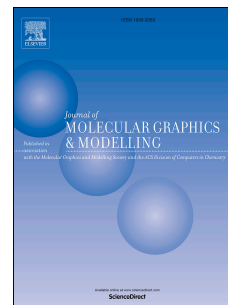


Accepted Manuscript

Morphing and docking visualisation of biomolecular structures using Multi-Dimensional Scaling

Ruth Veevers, Steven Hayward



PII: S1093-3263(18)30198-0

DOI: [10.1016/j.jmgm.2018.04.013](https://doi.org/10.1016/j.jmgm.2018.04.013)

Reference: JMG 7158

To appear in: *Journal of Molecular Graphics and Modelling*

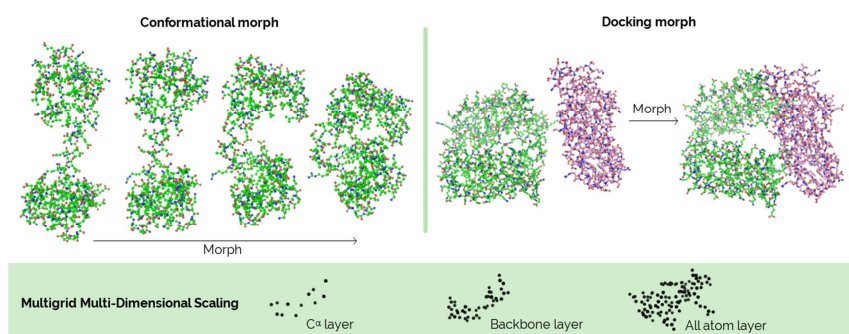
Received Date: 28 March 2018

Revised Date: 20 April 2018

Accepted Date: 22 April 2018

Please cite this article as: R. Veevers, S. Hayward, Morphing and docking visualisation of biomolecular structures using Multi-Dimensional Scaling, *Journal of Molecular Graphics and Modelling* (2018), doi: 10.1016/j.jmgm.2018.04.013.

This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.



Morphing and Docking Visualisation of Biomolecular Structures using Multi-Dimensional Scaling

Ruth Veevers¹ and Steven Hayward^{1,*}

¹Computational Biology Laboratory, School of Computing Sciences, University of East Anglia, Norwich, NR4 7TJ, UK.

*To whom correspondence should be addressed.

Keywords: Conformational change; SMACOF; Multigrid methods; MolProbity

Abstract

Protein structures are often solved at atomic resolution in two states defining a functional movement but intervening conformations are usually unknown. Morphing methods generate intervening conformations between two known structures. When viewed as an animation using molecular graphics, a smooth, direct morph enables the eye to track changes in structure that might be otherwise missed. We present a morphing method that aims to linearly interpolate interatomic distances and which uses SMACOF (Scaling by MAjorisation of COmplicated Function) and multigrid techniques with a cut-off distance based weighting that optimizes the MolProbity score of intervening structures. The all-atom morphs are smooth, move directly between the two structures, and are shown, in general, to pass closer to a set of known intermediates than those generated using other methods. The techniques are also used for docking by putting the unbound structures in a “near-approach pose” and then morphing to the bound complex. The resulting GPU-accelerated tools are available on a webserver, Morphit_Pro, at <http://morphit-pro.cmp.uea.ac.uk/> and more than 5000 domains movements available at the DynDom website can now be viewed as morphs <http://morphit-pro.cmp.uea.ac.uk/dyndom/>.

1 INTRODUCTION

Conformational change and protein function are intimately linked. Receptors undergo conformational change upon ligand binding (Remy, et al., 1999) and enzymes change conformation upon formation of the enzyme-substrate complex (Hammes, 2002). A common type of conformational change upon ligand binding is a domain movement (Amemiya, et al., 2011; Amemiya, et al., 2012; Bennett and Huber, 1984; Gerstein, et al., 1994; Hayward, 1999; Hayward, 2004; Qi and Hayward, 2009; Schulz, 1991), whereby a ligand binds to an open-domain conformation inducing a closed-domain conformation with the ligand buried in the interdomain cleft.

Structures deposited in the Protein Data Bank (PDB) (Berman, et al., 2000) provide a rich source of information on functional movements. Recent advances in cryo-electron microscopy have revealed atomic resolution snapshots of the ribosome during the process of translation (Loveland, et al., 2017) and the high time resolution of X-ray free electron lasers also promises to provide snapshots of biomolecules undergoing functionally related conformational change (Kupitz, et al., 2017). This suggests that in the near future multiple conformations representing functionally relevant states of a biomolecule will be deposited in the PDB. However, static structures do not directly reveal the path taken between them.

Morphing techniques provide structures on the path between two known structures. The optimal way to computationally determine the path is to use Molecular Dynamics (MD) simulation. However, this is normally not practicable as it is computationally expensive, often taking weeks or months to generate trajectories and is consequently only viable for long-term, in-depth research on a particular target protein. Coarse-graining methods can be used to overcome this limitation, in particular Elastic Network Models (ENM) where only C^α atoms are included. In the "Plastic Network Model" by Maragahis and Karplus (Maragakis and Karplus, 2005) the intermediate structures are those on the minimum energy path between the two conformations where the energy is constructed from two ENMs, one for each structure. In a related method, "MinActionPath" (Franklin, et al., 2007) the path taken minimizes the Onsager and Machlup action.

An obvious approach to morphing is to use linear interpolation of the Cartesian coordinates as first reported by Vonrhein et al. (Vonrhein, et al., 1995). It is used at the MolMovDB webserver (Krebs and Gerstein, 2000) and by MORPH-PRO

(Castellana, et al., 2013). Linear interpolation of Cartesian coordinates grants a quick result but intermediate structures are often infeasible; atoms can pass through each other and bonds can be compressed or stretched beyond reasonable limits making the resulting energy improbably high. MolMovDB, an all-atom method, overcomes this by performing energy minimization at each step and MORPH-PRO by applying a correction to keep consecutive C^α to C^α distances close to 3.8 Å. A further limitation of Cartesian coordinate interpolation is that results depend on the relative orientations of the two structures.

Linear interpolation of internal coordinates can overcome some of these limitations and has been implemented in the LSQMAN program (Kleywegt, 1996). However, these methods still produce high-energy distortions. For example, in a loop with fixed end positions, linear interpolation of the ϕ -, ψ -angles will cause the ends to move and it requires inverse kinematics techniques to keep them fixed (Hayward and Kitao, 2010).

Interpolation of interatomic distances is independent of the relative orientation of the two structures and also has the advantage of being easy to implement as it does not require identification of the covalent topology of the molecules – in terms of implementation chain breaks cause no problems and there is no difference between monomeric and oligomeric structures. Interatomic distances are used in CLIMBER (Weiss and Levitt, 2009), an all-atom method, which at each step minimizes the sum of two energy terms, one based on the difference between C^α - C^α distances in the current and the target structure, the other the total internal energy evaluated using a standard force-field. Kim et al. (Kim et al., 2002) describe a method in which C^α - C^α distances are interpolated linearly between the start and end structures with an ENM-based “cost” function being minimised at each step to construct intermediate structures. This method has been implemented at the NOMAD-REF webserver (Lindahl, et al., 2006).

An alternative approach is taken by FATCAT (Ye and Godzik, 2004) which has as its aim the optimal structural alignment of the two structures achieved by rigid body rotations of substructures about a minimal number of hinge points. It is these rigid body rotations that are interpolated and available from the FATCAT webserver.

