Biophysical Chemistry 190–191 (2014) 50–55

Contents lists available at ScienceDirect

Biophysical Chemistry

journal homepage: http://www.elsevier.com/locate/biophyschem

Letter to the editor

A novel protein distance matrix based on the minimum arc-length between two amino-acid residues on the surface of a globular protein

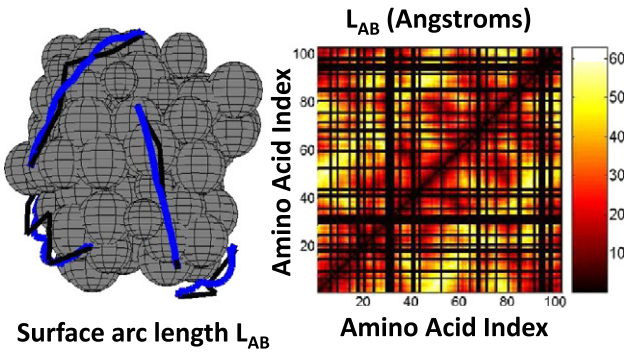
Damien Hall [a,b,*], Songling Li [b], Kazuo Yamashita [b], Ryuzo Azuma [b], John A. Carver [a], Daron M. Standley [b,**]

[a] Research School of Chemistry, Australian National University, Canberra ACT 0200, Australia
[b] Immunology Frontier Research Center (IFReC), Section on Systems Immunology. Osaka University, 3-1 Yamada-oka, Suita, Osaka 565-0871, Japan

HIGHLIGHTS

- A novel protein distance matrix based on surface residue arc-length is discussed.
- Two methods for calculating the protein distance matrix for globular proteins are presented.
- The simpler of the two methods, based on a rule-based algorithm, was found to yield suitable results.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 5 December 2013
Received in revised form 20 January 2014
Accepted 23 January 2014
Available online 13 February 2014

Keywords:
Distance matrix
Surface arc-length
Algorithm
Protein–polymer interaction

ABSTRACT

We present a novel protein distance matrix based on the minimum line of arc between two points on the surface of a protein. Two methods for calculating this distance matrix are developed and contrasted. The first method, which we have called TOPOL, is an approximate rule based algorithm consisting of successive rounds of vector addition. The second method is adapted from the graph theoretic approach of Dijkstra. Both procedures are demonstrated using cytochrome c, a 12,500 Da protein, as a test case. In respect to computational speed and accuracy the TOPOL procedure compares favorably against the more complex method based on shortest path enumeration over a surface manifold grid. Some potential uses of the algorithmic approaches and calculated surface protein distance measurement are discussed.

1. Introduction

When proteins interact with flexible docking partners [1,2], or the protein surface facilitates the transfer of a reactant or product between active and regulatory sites [3–5], the minimum connecting surface arc length[1] between surface residues A and B, L_{AB}, (Eq. ((1a)), may be a greater determinant of A and B's role in facilitating an interaction than their simple Euclidean distance, d_{AB} (Eq. ((1b)).

$$L_{AB} = \int_{t=A}^{t=B} \sqrt{\left(\frac{dx}{dt}\right)^2 + \left(\frac{dy}{dt}\right)^2 + \left(\frac{dz}{dt}\right)^2} \cdot dt. \quad (1a)$$

* Correspondence to: D. Hall, Research School of Chemistry, Australian National University, Canberra, ACT 0200, Australia.
** Corresponding author.
E-mail addresses: damien.hall@anu.edu.au (D. Hall), standley@ifrec.osaka-u.ac.jp (D. Standley).

[1] The minimum total surface distance connecting two points is formally referred to as the geodesic of those two points. In this paper we prefer total arc length.

http://dx.doi.org/10.1016/j.bpc.2014.01.005
0301-4622/

$$d_{AB} = \sqrt{(x_B - x_A)^2 + (y_B - y_A)^2 + (z_B - z_A)^2}. \quad (1b)$$

Although a value for the arc length could, in principle, be obtained from an analysis of a differentiable function based representation of the surface [6,7], evaluation of a large number of directed random walks [8], or by brute force enumeration of all possible paths on a discretized surface manifold, in practice all these processes are computationally heavy, leaving few easy options available for estimating this quantity's minimum for a non-determined or non-regular surface shape. Here we present two methods for calculating the protein surface arc-length. The first method, suitable for approximately globular proteins, involves a novel rule based algorithm we have developed called TOPOL. TOPOL utilizes an identity associated with the geometry of curved surfaces to rapidly estimate the surface arc. Although yielding only a semi-approximate solution, TOPOL, is suitably computationally rapid to constitute a first-pass filtering step over large data sets. The second method is based upon Dijkstra's graph theoretic approach for shortest path estimation [9,10]. Dijkstra's method, which is potentially semi-exact, generates a shortest path via continuous selective sampling of nodes from a discretized surface manifold. In the next sections we discuss these two methods for determining the novel protein distance matrix based on arc length, along with possible uses of the matrix.

