2012 2nd International Conference on Biomedical Engineering and Technology
IPCBEE vol. 34 (2012) © (2012) IACSIT Press, Singapore

A Normalized Weighted RMSD for Measuring Protein Structure Superposition

Xueyi Wang[1] and Jianmin Dong[2]

[1]Northwest Nazarene University, Nampa, USA
[2] Tibet University for Nationalities, Xianyang, China

Abstract. Root-mean-square-deviation (RMSD) is the most widely used measure of the similarity of superimposed protein structures, but it is sensitive to outliers and a smaller RMSD value may not correspond to a better structure superposition. Many alternative measures have been proposed to overcome the deficiency in RMSD. In this paper, we extend the RMSD to normalized weighted RMSD (nwRMSD) to measure the quality of superimposed structures, where the nwRMSD assigns a normalized weight to each superimposed position. We present an iterative algorithm to minimize nwRMSD for structure superposition and propose a new weight function for structure superposition. We show that NMR ensembles minimized by the nwRMSD measure can clearly display structurally conserved and flexible regions, which are better than the superposition in original structures.

Keywords: normalized weighted RMSD; structure superposition; position weight

1. Introduction

Protein structure comparison is an important research topic in the area of bioinformatics and computational biology. It can be classified into two categories: structure alignment, which compares the similarity of different protein structures, and structure superposition, which compares the similarity of different conformations of a protein structure. Protein structure alignment is useful in classifying protein 3D structures, identifying structure conserved regions, and disclosing evolutionary relationship of proteins [14, 15]. Protein structure superposition is useful for evaluating the quality of predicted protein models [11], assessing the precision of NMR ensembles [22], and identifying structurally conserved or flexible regions [7].

Root-mean-squared deviation (RMSD) is the most widely used measure for comparing protein structures. In structure superposition, we usually assume structures as rigid bodies, minimize the RMSD by translating and rotating the structures in 3D space, and then measure the similarity. For a pair of superimposed structures, we measure the average distance of all point pairs. For multiple superimposed structures, we measure the average distance of point pairs in all structure pairs, where there are $n(n–1)/2$ pairs for n structures.

One deficiency with RMSD is its sensitivity to outliers, in which case a single outlier may cause a significant increase of the RMSD. The outliers may be caused by experimental errors or conformational changes of structural flexible regions. For structure superimposition, in the former case we definitely want to remove the effects of the outliers, and in the latter case we still want to reduce the effects of outliers in order to identify structurallyconserved or flexible regions.

Many RMSD-based or alternative methods have been proposed to overcome the deficiency in RMSD. Depending on criteria used for minimizing the structure superposition, these methods can be classified into four categories:*distance-cutoff*methods, *number-cutoff*methods, position weights based methods, and others. Distance-cutoff methods minimize superimposed structures by considering only the point pairs whose distances are below a threshold value. For example, GDT algorithm (Global Distance Test) [23] uses an exhaustive search algorithm and finds a largest set of point pairs whose distances are within a threshold value

x and records the RMSD of the set of point pairs as GDT_Px. The GDT_TS score, (GDT_P1 + GDT_P2 + GDT_P4 + GDT_P8) / 4, becomes a standard in comparing the similarity of a predicted protein model to an experimentally determined structure. MaxSub [19] minimizes the superposition of various continuous segments of L pairs, then extends each segment to include other point pairs within a threshold value, and finally outputs a superposition with the most point pairs. Snyder and Montelione [20] identify a set of core atoms, partition the core atom set into several RMSD-stable domains, minimize each RMSD-stable domain, and calculate the RMSD for each domain. Remington and Matthews [17] and Aleksandrov *et al*. [1] analyze the statistical distribution of RMSD and choose distance-cutoff ranges that are statistically significant to the lengths of proteins. Maiorov and Crippen [13] choose a distance-cutoff range when the original RMSD is smaller than the RMSD with one structure reflected.

Number-cutoff methods minimize structure superposition by choosing a fixed number of point pairs with the smallest distances. For example, rmsd_L [4] calculates the smallest RMSD for L point pairs.The algorithm suggests to use rmsd_{100} as a normalized and size-independent measurement, where100 is the mean number of amino acids per domain.

