again, traveling most directly to the final conformation [44].

In the above treatment, the chain has zero thickness. A tube thickness ρ can be straightforwardly incorporated into the treatment by letting $r(s)$ $\mathbf{r}(s) \rightarrow \mathbf{r}(s) + \rho\hat{e}_\rho$ in Eq. (12), and then integrating over the surface of the cylinders which compose the resulting piece-wise tube.

Another modification that can be made to the Lagrangian is one involving curvature constraints. In the current treatment, the angle between two consecutive links of the chain can have any value, whereas in real protein chains angles defined by bonds between atoms or residues are restricted. We will not discuss these aspects in this manuscript.

The Minimal Distance Between Protein Fragments

In Ref. 44, protein fragments such as an alpha helix and beta hairpin were considered for purposes of calculating the minimal distance. An extended strand was aligned to the respective structures by minimizing either RMSD or MRSD, and the distance D was subsequently calculated for the aligned structural pairs. Both real and idealized protein fragments were considered. Most pairs of structures had smaller distance minimal pathways when aligned using MRSD as the cost function. In some cases, however, the smaller distance minimal pathway was obtained when the boundary conformations were aligned using RMSD as the cost function.

For example, the straight line conformation in Figure 10 was aligned to an idealized β-hairpin structure also shown in that figure. The alignment was performed by both minimizing the MRSD between the structures [Fig. 10(A)], and by minimizing RMSD between the structures [Fig. 10(B)]. In each instance, the minimal distance D between the structural pairs was calculated after alignment. The resulting aligned straight-line structures have significantly different position/orientation depending which cost function was used, MRSD or RMSD: the MRSD between the two straight-line structures is in fact larger than the MRSD between each and the hairpin structure [44].

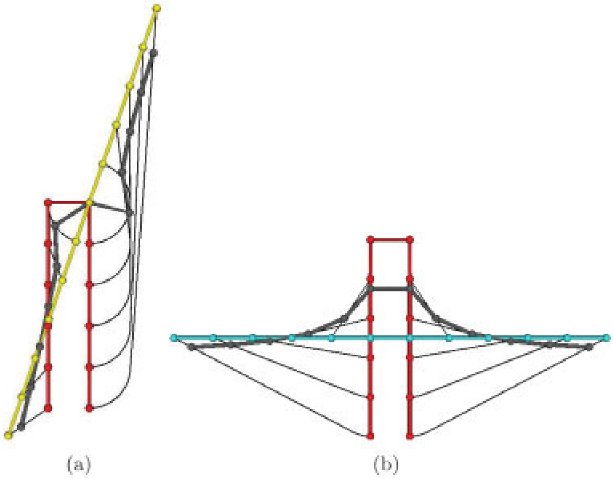


FIGURE 10. (color) D minimizing transformations for MRSD aligned (yellow) and RMSD (cyan) aligned hairpins. Intermediate state is shown in gray. The distances for each transformation, in units of link length squared, are 3.20 for MRSD-aligned and 3.22 for RMSD-aligned structures. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

Both transformations are minimal transformations but are subject to different boundary conditions and thus yield different pathways and D's. The question remains as to how to align the structures to obtain the minimum of all minimal transformations, that is, the minimum minimal distance D. To calculate this quantity, D itself must be used as the cost function for alignment.[3]

In this article, we align structures using D as a cost function to obtain for the first time the minimum of all minimal transformations. The structures

[3] In the limit of a large number of residues (N), the distance converges to the N times the MRSD: $D \rightarrow N \times \text{MRSD}$, so for long chains, MRSD can be considered a first step toward optimal alignment. However, ideally, one wants to align the two structures using D itself as a cost function.

that we consider are idealized straight-line segments with varying number of links, which are then aligned to idealized beta hairpins using D as a cost function. The alignment and resulting distance D are compared with the alignments and distances of RMSD and MRSD. This is a first step toward aligning more complex structures using D as a cost function. We will also see that there exist high order approximations which capture much of the properties of a true D alignment. Applying these approximate metrics to align structures such as a full protein is a topic for future research.