2. TOPOL approach

The TOPOL approach represents a simple geometrical/computational solution to the potentially complex problem of arc length determination. The basic *ansatz* of the TOPOL approach is drawn from a fundamental tenet of Riemannian geometry [11], namely that the shortest arc length between any two points on a sphere is given by the arc of the great circle connecting those two points and the sphere's center [5–7,11]. Given that globular proteins are not spheres we allow the protein center to vary in a semi-random fashion and compute the length of total arc for each choice of center, selecting the minimum as the best estimate. The major computational features of the TOPOL procedure are described in Fig. 1. Algorithmic flow involves the following steps. (*A*) *Protein Selection*: Protein 3D atomic coordinate file selected from Protein Data Bank (in PDB file format). (*B*) *Amino-Acid United Atom Model*: Conversion of an all atom protein structure to a reduced amino acid pseudo-atom model is achieved by finding the heavy atom weighted center of mass for each amino acid and then calculating an equivalent amino acid spherical radius. (*C*) *Surface Amino Acid United Atom Model*: Internal amino acids are discarded to leave N surface residues in the data set. Hollowing out the protein is achieved by determining which of the reduced amino acid atoms can accommodate a small test probe repeatedly inserted at a constant distance R_P from their surface in a random direction. (*D*) *Multiple Coordinate File Representations*: Q surface amino acid pseudo-atom coordinate files are generated in relation to a center of mass that can vary from its true value by plus or minus one-third of the average radius, R_{av}, of the protein. Seven 'fixed' center points located at the true center [0,0,0] and along each of the Cartesian axes [$\pm R_{av}/3$,0,0], [0,$\pm R_{av}/3$,0] and [0,0,$\pm R_{av}/3$]. A further Z number of random center points selected within a box of side length $\pm 2R_{av}/3$ with its center located at the origin can be included to increase sampling proficiency. Each of the Q coordinate files specifies an N × 9 element matrix, $\mathbf{C}_{\{Q\}}$. For a given amino acid row the nine column positions of $\mathbf{C}_{\{Q\}}$ respectively describe [1] the surface amino acid file running index [2–4], the surface amino acid Cartesian coordinates with respect to a particular designated center (x',y',z') [5–7], the amino acid's spherical coordinates with respect to a designated center (r',θ',ϕ') [8], the amino acid radius and [9] the originally assigned PDB amino acid running index. (*E*) *Selection of Surface Boundary Points*: Two surface points (designated A and B) are defined by projecting along the direction of each amino acid's positional vector (originating from the true center of mass) by a total magnitude equal to the sum of the magnitude of the positional vector, the particular amino acid radius, $R_{\alpha\alpha i}$, and the radius of a test molecule in contact with the surface, R_T. (*F*) *Cycle of Vector Addition/Resultant Vector Modification*: Positional vectors for points A and B are defined for each coordinate file in relation to the semi-random location of the protein origin/center of mass. For a particular coordinate file these vectors are added and the magnitude of the resultant modified until it lies a distance R_T above the surface along its unit direction. This process is then repeated by selecting the modified resultant vector as either the left or right boundary vector in a new cycle of vector addition and resultant modification. The process is continued for either a set number of division cycles or until the magnitude of the minimum line element corresponding to the difference of the two boundary vectors is approximately equal to a specified minimal distance segment (set as $R_T/2$). (*G*) *Arc Length Estimation*: The irregular arc length for each choice of center is calculated by summing the magnitudes of the various line elements with the shortest line of arc selected as the TOPOL estimate.

Fig. 2 describes the application of the TOPOL routine to cytochrome c, a relatively small 104 amino acid globular protein of ~12,500 Da in mass [12] with 71 amino acids defined as surface located. The 104 × 104 element matrix describes the TOPOL estimate of the minimum arc length connecting each surface amino-acid pairing at a height of 2.5 Å above the surface. The black lines extending vertically or horizontally across the graph represent internal amino acids not included in the measurement. In any practical usage the heat map designations shown in Fig. 2 will be replaced by searchable numerical table generated by the routine. Like other rule based/heuristic analytical methods [13,14], the TOPOL procedure lacks a supporting theoretical basis, although such a theoretical proof does exist for the limiting case of a spherical protein (provided in Appendix A).

