5-10-2010

Intrinsic Contact Geometry of Protein Dynamics

Yosi Shibberu
Rose-Hulman Institute of Technology, shibberu@rose-hulman.edu

Allen Holder
Rose-Hulman Institute of Technology, holder@rose-hulman.edu

David Cooper
Rose-Hulman Institute of Technology

MSTR 10-03

Intrinsic Contact Geometry of Protein Dynamics

Y. Shibberu, M. E. Brandt, and D. R. Cooper

Mathematical Sciences Technical Report Series
MSTR 10-03

May 9, 2010

Department of Mathematics
Rose-Hulman Institute of Technology
http://www.rose-hulman.edu/math

Fax (812)-877-8333 **Phone (812)-877-8193**

Intrinsic Contact Geometry of Protein Dynamics

Y. Shibberu[1], M. E. Brandt[2], and D. R. Cooper[2]
[1]Department of Mathematics
[2]Department of Chemistry & Biochemistry
Rose-Hulman Institute of Technology
Terre Haute, IN 47803

May 9, 2010

Abstract

We introduce a new measure for comparing protein structures that is especially applicable to analysis of molecular dynamics simulation results. The new measure generalizes the widely used root-mean-squared-deviation (RMSD) measure from three dimensional to n-dimensional Euclidean space, where n equals the number of atoms in the protein molecule. The new measure shows that despite significant fluctuations in the three dimensional geometry of the estrogen receptor protein, the protein's intrinsic contact geometry is remarkably stable over nanosecond time scales. The new measure also identifies significant structural changes missed by RMSD for a residue that plays a key biological role in the estrogen receptor protein.

Keywords: RMSD, superposition, contact map, molecular dynamics, estrogen receptor

1 Introduction

Proteins play an important role in most biochemical processes. Since the three dimensional structure of a protein is crucial for protein function, methods for describing and comparing protein structures are of considerable interest [1].

A protein is a long polymeric chain of twenty different types of smaller molecules derived from amino acids. Under biological conditions, most protein chains adopt tightly packed, well-defined, globular structures. The

folded structure of a protein can be deduced from experimental data and is typically reported as a list of three-dimensional coordinates for each atom in the protein. The protein structure can then be classified as one of a relatively small number of protein "folds" of overall similar geometry. The protein fold is determined by comparing features of the overall structure to those of other proteins.

Root-mean-squared-deviation (RMSD) is a widely used measure for quantifying the differences between two protein structures [2]. One shortcoming of RMSD is that it treats protein structures as rigid objects. In reality, most proteins exhibit dynamic structural fluctuations. In comparing structures and in evaluating the output from molecular dynamics simulations, a need exists for a measure which takes this large degree of flexibility into account [3].

Contact map overlap (CMO) is an alternative measure that overcomes some of the shortcomings of RMSD. A contact map is a matrix of ones and zero that specifies which pairs of atoms in a protein structure are in close proximity or in "contact" with one another. A typical criterion for contact defines contact as a pair of atoms that are within 4 Angstroms of one another. A contact map of an entire protein contains enough information to reconstruct the overall geometry of a protein's structure up to reflections. Since contact maps are obtained from only local information, they are insensitive to changes in distance between atoms that are far apart from one another, allowing more flexibility than RMSD. Because of the discrete nature of contact maps, they have the shortcoming of being insensitive to small changes in protein structures.

In this article, we introduce a new measure, mean-direction-cosine-deviation (MDCD), which avoids some of the limitations of RMSD and CMO. For each atom in a protein molecule, we assign an intrinsic contact vector in n-dimensional Euclidean space, where n equals the number of atoms in the protein molecule. The cosine of the angle between pairs of intrinsic contact vectors equals the contact between the corresponding pairs of atoms.

2 Intrinsic Contact Geometry

The distance matrix of a protein's structure is the matrix $D = \left(d_{ij}\right)$ where d_{ij} equals the distance between atoms i and j. We can use a cutoff function like the one defined below (see Figure 1)

$$s(d) = \begin{cases} 1 - \frac{1}{4}d & 0 \leq d \leq 4 \\ 0 & \text{otherwise} \end{cases}$$

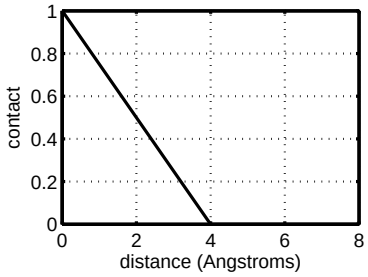

Figure 1: Graph of the cutoff function $c = s(d)$.

to construct a smoothed contact matrix $C = \big(c_{ij}\big)$, where $c_{ij} = s(d_{ij})$. It is possible to choose the cutoff function so that the contact matrix C will be a positive-definite matrix. This will be important for the method described below.