The As-Rigid-As-Possible (ARAP) approach (Nguyen, et al., 2017) applies mesh distortion techniques used in computer graphics. A topology is created based

on atomic bonding, and the rotation of each atom and its connected neighbours, or “cell”, is calculated by minimizing the resulting cell energy. The rotation in each frame is derived using spherical linear interpolation, and then after linearly interpolating the position of an arbitrarily chosen atom between its start and end position, the position of each atom is constructed by minimizing the total ARAP energy.

Here we use Multi-Dimensional Scaling (MDS) methods to construct structures from linearly interpolated interatomic distances. MDS has a long-standing history and is primarily used to construct points in a 2D space for visualization of objects for which only a set of pairwise dissimilarities are known (Cox and Cox, 2008). In protein research, MDS methods are used in Nuclear Magnetic Resonance spectroscopy (NMR) to determine structure from a set of interatomic distances (Havel, 1991). Although an MDS method, the approach of Kim et al. (Kim, et al., 2002) was put in the context of an ENM and usual MDS techniques were not used. Here we apply MDS to construct intermediate structures from a linear interpolation of interatomic distances.

Conformational changes frequently accompany the process of biomolecular interaction during the formation of complexes. In these cases it can be informative to visualise the intramolecular changes of each molecule in relation to the other. In addition to conformational morphing, we also present a morphing application tailored to morphs depicting the process of protein docking. Docking refers to the computational endeavour to predict the binding pose of two biomolecules that are known to form a specific complex given the structures of two unbound molecules (Huang, 2015). Whilst treating molecules as rigid bodies is computationally convenient, flexibility of both backbone and side chain atoms has been shown to be important to the success of docking even when the conformational changes that occur are very small (Ehrlich, et al., 2005). Therefore docking techniques attempt to incorporate flexibility in various ways (Bonvin, 2006). Irrespective of the methodology employed it is apparent that modelling flexibility is a goal for current docking prediction techniques. Candidate structures found by docking methods will therefore vary not only in the relative pose of the two molecules but also in their conformations. This would make the visualisation of the conformational changes that occur upon docking particularly instructive. Our docking server produces animations showing each constituent moving into its docked configuration as well as

intramolecular conformational changes. As with the conformational morphing, docking visualisation is useful in that it allows the user to track movement they might otherwise have missed by showing each atom move from its start point to end point, and showing areas where parts of the proteins would have to move out of the way. Protein visualisation and animation software that currently exists such as UCSF Chimera (Pettersen, et al., 2004) typically uses a rigid-body linear interpolation of start and end positions provided by the users, with any morphing happening in a separate process. The web server MovieMaker (Maiti, et al., 2005) automates the position and trajectory of the constituents from an input docked complex, but operates entirely rigidly, offering no input for the undocked constituents. It is our hope that by providing protein docking teams with visualizations of potential docking trajectories, whether of experimentally observed structures, successfully predicted near-native conformations, or of false positive (or decoy) conformations, these tools will help teams to further improve their methods. Furthermore, animated, interactive visualisations are also helpful for demonstration purposes.

Our all-atom morphs are smooth, direct, and overall achieve a better improvement score on Weiss and Levitt's set of intermediate structures than other methods. Using multigrid methods and GPU-acceleration techniques, our implementation is fast enough for it to be made available on a webserver, Morphit_Pro.

2 METHODS

2.1 Interatomic Distance Interpolation

The aim is to interpolate between two known structures A and B (referred to as start and end structures, respectively). At time $t=0$ let the protein be at known structure A and at time $t=T$ at known structure B . Let $\lambda_{ij}(t)$ represent continuous functions of t such that $\lambda_{ij}(0)=0$ and $\lambda_{ij}(T)=1$, for all $i=1, n$ and $j=1, n$; $i \neq j$, where n is the number of atoms. Any path between A and B can be expressed in terms of the interatomic distances in the form:

$$d_{ij}(t) = (1 - \lambda_{ij}(t)) d_{ij}^A + \lambda_{ij}(t) d_{ij}^B \quad (1)$$

where d_{ij}^A and d_{ij}^B are the atomic distances between atoms i and j in structures A and B , respectively. Let us divide this time period into N intervals to give a time interval, $\Delta t = T/N$. Frames at times $t = k\Delta t$, are indexed $k = 0..N$; that is from A to B inclusively. Following Kim et al (Kim, et al., 2002), we presume a linear path where $\lambda_{ij}(t) = \lambda = k\Delta t/T$, for all i, j . A linear path in interatomic distances would, if it were possible to achieve, keep distances between atoms within their values in structures A and B , preventing them from clashing, and keeping bond lengths and bond angles within reasonable limits. At each frame, k , MDS allows one to construct the atomic coordinates $\mathbf{r}(k)$ from the set of linearly interpolated interatomic distances, $d_{ij}(k)$. There are a number of variants to MDS but here we use classical MDS and metric MDS utilizing the SMACOF (Scaling by MAjorisation of COmplicated Function) algorithm and multigrid methods.

2.2 Classical MDS

MDS is a technique commonly used to visualize dissimilarities among sets of items by representing them as points in space, often a 2D plane. The distance between each pair of items in this space reflects the dissimilarity between them. The goal of MDS is to take a matrix of dissimilarity values and construct a set of points with inter-point distances matching as close as possible the dissimilarities. Here $d_{ij}(k)$ represent the dissimilarities at frame k on which MDS is performed to determine the atomic coordinates, $\mathbf{r}(k)$.

Classical MDS (Cox and Cox, 2008) constructs an inner product matrix from the $d_{ij}(k)$ and performs an eigenvalue decomposition to determine the coordinates $\mathbf{r}(k)$. One can judge how well the constructed coordinates reproduce the desired $d_{ij}(k)$ by calculating the “strain”. If the interatomic distances are from a real structure, as would be the case for structures A or B , then the strain is zero and the constructed structure would have a Root Mean Square Deviation (RMSD) of zero with the real structure (apart from when it is possibly the enantiomeric structure, see below). However, $d_{ij}(k)$ at intervening frames are not from a real structure and the strain is not likely to be zero meaning that in the constructed structure not all interatomic distances can be simultaneously satisfied. We noticed that some of the structures resulting from Classical MDS had distorted covalent structures, e.g. compressed or stretched bonds. This is due to the solution being a compromise

between satisfying short-range and long-range distances. This suggests that shorter distances should be weighted more in order to maintain the covalent structure and prevent interatomic clashes. Weighting can be implemented in “Metric MDS”.

2.3 Metric MDS

Metric MDS adjusts each atom’s position to minimise the “stress”, $\sigma(\mathbf{r}(k))$ at frame k , which is given by:

$$\sigma(\mathbf{r}(k)) = \sum_{i < j}^n w_{ij} \left(|\mathbf{r}_i(k) - \mathbf{r}_j(k)| - d_{ij}(k) \right)^2 \quad (2)$$

where $\mathbf{r}_i(k)$ is the position vector for atom i , $\mathbf{r}_j(k)$ is the position vector for atom j , and w_{ij} is the weight applied to the pair i,j . Using a cut-off distance r_C we determine the set of atom pairs in structure A and the set of atom pairs in structure B that are within this cut-off distance. Atom pairs in the union of these two sets have $w_{ij}=1$, otherwise $w_{ij}=0$. The value of r_C is to be determined by optimizing a morph’s MolProbit score. We use the coordinates from Classical MDS as the starting coordinates for metric MDS.

$\sigma(\mathbf{r}(k))$ is a non-linear function of the coordinates. The SMACOF algorithm uses de Leeuw’s iterative majorization process (de Leeuw, 1988), which has been proven to be an efficient algorithm that decreases stress monotonically. At each iteration of the SMACOF process a so-called Guttman transformation is solved by using the Moore-Penrose inverse. A GPU-based multigrid acceleration approach was implemented for speed improvements.