3. A modification of Dijkstra's approach

As Dijksta's [9] procedure involves a partial enumeration of pathway space, we shall call the shortest path length estimated in this fashion, $L_{AB\{ENUM\}}$. In traveling from the starting position to a particular node, a myriad of potential path edges are possible. Indeed a complete enumeration of all possible connecting paths between residues A and B on a protein surface manifold with P total number of nodal points for a pre-designated K number of straight-line steps would require P!/[P − (K − 1)]! total path evaluations. Of course with variable step size between nodal points the shortest path may be reached with either K + 1, K + 2, K + 3 … etc. number of steps. The key aspects of Dijkstra's procedure which help to reduce this potentially large search space[2] are,

(1) transitions from one node to another on a discretized surface manifold, are limited to the range of neighbors meeting some criterion, typically the surrounding vertices of a regular tessellation of space,
(2) for each path estimation, only non-previously selected nodes may be chosen as the next potential nodal connection, such that there can be no multiple selection of a node along a single path,
(3) for each path originating from the starting position the total distance traveled is recorded. If a pathway leading to an intermediate node is found to have a shorter distance than the previous shortest pathway then the new pathway is selected as the preferred pathway to that node. In this way each node has an associated pathway register that is updated,
(4) paths having an intermediate cumulative distance greater than an already completed solution are discarded before reaching their target.

[2] In the present case with P = 71 and K = 7 this would involve approximately 1×10^{13} path evaluations.

The most appropriate value for K will be a joint function of the density of points used to discretize the surface manifold, along with the actual shortest solution path for the particular protein. This second aspect makes the selection of K somewhat implicit in nature since knowledge of the solution is required in order for its determination. To improve the utility of the Dijkstra procedure we included an additional step in which an educated guess of the appropriate value of K was made as follows,

(5) utilizing the $L_{AB\{TOPOL\}}$ derived arc length result as a first estimate we determined an approximately optimum value for K as the rounded up integer of $0.7 \times L_{AB\{TOPOL\}}/\langle$grid spacing$\rangle$, where $\langle$grid spacing $\rangle$ implies the average distance between nodes over the entire surface manifold.

A consequence of imposing a fixed value for K on the search process is the required re-specification of nodal connectivity. We tackled this problem by re-defining rule 1 such that a potential connecting path between nodal surface points must lie within a distance range greater than $0.35L_{AB\{TOPOL\}}/K$ and smaller than $2L_{AB\{TOPOL\}}/K$. To apply the modified Dijkstra procedure to the estimation of shortest line of irregular arc between regions of surface above cytochrome c it was necessary to construct a suitable discretization of protein surface space. Such a discretization was generated by connecting the centers of spheres, of radius R_T, placed in closest hard contact with each surface amino acid, along the line extending from the protein center of mass through the amino acid center. Applying the modified Dijkstra procedure to this discretized surface manifold we determined the value of $L_{AB\{ENUM\}}$ and compared it against the $L_{AB\{TOPOL\}}$ derived from the TOPOL based procedure (Table 1). One immediately notable feature of the Dijkstra approach is that the generated line of arc is irregular along both the angular and radial dimensions (Fig. 3). A range of estimates of $L_{AB\{ENUM\}}$ across seven different values of K were evaluated and compared against the $L_{AB\{TOPOL\}}$ estimate (Table 1). In general the TOPOL distance measure is typically ~15% better than the potentially more accurate, yet more computationally expensive enumeration procedure. In theory the enumeration procedure could be made virtually exact for the situation of $P \rightarrow \infty$, and $K \sim L_{AB\{ENUM\}}/\langle$grid Spacing$\rangle$. However, approach to the limit would significantly increase the computational complexity. In the present case we have chosen P to be equal to N, the number of surface amino acids, a situation corresponding to a relatively coarse discretization of the system. To some extent much of the difference between the two procedures shown in Table 1 is to be expected in the limit of such a low-density discretization of the surface manifold. For this situation observations of $L_{AB\{TOPOL\}} < L_{AB\{ENUM\}}$ will most likely be due to fundamental limits imposed on the path length minimization by a surface point density far from the continuum limit (a situation which could be improved by 'off-lattice' mid-point corrections of the type suggested by Kimmel and Sethian [10]). Similarly, coarse discretization will also produce cases for which $L_{AB\{TOPOL\}} > L_{AB\{ENUM\}}$ may sometimes occur due to a higher density of sampling of the surface arc in the TOPOL procedure, thereby inducing a limited fractal-like quality to the comparison of TOPOL and enumeration estimates.