Structural Alignment of Protein Fragments Using the Distance *D*

In principle, minimal pathways can be computed for any initial and final configurations, just as RMSD can be computed between any two configurations. However, it is of special significance to anneal the configurations allowing translations and rotations, until the minimal distance transformation is achieved (i.e. the minimum of minimal distance transformations). This is analogous to the usual procedure of using RMSD or MRSD as a cost function between two structures and minimizing with respect to translations and rotations. Although the minimization procedure is particularly straightforward for RMSD and involves the inversion of a matrix, the minimization using the distance D as a cost function involves a simplex or conjugate gradient minimization and so is more computationally intensive.

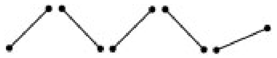
In short, the boundary conformations are allowed to translate and rotate in 3D space. Their position and orientation are modified to produce a pathway with minimal length, when compared with all other minimal pathways that can be obtained by positioning and orienting the same two structures in 3D space.

Method and Results

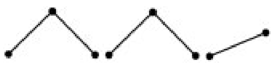
For the purpose of generating accurate initial guesses for the minimal distance aligned structure, we introduce the following hierarchy:

$$D_0 = N \times \text{MRSD} \qquad (13a)$$

$$D_1 = \sum_{i=1}^{N-1} D(l_i^{(A)}, l_i^{(B)}) \qquad (13b)$$

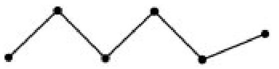


$$D_2 = \sum_{i=1}^{\lfloor((N-1)/2\rfloor} D(\{l_i^{(A)}\},\{l_i^{(B)}\}) + D_1^{(\text{end link})} \qquad (13c)$$



$$\vdots$$

$$D_N = D. \qquad (13d)$$



In this hierarchy, the D_α has the following interpretation: D_0 is the cumulative distance between the sets of points comprising the residue locations of conformations A and B, D_1 is the cumulative distance between the sets of single links, l_i, comprising configurations A and B, D_2 is the cumulative distance between the sets of double links, $\{l_i\}$, comprising configurations A and B plus any single-link remainder if one exists, and so on. That is, at level α the polymer chain is divided up into subsegments each of link-length α, plus one segment constituting the remainder. When $\alpha = N$, the chain as a whole is considered, which is the true distance D. This procedure is also illustrated schematically adjacent to each equation above.

We observed that D_1 was a good approximation to the total D between two chains, was much easier in practice to calculate, and could be automated in a robust way. For these reasons, we used it to generate initial guesses for minimal distance aligned structures. After the initial alignment using D_1 the chains were further aligned using the full distance D. At this stage, the general form of the transformation is established and the computation can be automated. We used a Nelder-Mead simplex method in our algorithm to find the minimal distance alignment.

Figure 11 shows the aligned structures using RMSD, MRSD, D_1, and D, for increasing numbers of links. Several points can be observed. For the smallest number of links (three), MRSD, D_1, and D all give the same alignment [Fig. 11(a)], but differ from RMSD. For five or more links, the MRSD- and D_1 aligned structures break symmetry by choosing particular diagonal direction, while the D-aligned