Position weight based methods assign a weight to each superimposed position and iteratively adjust the weights until the superposition converges. For example, Gaussian-weighted RMSD [6] presents a weight function $w_k = e^{-(d_k)^2/c}$, where d_k is the distance of a point pair at a superimposed position k and c is an arbitrary scaling factor. Wang and Snoeyink [21] present a weight function inverse to the average distance of all point pairs in superimposed positions.

Some methods use other measures to replace the RMSD. For example, AL0 [23] calculates the number of atom pairs within a 5Å distance threshold using a LGA sequence independent superposition algorithm. MAMMOTH z-score [16] finds the largest subset of similar local structures by the MAMMOTH alignment algorithm and calculates the probability of proportion of superimposed residues. Similarly, Dali z-score [8] finds the largest subset of similar local structures by the DALI algorithm and calculates the z-score. TM-score function [10] measures the inverse of the squared distance of an atom pair and generates a structure similarity score.Least median of squares regression [12, 18] minimizes the median of squared distances instead of the sum of squared distances in the RMSD.

In this paper, we introduce a new measure called normalized weighted RMSD (nwRMSD), which is extended from weighted RMSD with position weights [21], to minimize the structure superposition. Although the weighted RMSD with position weights can reduce or remove the effects of outliers, it fails at quantifying the similarity of superimposed structures on an absolute scale. For the multiple structure superposition, we present an efficient iterative algorithm, which is extended from Wang and Snoeyink [21], to minimize the nwRMSD given any convergent weight function.

Furthermore, we propose a new weight function for quantifying the similarity of superimposed structures. We test on NMR ensembles and compare it with the ensembles optimized by standard RMSD and those used by the Protein Data Bank (PDB) [3]. The results show nwRMSD performs better in displaying structurally conserved or flexible regions of NMR ensembles than standard RMSD andoriginal PDB ensembles.

2. Methods

In this section, we present the normalized weighted RMSD and its properties, an algorithm to minimize superimposed structures given fixed position weights, and an algorithm to minimize superimposed structures given a convergent weight function.

2.1. Properties of the normalized weighted RMSD

We assume there are n structuresS_i for ($1 \le i \le n$). Each structure S_i has m points (atoms)$p_{i1}, p_{i2}, \ldots, p_{im}$. For a fixed position k,we assume the n points p_{ik} for ($1 \le i \le n$) correspond. We assign a position weight $w_k \ge 0$ to each superimposed position k that $\sum_{k=1}^{m} w_k > 0$ and define a normalized position weight $\hat{w}_k = m w_k / \sum_{k=1}^{m} w_k$ (note that $\sum_{k=1}^{m} \hat{w}_k = m$). We define a weighted average structure $\bar{S}$ to have points $\bar{p}_k = \sum_{i=1}^{n} \hat{w}_k p_{ik} / \sum_{i=1}^{n} \hat{w}_k$ for ($1 \le k \le m$). Since $\sum_{i=1}^{n} \hat{w}_k p_{ik} / \sum_{i=1}^{n} \hat{w}_k = \sum_{i=1}^{n} p_{ik} / n$, the weighted average structure is the same as an average structure.

We define a normalized weighted root mean squared deviation (nwRMSD) as a square root of normalized weighted sum of all squared distances of structures: $\sqrt{\sum_{i=2}^{n}\sum_{j=1}^{i-1}\sum_{k=1}^{m}\widehat{w}_k\left\|p_{ik}-p_{jk}\right\|^2/\left(\frac{n(n-1)}{2}\sum_{k=1}^{m}\widehat{w}_k\right)} = \sqrt{\frac{2}{mn(n-1)}\sum_{i=2}^{n}\sum_{j=1}^{i-1}\sum_{k=1}^{m}\widehat{w}_k\left\|p_{ik}-p_{jk}\right\|^2}$, where there are $n(n-1)/2$ structure pairs in total and each pair has a normalized weighted sum of squared distances: $\sum_{k=1}^{m}\widehat{w}_k\left\|p_{ik}-p_{jk}\right\|^2/\sum_{k=1}^{m}\widehat{w}_k = \sum_{k=1}^{m}\widehat{w}_k\left\|p_{ik}-p_{jk}\right\|^2/m$. Since m and n are fixed and the square root function is monotonically increasing, we use the weighted sum of all squared pairwise distances $\sum_{i=2}^{n}\sum_{j=1}^{i-1}\sum_{k=1}^{m}\widehat{w}_k\left\|p_{ik}-p_{jk}\right\|^2$ instead of nwRMSD.