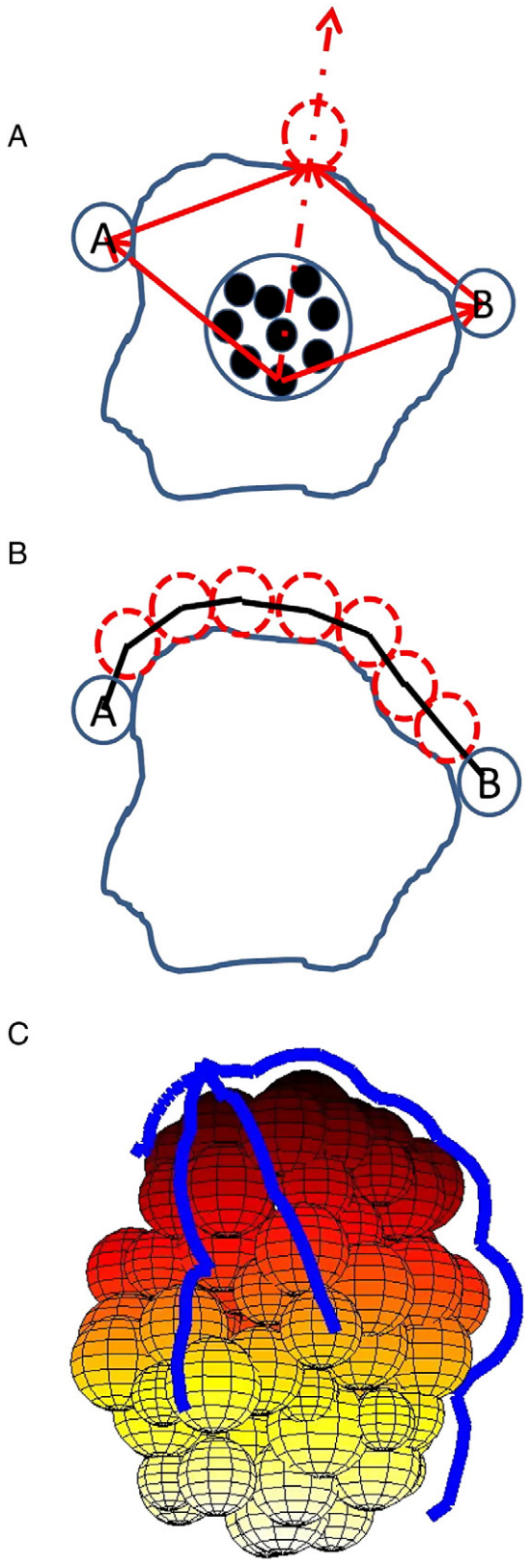


Fig. 1. Major points in the TOPOL arc length measurement procedure: (A) initial boundary point vectors (circles on the surface denoted by A and B) are defined in relation to a variable protein center (indicated by the black balls) and are added together. The resultant vector is then extended or contracted until a test probe makes contact with the protein surface (dotted red ball). (B) Prior resultant vector is used as the new left/right boundary vector and the procedure described in A is repeated through a number of cycles until a trajectory of surface arc is established for a particular protein center. (C) Four examples of the TOPOL procedure applied to a united amino acid atom model of horse heart cytochrome c (PDB file code 1HRC [12]). Line arcs denote the TOPOL determined shortest arc lengths between Lysine 8 (PDB designation) and from right to left Glycine 56, Asparagine 103, Lysine 22 and Valine 11. Colors approximately correspond to the distance heat map representation in Fig. 2.

4. Discussion

In this short note we have introduced a new type of protein distance matrix describing the minimum irregular line of arc between all pairwise combinations of amino acids. Each L_{AB} measurement refers to

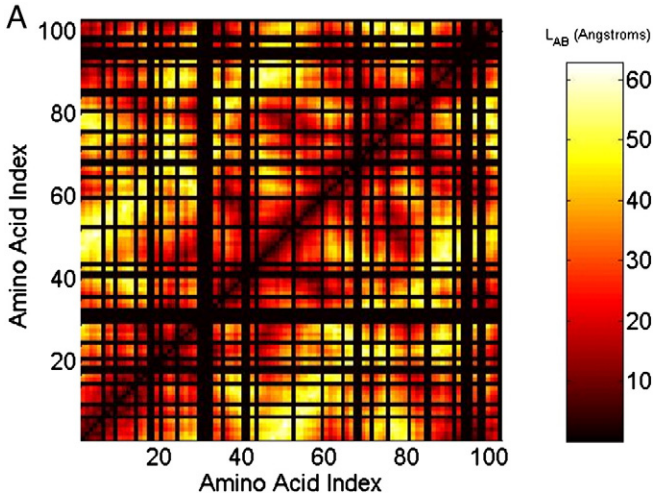


Fig. 2. Systematic application of the TOPOL procedure to cytochrome c: Two-dimensional heat map representation of TOPOL estimate of minimum connecting arc length for all possible combinations of surface amino acids. Black lines extending across the graph represent non-surface amino acids.