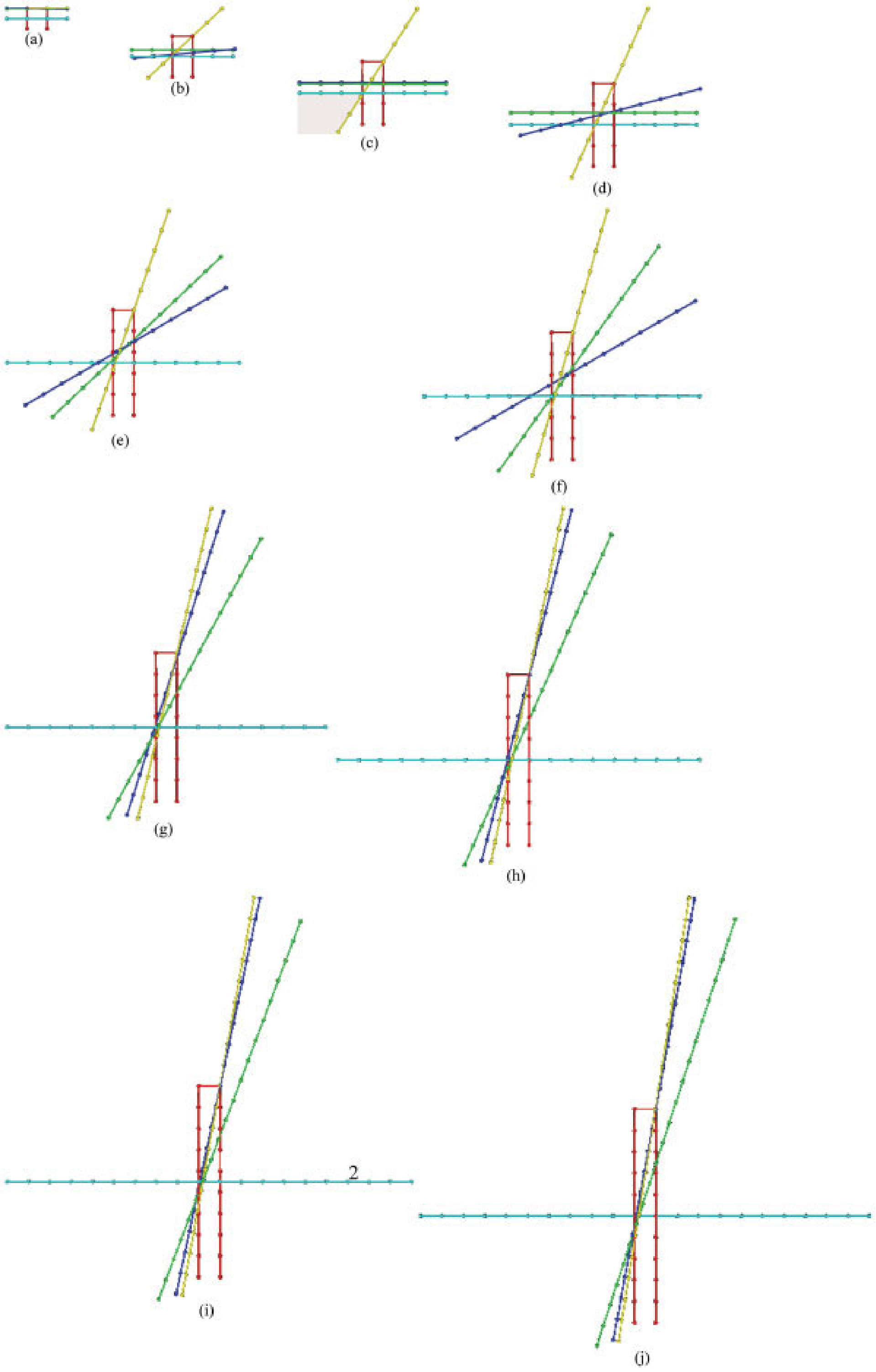


FIGURE 11. Alignments with different cost functions. The Hairpin is shown in red. *D* alignment in green, D_1 in blue, MRSD in yellow, and RMSD in cyan.

TABLE I

***D/N* (in units of link length squared) between the aligned structures in Figure 11.**

	Alignment cost function			
N	*D*	*D*1	MRSD	RMSD
4	0.785	0.785	0.785	0.822
6	1.391	1.404	1.473	1.419
8	1.974	1.975	2.085	2.014
10	2.559	2.562	2.654	2.615
12	3.127	3.155	3.197	3.216
14	3.674	3.730	3.726	3.817
16	4.207	4.240	4.247	4.418
18	4.732	4.760	4.762	5.019
20	5.252	5.271	5.272	5.620
22	5.767	5.780	5.783	6.221

Each of the four columns represents the structural pairs for the cost function labeled. For example, column 3 gives *D/N* for structural pairs in figure 11 aligned using MRSD.

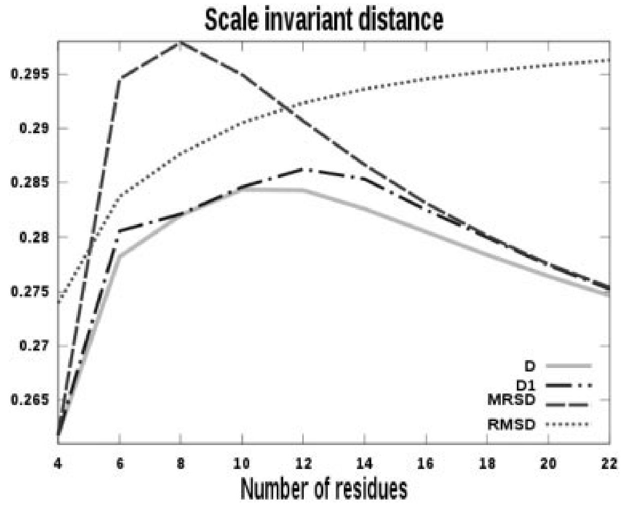


FIGURE 12. Scale invariant distance resulting from different alignments with different cost functions, for alignment of a straight line segment to an idealized β-hairpin.