2.4 Multigrid acceleration

Our multigrid acceleration was based on the implementation provided by the Toolbox for Surface Comparison and Analysis (Bronstein, et al., 2006; Rosman, et al., 2008) with changes to the code in order to implement a three-level cut-off based weighting and GPU acceleration.

The multigrid acceleration method creates a hierarchy of points, where each level of the hierarchy has a lower resolution than its predecessor. The multigrid method as detailed by Bronstein et al. (Bronstein, et al., 2006) uses this hierarchy,

along with matrices referred to as the interpolation and restriction operators to construct a MDS solution for $r(k)$.

The restriction operator for each level l is a sparse $n_{l+1} \times n_l$ matrix, P_l^{l+1} , that describes how points are restricted to a coarser level ($n_{l+1} < n_l$). The term at (i,j) will be 1 if the j^{th} point in level l is restricted to the i^{th} point in coarser level $l+1$, otherwise 0.

The interpolation operator, P_l^{l-1} is an $n_{l-1} \times n_l$ matrix that describes how points are interpolated to a finer layer. The term at (i,j) will be 1 if the j^{th} point in level l is interpolated to the i^{th} point in level $l-1$, otherwise 0.

In the conformational morph, and when proteins are input for docking morphs, we construct our hierarchy from the three level hierarchy inherent in protein structure. Level 1 comprises all atoms (finest level of detail), level 2 comprises backbone atoms only, and level 3, C^α atoms only. The interpolation matrix, P_3^2 interpolates from each residue's C^α atom to its backbone atoms, and P_2^1 interpolates from each residue's backbones atoms to all of the corresponding residue's atoms. The restriction operators are given by $P_1^2 = (P_2^1)^T$ and $P_2^3 = (P_3^2)^T$ where T denotes the transpose. Each level of the hierarchy can be assigned an appropriate cut-off distance, r_c^l , for weighting.

The multigrid algorithm uses so-called "V-cycles" (Bronstein, et al., 2006). Starting at the all-atom grid level the SMACOF result for $r(k)$ is restricted to the backbone level where further SMACOF iterations are performed. This is repeated for the backbone to C^α level after which the results are interpolated from coarse to fine grid levels again performing SMACOF iterations at each level. This constitutes one V-cycle.

The many matrix multiplications required made the process slow for large proteins. By moving these calculations onto the GPU, the process could be accelerated.

The values of r_c^l , $l = 1,2,3$ are to be determined by optimizing the morphs' MolProbity scores.

2.5 Docking Morphs

The docking morphs are produced using the same multigrid SMACOF approach to MDS, but the pre-processing and weighting steps are tailored to the problem of two molecules coming together.

Morphs are constructed using three structures: the unbound “receptor”, the unbound “ligand”, and the bound complex containing both. The unbound structures are initially superposed onto their respective bound structures in the complex and then the structure of the unbound ligand is moved away from the receptor along the line joining their centres of mass until no intermolecular atomic distance is less than the cut-off distance, r_c^l , ($l = 1$ for all-atom and $l = 3$ for C $^\alpha$ -atom only docking) ensuring that there are no clashes at the start of the morph. We call this conformation the “near-approach pose”. Thus the start structure is the unbound structures in the near-approach pose and the end structure is the complex structure. From the near-approach pose there is only a small movement to the docked pose allowing one to focus on the intramolecular conformational changes that occur.

Atom pairs for the three-level multigrid weighting scheme are selected for both of the unbound structures as the start structure and the complex structure as the end structure. Intermolecular atomic distances of the start structure in near-approach pose are not included to avoid artefacts that might arise from it not being a true structure. In addition to proteins, for which the three-level hierarchy is constructed as above, the docking morph server supports RNA, DNA and other types of molecules as input. For RNA and DNA molecules a similar hierarchy is constructed in which level 1, the finest level of detail, comprises all atoms, level 2 comprises each residue’s sugar-phosphate backbone, and level 3 comprises only the backbone’s 5’ carbons. For all other residues, if either set of expected backbone atoms are present then the residue’s atoms are included in the hierarchy as if it were an amino acid or base. Otherwise, the first atom in the residue is included at the coarsest level and all other atoms are included at levels 1 and 2.

2.6 Structural Alignment

The structures resulting from MDS are constructed to best reflect the distances between atoms, but coordinates are not fully determined in that all translated, rotated and enantiomeric (mirrored) structures are equally valid. The morph structures are brought into structural alignment for two reasons: first, the structures must be appropriately aligned between frames so that the animation appears smooth and second, chirality must be maintained.

The method used is a least-squares best-fitting procedure (this is also known as the “Procrustes analysis”) commonly used in structural bioinformatics but also

includes, in addition to translation and rotation, mirror inversion. This process brings each intermediate frame into alignment with the starting coordinates. For conformational morphs superposition is over the whole protein, whereas for morphs displayed at the DynDom website, superposition is on the fixed domain to clearly demonstrate the domain motion. For docking morphs superposition is on the receptor.

2.7 Server

Morphit_Pro is available to run from a GPU-server <http://morphit-pro.cmp.uea.ac.uk/> and the results displayed using the molecular graphics program, JSmol. In addition, a database of 5,251 morphs has been constructed from pairs of protein conformations taken from the non-redundant database (Qi, et al., 2005) and the user-created database (Lee, et al., 2003). These can be viewed at the new DynDom website, currently available at <http://morphit-pro.cmp.uea.ac.uk/dyndom/>. Furthermore, the morphing software has also been integrated into the new DynDom protein domain movement analysis webserver (Lee, et al., 2003), producing a morph viewable using JSmol whenever a pair of protein structures are successfully analysed by DynDom using the Run App/Run DynDom option.

The MDS method requires two sets of coordinates for each atom, one from each structure. In order to achieve the necessary equivalence at the atomic level, the server performs alignment of the amino acid/nucleotide sequences removing atoms from inserted residues and in turn alignment of atom types within each matched residue, removing atoms with atom types that are not common to both.

As well as the conformational morphing method, the web server has a protein docking morph application and associated database. This application takes as input one PDB file containing the complex structure, and two PDB files each containing one of the corresponding unbound structures. The server first identifies all molecules within the complex structure and the unbound structure files allowing the user to identify the corresponding molecules for morphing. Molecules can be proteins, DNA, RNA, or indeed any ligand as identified by "HETATM" in the PDB file format.

By default, both morphs and docking morphs are calculated using all atoms, but the user may request a C^α atom only morph for speed, or it may be necessary when the input is very large (over 10,000 atoms) due to memory limitations on the

GPU. The C^α atom only morph uses the SMACOF method without multigrid acceleration, with a cut-off distance from the C^α atom layer, r_C^3 , of the all atom multigrid approach.

Conformational morphs and docking morphs can be downloaded in PDB format for display with other molecular graphics programs such as ProteinViewer (Matthews, et al., 2017) or Pymol for high-quality rendering for presentation purposes.

The server runs CPU calculations on an Intel Core i7-7700 CPU @ 3.60GHz processor with 32GB RAM and GPU calculations on a GeForce GTX 1080 Ti.

3 RESULTS

3.1 MolProbity-tuned cut-off distance for weighting

Among the values calculated by the MolProbity validation tool (Chen, et al., 2010) is a single score which can be used as an overall measure of quality for the structures generated in a morph. This MolProbity score, $S_{molprob}$, is given as:

$$S_{molprob} = 0.43 \ln(1 + clash) + 0.33 \ln(1 + \max(0, rota - 1)) + 0.25 \ln(1 + \max(0, 100 - rama - 2)) + 0.09 \quad (3)$$

where *clash* is the number of atoms that overlap by at least 0.4Å per 1,000 atoms, *rota* is the percentage of sidechain rotamers classed as outliers, and *rama* is the percentage of Ramachandran conformations outside favoured regions. The lower $S_{molprob}$ the better the quality.

