METHOD ARTICLE

Modelling structural rearrangements in proteins using Euclidean distance matrices [version 1; peer review: 2 approved with reservations]

Aleix Lafita , Alex Bateman

European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Hinxton, CB10 1SD, UK

V1 First published: 20 Jul 2020, 9:728
<https://doi.org/10.12688/f1000research.25235.1>
Latest published: 20 Jul 2020, 9:728
<https://doi.org/10.12688/f1000research.25235.1>

Abstract

Proteins undergo large structural rearrangements such as circular permutations, dimerisation via domain swapping, and loss of core secondary structure elements in domain atrophy, among others. These structural changes can be naturally represented as distance matrix transformations, exploiting their conserved native residue contacts at the protein core. Here we present an homology modelling approach to formulate structural rearrangements as a Euclidean distance matrix (EDM) problem and use it to build their 3D structures. This modelling approach aims to be lightweight, flexible and fast, suitable for large-scale analyses. Models are typically coarse-grained and solely based on protein geometry. We demonstrate various applications of EDM-based modelling for protein structure analysis and release an open repository with the source code at: <https://github.com/lafita/protein-edm-demo>.

Keywords

protein modelling, structural rearrangements, circular permutation, domain swapping, domain atrophy, Euclidean distance matrices

Open Peer Review

Approval Status ? ?

	1	2
version 1	?	?
20 Jul 2020	view	view

1. **Sergey Ovchinnikov**, Harvard University, Cambridge, USA
2. **Henry Wolkowicz**, University of Waterloo, Waterloo, Canada
Haesol Im, University of Waterloo, Waterloo, Canada

Any reports and responses or comments on the article can be found at the end of the article.

This article is included in the **EMBL-EBI** collection.

Corresponding authors: Aleix Lafita (aleixlafita@ebi.ac.uk), Alex Bateman (AGB@ebi.ac.uk)

Author roles: **Lafita A:** Conceptualization, Data Curation, Formal Analysis, Investigation, Methodology, Software, Writing – Original Draft Preparation; **Bateman A:** Conceptualization, Supervision, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: This work was funded by the European Molecular Biology Laboratory. AL further acknowledges support by the EMBL International PhD Programme.

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Copyright: © 2020 Lafita A and Bateman A. This is an open access article distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: Lafita A and Bateman A. **Modelling structural rearrangements in proteins using Euclidean distance matrices [version 1; peer review: 2 approved with reservations]** F1000Research 2020, 9:728 <https://doi.org/10.12688/f1000research.25235.1>

First published: 20 Jul 2020, 9:728 <https://doi.org/10.12688/f1000research.25235.1>

Introduction

Most proteins fold into compact globular domains stabilised by hydrogen bonds and a hydrophobic core. Protein domains typically allow for sequence mutations and short indels, which have a minor effect on the domain core. Some domains additionally allow for larger structural rearrangements that conserve the majority of core interactions. Some examples are domain swapping, where two independent domains exchange secondary structures to form an intertwined dimer¹ circular permutations, where the sequence order of the domain is rearranged, connecting the N and C-termini and introducing new termini at a specific "cut" position² and domain atrophy, where complete core secondary structures of a domain are deleted³.

These structural rearrangements have been extensively studied both experimentally and computationally due to their importance for protein stability and function^{4–6}. A number of methods have been devised to predict domains susceptible to circular permutations and domain swaps^{7–9}. However, modelling the 3D structure of the alternate conformations proves challenging, often requiring complicated software pipelines and long simulations^{8,10}. We present an alternative modelling approach that consists in representing domains as Euclidean distance matrices (EDMs) and generating rearranged structures by applying a series of matrix operations (Figure 1).

Figure 1. Protein modelling approach based on Euclidean distance matrices (EDMs). Atomic coordinates of the native SH3 domain (PDB:1SHG) structure (top left) are converted into a distance matrix (DM), shown at the bottom left as a heatmap of C-beta distances in a viridis color scale. A matrix transformation operation corresponding to the structural rearrangement is applied to the distance matrix (bottom middle), here shown as a red arrow for the circular permutation cut position and red boxes encapsulating conserved regions of the matrix. Values for unknown entries in the transformed distance matrix, shown as white stripes, are filled using an EDM completion algorithm to generate a complete EDM (bottom right). The distance matrix is finally converted to points in 3D corresponding to the atomic coordinates of the rearranged domain structure (top right).