structure retains horizontal symmetry but shifts in position along the hairpin [Fig. 11(b)]. Interestingly, for seven links the D_1-aligned structure becomes horizontal again and nearly identical with the *D*-aligned structure, while the MRSD-aligned structure remains diagonal. The deviation from MRSD and *D* is a finite-size effect [57], so we know that the two alignments must eventually converge as *N* is increased. At nine links [Fig. 11(d)], the D_1-alignment breaks symmetry again, in the same fashion as MRSD, while the *D*-alignment remains similar to RMSD. By 11 links [Fig. 11(e)], the *D*-aligned structure has broken symmetry as well, with a larger angle to the horizontal than D_1 but smaller angle than MRSD. This situation persists for 13 links [Fig. 11(f)]. As *N* is further increased beyond 15 links [Fig. 11(g)], the D_1 and MRSD aligned structures quickly converge, while the angle with respect to the horizontal of the *D*-aligned structure continues to lag behind that of either MRSD and D_1 aligned structures, converging slowly as *N* continues to increase [Figs. 11(g)–(j)]. The RMSD-aligned structure remains horizontal for all hairpin lengths.

Average lengths of β-hairpins in databases constructed from the PDB are about 17 residues [64], most consistent with Figure 11(h). From this figure, we see that hairpins of this length have a globally different structural alignment with extended structures depending on whether *D* or RMSD is used.

Table I and Figure 12 summarize the results for the minimal distance transformations from the aligned structures. Table I gives the numerical value of the distance *D* for each aligned structure, aligned using the various cost functions listed: *D*, D_1, MRSD, and RMSD. Note that the distance *D* is always minimized for the distance-aligned structure, and tends to increase as one considers the D_1, MRSD and then RMSD-aligned structures for a given number of links.

For comparison, in Table II the corresponding

TABLE II

MRSD (in units of link length) between the aligned structures in Figure 11 using the four cost functions we considered.

	Alignment cost function			
N	*D*	*D*1	MRSD	RMSD
4	0.707	0.707	0.707	0.809
6	1.375	1.393	1.337	1.412
8	1.961	1.960	1.899	2.008
10	2.547	2.545	2.436	2.610
12	3.062	3.108	2.959	3.211
14	3.575	3.675	3.475	3.813
16	4.081	4.004	3.987	4.414
18	4.585	4.506	4.495	5.015
20	5.088	5.008	5.002	5.616
22	5.591	5.511	5.508	6.218

For example, column 1 gives MRSD for structural pairs in figure 11 aligned using the distance *D*.

values of MRSD are given for the aligned structures using each cost function. Note in each table that as $N \rightarrow \infty$, D tends to converge to MRSD.

The distance traveled per residue, in units of link length is D/Nb. Dividing this measure by the chain length $(N - 1)b$ gives a scale-invariant measure of the distance: $\tilde{D} = D/(N(N - 1)b^2)$. This quantity is plotted in Figure 12. We can see from the plot that the D_1-aligned structure generally gives a good approximation to the true D-aligned structure. Moreover, MRSD, D_1 and D all converge to the same value while RMSD converges to a dissimilar value.

Conclusions and Discussion

In this article, we reviewed the concept of the generalized distance D, and then used it as a cost function to align unfolded idealized strands of various sizes to their corresponding idealized β-hairpin structures. This is the first time that the true Euclidean distance has been used as a cost function for structural alignment. The distance D for the minimal transformation between aligned structural pairs was compared for various alignment cost functions: RMSD, MRSD, D_1, and D itself. D_1 is the distance between conformational pairs if the chain were decimated to single links and the distance of all single-link transformations was summed.