Our aim here is to determine the set of cut-off distances, r_C^l , $l=1,2,3$, that optimizes $S_{molprob}$ over the morphs. To do this we morphed a sample of 100 proteins taken from the non-redundant database of protein domain movements (Qi, et al., 2005) (see Supplementary material for the list of PDB structures used). It became apparent that the time required to find optimal values for r_C^l at all three levels far exceeded practical limits so we set $r_C^l = lr_C^1$, $l=2,3$; that is we imposed a linear relationship based on the parameter r_C^1 only. A series of values for r_C^1 between 2Å and 10Å at an interval of 0.5 Å were tried and for each value of r_C^1 a thirteen-frame

morph ($N=12$) was produced. At each frame of the morph, $S_{molprob}$, was calculated to determine the quality of the structure.

Figure 1 shows a typical trajectory for $S_{molprob}$ for the protein calmodulin (PDB: 1c1l chain A, PDB: 1cm1 chain A). For each protein, r_C^1 , was recorded that gave the lowest peak score along the morph. Plotting a frequency distribution for the optimal values of r_C^1 , as in Figure 2, allowed us to identify a value of r_C^1 between 3.5 Å and 4.5 Å that most commonly gave the lowest peak score. We re-ran this experiment with the same 100 proteins, focusing in on r_C^1 between 3.5 Å and 4.5 Å with the smaller interval of 0.1 Å. We identified $r_C^1=4.0$ Å to be the most commonly optimal cut-off distance. Morphs produced with $r_C^1=4$ Å will be referred to as “MDS_4”. This means that at the backbone level, 2, $r_C^2=8$ Å, and at level 3, the C $^\alpha$ level, $r_C^3=12$ Å. In this sense we have parameterized the cutoff distance at the C $^\alpha$ atom level and we will use a cut-off of 12 Å for morphs that use C $^\alpha$ -atoms only, including docking morphs. These morphs will be referred to as MDS_CA12.

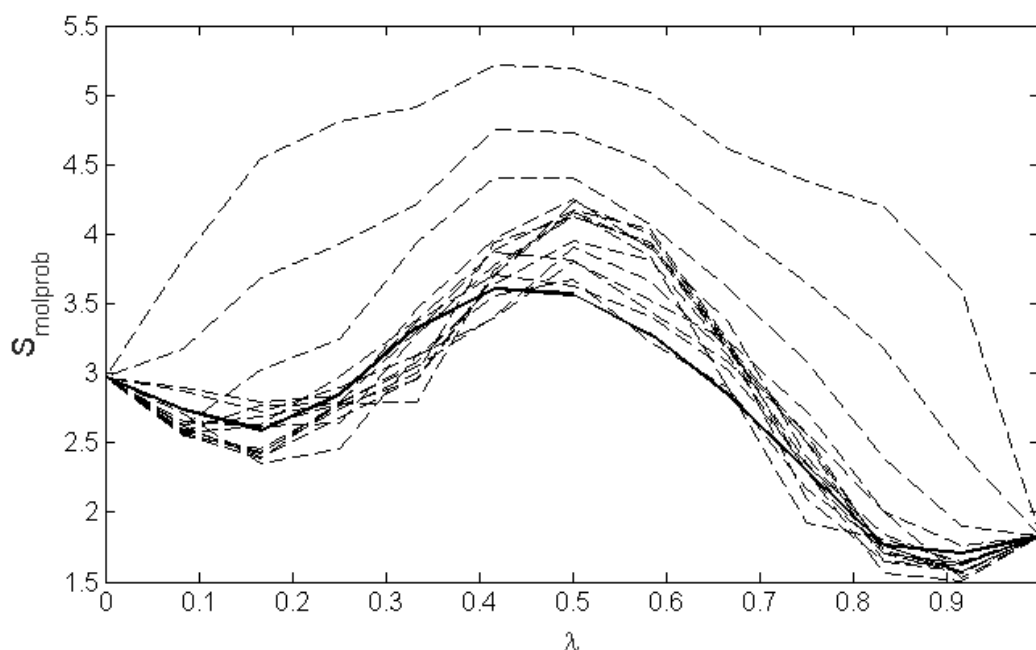


Figure 1: Plots of the MolProbity score, $S_{molprob}$, versus λ for r_C^1 in the range 2-10 Å at intervals of 0.5 Å for the morph of calmodulin (PDB codes 1CLL [A] to 1CM1 [A]). The thick line has the lowest peak value which occurs when $r_C^1 = 4.0$ Å.

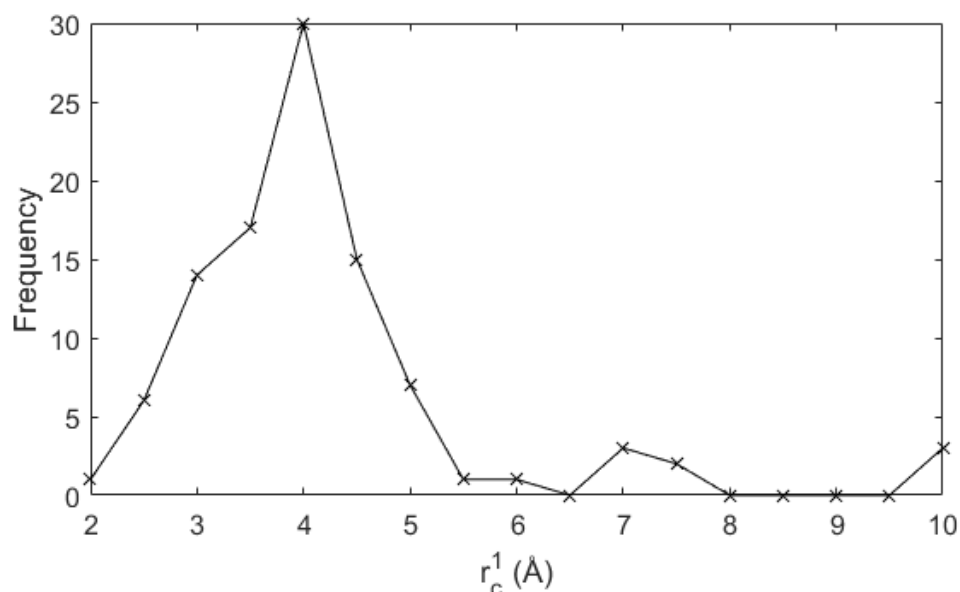


Figure 2: For 100 protein sample, the number of proteins is plotted against the value of r_c^1 that yielded the best (lowest) MolProbity peak.

3.2 Directness of morphs

In the absence of any other information, the path the morph takes between the two structures should be as direct as possible without violating structural constraints. This would avoid arbitrary detours that could make visual tracking of the various changes that take place more difficult. In order to judge the directness of the morphs, we evaluated $\text{RMSD}(k,A)$ and $\text{RMSD}(k,B)$ at each frame k using the all-atom method, MDS_4. Figure 3 shows the plot of $\text{RMSD}(k,A)$ and $\text{RMSD}(k,B)$ against λ ($\lambda=k/N$) for the 100 protein samples, with $N=12$, removing cases in which both structures were divided into parts separated by a distance greater than the 4 Å cut-off distance due to missing atoms or excised insertions from the structural alignment. The resulting figure shows that the morphs generally take a direct path. This means that with a value of $\lambda=0.5$, the method gives a structure that is approximately halfway between structures A and B in terms of RMSD.