Euclidean distance matrices (EDMs)

EDMs are matrices of squared distances between points in an m -dimensional Euclidean space¹¹. By definition, they are square and symmetric matrices with zero diagonal and non-negative off-diagonal values, fulfilling the triangle inequality. Thanks to their many useful properties they have been used for applications in sensor localisation, molecular conformation and dimensionality reduction, among others. In proteins, EDM-based algorithms have been previously used to model structures from distance constraints derived from nuclear magnetic resonance (NMR) experiments¹². EDMs are especially well suited for modelling protein structural rearrangements, since most of the distances at the domain core are conserved and can therefore be used as known entries in the matrix. Here we describe in detail a modelling approach based on EDMs and demonstrate its applications for studying circular permutations, domain swaps and domain atrophy in proteins.

Methods

Protein distance matrices

Given an input protein structure, its distance matrix can be computed as the Euclidean norm between all pairs of atoms. The distance matrix of a protein with n atoms will be n columns by n rows for a total of n^2 entries. The granularity of the distance matrix can be easily adjusted by changing the number and type of atoms included from each protein residue. For example, a coarse-grained representation of a protein can be achieved using only C-alpha atoms.

Distance matrices are invariant to point rotations, translations, and mirror images. This means that proteins rotated and translated in space have the same distance matrix. To the advantage of protein structure analysis, distance matrices of similar proteins can be directly compared without the need to superpose them in space. On the other hand, a protein and its non-natural mirror image are also indistinguishable from their distance matrix. Distance matrices can be inverted into a set of points in an embedding space of m dimensions using a technique known as multidimensional scaling (MDS). Distance matrix inversions can be challenging for noisy and incomplete matrices. Thanks to the degeneracy in low-rank protein distance matrices, where the number of points is much higher than the number of dimensions (three), MDS can handle small levels of noise and produce accurate reconstructions. Atomic coordinates of proteins can be generated from distance matrices using MDS in three dimensions. If the MDS point reconstruction results in a non-natural mirror image of a protein, its natural stereoisomer can be obtained by simply changing the sign of the z dimension of all points.

EDM completion

Partial or incomplete EDMs, where only a subset of distances in the matrix are known, are a common problem in many applications. Finding values for the missing entries such that the resulting matrix is still an EDM is known as the EDM completion (EDMC) problem. Several different algorithms formulated as optimisation problems have been devised to complete EDMs, ranging from semidefinite programming (SDP) to

numerical optimization, each with different loss functions and constraints.

EDMC algorithms only converge to an exact solution with zero loss, in which the resulting matrix is an EDM. In some cases, there are multiple exact solutions to the completion problem and algorithms will randomly converge to one of the solutions. In other cases, convergence will not be possible and algorithms will return a solution with minimal loss instead, corresponding to the closest matrix to an EDM.

For the purpose of protein modelling, the dissimilarity parameterization formulation (DPF) algorithm¹³ offers a suitable solution. The dimensions of the EDM embedding space and upper and lower bounds for unknown distances are built into the loss function and can be set as input parameters. Thus, atomic distance bounds can be used to avoid clashes and unrealistic conformations.

Protein distance geometry

To obtain realistic protein conformations in EDM models, we use atomic distances extracted from experimental structures in order to set upper and lower bounds on unknown entries in protein distance matrices. We selected one manual representative domain from each architecture type of the Evolutionary Classification of Protein Domains (ECOD) database¹⁴. The set of 20 domain structures is of diverse length, secondary structure content and fold topology, and determined by high-resolution X-ray crystallography (below 2Å).

In total, we calculated over three million distances between backbone atoms (N, CA, CB, C, O) and grouped them by atom types and residue number separation along the peptide chain. Separations above five residues were grouped together into one. For each of the 270 pairs of atom types and residue separation groups, we constructed a lookup table of with the average value, the minimum value (lower bound) and the maximum value (upper bound). Outliers, defined as distances outside the 1/1,000 percentile, were discarded in the upper and lower bound calculations, for robustness. Given a partial protein distance matrix, lower and upper bounds for unknown distances are set to the corresponding values in the lookup table.

The average, upper and lower bounds for C-alpha and C-beta distances are shown in Figure 2A. As expected, the uncertainty around the average distance increases with the residue separation. The lower bound, however, reaches a minimum value of 3.70Å for C-alpha and 3.53Å for C-beta contacts. Values below 3Å are observed for C-alpha atoms of adjacent residues (separation equal to 1), which correspond to cis-Proline conformations. In order to simplify the lookup table of atomic distance bounds, we discarded cis-Prolines. A subset of the final distance bounds lookup table for two adjacent residues is shown in Figure 2B. Dark colours in the table represent distances with low variation, for example those between atoms that form the peptide bond plane (CA0, C0, O0, N1, CA1).

Figure 2. Atomic distance bounds extracted from experimental domain structures. **A** C-alpha (CA) and C-beta (CB) interatomic distances as a function of residue separation along the peptide chain. The average distances are shown as a black line with a grey area for lower and upper bounds. Upper bounds trimmed at 20Å. **B** Table of average backbone interatomic distances between adjacent residues (0,1), excluding cis-Prolines. Colours indicate amount of variability as the difference between upper and lower bounds in Å, with darker colours for smaller variations.