the irregular line of arc traced over a protein surface by a spherical tracer particle of set radius, R_T, and hence, in this regard, the L_{AB} measurement is a relative value. Nevertheless, despite this lack of an absolute nature, the information contained within the shortest arc estimate, L_{AB}, whether determined by the modified Dijkstra method, the TOPOL procedure, or an entirely different approach, presents itself as a potentially interesting variable for further exploration. We have identified four areas in which we think the L_{AB} measurement may play an important role.

(1) *L_{AB} in Evolution*: Analysis of the evolutionary conservation of L_{AB}, or lack thereof, could be examined using protein structural evolution methodologies [15,16]. Of particular interest in regard to this point would be the evolutionary maintenance/divergence of the L_{AB} measurement in members of a globular protein superfamily evolved to interact with structurally disordered polymers, such as might be constituted by unfolded proteins, flexible carbohydrates and/or single stranded nucleic acids.[3]

(2) *L_{AB} for reactant/product surface highways*: Substrate channeling is a phenomenon in which surface reactive regions of a protein help to steer the substrate to an enzyme active site [3–5]. Searching for correlations between regions of minimum arc and uniform surface amino acid properties e.g. charge density, may help to identify putative channeling areas.

(3) *L_{AB} in Ligand Docking*: Another potential area of interest for the L_{AB} distance matrix lies in the prediction of flexible polymer binding sites on folded proteins. Present day docking algorithms based on Monte Carlo and molecular dynamic search processes [17,18] work most efficiently when both partners are essentially structurally rigid. However these procedures are not generally applicable to interactions involving a large flexible polymer due to the added computational cost of exploring the interaction for each of its multiple configurations. Utilizing the L_{AB} distance matrix to identify potential structural limits to polymer binding (for each configuration) will provide an invaluable tool for reducing search space [1,2]. Potential coupling of the estimate of the equivalent solvent accessible surface area occluded by the bound polymer in the equivalent minimum arc configuration could yield information on the strength of the interaction.

(4) *L_{AB} in Protein Folding Studies*: Analysis of time dependent changes in the L_{AB} distance matrix computed for protein simulations at each stage of the folding/unfolding reaction may provide a unique insight into the kinetics of protein structure formation. As such it will yield alternative means for examining both the structure and function of natively unfolded polymers existing within the cell [1,19].

In conclusion we note that calculation of the surface arc is certainly a more complex undertaking than the simple calculation of the Euclidean distance. In this short note aside from introducing a novel distance matrix based on arc length we have described two means for calculating this quantity, a modification of the approach developed by Dijkstra [9] and a novel procedure called TOPOL, reliant upon a mathematical identity associated with spherical geometry. The more complex modified Dijkstra procedure should, in principle, be the more robust of the two procedures. Somewhat surprisingly we found that TOPOL, the computationally simpler of the two methods, was the more effective procedure, when applied to the analysis of cytochrome c at a relatively coarse level of surface discretization. While obviously not representing an inductive proof of the method, the close congruency (and indeed better performance) observed between estimates derived from the TOPOL and the modified Dijkstra procedure (Table 1) does suggest a role for the TOPOL routine, perhaps as a coarse filter in examining extremely large data sets [15], or, as used in this paper, as a way of providing a first estimate of the total number of connecting paths (K) for use in a modified Dijkstra procedure. Indeed the TOPOL approach proffers a number of advantages with respect to more computationally complex procedures for evaluating the total surface arc length. TOPOL's chief advantage relates to its numerical simplicity, which helps at the implementation stage and reduces computational cost. Another advantage of TOPOL is the straightforward manner in which it can cater for the finite size and shape of a tracer molecule in contact with the surface (shown in Fig. 1A) without requiring the re-evaluation of points on a surface manifold, or alternatively re-specification of an analytic function, representing the protein-tracer surface contact features. The main disadvantage of the TOPOL approach is its restriction to globular proteins[4] due to its reliance on the geometrical concept of the great circle [5–7,11]. As such the TOPOL approach should become less accurate as protein shape asymmetry increases. In this situation the modified Dijkstra procedure should come to the fore despite its greater relative demand for computer time. Irrespective of their individual shortcomings the two methods have served as vehicles to introduce the novel concept of a distance matrix based on surface arc. We plan on further exploring the utility of this analytical device particularly in the area of predicting protein–single stranded ribo-nucleic acid interactions.