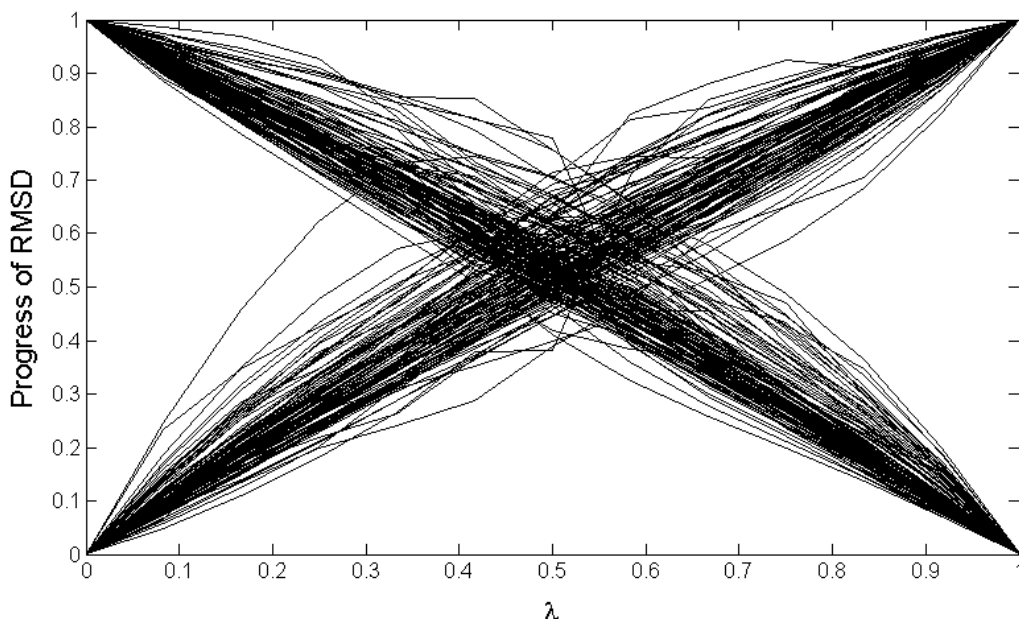


Figure 3: For 100 protein sample, the RMSD between the structure at each frame and the start and end structures is plotted against λ .

3.3 Comparison with other methods using Weiss and Levitt set of intermediate structures

Weiss and Levitt (Weiss and Levitt, 2009) identified a set of five proteins, Myosin, Ribose-Binding Protein, RNase III, 5'-Nucleotidase, and Ca^{2+} ATPase, for which a crystallographic intermediate structure, I , is available. Figure 4 shows trajectories of the RMSD between each MDS_4 morph structure and the intermediate structure for each of the five examples showing that for all but Ca^{2+} ATPase (Fig4(E)) a structure along the morph is closer to the intermediate than both start and end structures.

Weiss and Levitt devised a measure of quality of each morph based on how close the morph passes to I . The so-called “improvement score”, S_{impr} is given by:

$$S_{impr} = \frac{\min[\text{RMSD}(A,I), \text{RMSD}(B,I)] - \min[\text{RMSD}(k,I)]}{\min[\text{RMSD}(A,I), \text{RMSD}(B,I)]} \times 100 \quad (4)$$

where RMSDs are calculated using C^{α} atoms only. We compare the improvement scores of MDS_4 and MDS_CA12 with those of five other methods (using all atom

versions of the methods if available and C^α only otherwise): Climber (all atom), FATCAT (C^α), MinActionPath (C^α), MolMovDB (all atom) and Nomad Ref (C^α). Figure 5 shows the improvement scores for these five methods taken from the Weiss and Levitt paper, plus the improvement scores for MDS_4 and MDS_CA12. Table 1 gives the number of proteins for which the morphing method indicated by the row has a higher value for S_{impr} than the method indicated by the column. When compared against each other method, both MDS_4 and MDS_CA12 achieve a higher improvement score on the majority of proteins. Unexpectedly, MDS_CA12 does better than MDS_4.

	MDS_4	MDS_CA12	Climber	FATCAT	MinActionPath	MolMovDB	Nomad-Ref
MDS_4	-	1	3	3	4	4	3
MDS_CA12	4	-	4	4	4	4	4
Climber	1	1	-	2	4	3	1
FATCAT	2	1	3	-	3	3	3
MinActionPath	1	1	1	1	-	3	2
MolMovDB	1	1	1	1	1	-	2
Nomad-Ref	2	0	3	2	3	2	-

Table 1: The number in the cell gives the number of proteins for which the improvement score S_{impr} was better for the method in the row than the method in the column.

None of the methods performed well on Ca^{2+} ATPase; the highest improvement score was reported by Climber (11.6%, 14% or 16% depending on the number of cycles selected) and our MDS methods performed poorly in comparison. It would be instructive to know the reason. The Ca^{2+} ATPase motion is described as the movement of cytoplasmic domains A (actuator), N (nucleotide binding) and P (phosphorylation) (Toyoshima and Mizutani, 2004). These domains are separated in the start structure (PDB:1SU4) and undergo domain movements to reach the compact end structure (PDB:1IWO). Although the intermediate structure (PDB:1VFP) is also compact, the arrangement of the domains is different to those in the end structure. Thus the movement from start to intermediate is quite different to that from intermediate to end. In the development of morphing methods it is necessary to

assume a-priori a direct path subject to structural constraints. Clearly for morphing this intermediate is “off-path”. This is the reason for the RMSD trajectory not having a minimum along the morphing path in Figure 4(E) and explains why all methods perform badly. Given our method is strongly on-path in its basic design this may explain why it performs badly on this particular example.