If the cutoff function used to construct the contact matrix C is chosen so that C is positive-definite, we can define a corresponding n-dimensional Euclidean space by defining the generalized inner product

$$\langle \mathrm{u}, \mathrm{v} \rangle_C = \mathrm{u}^\top C \mathrm{v}.$$

This n-dimensional Euclidean space has a useful interpretation. First, let atom i be assigned the standard unit vector $\mathrm{e}_i = (0, \ldots, 0, 1, 0, \ldots, 0)$. Then, c_{ij}, the contact between atoms i and j, is given by the inner product below.

$$\langle \mathrm{e}_i, \mathrm{e}_j \rangle_C = \mathrm{e}_i^\top C \mathrm{e}_j = c_{ij}$$

Thus, the contact between atoms i and j equals the cosine of the angle between the standard unit vectors e_i and e_j *in the Euclidean geometry defined by the inner product* $\langle \mathrm{u}, \mathrm{v} \rangle_C = \mathrm{u}^\top C \mathrm{v}$.

Before we can compare the Euclidean spaces of two protein structures, we must first transform their respective Euclidean geometries to the standard Euclidean geometry with inner product $\langle \mathrm{u}, \mathrm{v} \rangle = \mathrm{u}^\top \mathrm{v}$. Under the transformation which standardizes the inner products, the standard unit vectors e_i are transformed to the unit vectors r_i which we will refer to as the *intrinsic contact vectors* of a protein's structure. The intrinsic contact vectors r_i are the appropriate vectors to compare when measuring the differences between the intrinsic contact structures of two protein molecules.

Before we measure the difference between the intrinsic contact vectors of two protein structures, we first align the vectors with a procedure similar

to the Kabsch procedure used in calculating RMSD [4], but carried out in n-dimensional Euclidean space instead of three dimensional Euclidean space. The extra dimensions allow for additional flexibility when comparing structures. The mathematical details of the procedure outlined above are described below.

Since C is a positive definite matrix, it can be factored as $C = VDV^\top$ where the columns of V are eigenvectors of C and D is a diagonal matrix with the eigenvalues of C (which are positive) on the main diagonal. Since C is a symmetric matrix, V is an orthonormal matrix. Define the matrix $R = \sqrt{D}V^\top$. The matrix R defines the linear transformation that transforms the non-standard Euclidean geometry of a protein's structure to the standard Euclidean geometry. Why this is so is explained as follows. Define $\mathrm{r}_i = R\mathrm{e}_i$. Then, in standard Euclidean geometry,

$$\begin{aligned}
\langle \mathrm{r}_i, \mathrm{r}_j \rangle &= \mathrm{r}_i^\top \mathrm{r}_j \\
&= (R\mathrm{e}_i)^\top (R\mathrm{e}_j) \\
&= \mathrm{e}_i^\top R^\top R \mathrm{e}_j \\
&= \mathrm{e}_i^\top \left(\sqrt{D}V^\top\right)^\top \left(\sqrt{D}V^\top\right) \mathrm{e}_j \\
&= \mathrm{e}_i^\top V D V^\top \mathrm{e}_j \\
&= \mathrm{e}_i^\top C \mathrm{e}_j \\
&= c_{ij}.
\end{aligned}$$

Thus, the standard inner product of two intrinsic contact vectors r_i and r_j equals the contact between their corresponding atoms. Note that since $\langle \mathrm{r}_i, \mathrm{r}_i \rangle = c_{ii} = 1$ (the distance between an atom and itself is zero) the intrinsic contact vectors r_i are unit vectors.

We define the mean-direction-cosine-deviation (MDCD) between two structures to equal the average deviation from one for the inner products of the intrinsic contact vectors of the corresponding pairs of atoms in the two structures. The structures must first be superimposed in n-dimensional Euclidean space using a Kabsch type procedure [4] as follows. Let the columns of R_1 and R_2 equal the intrinsic contact vectors of protein structures 1 and 2. Performing a singular value decomposition, we have that $R_1 R_2^\top = UDV^\top$. Define $T = UV^\top$. Then the MDCD measure of the difference between protein structures 1 and 2 equals one minus the average of the diagonal elements of the matrix $R_1^\top T R_2$. Specifically,

$$\mathrm{MDCD} = 1 - \mathrm{ave}(\mathrm{diag}(R_1^\top T R_2)).$$

To test the utility of MDCD for measuring the structural differences between two protein structures, we applied it to measure the average structural changes occurring during 5 nanoseconds of equilibrated dynamics of the estrogen receptor protein monomer (in vacuum). We compared consecutive protein structures in the trajectory at 0.1 nanosecond time intervals and computed average changes in RMSD and MDCD for each individual atom and for each structure as a function of time. The details are covered in the next section.