Acknowledgments

We would like to acknowledge helpful discussions with Profs. H. Nakamura, Y. Goto and A.R. Kinjo. We would like to point out that the suggestion for potential coupling of the minimum arc measurement with solvent accessible surface area measurements was suggested by the three anonymous Reviewers. This work was supported by the Special Coordination Funds of the Japanese Ministry of Education, Culture, Sports, Science and Technology, and the Ministry of Health, Labor and Welfare in Japan, and the Japan Society for the Promotion of Science

[3] Indeed an ongoing motivation for this work has been the investigation of evolutionary divergence among members of specific single stranded RNA binding proteins involved in regulating the immunological response across the higher primates.

[4] A possible extension of the basic TOPOL approach to asymmetric protein shapes might arise from considering the protein pseudo-center as an internal ellipsoid, centered around the volume averaged center, with the ellipsoid's axes defined by the three major structural eigenvectors of the protein.

Table 1
Surface arc length measurements for TOPOL and enumeration procedures.[#]

Boundary amino acids [A,B] – PDB index	K	Transition [A→B]–PDB index	$L_{\{AB\}ENUM}$ (Ångstrom)	$L_{\{AB\}TOPOL}$ (Ångstrom)	Ratio $L_{\{AB\}TOPOL}/L_{\{AB\}ENUM}$
[26,92]	7	[26→19→21→101→100→96→1→92]	59.3	51.8	0.87
[11,51]	6	[11→12→13→82→78→77→51]	51.0	43.0	0.84
[72,92]	5	[72→70→69→47→63→66]	41.7	34.8	0.84
[1,82]	4	[1→92→89→90→96→82]	39.3	33.7	0.86
[5,103]	4	[5→2→3→100→103]	30.3	28.7	0.95
[38,62]	4	[38→37→58→60→62]	31.0	26.3	0.84
[2,36]	3	[2→96→99→36]	25.5	24.9	0.98
[62,87]	3	[62→65→88→87]	25.5	23.5	0.92
[1,103]	3	[1→96→99→103]	23.9	21.0	0.88
[11,27]	3	[11→10→15→27]	23.7	18.8	0.80
[5,11]	3	[5→8→12→11]	23.6	17.7	0.75
[28,46]	2	[28→47→46]	18.6	12.8	0.69
[37,38]	1	[37→38]	7.3	7.4	1.01
[3,4]	1	[3→4]	6.5	6.9	1.10
[57,58]	1	[57→58]	6.6	6.7	1.01
[88,89]	1	[88→89]	7.4	7.4	1.00

[#]Colors approximately correspond to the heat map representation of distances shown in Figs. 1 and 2.
[*]K refers to the number of sub-steps in the arc length determination used for the enumeration procedure.

through Funding Program for World-Leading Innovative R&D on Science and Technology (FIRST Program).

Appendix A. Derivation of formula for shortest arc length between two points on a sphere

Here we employ an approach based on the calculus of variations [A1, A2] to prove the basic starting postulate of the TOPOL approach, i.e. the shortest connecting arc between two points lying on the sphere's surface is that corresponding to an intersecting great circle which is a circle joining the two surface points with the center of the sphere. We start by describing the equation for a sphere of radius F, whose center is located at the origin, in Cartesian and polar coordinates.

$$x^2 + y^2 + z^2 = F^2. \tag{A1}$$

$$x = F\cos\theta \cdot \sin\phi. \tag{A2a}$$

$$y = F\sin\theta \cdot \sin\phi. \tag{A2b}$$

$$z = F\cos\phi. \tag{A2c}$$

As per Eq. (1) the total arc length is the square root of the sum of the incremental squared differences in x, y and z i.e. dx, dy and dz which for the case of a sphere of constant radius we may express in terms of variations in the azimuthal and polar angles, θ and ϕ (Eq. (A3)).

$$dx^2 = \left(\frac{\partial x}{\partial\theta}\right)^2 \cdot d\theta^2 + 2\left(\frac{\partial x}{\partial\theta}\right)\left(\frac{\partial x}{\partial\phi}\right)d\theta \cdot d\phi + \left(\frac{\partial x}{\partial\phi}\right)^2 \cdot d\phi^2 \tag{A3a}$$

$$dy^2 = \left(\frac{\partial y}{\partial\theta}\right)^2 \cdot d\theta^2 + 2\left(\frac{\partial y}{\partial\theta}\right)\left(\frac{\partial y}{\partial\phi}\right)d\theta \cdot d\phi + \left(\frac{\partial y}{\partial\phi}\right)^2 \cdot d\phi^2. \tag{A3b}$$

$$dz^2 = \left(\frac{\partial z}{\partial\theta}\right)^2 \cdot d\theta^2 + 2\left(\frac{\partial z}{\partial\theta}\right)\left(\frac{\partial z}{\partial\phi}\right)d\theta \cdot d\phi + \left(\frac{\partial z}{\partial\phi}\right)^2 \cdot d\phi^2. \tag{A3c}$$