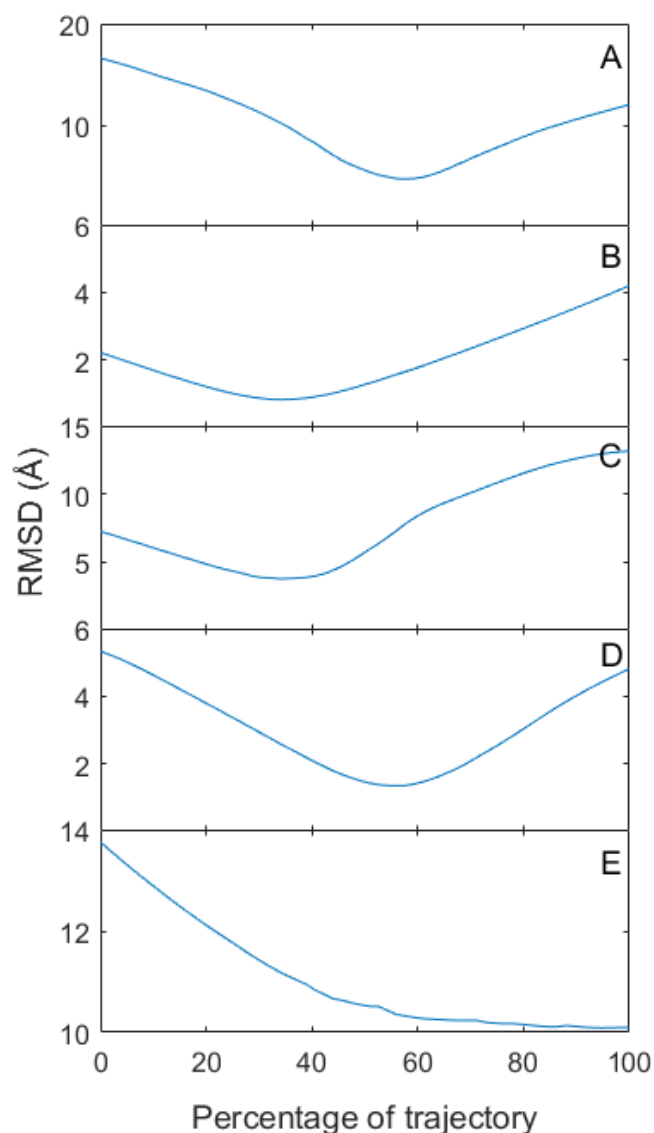


Figure 4: Trajectory of C^α-atom RMSDs of MDS₄ morphs from intermediate during the five example morphs. (A) Myosin (PDB codes 1QVI [A] to 1KK8 [A], intermediate: 1KK7 [A]). (B) Ribose-Binding Protein (1BA2 [A] to 2DRI [A], intermediate: 1URP [D]). (C) RNase III (1YYO [AB] to 1YYW [AB], intermediate: 1YZ9 [AB]). (D) 5'-Nucleotidase (1OID [A] to 1HPU [D], intermediate: 1OI8 [B]). (E) Ca²⁺ ATPase (1SU4 [A] to 1IWO [A], intermediate: 1VFP [A]).

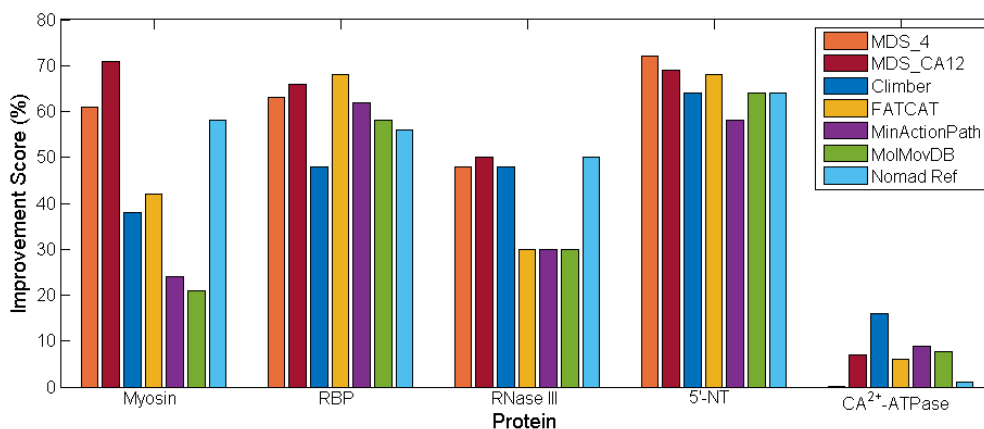


Figure 5: A comparison of the improvement scores of MDS_4 compared to the results of other protein morphing methods (Weiss and Levitt, 2009).

3.4 Docking Results

The $S_{molprob}$ values resulting from the MDS docking morph method were compared to a simple linear interpolation of Cartesian coordinates for 189 example cases taken from the Protein-Protein Docking Benchmark 5.0 (Vreven, et al., 2015). For each example, starting from the near-approach pose and morphing to the complex structure we calculated $S_{molprob}$ using linear interpolation of Cartesian coordinates and MDS docking morphing. Figure 6 clearly shows that MDS outperforms linear Cartesian coordinate interpolation in almost every case. In fact only in two cases did the MDS method perform marginally worse than the linear method (with a difference in scores of 0.06 and 0.03). Statistical significance was tested using a paired t-test which gave a P value of less than 0.0001. Given that $S_{molprob}$ is logarithmic, the significance of this result is in fact even greater.

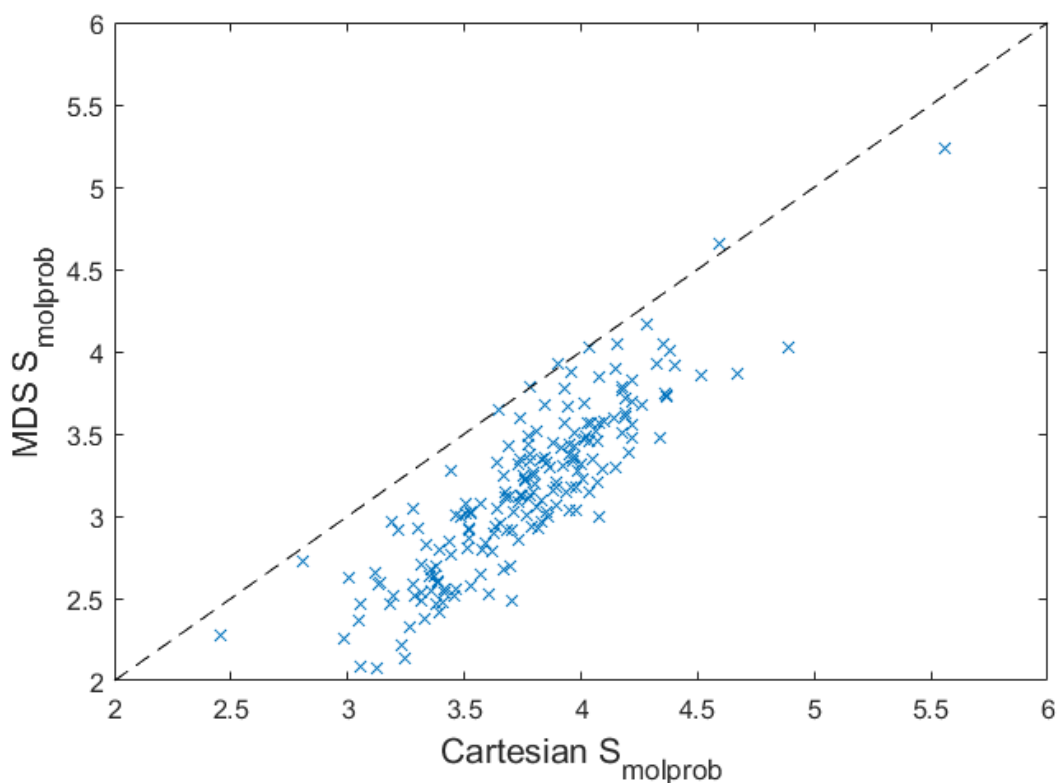


Figure 6: A comparison of the poorest MolProbity scores achieved by the MDS docking morph and a linear interpolation for 189 proteins found in the Protein-Protein Docking Benchmark 5.0 (Vreven et al., 2015).

The Protein-Protein Docking Benchmark 5.0 classifies docking interactions as “rigid-body”, “medium difficulty” or “difficult” based on the RMSD of their interface residues, affinities, and the fraction of non-native contacts in the complex structure. The increasing difficulty in these categories correlates with increasing mean $S_{molprobability}$ values for both MDS (3.04, 3.34, and 3.55, with respective standard deviations 0.473, 0.300, and 0.580) and Cartesian (3.67, 3.90, 4.07, with standard deviations 0.365, 0.294, and 0.471) morph methods.

3.5 Multigrid and GPU Acceleration

Table 2 shows the runtimes of the multigrid and pure SMACOF MDS methods when run on five morphs of proteins spanning a range of sizes. SMACOF was run until either the limit in iterations or a minimum resulting stress value was reached. The

aim was to determine the acceleration provided by the multigrid SMACOF method over the pure SMACOF method and the degree of speedup provided by GPU processing.