Structural rearrangement examples

Proteins with known pairs of native and rearranged structures have been selected from the Protein Data Bank (PDB) and used as modelling examples. The use of pairs has allowed the comparison of EDM models to their experimentally determined structures. For domain atrophy, a pair of close homologs were selected instead due to the lack of domain atrophied structures from the same protein.

The procedure to model structural rearrangements using EDMs is schematically shown in Figure 1 and starts by parsing the atomic coordinates of the domain structure from a PDB file and extracting a subset of atoms, either C-alpha or backbone atoms. Next, the distance matrix is calculated and its entries are rearranged, corresponding to the structural changes to be modelled. Unknown distances are bounded using the lookup table of atomic distance bounds and filled using EDM completion. Finally, the atomic coordinates of the rearranged structure are obtained using multidimensional scaling and substituted into the original PDB file, which is saved as the new model.

Implementation

The EDM protein modelling approach presented here has been implemented in the R programming language, version 3.6. Protein structures are parsed and manipulated using the latest

version (2.4) of the *bio3d* package¹⁵. The *edmc* package version 1.0¹⁶ is used for EDM completion and multidimensional scaling. The implementation requires minimal dependencies and only a few lines of code, and runs on a single CPU. A general script for each type of structural rearrangement has been created so that it can be applied to any query protein by simply changing the path to the input file.

Results and Discussion

Circular permutations

A circular permutation of a protein domain is the alteration of its amino acid sequence order so that the N and C-terminal regions are interchanged at a specific “cut” position. The result is a new domain with a similar overall 3D structure and topology but different termini².

The core of the domain remains largely the same, so atomic distances can be rearranged in the matrix, as shown in [Figure 1](#). For a given native distance matrix D with n rows and columns, the distance matrix D_{cp} of a circular permutation at position k corresponds to:

$$D_{cp}[i, j] = D[(i + n - k) \bmod n, (j + n - k) \bmod n] \\ \text{FOR } i \in [1..n], j \in [1..n]$$

Since the old N- and C-terminal residues are now connected in the circular permutation, their distances to other atoms are set to unknown in order to remove their constraints and let them to freely adopt a new conformation, effectively unfolding them. A linker of any length between the old domain termini can also be added to the circular permutant by inserting columns and rows of unknown distances in the matrix. Next, upper and lower bounds for all unknown distances are set using the lookup table of distance bounds. Finally, the matrix is completed and the coordinates of the atomic 3D points are reconstructed as described in the Methods section.

The EDM completion algorithm will not converge to a solution if joining the domain termini is not geometrically possible. In that case, a longer linker or more unfolded terminal residues might have to be considered, as is the case for YibK methyltransferase. The distance between the N and C-terminal residues is 20Å, which means a linker is required to connect them. An experimental structure of a circular permuted YibK has five extra residues forming the termini linker loop¹⁷. Modelling the same circular permutation at position 82 with EDMs using only two unfolded terminal residues does not converge to a valid solution, with a large loss value over 20K and distance errors in the terminal loop region above 8Å. An EDM model with four unfolded residues neither converges to a solution, but has a smaller loss of 2K and errors on the order of 2Å. Using six unfolded terminal residues finally converges to an exact solution, shown in [Figure 3A](#).

Domain swapping

Domain swapping refers to the reciprocal exchange of equivalent secondary structure elements between protein domains⁶. The swapping mechanism requires the extension of a short

region of the native domain to form what is known as the “hinge loop” that bridges the two interacting domains. Although the core of the domain in a domain swap is formed by inter-domain interactions, as opposed to intra-domain interactions in the non-swapped conformation, its overall 3D structure is similar everywhere else apart from the “hinge loop” region. Therefore, like for circular permutations, atomic distances in the matrix can be rearranged to alter the pattern of residue interactions.

For a given distance matrix D from a monomeric domain with n atoms, the matrix D_{sw} for a domain swap with “hinge loop” centred at position k corresponds to a duplicated matrix with $2n$ columns and rows. Intuitively, interactions between atoms at each side of the “hinge loop” are changed to be inter-domain and interactions involving residues within each side are kept intra-domain, as shown in [Figure 3E](#). This corresponds to the following matrix index rearrangements:

$$D_{sw}[i, j] = D[i \bmod n, j \bmod n] \\ \text{IF } ([i, j] \bmod n) < k \text{ OR } ([i, j] \bmod n) > k \\ \text{OR } : [i, j] < k \text{ AND } [i, j] > (n + k) \\ \text{OR } : n < [i, j] < (n + k) \text{ AND } k < [i, j] < n \\ \text{FOR } i \in [1..2n], j \in [1..2n]$$

Similarly to the circular permutation case, distances involving residues that form the “hinge loop” have to be set as unknown to allow enough molecular flexibility to extend the loop and permit the interaction between the two domains. The number of “unfolded” residues needed as a “hinge loop” to make the domain swap geometrically possible is a parameter that has to be set before completing the matrix and can vary across domains.