We found that D_1-aligned structures generally gave a distance that was close to the true D-aligned structure, and in this sense was a good approximation. However, the aligned structures were noticeably different depending on the cost function, for the finite values of N that we studied. Our largest value of N was 22 residues, while the average length of β-hairpins is about 17 residues. For these average hairpin lengths, the minimal D aligned structure is globally different from the RMSD structure. Whether this discrepancy is generally true for larger structures or whole proteins remains to be determined, but we feel it is likely, in particular when aligning extended structures to proteins. It is not yet clear at this point whether alignment using distance will yield more accurate predictions for such problems as protein structure prediction or ab initio drug design. What is clear is that the best-aligned structures using a reasonable alignment metric such as the true distance give very different results than RMSD, even for relatively simple structures such as the β-hairpin.

ACKNOWLEDGMENTS

S.S.P. gratefully acknowledges support from the Natural Sciences and Engineering Research Council, and the A. P. Sloan Foundation.

References

1. Anfinsen, C. B.; Haber, E.; Sela, M.; White, F., Jr. Proc Natl Acad Sci USA 1961, 47, 1309.
2. Wolynes, P. G. In Spin Glasses and Biology; Stein, D., Ed.; World Scientific: Singapore, 1992; p 225.
3. Bryngelson, J. D.; Onuchic, J. N.; Socci, N. D.; Wolynes, P. G. Proteins: Struct Funct Genet 1995, 21, 167.
4. Wolynes, P. G.; Onuchic, J. N.; Thirumalai, D. Science 1995, 267, 1619.
5. Onuchic, J. N.; Luthey-Schulten, Z.; Wolynes, P. G. Annu Rev Phys Chem 1997, 48, 545.
6. Fersht, A. R. Structure and Mechanism in Protein Science; W. H. Freeman and Co.: New York, 1999.
7. Dobson, C. M. Nature 2003, 426, 884.
8. Plotkin, S. S.; Onuchic, J. N. Q Rev Biophys 2002, 35, 111.
9. Plotkin, S. S.; Onuchic, J. N. Q Rev Biophys 2002, 35, 205.
10. Bryngelson, J. D.; Wolynes, P. G. J Phys Chem 1989, 93, 6902.
11. Šali, A.; Shakhnovich, E.; Karplus, M. Nature 1994, 369, 248.
12. Plotkin, S. S.; Wang, J.; Wolynes, P. G. J Chem Phys 1997, 106, 2932.
13. Du, R.; Pande, V. S.; Grosberg, A. Y.; Tanaka, T.; Shakhnovich, E. S. J Chem Phys 1998, 108, 334.
14. Onsager, L. Phys Rev 1938, 54, 554.
15. García, A. E. Phys Rev Lett 1992, 68, 2696.
16. Chan, H. S.; Dill, K. A. J Chem Phys 1994, 100, 9238.
17. Plotkin, S. S.; Wolynes, P. G. Phys Rev Lett 1998, 80, 5015.
18. Hummer, G.; García, A. E.; Garde, S. Proteins 2001, 42, 77.
19. Baumketner, A.; Shea, J.-E.; Hiwatari, Y. J Chem Phys 2004, 121, 1114.
20. Ma, A.; Dinner, A. R. J Phys Chem B 2005, 109, 6769.
21. Best, R. B.; Hummer, G. Proc Natl Acad Sci USA 2005, 102, 6732.
22. Das, P.; Moll, M.; Stamati, H.; Kavraki, L. E.; Clementi, C. Proc Natl Acad Sci USA 2006, 103, 9885.
23. Cho, S. S.; Levy, Y.; Wolynes, P. G. Proc Natl Acad Sci USA 2006, 103, 586.
24. Dokholyan, N. V.; Li, L.; Ding, F.; Shakhnovich, E. I. Proc Natl Acad Sci USA 2002, 99, 8637.
25. Chavez, L. L.; Onuchic, J. N.; Clementi, C. J Am Chem Soc 2004, 126, 8426.
26. Wang, J.; Zhang, K.; Lu, H. Y.; Wang, E. K. Phys Rev Lett 2006, 96, 168101.
27. Sega, M.; Faccioli, P.; Pederiva, F.; Garberoglio, G.; Orland, H. Phys Rev Lett 2007, 99, 118102.
28. Beck, D. A. C.; Daggett, V. Biophys J 2007, 93, 3382.
29. Krivov, S. V.; Muff, S.; Caflisch, A.; Karplus, M. J Phys Chem B 2008, 112, 8701.