The multigrid timings listed in Table 2 were run using 3 V-cycles of 16 SMACOF iterations per frame of the morph and the pure SMACOF results were determined using 20 iterations per frame. After each V-cycle the stress was compared to the stress resulting from the pure SMACOF process. In order to provide a comparison the multigrid run was stopped if the stress dropped below the stress calculated by the pure SMACOF method. The whole process was repeated on the GPU. Table 2 shows that on large proteins the multigrid method provides considerable speedup over pure SMACOF on the GPU although not on the CPU. We believe that the apparent lack of improvement on the CPU is due to the additional steps added by the multigrid method. While it reaches the same stress in fewer iterations, the multigrid must perform additional operations at each point in the V-cycle including calls to the function that calculates $|r_i(k) - r_j(k)|$. This calculation is a large part of runtime on the CPU but can be greatly accelerated by performing it on the GPU. The choice in number of iterations also effects the speed comparison as the multigrid method must complete a full cycle before comparing the stress to that of the pure SMACOF method, even if the target stress would have been reached by fewer iterations. In most frames of the morph, the Fab frag 7G12 multigrid morph reached the target stress in the first cycle. When rerunning the experiment on the CPU using 30 or 40 iterations for the pure SMACOF method (resulting in a longer runtime) but the same parameters for the multigrid morph, the multigrid method still completed most frames in the first cycle, yielding a comparatively better performance.

Protein	Number of atoms	Average MDS runtime per frame (seconds)			
		SMACOF (CPU)	SMACOF (GPU)	Multigrid (CPU)	Multigrid (GPU)
Kallikrein	350	0.09	0.06	0.17	0.61

Fab frag 7G12	1,638	2.92	0.67	5.32	0.71
Pyruvate kinase	3,300	13.89	4.73	23.40	1.93
Phosphoenolpyruvate carboxykinase	4,844	33.60	14.16	64.14	4.23
Glycogen phosphorylase B	6,656	74.65	35.95	112.17	8.83

Table 2: Runtimes for 24-frame morphs: pyruvate kinase (PDB codes 1ET0 [chain A] to 1E0U [chain A]), glycogen phosphorylase b (1GBP [A] to 1GPA [C]), kallikrein (1HIA [L] to 1BX7 [A]), Fab frag 7G12 (1N7M [H] to 1NGY[A]), and phosphoenolpyruvate carboxykinase (2RKA [C] to 2QF2 [A]). The SMACOF implementation used 20 iterations per frame. The multigrid implementation repeated cycles of 16 iterations until reaching the final stress from the SMACOF method. CPU calculations were performed on an Intel Core i7 870 @ 2.93GHz processor with 16GB RAM, and GPU calculations on a NVIDIA Titan X.

4 DISCUSSION

We have used advanced MDS methods for protein morphing and shown that these methods are able to efficiently produce all-atom morphs that are direct, smooth and, in general, pass closer to known intermediates than other methods. MDS comprises a set of methods that have been developed by computer scientists over decades mainly to visualize objects for which pairwise dissimilarities are known. The MDS approach taken here is one that has been taken by others (de Leeuw and Mair, 2009), namely to start with classical MDS and then to use metric MDS using the SMACOF algorithm. However, the approach taken is one that is tailored specifically to biomolecules using a fast multigrid method.

The logic of linearly interpolating the interatomic distances, first proposed by Kim et al (Kim, et al., 2002), is that *if* one could create structures that satisfied the interpolated distances, then all interatomic distances in interpolated structures would remain within the bounds of the start and end structures, thus keeping bond lengths and bond angles within normal ranges and preventing atomic clashes. The problem is that there is no structure embedded in 3D space that can reproduce all linearly

interpolated distances. Thus the aim is to *minimize* the difference between interatomic distances in the constructed conformation and the interpolated distances. By weighting pairs of atoms within a cut-off distance through optimization of the MolProbity score our approach aims to avoid atomic clashes and undue distortions of the covalent structure. At the all-atom level, a cut-off distance of 4 Å proved to be the most common optimal value, a distance that is often used as an atomic contact distance between non-bonded atoms, e.g. as in the contact of two domains in a protein (Taylor, et al., 2013).

It is intriguing to realize that the stress in MDS is identical to the energy of an ENM and Kim et al framed their approach without referring to MDS – they refer to a “cost function”. For a real structure the stress or ENM energy is zero at the native state which is situated at the bottom of a single energy well. However, for linearly interpolated distances the stress function is a complex function of coordinates, possibly with multiple minima, but the SMACOF method deals with this by using a quadratic majorizing function. This is why the SMACOF method works so well, although it is not guaranteed to achieve the global minimum.

Our approach is not iterative along the morph (i.e. a structure along the morph is not constructed based on the structure of a previous structure) as by using classical MDS we can construct a start structure for metric MDS for any value of λ . In fact this is a strong feature of the method as we have shown that $\text{RMSD}(k,A)$ and $\text{RMSD}(k,B)$ are approximately linear in k , meaning that if the protein were so large that time constraints meant that calculation of only one intervening structure were desirable, then one could set $\lambda=0.5$ in the knowledge that $\text{RMSD}(1,A)/\text{RMSD}(A,B)\approx\text{RMSD}(1,B)/\text{RMSD}(A,B)\approx 0.5$. A further feature of the approach is that it is reversible, i.e. the morphs are the same irrespective of whether structure A is the start and structure B the end, or vice-versa.

The main beneficiaries of morphs are likely to be structural biologists, possibly X-ray crystallographers or NMR spectroscopists, who have solved a structure revealing a functional movement with a previously known structure. In highly frustrated systems such as proteins, local rearrangements are necessary for a global movement to occur. Thus a functional movement is a combination of global movements and local movements, e.g. a domain movement alongside sidechain

rotamer transitions. Therefore a good morph viewed using molecular graphics animation will enable the eye to track changes at all levels. It is important therefore in this regard that morphs are all-atom; those restricted to C^α atoms only limit usefulness. Other beneficiaries might include those using MD techniques for calculation of the potential of mean force. For example, in umbrella sampling all-atom starting structures along the reaction path are required.

Although superposition of two structures is a common method to determine differences, there are two reasons to prefer a morph as a comparison method. First, they engage more than superposition as they are a closer representation of what really happens. Second, if there is a large global movement such as a hinge-bending movement, corresponding residues in a superposition of the two structures are not co-located and differences will be difficult to appreciate. A good morph, however, will allow the viewer's eye to track the path of a particular residue whilst at the same time being aware of global changes. To serve this purpose a morph should be smooth, direct, and not distort substructures beyond their bounds in the two structures. Although we tried energy minimization on each frame to help improve the MolProbity score, we found that this produced a jerky morph defeating a primary objective. In viewing a morph it is important to bear in mind that the relative timings of events may be different in reality.

We have also implemented this for the docking of two biomolecules. We believe this will be useful for revealing the intramolecular conformational changes that occur upon complexation both for real structures and predicted structures that model flexibility. It will also be useful for demonstrating biomolecular complexation for presentation purposes.

There are a number of ways to approach morphing. Cartesian coordinate interpolation, although simple to implement, produces very poor morphs when there are large rotations. Energy minimization may correct the distortions that would occur in say the rotation of the ring in phenylalanine, but when it involves the large rotation of a whole domain such as occurs in calmodulin, energy minimization cannot correct completely for the distortions that arise. It also has the disadvantage of being dependent on the relative orientations of the two structures. Although linear interpolation of internal variables overcomes the latter, it is technically difficult to implement and also suffers from unnatural distortions that require the use of inverse kinematics techniques. The advantage of interpolating interatomic distances is that it

is relatively easy to implement and in performing a linear interpolation one is aiming to move on a direct path between the two structures, whilst at the same time aiming to preserve structural integrity. Functional movements whilst complex are likely to be direct and not make unnecessary detours. Perhaps the reason why the method presented here outperforms other current methods is due to its intrinsic parsimony.