The structures of Cyanovirin-N and EPS8 SH3 domains have been experimentally determined both as monomeric and domain swapped dimers. The EDM model of the Cyanovirin-N domain swap structure from its monomer closely resembles the experimental domain swap at position 51¹⁸, as shown in [Figure 3B](#). However, the EDM model of the EPS8 domain swap structure from its monomeric form converges to a valid solution, but has a different domain orientation to the experimental domain swap at position 34¹⁹, as shown in [Figure 3C](#). These results imply that the domain swap of Cyanovirin-N has a lower flexibility than EPS8. The relative orientation between Cyanovirin-N domains is constrained and cannot change without disrupting natural protein geometry, while EPS8 domains have a greater range of possible orientations.

To model domain swap dimers, atomic coordinates can be directly assigned to two different chains. Another case is that of tandem domain swaps, where the two domains forming the swapped conformation are part of the same protein chain. In tandem domain swaps, an additional loop is required to connect the termini of the central domain at indices $(n, n + 1)$, similar to the circular permutation case. The TAndem DOrain Swap Stability predictor (TADOSS) method⁹ already reports predictions for the “hinge loop” position and linker lengths required

Figure 3. Protein structure models generated using Euclidean distance matrices (EDMs). **A**) C-alpha model of the circular permutation of YibK methyltransferase as a rainbow backbone, superposed to the native structure (PDB: 1MXI) as transparent cartoon; **B**) Domain swap dimer model of Cyanovirin-N (PDB: 2EZM) as a rainbow backbone, superposed to its experimental domain swap dimer (PDB: 3EZM) as transparent cartoon; **C**) Domain swap dimer model of EPS8 SH3 domain (PDB: 110C) as a rainbow backbone, superposed to the left-most domain of its experimental domain swap dimer (PDB: 1107) as transparent cartoon; **D**) Backbone EDM model of domain atrophied Rib domain (PDB: 6SX1) from the structure of Rib Long (PDB: 6S5W); and **E**) C-beta distance matrices of monomeric EPS8 domain (PDB: 110C), its experimental domain swap dimer (PDB: 1107) and its model generated with EDMs. Conserved intra-domain contacts in the monomer distance matrix are highlighted as red boxes, while contacts swapped to inter-domain are highlighted as orange boxes.

for the formation of tandem domain swaps. Hence, EDM modelling can be directly used to generate 3D models for TADOSS predictions. This feature has been implemented in the latest version of the method. The possibility to generate models for TADOSS predictions not only improves result visualisation and interpretability, but offers a new avenue to improve its accuracy.

Domain swapping is also found as a crystallisation artefact^{20–22}. By inverting the domain swap matrix rearrangement, it is possible to “unswap” these structures using EDMs, facilitating its interpretation and comparison to other known domain structures.

Domain atrophy

Domain atrophy is the abrupt loss of entire secondary structures in the core of a protein domain, resulting in a shorter version of the domain³. Although it is a rare evolutionary event, it has been described for some of the most widespread and conserved domain folds, like immunoglobulins (Ig) and TIM barrels^{3,23}.

A structural deletion between positions k and l in a domain can be modelled using EDMs by removing all columns and rows

between indices k and l in the native distance matrix. Since residues k and l are connected through a peptide bond in the atrophied domain, constraints for adjacent residues from the lookup table of distance bounds have to be applied to the distance matrix accordingly.

Opposite to circular permutations and domain swaps, domain atrophy can have a more significant effect on the core of the domain, so it might have an impact on a higher fraction of native distances. As a result, the EDM completion is more likely to fail to converge to an exact solution. Proposed workarounds are to either “unfold” the residues around the deletion or to “relax” the native distances to intervals. Instead of using the exact distances, bounds can be set to a 10% interval of their native distances, for example. That way, the structure is given enough flexibility to adapt to the new conformation, while still conserving its overall 3D shape.

The domain atrophy in Rib domains involved the loss of a pair of beta strands core to its beta-sandwich topology²³. The distance between the two ends of the deleted segment between positions 55 and 72 is, however, close enough so that a short linker can connect them, as captured by the EDM atrophy model shown in Figure 3D. The atrophy leaves, however, part of the

domain core exposed to the surface. This gap is filled in the Rib domain by another part of the protein which adopts a helical turn conformation. This type of structural compensation is common in domain atrophy cases. Although such detailed structural consequences of domain atrophy can not be accounted for by EDM models, they provide a useful starting point to investigate them further.

