Extending the incorporation of superfamily structural information in the process of flexible fitting in 3D-EM

Velázquez-Muriel, J.A., Carazo, J.M
Centro Nacional de Biotecnología, Campus Univ. Autónoma de Madrid, 28049, Madrid,
España. Tel.: 91 5854510
javi@cnb.uam.es, carazo@cnb.uam.es (*corresponding author)*

Abstract

A new procedure to increase the number of aminoacids that are taken into account in the structural alignment core when performing flexible fitting with superfamily information in three-dimensional electron microscopy (3D-EM) is described. We propose to use incremental singular value decomposition (ISVD) instead of regular SVD to compute the principal components of the alignment core, in this way allowing to add new data and to estimate missing values. The results proof that positions of the alignment with gaps contain variational information that improves the models built for the fitting step of the procedure.

1. Introduction

Rigid body fitting is the usual way of interpreting the information contained in a medium resolution 3D-EM density map of a protein. The atomic resolution structure of the protein, typically obtained by X-ray crystallography, is fitted into the map by exploring all the possible rotations and translations. The computational solution of the problem consists on finding the best orientation by maximizing a measure of fitness. The most popular one is the cross-correlation coefficient (CCC) and its variants: local cross-correlation coefficient (LCCC) and the rotational correlation function.

Flexible fitting is useful either when none of the domains of the structure contained in the map have been solved to atomic resolution, but some related structure has, or in those other cases in which the complex solved by 3D-EM is in a different conformational state than the related data solved at atomic resolution. Molecular dynamics [1], Normal Mode Analysis [2] and homology modeling [3] have been proposed to deal with this case, allowing a modification of the atomic resolution structure in order to increase the measure of fitness. In previous works we have developed an alternative to these methods [4] based on deforming the structure of a domain following the conformational variations observed among domains within its structural superfamily. The variational space of the superfamily is then decomposed with SVD into its principal components, and the linear combination of the three first is used to represent the main variational trends within the superfamily. In this way, it is guaranteed that the deformations are biologically meaningful. This idea comes from recent development in the homology modeling field [5]. Afterwards, we use an articulated model to apply the deformations so that they are also chemically correct, respecting the bond lengths and torsion angles: Rigid body transformations are considered for the secondary structure elements (SSEs) of the domain, and the loops are closed with the cyclic coordinate descent (CCD) algorithm [6]. Finally, all the structures are fitted into the 3D-EM map using the program COLORES [7], and the best one is chosen based on cross-correlation measures.

Here we propose the use of ISVD [8] instead of SVD as a way to increase the number of aminoacids that are taken into account in the structural alignment core previous to the decomposition of the variational space. The new approach in shown to improve fitting results and two examples are presented.

2. Theory

To build the variatonal space of a superfamily, we structurally align all the members of the superfamily with the reference domain (the domain that is going to be fitted), using MAMMOTH [9]. The structural alignment, and derived sequence alignment, pairs similar aminoacids in all structures.

Then, coordinate displacement vectors (*cdv*) for the backbone atoms (C_α, O, N and C) of the reference domain are computed. For each atom I, the coordinate displacement vector cdv_i is defined as the difference between its coordinates and the coordinates of the same atom j in the aminoacid with which it is aligned:

$$cdv_i = (x_j - x_i, y_j - y_i, z_j - z_i)$$

Each matrix CDV contains the coordinate displacement vectors *cdv* of all the participating atoms of an aligned domain, and a final **CDV** matrix is built with all the CDVs.

If the **CDV** matrix is to be decomposed with the SVD algorithm to capture the principal components of variation, it must not have any unknown value. This implies using the core region, composed by the alignment positions with no gaps in none of the sequences because only in the core the variational information is complete. However, the use of ISVD instead of SVD to decompose **CDV** allows to incrementally introduce unknown values. The domain with the highest number of aligned aminoacids with the reference domain is considered by ISVD in the first place. Then, all the other domains, in the order given by the number of aligned aminoacids. All the positions with unknown values (because there are no aligned aminoacis for this domain) are estimated so that the sum of squared errors for the known values is minimum.