We can easily obtain the standard RMSD from nwRMSD by assigning all position weights to 1, i.e. $w_k = \widehat{w}_k = 1$ for ($1 \le k \le m$). By using weights, we can reduce or eliminate the effects of outliers by assigning higher weights to those structurally inflexible regions and lower weights to structural flexible regions or outliers.

2.2. Algorithm for minimizing nwRMSD with fixed position weights

We take each structure as a rigid body and allow translating and rotatingall the structures in minimizing nwRMSD. For each structure S_i, we define R_i as a 3×3 rotation matrix and T_i as a 3×1 translation vector. We use the target function $\text{argmin}_{R,T}\left(\sum_{i=2}^{n}\sum_{j=1}^{i-1}\sum_{k=1}^{m}\widehat{w}_k\left\|R_ip_{ik}-R_jp_{jk}-T_i-T_j\right\|^2\right)$ and aim to find the optimal T_i and R_i for each structure that minimize the function.

To minimize the superposition of two structures by nwRMSD, we can extend Horn's method [9] to translate a weighted centroid of each structure to the origin and then apply an optimum rotation for one structure. We leave the proof out as it can be easily obtained from Horn's analysis by adding weights to all terms.

We present an iterative algorithm that minimizes the nwRMSD for multiple structures, which is also extended from Wang and Snoeyink [21]. The algorithm repeatedly minimizes nwRMSD from each structure to the average and recalculates a new average until the nwRMSD converges to a local minimum.

Algorithm 1. Given $n$aligned structuresS_i for ($1 \le i \le n$), where each structure has m points (atoms) and each aligned position has a $w_k$for ($1 \le k \le m$), minimize the nwRMSD to within a threshold value ε (e.g. $\varepsilon = 1.0\times10^{-5}$).

1. Translate a weighted centroid of each structure S_i for ($1 \le i \le n$) to the origin.

2. Calculate the average structure $\bar{S}$ with points $\bar{p}_k = \sum_{i=1}^{n}p_{ik}/n$ and deviation $SD = \sum_{i=1}^{n}\sum_{k=1}^{m}\widehat{w}_k\|p_{ik}-\bar{p}_k\|^2$.

3. For each S_i ($1 \le i \le n$), superimpose it to$\bar{S}$ using Horn's method that minimizes $\sum_{k=1}^{m}\widehat{w}_k\|R_ip_{ik}-\bar{p}_k\|^2$with an optimum rotation matrix R_i. Replace $S_i^{\text{new}} = R_i\times S_i$.

4. Calculate a new average $\bar{S}^{\text{new}}$ and deviation $SD = \sum_{i=1}^{n}\sum_{k=1}^{m}\widehat{w}_k\|p_{ik}{}^{\text{new}}-\bar{p}_k{}^{\text{new}}\|^2$.

5. If $SD - SD^{\text{new}} < \varepsilon$, then the algorithm terminates; otherwise, set $SD = SD^{\text{new}}$ and $\bar{S} = \bar{S}^{\text{new}}$ and go to step 3.

Similar to the analysis in Wang and Snoeyink [21], from Horn [9], in step 3 we have:$\sum_{i=1}^{n}\sum_{k=1}^{m}\widehat{w}_k\|p_{ik}{}^{\text{new}}-\bar{p}_k\|^2 \le \sum_{i=1}^{n}\sum_{k=1}^{m}\widehat{w}_k\|p_{ik}-\bar{p}_k\|^2 = SD$.