Runtime and scalability

The size of a distance matrix scales quadratically with the number of atoms in the input structure. This means that the running time and memory of EDM algorithms is expected to be at least $O(n^2)$.

Calculation of distance matrices and multidimensional scaling operations are very fast, and take less than a second for the modelling examples presented in Figure 3. The matrix completion step is, however, the bottleneck, varying from just a few seconds to few minutes, heavily dependent on the size of the input matrix, but also on the number and complexity of unknown entries and a random convergence factor.

Running times for modelling structural rearrangements of several domain structures is shown in Figure 4. C-alpha models take less than a minute for up to 200 residues, which is within the size of an average protein domain and suitable for large-scale analysis. C-alpha models of domains up to 400 residues run in less than ten minutes. In comparison, backbone models do not scale particularly well, with an expected factor of 25 times slower than C-alpha models due to their five atoms per residue. Hence, backbone models for domains over 200 residues are possible but not practical for most applications.

Running times were calculated on a single standard CPU with 8GB of RAM. The small computational resources needed to create a model allows multiple modelling jobs to run concurrently in large-scale analyses.

Conclusions

We have described a new homology modelling approach for protein structures based on their distance matrices. The approach aims to be intuitive, flexible, light-weight and fast, being an alternative to more traditional modelling techniques based on simulation. Through a series of examples, we have demonstrated how EDMs can be handled to suit different modelling needs and its applications for protein structure analysis. We are also releasing an open code repository with an implementation of the method to the community, hoping that other researchers can explore the modelling cases presented in this study and extend the code for other applications.

Modelling protein structures with EDMs is well suited to tasks where a large proportion of atomic distances are known close to their exact values. As we have shown here, this is the case for structural rearrangements, where only a small subset of distances in the native matrix are perturbed. There are a number

Figure 4. Running time scaling of protein modelling using Euclidean distance matrices as a function of domain size.

Time calculated for backbone and C-alpha models of circular permutations (CP) and domain swap dimers (DSD) of the 20 ECOD representative domains. Backbone models for domains over 300 residues are omitted. Quadratic fits of the form $y = Ax^2$ for each model granularity are shown as lines with uncertainty grey shades. Running time shown as a square root scale.

of other applications in protein modelling where this condition holds and that we have not explored here, including modelling of multidomain proteins by concatenating tandem or nested domains, docking protein subunits based on a subset of interface contacts, modelling allosteric conformational changes, and incorporating distance constraints from experimental techniques (NMR, cross-linking) and sequence co-evolution inferences.

Modelling protein structures using EDMs is far behind state-of-the-art techniques based on atomic energies and simulations, and therefore has its own limitations. EDM models are solely based on geometry, and it is not trivial how to incorporate other more complex energetic terms essential to protein structures. Additionally, there is a lack of specialised software libraries for EDM-based protein modelling, making it difficult to write efficient code for more sophisticated applications.

The current implementation in R does not scale well and proves impractical for inputs over 600 atoms, mainly due to the matrix completion bottleneck. This means most of the models are restricted to be coarse-grained (C-alpha or backbone only), even though the same approach described here can be used to generate full side-chain models. Adding residue side-chains to backbone models is, however, a common task in protein modelling and good tools already exist, for example Chimera's Dock Prep²⁴. The actual computational costs of EDM modelling can be reduced much further. On the software implementation side, matrix operations in R are two orders of magnitude slower than in other programming languages like C++ and Java. The use of an alternative library for faster EDM completion will have a big impact on the running time, but we are unaware of

its existence. Algorithmically, the size of EDMs can be reduced exploiting rigid “cliques” of fully connected points. This technique is known as facial reduction and has been previously used to reduce the size of protein distance matrices¹².

Despite its current limitations, several protein modelling tasks can benefit from the EDM modelling approach described in this article. As shown for structural rearrangements, EDM models prove to be another valuable tool in the protein structure analysis toolbox.

Data availability

Underlying data

Zenodo: lafita/protein-edm-demo: Initial public release. <https://doi.org/10.5281/zenodo.3945383>²⁵ This project contains the following underlying data:

- protein_bounds.tsv (atomic distance bounds in Figure 2)
- runtimes_ecod.tsv (running time data underlying Figure 4)

Data are available under the terms of the [Creative Commons Attribution 4.0 International license](#) (CC-BY 4.0).

Code availability

Protein EDM modelling implementation in R.