3 Estrogen Receptor Protein Dynamics

The estrogen receptor is a widely studied protein due to its crucial role in mediating normal development and reproduction in all vertebrate species, and due to its role in the formation and growth of breast cancer and other reproductive tract cancers in humans [5]. The estrogen receptor exerts its biological function as a result of structural changes to the protein in a region termed the ligand-binding domain (LBD) elicited by biological ligands such as estradiol, and by drugs such as tamoxifen. Understanding these structural changes is a major area of biochemical research [5, 6].

The increasing power of readily available computer technology combined with an improved understanding of protein structure has allowed computer simulations to generate useful predictions for protein behavior. However, evaluating the large amounts of information generated during computer simulations presents some problems.

Before viewing a protein dynamics trajectory, a common procedure (the Kabasch procedure) is to align the protein trajectory to the initial or final structure in the trajectory to remove any net rotations and translations of the trajectory. We can extend this idea to consecutive structures in a trajectory to explore the structural deformations the protein undergoes. Measuring structural deformations using RMSD has the shortcoming of treating protein structures as rigid objects in three dimensional Euclidean space. Proteins are in fact quite flexible. We believe the intrinsic contact geometry we outlined in the previous section is a more natural description of protein structures.

To evaluate the usefulness of our intrinsic contact geometry, we used the molecular dynamics package NAMD [7] to simulate apo estrogen receptor α monomer (PDB ID 1G50 [8]) in vacuum for 5 nanoseconds to equilibrate the structure and an additional 5 nanoseconds to measure RMSD and MDCD structural changes between consecutive structures separated by 0.1

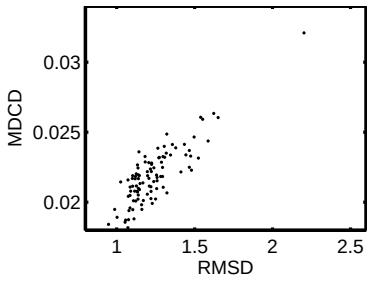

Figure 2: Scatter plot of RMSD vs MDCD for 50 consecutive structures of equilibrated estrogen receptor α monomer dynamics. Structures are separated by 0.1 nanoseconds in time.

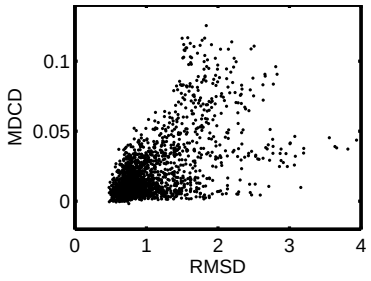

Figure 3: Scatter plot of average atom RMSD vs average atom MDCD between consecutive structures of equilibrated estrogen receptor α monomer. Comparisons are made between atom positions separated by 0.1 nanoseconds in time.

nanoseconds. Figure 2 is a scatter plot of RMSD vs MDCD for individual structures. Figure 2 suggest that for the small structural changes which occur in equilibrated dynamics over short periods of simulated time, RMSD and MDCD are correlated. Figure 3 is a scatter plot for RMSD vs MDCD for individual atoms of the estrogen receptor averaged over the equilibrated portion of the trajectory. Although, once again, small changes are well correlated, larger changes exhibit much less correlation. Observe from both figures that while RMSD is as large as approximately 2 Angstroms RMSD for structures and 4 Angstroms RMSD for atoms, the corresponding MDCD values are below approximately 0.027 and 0.12 respectively. From Figure 1, we see that a difference in contact value of 0.027 and 0.12 is equivalent to 0.00675 and 0.03 Angstroms respectively, indicating that the intrinsic con-

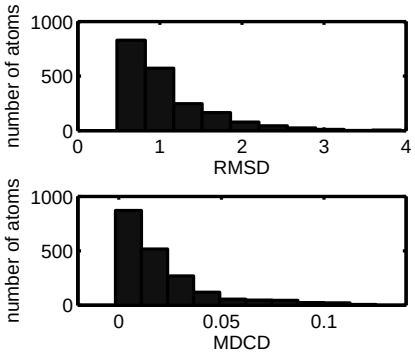

Figure 4: Distribution of average RMSD and MDCD for atoms separated by 0.1 nanosecond consecutive structures of equilibrated estrogen receptor α monomer dynamics.

tact structure of the estrogen receptor protein, over time scales of nanoseconds, is remarkably stable. RMSD overemphasizes the global structural fluctuations that do not affect the local interatomic interactions responsible for protein structural stability.