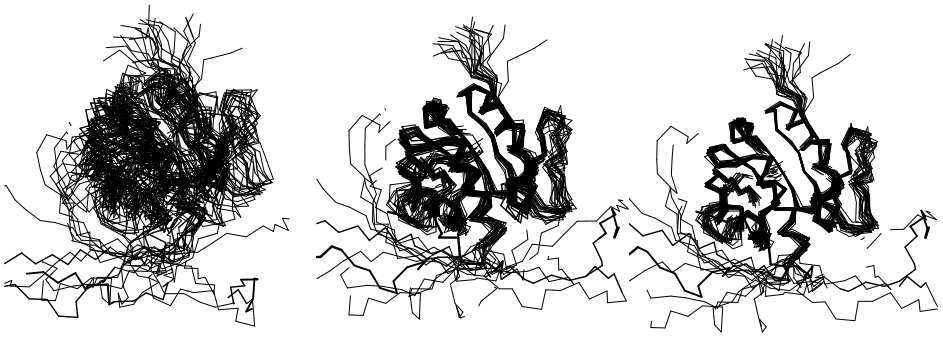
In step 4 we have:$SD^{\text{new}} = \sum_{i=1}^{n}\sum_{k=1}^{m}\widehat{w}_k\|p_{ik}{}^{\text{new}}-\bar{p}_k{}^{\text{new}}\|^2 \le \sum_{i=1}^{n}\sum_{k=1}^{m}\widehat{w}_k\|p_{ik}{}^{\text{new}}-\bar{p}_k\|^2$

So $SD^{\text{new}} \le SD$ andSD decreases in each iteration. The algorithm stops when $SD - SD^{\text{new}}$ is less than the threshold valueε, which means SD reaches a local minimum.

2.3. Algorithm for optimizing structure superposition

Algorithm 1 minimizes nwRMSD if all position weights are fixed, but one bigger problem is how to minimize nwRMSD if position weights change. If we already know a weight function $f(k)$ for ($1 \le k \le m$) that assigns higher weights to better superimposed positions and lower weights to outliers, then we could use the following heuristic algorithm to optimize structure superposition.

Algorithm 2. Given $n$aligned structuresS_i for ($1 \le i \le n$), where each structure has m points (atoms), optimize the structure superposition based on weight function $f(k)$ for ($1 \le k \le m$).



a. standard RMSD b. PDB c. nwRMSD

Fig. 1: The superposition of NMR structure target T0472 (2K4M) by standard RMSD, original PDB, and nwRMSD.

1. Set all $w_k = 1$ for ($1 \le k \le m$) and minimize SD of the protein structures using the Algorithm 1.

2. For each aligned position k, calculate and set $w_k^{new} = f(k)$ and minimize SD^{new} using Algorithm 1.

3. If $SD - SD^{new} < \varepsilon$ (e.g. $\varepsilon = 1.0 \times 10^{-5}$), then the algorithm terminates; otherwise set $SD = SD^{new}$ and go to step 2.

As $SD \ge 0$ and SD decreases in steps 2 and 3, Algorithm 2 will eventually stop.

Note that convergence of Algorithm 2 depends on the weight function. If a weight function tends to assign higher weights to outliers, then although the $SD^{new} > SD$ in step 2 and the algorithm exits in step 3, the nwRMSD never converges. So we should carefully choose a weight function to make the convergence for Algorithm 2.

3. Results and Discussion

3.1. nwRMSD weight function examples and comparison

We choose a weight function $w_k = 1/\left(log\left({d_k}^3 + 1\right) + c\right)$, where c is a non-negative constant ($c = 0.2$ in the experiments), and use it to optimize structure superposition. The inverse of d_k^3 allows us to assign higher weights to better superimposed positions and the constant c allows us to avoid the influence of certain positions with extremely small d_k^3. We use Algorithm 2 to minimize the nwRMSD.

We test the weight function on 14 NMR structure targets in CASP8 and compare the results to both the superposition by standard RMSD and the one used by the PDB [3].Since both PDB and nwRMSD emphasize the superpositionof structural conserved regions and ignore flexible regions to remove the effects of outliers, the RMSD values by the PDB and nwRMSD are similar to each other and are significantly higher than standard RMSD. Figure 1 shows the NMR ensemble of T0472 (2K4M). We can see that the ensembles optimized by the PDB and nwRMSD are significantly better than optimized by RMSD and the ensemble by the nwRMSD is slightly better than the one by the PDB.

4. Conclusions

In this paper, we present a measure called normalized weighted RMSD, which allows us to directly compare different nwRMSD values in structure superposition, and propose a new weight function. The results show that the nwRMSD with the weight function performs well. The rankings of predicted structure models in CASP7 and CASP8 targets are comparable to expert rankings and are better than most of existing measures.

5. Acknowledgement

This work is supported by NIH grant #P20 RR016454.