- Source code available from: <https://github.com/lafita/protein-edm-demo>
- Archived source code at the time of publication: <https://doi.org/10.5281/zenodo.3945383>²⁵
- License: MIT license

TADOSS method incorporating automatic generation of tandem domain swap models.

- Source code available from: <https://github.com/lafita/tadoss>
- Archived source code at the time of publication: <https://doi.org/10.5281/zenodo.3939150>²⁶
- License: MIT license

References

- Rousseau F, Schymkowitz J, Itzhaki LS: **Implications of 3D domain swapping for protein folding, misfolding and function.** *Adv Exp Med Biol.* 2012; **747**: 137–52. [PubMed Abstract](#) | [Publisher Full Text](#)
- Uliel S, Fliess A, Unger R: **Naturally occurring circular permutations in proteins.** *Protein Eng.* 2001; **14**(8): 533–42. [PubMed Abstract](#) | [Publisher Full Text](#)
- Prakash A, Bateman A: **Domain atrophy creates rare cases of functional partial protein domains.** *Genome Biol.* 2015; **16**(1): 88. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Iwakura M, Nakamura T, Yamane C, et al.: **Systematic circular permutation of an entire protein reveals essential folding elements.** *Nat Struct Biol.* 2000; **7**(7): 580–585. [PubMed Abstract](#) | [Publisher Full Text](#)
- Yang S, Cho SS, Levy Y, et al.: **Domain swapping is a consequence of minimal frustration.** *Proc Natl Acad Sci U S A.* 2004; **101**(38): 13786–13791. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Lafita A, Tian P, Best RB, et al.: **Tandem domain swapping: determinants of multidomain protein misfolding.** *Curr Opin Struct Biol.* 2019; **58**: 97–104. [Publisher Full Text](#)
- Lo WC, Wang LF, Liu YY, et al.: **CPred: A web server for predicting viable circular permutations in proteins.** *Nucleic Acids Res.* 2012; **40**(Web Server issue): W232–W237. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Ding F, Prutzman KC, Campbell SL, et al.: **Topological determinants of protein domain swapping.** *Structure.* 2006; **14**(1): 5–14. [PubMed Abstract](#) | [Publisher Full Text](#)
- Lafita A, Tian P, Best RB, et al.: **TADOSS: computational estimation of tandem domain swap stability.** *Bioinformatics.* 2019; **35**(14): 2507–2508. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Tian P, Best RB: **Structural Determinants of Misfolding in Multidomain Proteins.** *PLoS Comput Biol.* 2016; **12**(5): e1004933. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Dokmanic I, Parhizkar R, Juri R, et al.: **Euclidean Distance Matrices: Essential theory, algorithms, and applications.** *IEEE Sign proc magaz.* 2015; **32**(6): 12–30. [Publisher Full Text](#)
- Alipanahi B, Krislock N, Ghodsi A, et al.: **Determining protein structures from NOESY distance constraints by semidefinite programming.** *J Comput Biol.* 2013; **20**(4): 296–310. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Trosset MW: **Distance matrix completion by numerical optimization.** *Comput Optim Appl.* 2000; **17**(1): 11–22. [Publisher Full Text](#)
- Cheng H, Schaeffer RD, Liao Y, et al.: **ECOD: An Evolutionary Classification of Protein Domains.** *PLoS Comput Biol.* 2014; **10**(12): e1003926. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Grant BJ, Rodrigues APC, ElSawy KM, et al.: **Bio3d: An R package for the comparative analysis of protein structures.** *Bioinformatics.* 2006; **22**(21): 2695–2696. [PubMed Abstract](#) | [Publisher Full Text](#)
- Rahman A, Oldford W: **edmc - Euclidean Distance Matrix Completion in R.** *J Stat Softw.* 2016; 1–40. [Reference Source](#)
- Chuang YC, Hu IC, Lyu PC, et al.: **Untying a Protein Knot by Circular Permutation.** *J Mol Biol.* 2019; **431**(4): 857–863. [PubMed Abstract](#) | [Publisher Full Text](#)
- Yang F, Bewley CA, Louis JM, et al.: **Crystal structure of cyanovirin-N, a potent HIV-inactivating protein, shows unexpected domain swapping.** *J Mol Biol.* 1999; **288**(3): 403–412. [PubMed Abstract](#) | [Publisher Full Text](#)
- Kishan KV, Newcomer ME, Rhodes TH, et al.: **Effect of pH and salt bridges on structural assembly: Molecular structures of the monomer and intertwined dimer of the Eps8 SH3 domain.** *Protein Sci.* 2001; **10**(5): 1046–55. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Xiaozhen Hu, Wang H, Ke H, et al.: **High-resolution design of a protein loop.** *Proc Natl Acad Sci U S A.* 2007; **104**(45): 17668–73. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Schlesier J, Rohde M, Gerhardt S, et al.: **A Conformational Switch Triggers Nitrogenase Protection from Oxygen Damage by Shethna Protein II (FeSII).** *J Am Chem Soc.* 2016; **138**(1): 239–47. [PubMed Abstract](#) | [Publisher Full Text](#)
- Hara K, Taharazako K, Ikeda M, et al.: **Dynamic feature of mitotic arrest deficient 2-like protein 2 (MAD2L2) and structural basis for its interaction with chromosome Alignment-maintaining phosphoprotein (CAMP).** *J Biol Chem.* 2017; **292**(43): 17658–17667. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Whelan F, Lafita A, Griffiths SC, et al.: **Defining the remarkable structural malleability of a bacterial surface protein Rib domain implicated in infection.** *Proc Natl Acad Sci U S A.* 2019; **116**(52): 26540–26548. [Publisher Full Text](#)
- Pettersen EF, Goddard TD, Huang CC, et al.: **UCSF Chimera - A visualization system for exploratory research and analysis.** *J Comput Chem.* 2004; **25**(13): 1605–1612. [PubMed Abstract](#) | [Publisher Full Text](#)
- Lafita A: **lafita/protein-edm-demo: Update license.** 2020. <http://www.doi.org/10.5281/zenodo.3945383>
- Lafita A: **lafita/tadoss: Tadoss v1.1.** 2020. <http://www.doi.org/10.5281/zenodo.3939150>