Eq. (1) may be profitably rewritten in the following form (Eq. (A4)).

$$L = \int_A^B \sqrt{P + 2Q\phi' + R\cdot\phi'^2}\cdot d\theta \tag{A4}$$

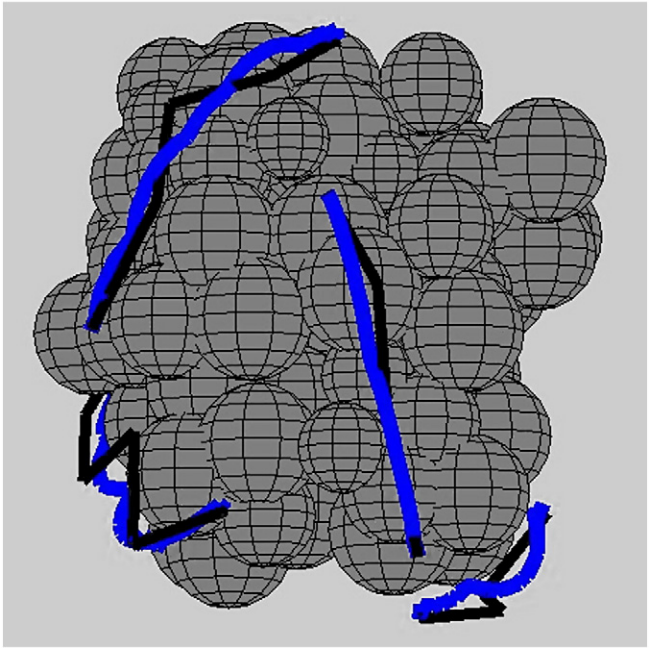


Fig. 3. Comparison of minimal arc estimates generated by the TOPOL and modified Dijkstra based enumeration procedures. Blue lines — TOPOL estimate, black lines — enumeration estimate for arc lengths corresponding to limited set of transitions listed in Table 1 (residue [2 → 36], [38 → 62], [11 → 27] and [72 → 92]).

where $P = \left(\frac{\partial x}{\partial \theta}\right)^2 + \left(\frac{\partial y}{\partial \theta}\right)^2 + \left(\frac{\partial z}{\partial \theta}\right)^2$

$Q = \left(\frac{\partial x}{\partial \theta}\right)\left(\frac{\partial x}{\partial \phi}\right) + \left(\frac{\partial y}{\partial \theta}\right)\left(\frac{\partial y}{\partial \phi}\right) + \left(\frac{\partial z}{\partial \theta}\right)\left(\frac{\partial z}{\partial \phi}\right)$

$R = \left(\frac{\partial x}{\partial \phi}\right)^2 + \left(\frac{\partial y}{\partial \phi}\right)^2 + \left(\frac{\partial z}{\partial \phi}\right)^2$

$\varphi' = \left(\frac{\partial \phi}{\partial \theta}\right).$

Evaluating the partial derivatives shown in A4, we can determine P, Q and R as follows.

$$P = F^2 \sin^2\theta; \quad Q = 0; \quad R = F^2 \tag{A5}$$

Eq. (A4) will be at its minimum value when the Euler–Lagrange condition is satisfied (Eqs. (A6)–A1)

$$\text{For } L = \int_{\theta_1}^{\theta_2} f(\theta, \phi, \phi').d\theta \tag{A6a}$$

$$\min(\mathrm{L}) \Rightarrow \frac{\partial f}{\partial \phi} - \frac{d}{d\theta}\left(\frac{\partial f}{\partial \phi'}\right) = 0 \tag{A6b}$$

The solution of the Euler Lagrange stationary point equation for the arc length in terms of $\theta = g(\phi)$ (given as Eq. ((A7)) can be rearranged by use of a trigonometric identity, to yield Eq. ((A8) (note C_1 and C_2 represent constants of integration).

$$\theta = -\sin^{-1}\left(\frac{\cot\phi}{\sqrt{\left(\frac{F}{C_1}\right)^2 - 1}}\right) + C_2. \tag{A7}$$

$$F\cos\theta\sin\phi\sin C_2 + F\sin\theta\sin\phi\cos C_2 - \frac{F\cos\phi}{\sqrt{\left(\frac{F}{C_1}\right)^2 - 1}} = 0. \tag{A8}$$