6. References

[1] Aleksandrov NI, Takahashi K, Go N: *Common spatial arrangements of backbone fragments in homologous and non-homologous proteins.*J. Mol. Biol.1992, 225:5-9

[2] Ben-David M, Noivirt-Brik O, PazA, Prilusky J, Sussman JL, Levy Y: *Assessment of CASP8 structure predictions for template free targets.Proteins* 2009,S9:50-65

[3] Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE: *The Protein Data Bank.Nucleic Acids Res*. 2000,28:235-242

[4] Carugo O, Pongor S: A normalized root-mean-square distance for comparing protein three-dimension structures.Protein Science 200110:1470-1473

[5] Cozzetto D, Kryshtafovych A, Fidelis K, Moult J, Rost B, Tramontano A: *Evaluation of template-based models in CASP8 with standard measures.Proteins* 2009,S9:18-28

[6] Damm KL, Carlson HA: Gaussian-Weighted RMSD Superposition of Proteins: *A Structural Comparison for Flexible Proteins and Predicted Protein Structure.Biophysical* Journal 2006,90:4558-4573

[7] Hilser VJ, Dowdy D, Oas TG, Freire E: *The structure distribution of cooperative interactions in proteins: analysis of the native state ensemble.PNAS* 1998, 95(17):9903-9908

[8] Holm L, Kääriäinen S, Rosenström P, Schenkel A: *Searching protein structure databases with DaliLite v.3.Bioinformatics* 2008,24:2780-2781

[9] Horn BKP: Closed-form solution of absolute orientation using unit quaternions.Journal of the Optical Society of America A 1987,4:629-642

[10] Jauch R, Yeo HC, Kolatkar PR, Clarke ND: *Assessment of CASP7 structure predictions for template free targets.Proteins* 2007,69(S8):57-67

[11] Kryshtafovych A, Prlic A, Dmytriv Z, Daniluk P, Milostan M, Eyrich V, Hubbard T, Fidelis K: *New tools and expanded data analysis capabilities at the protein structure prediction center.Proteins* 2007, 69(S8):19-26

[12] Liu Y, Fang Y, Ramani K: *Using least median of squares for structural superposition of flexible proteins.BMC Bioinformatics* 2009,10:29

[13] Maiorov VN, Crippen GM: *Significance of Root-Mean-Square Deviation in Comparing Three-dimensional Structures of Globular Proteins.*J. Mol. Biol. 1994, 235:625-634

[14] Malmstrom L, Riffle M, Strauss CEM, Chivian D, Davis TN, Bonneau R, Baker D: *Superfamily assignments for the yeast proteome through integration of structure prediction with the gene ontology.PLoS Biol*. 2007,5(4):e76

[15] Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, Thornton JM: CATH: *A hierarchical classification of protein domain structures.Structure* 1997, 5(8):1093-1108

[16] Ortiz AR, Strauss CE, Olmea O: MAMMOTH (Matching molecular models obtained from theory): *An automated method for model comparison.Protein Sci*. 2002,11(11):2606-2021

[17] Remington SJ, Matthews BW: *A systematic approach to the comparison of protein structures.J.* Mol. Biol. 1980, 140:77-99

[18] Rousseeuw PJ: *Least Median of Squares Regression.Journal of the American Statistical Association* 1984, 79(388):871-880

[19] Siew N, Elofsson A, Rychlewski L, Fischer D: *MaxSub: an automated measure for the assessment of protein structure prediction quality.Bioinformatics* 2000, 16(9):776-785

[20] Snyder DA, Montelione GT: *Clustering Algorithms for Identifying Core Atom Sets and for Assessing the Precision of Protein Structure Ensembles.Proteins* 2005, 59:673-686

[21] Wang X, Snoeyink J: *Multiple Structure Alignment by Optimal RMSD Implies that the Average Structure is a Consensus. In Proceedings on 2006 LSS Computational Systems Bioinformatics Conference*, 2006, 79-87

[22] Wüthrich K: *NMR — this other method for protein and nucleic acid structure determination.Acta Crystallogr D* 1995, 51:249-270

[23] Zemla A: LGA: *a method for finding 3D similarities in protein structures.Nucleic Acids Research* 2003, 31(13):3370-3374.