Open Peer Review

Current Peer Review Status: ? ?

Version 1

Reviewer Report 06 November 2020

<https://doi.org/10.5256/f1000research.27848.r67413>

© 2020 Wolkowicz H et al. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Henry Wolkowicz

Department of Combinatorics and Optimization, Faculty of Mathematics, University of Waterloo, Waterloo, ON, Canada

Haesol Im

University of Waterloo, Waterloo, Canada

Overview

This work focuses on how to formulate structural rearrangements of proteins as a Euclidean distance matrix completion (EDMC) problem with an homology modelling approach. The authors present three types of structural rearrangements that can be formulated as a EDMC problem; circular permutations, domain swapping and domain atrophy. These structural arrangements yield some missing entries in the Euclidean distance matrices. In order to fill in the missing entries, they adopt the dissimilarity parametrization formulation algorithm by Trosset.

For the purpose of obtaining realistic conformations using EDM model, the authors engage bounds on missing values, where the bounds are obtained from some experimental data. They also suggest some modelling heuristics in order to accommodate some difficulties that give rise to algorithmic failure.

The paper is well written with many interesting overviews of the underlying molecular problems, but with little of the mathematical background of EDM and of rigidity.

The main concern is that the numerics use one approach from 2000, i.e., one approach that does not allow readers to know if it is competitive with other approaches of the time and more recent methods. And, how does the noise in the data influence the results?

Questions and Suggestions

1. There are many papers on using EDM for molecular conformation. First there is a recent book on EDM¹ that emphasizes rigidity, a property that is important to exploit for molecular conformation and protein structure problems, as the backbone often is fixed and known, i.e., rigid.
2. The following is a quote from the submitted paper: "The actual computational costs of EDM modelling can be reduced much further. On the software implementation side, matrix

operations in R are two orders of magnitude slower than in other programming languages like C++ and Java". If you know that the implementation using R is slower than other programming languages, what is the reason for choosing this particular language?

3. The following is a quote from the submitted paper: "Distance matrices can be inverted into a set of points in an embedding space of m dimensions using a technique known as multidimensional scaling (MDS)". Some references would be nice so that readers who are not familiar with the subject can go find the relevant materials.
4. The only reason why the dissimilarity parametrization formulation (DPF) algorithm is chosen to solve the EDMC problem seems to engage the bounds on missing entries effectively. Are there any other reasons why the DPF is an appropriate choice of the algorithm? Are there any other algorithms or methods that appear in the literature that are able to handle this model?
The approach by Trosset is from 2000². There are many other approaches and many more recent. For example for noisy problems, see reference 3³ and the references therein, e.g., the papers by Alipanahi *et al.*,⁴.
5. The matrix rearrangement formula, appears in the middle of the second paragraph in domain swapping on page 4, is difficult to follow. A clearer and understandable description is required.
6. It is mentioned in Introduction that the structural rearrangements of interest have been extensively studied. However, in the numerical experiment presented in the submitted paper, no discussions on the comparative performances are made in order to address the strength of this new approach. Some discussions that engage the advantages of this approach over many other approaches are suggested.
 - Is the rationale for developing the new method (or application) clearly explained?
Yes. It is clearly explained but there is no proper rational for using only ONE method without any comparison with other methods. There are many other techniques around for (noisy) EDM completion problems
 - Is the description of the method technically sound?
Yes. The paper is well written and results/method well explained.
 - Are sufficient details provided to allow replication of the method development and its use by others?
Yes
 - If any results are presented, are all the source data underlying the results available to ensure full reproducibility?
Yes. The references for the data that is used is provided.
 - Are the conclusions about the method and its performance adequately supported by the findings presented in the article?
As mentioned above only one method (from 2000) is used. Moreover, the details of the influence of noise in the data is NOT clearly explained.