ACKNOWLEDGEMENTS

The Titan X Pascal used for this research was donated by the NVIDIA Corporation. We thank Gavin Cawley for helpful discussions and Russel Smith, Owen Dodd, Ben Wharnsby for help with the servers.

FUNDING

RV has been supported by a UEA studentship.

Conflict of Interest: none declared.

REFERENCES

- Amemiya, T., et al. Classification and Annotation of the Relationship between Protein Structural Change and Ligand Binding. *Journal of Molecular Biology* 2011;408(3):568-584.
- Amemiya, T., et al. PSCDB: a database for protein structural change upon ligand binding. *Nucleic Acids Research* 2012;40(D1):D554-D558.
- Bennett, W.S. and Huber, R. Structural and functional-aspects of domain motions in proteins. *Crc Critical Reviews in Biochemistry* 1984;15(4):291-384.
- Berman, H.M., et al. The Protein Data Bank. *Nucleic Acids Research* 2000;28(1):235-242.
- Bonvin, A.M. Flexible protein–protein docking. *Current opinion in structural biology* 2006;16(2):194-200.
- Bronstein, M.M., et al. Multigrid multidimensional scaling. *Numerical linear algebra with applications* 2006;13(2-3):149-171.
- Castellana, N.E., et al. MORPH-PRO: a novel algorithm and web server for protein morphing. *Algorithms for Molecular Biology* 2013;8.
- Chen, V.B., et al. MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallographica Section D-Biological Crystallography* 2010;66:12-21.
- Cox, M. and Cox, T. *Multidimensional Scaling*. Berlin, Heidelberg: Springer; 2008.

- de Leeuw, J. Convergence of the majorization method for multidimensional-scaling. *Journal of Classification* 1988;5(2):163-180.
- de Leeuw, J. and Mair, P. Multidimensional Scaling Using Majorization: SMACOF in R. *Journal of Statistical Software* 2009;31(3):1-30.
- Ehrlich, L.P., Nilges, M. and Wade, R.C. The impact of protein flexibility on protein–protein docking. *Proteins: Structure, Function, and Bioinformatics* 2005;58(1):126-133.
- Franklin, J., et al. MinActionPath: maximum likelihood trajectory for large-scale structural transitions in a coarse-grained locally harmonic energy landscape. *Nucleic Acids Research* 2007;35:W477-W482.
- Gerstein, M., Lesk, A.M. and Chothia, C. Structural mechanisms for domain movements in proteins. *Biochemistry* 1994;33(2):6739-6749.
- Hammes, G.G. Multiple conformational changes in enzyme catalysis. *Biochemistry* 2002;41(26):8221-8228.
- Havel, T.F. An evaluation of computational strategies for use in the determination of protein-structure from distance constraints obtained by nuclear-magnetic-resonance. *Progress in Biophysics & Molecular Biology* 1991;56(1):43-78.
- Hayward, S. Structural principles governing domain motions in proteins. *Proteins* 1999;36:425-435.
- Hayward, S. Identification of specific interactions that drive ligand-induced closure in five enzymes with classic domain movements. *Journal Molecular Biology* 2004;339:1001-1021.
- Hayward, S. and Kitao, A. The Effect of End Constraints on Protein Loop Kinematics. *Biophysical Journal* 2010;98(9):1976-1985.
- Huang, S.-Y. Exploring the potential of global protein–protein docking: an overview and critical assessment of current programs for automatic ab initio docking. *Drug discovery today* 2015;20(8):969-977.
- Kim, M.K., Jernigan, R.L. and Chirikjian, G.S. Efficient generation of feasible pathways for protein conformational transitions. *Biophysical journal* 2002;83(3):1620-1630.
- Kleywegt, G.J. Use of non-crystallographic symmetry in protein structure refinement. *Acta Crystallographica Section D-Biological Crystallography* 1996;52:842-857.
- Krebs, W.G. and Gerstein, M. Survey and summary: The morph server: a standardized system for analyzing and visualizing macromolecular motions in a database framework. *Nucleic acids research* 2000;28(8):1665-1675.
- Kupitz, C., et al. Structural enzymology using X-ray free electron lasers. *Structural Dynamics* 2017;4(4).
- Lee, R.A., Razaz, M. and Hayward, S. The DynDom database of protein domain motions. *Bioinformatics* 2003;19(10):1290-1291.

Lindahl, E., et al. NOMAD-Ref: visualization, deformation and refinement of macromolecular structures based on all-atom normal mode analysis. *Nucleic Acids Research* 2006;34(suppl_2):W52-W56.

Loveland, A.B., et al. Ensemble cryo-EM elucidates the mechanism of translation fidelity. *Nature* 2017;546(7656):113-+.

Maiti, R., Van Domselaar, G.H. and Wishart, D.S. MovieMaker: a web server for rapid rendering of protein motions and interactions. *Nucleic acids research* 2005;33(suppl_2):W358-W362.

Maragakis, P. and Karplus, M. Large amplitude conformational change in proteins explored with a plastic network model: Adenylate kinase. *Journal of Molecular Biology* 2005;352(4):807-822.

Matthews, N., et al. High quality rendering of protein dynamics in space filling mode. *Journal of Molecular Graphics and Modelling* 2017;78:158-167.

Nguyen, M.K., Jaillet, L. and Redon, S. As-Rigid-As-Possible molecular interpolation paths. *Journal of Computer-Aided Molecular Design* 2017;31(4):403-417.

Pettersen, E.F., et al. UCSF Chimera—a visualization system for exploratory research and analysis. *Journal of computational chemistry* 2004;25(13):1605-1612.

Qi, G., Lee, R.A. and Hayward, S. A comprehensive and non-redundant database of protein domain movements. *Bioinformatics* 2005;21(12):2832-2838.

Qi, G.Y. and Hayward, S. Database of ligand-induced domain movements in enzymes. *BMC Structural Biology* 2009;9.

Remy, I., Wilson, I.A. and Michnick, S.W. Erythropoietin receptor activation by a ligand-induced conformation change. *Science* 1999;283(5404):990-993.

Rosman, G., et al. Fast multidimensional scaling using vector extrapolation. *SIAM J. Sci. Comput* 2008;2.

Schulz, G.E. Domain motions in proteins. *Current Opinion in Structural Biology* 1991;1:883-888.

Taylor, D., Cawley, G. and Hayward, S. Classification of domain movements in proteins using dynamic contact graphs. *PloS one* 2013;8(11):e81224-e81224.

Vonrhein, C., Schlauderer, G.J. and Schulz, G.E. Movie of the structural changes during a catalytic cycle of nucleoside monophosphate kinases. *Structure* 1995;3(5):483-490.

Vreven, T., et al. Updates to the integrated protein–protein interaction benchmarks: docking benchmark version 5 and affinity benchmark version 2. *Journal of molecular biology* 2015;427(19):3031-3041.

Weiss, D.R. and Levitt, M. Can Morphing Methods Predict Intermediate Structures? *Journal of Molecular Biology* 2009;385(2):665-674.

Ye, Y. and Godzik, A. FATCAT: a web server for flexible structure comparison and structure similarity searching. *Nucleic acids research* 2004;32(suppl_2):W582-W585.

ACCEPTED MANUSCRIPT

- Outperforms other morphing techniques on set of known intermediates
- Method also used to visualise conformational changes that occur upon docking
- Conformational-morph and docking-morph tools available to run from server
- More than 5000 domain movements implemented as morphs

ACCEPTED MANUSCRIPT