30. Nymeyer, H.; Socci, N. D.; Onuchic, J. N. Proc Natl Acad Sci USA 2000, 97, 634.
31. Plotkin, S. S.; Onuchic, J. N. J Chem Phys 2002, 116, 5263.
32. Wind, A. F.; Kemp, J. P.; Ermoshkin, A. V.; Chen, J. Z. Y. Phys Rev E 2002, 66, 031909.
33. Cheung, M. S.; Garcia, A. E.; Onuchic, J. N. Proc Natl Acad Sci USA 2002, 99, 685.
34. Karanicolas, J.; Brooks, C. L., III. Protein Sci 2002, 11, 2351.
35. Snow, C.; Nguyen, H.; Pande, V.; Gruebele, M. Nature 2002, 420, 102.
36. Simmerling, C.; Strockbine, B.; Roitberg, A. J Am Chem Soc 2002, 124, 11258.
37. Garcia, A. E.; Onuchic, J. N. Proc Natl Acad Sci USA 2003, 100, 13898.
38. Hummer, G.; García, A. E.; Garde, S. Phys Rev Lett 2000, 85, 2637.
39. Veitshans, T.; Klimov, D.; Thirumalai, D. Fold Des 1996, 2, 1.
40. Baumketner, A.; Hiwatari, Y. Phys Rev E 2002, 66, 011905.
41. Cheung, M. S.; Thirumalai, D. J Mol Biol 2006, 357, 632.
42. Portman, J. J.; Takada, S.; Wolynes, P. G. J Chem Phys 2001, 114, 5069.
43. Portman, J. J.; Takada, S.; Wolynes, P. G. J Chem Phys 2001, 114, 5082.
44. Mohazab, A. R.; Plotkin, S. S. Biophys J 2008, 95, 5496.
45. Plotkin, S. S.; Onuchic, J. N. Proc Natl Acad Sci USA 2000, 97, 6509.
46. Lindberg, M.; Tangrot, J.; Oliveberg, M. Nature Struct Biol 2002, 9, 818.
47. Plotkin, S. S. Proteins: Struct Funct Genet 2001, 45, 337.
48. Plotkin, S. S.; Wolynes, P. G. Proc Natl Acad Sci USA 2003, 100, 4417.
49. Clementi, C.; Plotkin, S. S. Protein Sci 2004, 13, 1750.
50. Shoemaker, B. A.; Wang, J.; Wolynes, P. G. Proc Natl Acad Sci USA 1997, 94, 777.
51. Shoemaker, B. A.; Wang, J.; Wolynes, P. G. J Mol Biol 1999, 287, 675.
52. Munoz, V.; Eaton, W. A. Proc Natl Acad Sci USA 1999, 96, 11311.
53. Galzitskaya, O. V.; Finkelstein, A. V. Proc Natl Acad Sci USA 1999, 96, 11299.
54. Alm, E.; Baker, D. Proc Natl Acad Sci USA 1999, 96, 11305.
55. Baker, D. Nature 2000, 405, 39.
56. Levine, R. D.; Bernstein, R. B. Molecular Reaction Dynamics and Chemical Reactivity; Clarendon Press: Oxford, 1987.
57. Plotkin, S. S. Proc Natl Acad Sci USA 2007, 104, 14899.
58. Mohazab, A. R.; Plotkin, S. S. J Phys: Condens Matter 2008, 20, 244133.
59. Zwiebach, B. A First Course in String Theory; Cambridge University Press: New York, 2004.
60. Pontryagin, L. S.; Boltyanskii, V. G.; Gamkrelidze, R. V.; Mishchenko, E. F. The Mathematical Theory of Optimal Processes; Wiley Interscience: New York and London, 1962.
61. Cass, D. Rev Econ Stud 1965, 32, 233.
62. Gregory, J.; Lin, C. J Math Anal Appl 1994, 187, 826.
63. Gregory, J.; Lin, C. Constrained Optimization in the Calculus of Variations and Optimal Control Theory; Springer: New York, 2007.
64. de la Cruz, X.; Hutchinson, E. G.; Shepherd, A.; Thornton, J. M. Proc Natl Acad Sci USA 2002, 99, 11157.