References

1. Alfakih AY: Euclidean Distance Matrices and Their Applications in Rigidity Theory. *Springer*. 2018.
2. Trosset M *Computational Optimization and Applications*. 2000; **17** (1): 11-22 [Publisher Full Text](#)
3. Drusvyatskiy D, Krislock N, Voronin Y, Wolkowicz H: Noisy Euclidean Distance Realization: Robust Facial Reduction and the Pareto Frontier. *SIAM Journal on Optimization*. 2017; **27** (4): 2301-2331 [Publisher Full Text](#)
4. Alipanahi B, Krislock N, Ghodsi A, Wolkowicz H, et al.: Protein Structure by Semidefinite Facial Reduction. 2012; **7262**: 1-11 [Publisher Full Text](#)

Is the rationale for developing the new method (or application) clearly explained?

Yes

Is the description of the method technically sound?

Yes

Are sufficient details provided to allow replication of the method development and its use by others?

Yes

If any results are presented, are all the source data underlying the results available to ensure full reproducibility?

Yes

Are the conclusions about the method and its performance adequately supported by the findings presented in the article?

Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Convex optimization; numerical linear algebra.

We confirm that we have read this submission and believe that we have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however we have significant reservations, as outlined above.

Reviewer Report 20 August 2020

<https://doi.org/10.5256/f1000research.27848.r67901>

© 2020 Ovchinnikov S. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Sergey Ovchinnikov

John Harvard Distinguished Science Fellowship Program, Harvard University, Cambridge, MA, USA

For this manuscript, the authors start by reviewing how Euclidean distance matrices (EDM) can be used to represent protein structures, highlighting the various pros of the approach. As an example, they highlight how EDM can be easily used to express and model circular permutations. Finally, they provide opensource R code.

Concerns/Questions:

1. One reason why structural modelers avoid working with EDMs directly is due to Chirality. Traditionally, one would work with coordinates directly, and use EDMs to calculate pairwise decomposable energies. Though the authors mention that chirality can be easily resolved after reconstruction (assuming the entire structure is of the same chirality), they do not go into detail of how chirality is handled during matrix-completion. It seems without explicit term to promote the same chirality through the entire structure, matrix-completion followed by MDS could return a structure that alternates between different chiralities. Maybe due to the shortness of the loops modeled this is not an issue? One way to check would be to compute the backbone dihedrals. For example, we'd expect the most of the ϕ dihedral between atoms c-n-ca-c to be negative... if the new modeled loop is positive, this could indicate a problem.
2. In conclusion, the authors write: "Modelling protein structures using EDMs is far behind state-of-the-art techniques based on atomic energies and simulations, and therefore has its own limitations." Given distances (aka EDMs) are used to represent pairwise decomposable terms in state-of-the-art techniques... it doesn't seem it would be that difficult to incorporate these energies in a technique that optimizes distance matrix values as opposed to coordinate values. Some energy terms that use distances include Lennard-Jones van der Waals, Coulomb electrostatics. Even angles and dihedrals could potentially be approximated as distances between $i, i+2$ and $i, i+3$. See Rosetta Energy function for reference:
<https://pubs.acs.org/doi/10.1021/acs.jctc.7b00125>
3. For modeling circular permutations, it seems one could just modify cartesian coordinates directly. Circular permutations would be as easy moving xyz entries from beginning to end of the list. It's not clear what advantage EDM provides, besides enabling the use of some of the matrix-completion algorithms for determining the best atom placements. What are the specific advantages over simply minimizing/placing coordinates directly?

Is the rationale for developing the new method (or application) clearly explained?

Partly

Is the description of the method technically sound?

Partly

Are sufficient details provided to allow replication of the method development and its use by others?

Yes

If any results are presented, are all the source data underlying the results available to ensure full reproducibility?

Yes

Are the conclusions about the method and its performance adequately supported by the findings presented in the article?

Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: I work on developing methods for protein structure prediction.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

The benefits of publishing with F1000Research:

- Your article is published within days, with no editorial bias
- You can publish traditional articles, null/negative results, case reports, data notes and more
- The peer review process is transparent and collaborative
- Your article is indexed in PubMed after passing peer review
- Dedicated customer support at every stage

For pre-submission enquiries, contact research@f1000.com

F1000Research