To further evaluate the measures, we compared the distributions of atoms by RMSD and MDCD (Figure 4). Both distributions are similar with most atoms exhibiting small deviations in both RMSD and MDCD. Based on the data in Figure 3, the two measures identify different subsets of atoms exhibiting large changes.

To further compare MDCD to RMSD as an evaluation measure, we performed molecular dynamics simulations using two different starting structures. The first structure was the crystal structure described earlier. The second structure was a homology model described in more detail below. It is believed that activation of the estrogen receptor protein involves a significant change in the protein's structure induced by ligand binding. Because no estrogen receptor structures corresponding to the biologically relevant ligand-free conformation of the protein are available, we generated a homology model for this configuration.

Homology modeling is a protein structure prediction method in which the sequence of the protein of interest is threaded into the structure of a related protein [9, 10, 11]. Unlike the estrogen receptor LBD, the structure of the human RXR-α LBD has been experimentally determined both in the absence [12] and presence [13] of ligand. The human RXR-α LBD exhibits

about 26% sequence identity with the human estrogen receptor-α. While this relatively low sequence identity presents some problems in generating a direct alignment, the existence of structures for both the estrogen receptor and RXR in the presence of ligands allows the use of a structural alignment to guide the sequence alignment.

The sequence of the human estrogen receptor-α LBD was aligned to that of the RXR-α LBD by comparing the structural features of the two ligand-bound LBDs (PDB ID 1G50 for the estrogen receptor [8] and 1FBY for RXR [13]). Once the sequences were aligned, the SWISS-MODEL server [11] was used to generate a homology model for the ligand-free estrogen receptor using the ligand-free RXR-α (PDB ID 1LBD) as a template.

Using this homology model as our starting structure, we repeated the molecular dynamics simulations and RMSD and MDCD calculations described earlier. A comparison of the two simulations (crystal structure and homology model) using both RMSD and MDCD is shown in Figures 5 and 6.

In a recent paper, Celik et al. [14] described the pivotal role that the residue His524 (located in the ligand binding pocket) plays in maintaining the biologically active agonist conformation of the estrogen receptor protein. We compared the ability of RMSD and MDCD to detect the biologically significant role played by His524. The data in Figure 5 show that for the crystal structure simulations, MDCD indicates that His524 is one of the most structurally *stable* residues in the protein. In contrast, RMSD does not indicate that His524 is significantly different from any of the other residues. For the homology model simulations shown in Figure 6, we see that MDCD now indicates that His524 is one of the most structurally *unstable* residues in the protein. Once again, RMSD does not suggest His524 is structurally different from a typical residue in the protein.

If we assume the homology model is a reasonable model for the estrogen receptor protein in the absence of ligand, it is logical to assume that His524 will be flexible enough to accommodate ligand binding. However, in the crystal structure, the ligand is already bound and therefore, His524 is likely constrained in position. In comparing the data for His524 in Figures 5 and 6, the MDCD results support the contention that His524 changes its behavior in a manner which is consistent with the role of His524 in ligand binding. In contrast, His524 has similar RMSD values for both crystal structure and homology model simulations.

It appears that MDCD highlights relevant residues under conditions where RMSD does not. In conclusion, we believe that MDCD holds promise as a biologically significant measure of the structural changes occurring in molecular dynamics.

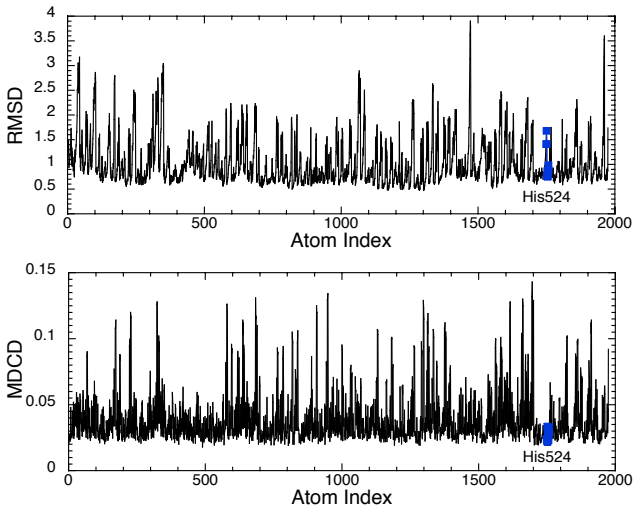

Figure 5: Average RMSD and MDCD plotted for each atom for the molecular dynamics simulation starting from the estrogen receptor-α LBD crystal structure. The atoms for residue His524 are highlighted.