3. Resuls

We illustrate the usefulness of introducing ISVD with two cases.

In the first case, we deal with all-alpha domains of CATH superfamily 1.10.238.10. Domain 1wdcB2 was forced to fit into the map of domain 1cll02 at 8 Å, using 20 domains to cover the variational space. In the second case, a more complex fold (CATH superfamily 2.80.10.50) of an all-beta type is treated. Domain 1jlxA1was forced to fit into the map of domain 1a8d02 at 8 Å, using 15 domains to cover the variational space.

After all the fittings, the RMSD between the structure to fit and the one that generated the map was measured. The lower the value, the better the reference structure was fitted. Results are shown in Table 1:

Table 1. Results of flexible fitting on two experiments, using SVD and SVD.

1wdcB2 fit	**Initial value**	**Final value, SVD**	**Final value, ISVD**
RMSD, all backbone atoms / Å	3.68	2.60	2.24
RMSD, SSE backbone atoms / Å	3.52	2.54	2.10
Core aminoacids	-	59	64
1jlxA1 fit	**Initial value**	**Final value, SVD**	**Final value, ISVD**
RMSD, all backbone atoms / Å	3.90	3.53	2.98
RMSD, SSE backbone atoms / Å	3.66	2.46	2.40
Core aminoacids	-	30	158

4. Discussion

We have introduced the use of ISVD algorithm as a new approach to assess the conformational valiability within a given protein superfamily. Doing so, we are able to take into account more aminoacids in the core of the alignment, taking advantage of those missing parts of one element that are present in another element of the superfamily. The results of our tests with different fold architectures show that this approach indeed increases the information captured by the decomposition algorithm, improving the quality of the flexible fitting.

Authors acknowlegde financial support to CICYT (proyect BFU2004-00217/BMC) and EU (proyects EU-512092 and FP6-502828).

5. References

[1] W. Wriggers and S. Birmanns, "Using situs for flexible and rigid-body fitting of multiresolution single-molecule data," J Struct Biol, vol. 133, pp. 193-202, 2001.

[2] F. Tama, O. Miyashita, and C. L. Brooks, 3rd, "Flexible multi-scale fitting of atomic structures into low-resolution electron density maps with elastic network normal mode analysis," J Mol Biol, vol. 337, pp. 985-99, 2004.

[3] M. Topf, M. L. Baker, B. John, W. Chiu, and A. Sali, "Structural Characterization of Components of Protein Assemblies by Comparative Modeling and Electron Cryo-Microscopy," J. Struct. Biol., vol. in press, 2005.

[4] J. A. Velazquez-Muriel and J. M. Carazo, "Flexible fitting in 3D-EM using superfamily information," Bioinformatics, vol. Proceedings of ECCB (September 2005) Madrid, pp. (submitted), 2005.

[5] B. Qian, A. R. Ortiz, and D. Baker, "Improvement of comparative model accuracy by free-energy optimization along principal components of natural structural variation," Proc Natl Acad Sci U S A, vol. 101, pp. 15346-51, 2004.

[6] A. A. Canutescu and R. L. Dunbrack, Jr., "Cyclic coordinate descent: A robotics algorithm for protein loop closure," Protein Sci, vol. 12, pp. 963-72, 2003.

[7] P. Chacon and W. Wriggers, "Multi-resolution contour-based fitting of macromolecular structures," J Mol Biol, vol. 317, pp. 375-84, 2002.

[8] M. E. Brand, "Incremental Singular Value Decomposition of Uncertain Data with Missing Values," Lecture Notes in Computer Science, vol. 2350, pp. 707-720, 2002.

[9] A. R. Ortiz, C. E. Strauss, and O. Olmea, "MAMMOTH (matching molecular models obtained from theory): an automated method for model comparison," Protein Sci, vol. 11, pp. 2606-21, 2002.