Recognition of the original spherical coordinate description of x, y and z in Eq. ((A8) produces Eq. (A9).

$$\sin C_2 x + \cos C_2 y - \frac{z}{\sqrt{\left(\frac{F}{C_1}\right)^2 - 1}} = 0. \tag{A9}$$

Closer inspection of Eq. (A9) reveals an equation for a plane (i.e. '$ax + by + cz = d$) which intersects the Cartesian axes at the origin i.e. $d = 0$. Therefore the minimum path length condition required by Eq. (A6) and shown in Eq. (A9) is that given by the intersection of a plane and a sphere i.e. a circle with its center located at the origin — a great circle.

References

[1] A.K. Dunker, M.S. Cortese, P. Romero, L.M. Iakoucheva, V.N. Uversky, Flexible nets: the roles of intrinsic disorder in protein interaction networks, FEBS J. 272 (2010) 5129–5148.
[2] N. Leulliot, G. Varani, Current topics in RNA-protein recognition: control of specificity and biological function through induced fit and conformational capture, Biochemistry 40 (2001) 7947–7956.
[3] S.A. Allison, G. Ganti, J.A. McCammon, Simulation of the diffusion-controlled reaction between superoxide and superoxide dismutase. I. Simple models, Biopolymers 24 (7) (1985) 1323–1336.
[4] E.W. Miles, S. Rhee, D.R. Davies, The molecular basis of substrate channeling, J. Biol. Chem. 274 (18) (1999) 12193–12196.
[5] R.M. Stroud, An electrostatic highway, Nat. Struct. Biol. 1 (1994) 131–134.
6 L. Euler, Concerning the shortest line on any surface by which any two points can be joined together, Comm: Ac. Scient. Petr. Tom. III1728. 110 (Translated & Annotated by Ian Bruce 17th, Century Mathematics Co.).
[7] T. Maekawa, Computation of shortest paths on free-form parametric surfaces, J. Mech. Des. Trans. ASME 118 (1996) 499–508.
[8] L. Shapira, A. Shamir, Local geodesic parametrization: an ant's perspective, Mathematical Foundations of Scientific Visualization, Computer Graphics, and Massive Data Exploration, Springer, Berlin Heidelberg, 2009, pp. 127–137.
[9] M. Dijkstra, A note on two problems in connexion with graphs, Numer. Math. 1 (1959) 269–271.
[10] R. Kimmel, J.A. Sethian, Fast marching methods on triangulated domains, Proc. Natl. Acad. Sci. 95 (1998) 8341–8435.
[11] T.H. Gronwall, On the shortest line between two points in non-Euclidean geometry, Ann. Math. Second Ser. 20 (1919) 200–201.
[12] G.W. Bushnell, G.V. Louie, G.V. Brayer, High resolution structure of horse heart cytochrome c, J. Mol. Biol. 214 (1990) 585–595.
[13] Huang, N. E. (1999). Computer Implemented Empirical Mode Decomposition Method, Apparatus and Article of Manufacture. U.S. Patent No. 5,983,162. Washington, DC: U.S. Patent and Trademark Office.
[14] C. Cortes, V. Vapnik, Support-vector networks, Mach. Learn. 20 (1995) 273–297.
[15] N. Tokuriki, D.S. Tawfik, Protein dynamism and evolvability, Science 324 (2009) 203–207.
[16] G. Celniker, G. Nimrod, H. Ashkenazy, F. Glaser, E. Martz, I. Mayrose, N. Ben-Tal, ConSurf: using evolutionary data to raise testable hypotheses about protein function, Israel J. Chem. 53 (2013) 199–206.
[17] A. Leach, Molecular modelling: principles and applications, Chapter 12 'The Use of Molecular Modelling and Chemoinformatics to Discover and Design New Molecules', 2nd ed., 2001.
[18] O. Trott, A.J. Olson, AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading, J. Comput. Chem. 31 (2010) 455–461.
[19] D. Hall, C.M. Dobson, Expanding to fill the gap: a possible role for inert biopolymers in regulating the extent of the 'macromolecular crowding' effect, FEBS Lett. *580 (2006) 2584–2590.*

Appendix References

[A1] R. Weinstock, Calculus of Variations, with Applications to Physics and Engineering, Dover, New York, 1974. (26-28 and 62-63).
[A2] M. Stone, P. Goldbart, Mathematics for Physicists, Cambridge University Press, Melbourne, 2010. pp. 10.