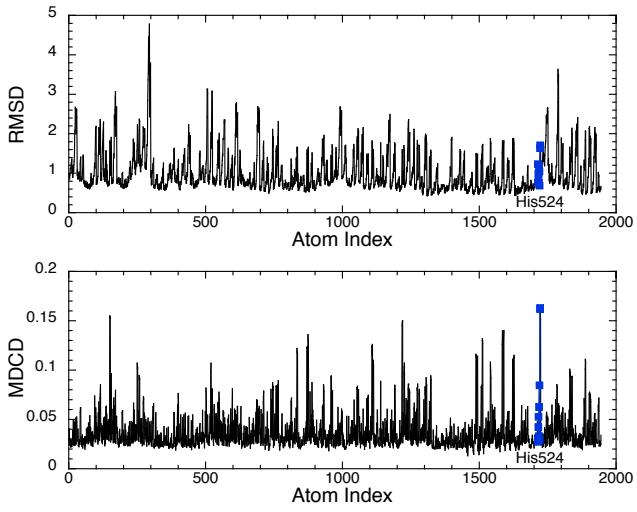

Figure 6: Average RMSD and MDCD plotted for each atom for the molecular dynamics simulation starting from the estrogen receptor-α LBD homology model. The atoms for residue His524 are highlighted.

References

[1] Mark T Oakley, Daniel Barthel, Yuri Bykov, Jonathan M Garibaldi, Edmund K Burke, Natalio Krasnogor, and Jonathan D Hirst. Search strategies in structural bioinformatics. *Curr Protein Pept Sci*, 9(3):260–274, Jun 2008.

[2] Evangelos A Coutsias, Chaok Seok, and Ken A Dill. Using quaternions to calculate RMSD. *J Comput Chem*, 25(15):1849–1857, Nov 2004.

[3] Matthew Menke, Bonnie Berger, and Lenore Cowen. Matt: local flexibility aids protein multiple structure alignment. *PLoS Comput Biol*, 4(1):e10, Jan 2008.

[4] W Kabsch. A discussion of the solution for the best rotation to relate two sets of vectors. *Acta Crystallog A*, 34:827–828, 1978.

[5] Matthew H Herynk and Suzanne A W Fuqua. Estrogen receptor mutations in human disease. *Endocrine Reviews*, 25(6):869–98, Dec 2004.

[6] Jean-Pierre Changeux and Stuart J Edelstein. Allosteric mechanisms of signal transduction. *Science*, 308(5727):1424–8, Jun 2005.

[7] James C Phillips, Rosemary Braun, Wei Wang, James Gumbart, Emad Tajkhorshid, Elizabeth Villa, Christophe Chipot, Robert D Skeel, Laxmikant Kale, and Klaus Schulten. Scalable molecular dynamics with NAMD. *J Comput Chem*, 26(16):1781–1802, Dec 2005.

[8] S Eiler, M Gangloff, S Duclaud, D Moras, and M Ruff. Overexpression, purification, and crystal structure of native ER alpha LBD. *Protein Expression and Purification*, 22(2):165–73, Jul 2001.

[9] Konstantin Arnold, Lorenza Bordoli, Jürgen Kopp, and Torsten Schwede. The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling. *Bioinformatics*, 22(2):195–201, Jan 2006.

[10] Florian Kiefer, Konstantin Arnold, Michael Künzli, Lorenza Bordoli, and Torsten Schwede. The SWISS-MODEL repository and associated resources. *Nucleic Acids Res*, 37(Database issue):D387–92, Jan 2009.

[11] M C Peitsch. Promod and SWISS-MODEL: Internet-based tools for automated comparative protein modelling. *Biochem Soc Trans*, 24(1):274–9, Feb 1996.

[12] W Bourguet, M Ruff, P Chambon, H Gronemeyer, and D Moras. Crystal structure of the ligand-binding domain of the human nuclear receptor RXR-alpha. *Nature*, 375(6530):377–82, Jun 1995.

[13] P F Egea, A Mitschler, N Rochel, M Ruff, P Chambon, and D Moras. Crystal structure of the human RXR-alpha ligand-binding domain bound to its natural ligand: 9-cis retinoic acid. *EMBO J*, 19(11):2592–601, Jun 2000.

[14] Leyla Celik, Julie Davey Dalsgaard Lund, and Birgit Schiott. Conformational dynamics of the estrogen receptor alpha: molecular dynamics simulations of the influence of binding site structure on protein dynamics. *Biochemistry*, 46(7):1743–1758, Feb 2007.