FLEXIBLE IMAGE REGISTRATION FOR THE IDENTIFICATION OF BEST FITTED PROTEIN MODELS IN 3D-EM MAPS

Laura Fernández-de-Manuel[1], *María J. Ledesma-Carbayo*[1*], *Julián Atienza-Herrero*[2],
Carlos O. S. Sorzano[2,3], *José-María Carazo*[2], *Andrés Santos*[1]

[1] ETSI Telecomunicación, Universidad Politécnica de Madrid, Spain
[2] Centro Nacional de Biotecnología - CSIC, Madrid, Spain
[3] Escuela Politécnica Superior, Univ. San Pablo – CEU, Madrid, Spain

ABSTRACT

In this work we propose an image registration algorithm to automatically fit protein atomic domain models into medium-resolution three-dimensional electron microscopy reconstructions (3D-EM map). The approach employs a flexible registration algorithm whose optimizer controls the generation of stereo-chemically correct models from a given reference domain belonging to a super-family of proteins. The proposed algorithm generates models automatically in the correct direction until the optimizer converges to the best fitted model. A local gradient optimizer is employed for this task. Mutual Information is used as an alternative to Cross Correlation Coefficient. An additional rigid registration that uses a local gradient optimization is applied between each generated model and the 3D-EM map.

Index Terms— Proteins, image registration, flexible structures, biomedical image processing, electron microscopy.

1. INTRODUCTION

The determination of the precise atomic information of a protein is essential in the understanding of macromolecular functional interactions. Three-dimensional electron microscopy (3D-EM) is a technique to obtain information about macromolecular structures. In essence, it deals with obtaining three dimensional information of the specimen under study from a set of its X-ray projection images; typical resolutions range between 1/6 and 1/25 Å^{-1}. As this resolution range is not high enough to access the atomic information, it is of great value to expand the medium-resolution information with high resolution data coming from other techniques such as X-ray diffraction or Nuclear Magnetic Resonance (NMR). These techniques allow to obtain high resolution crystallographic data of some protein domains. However, not all the domains are suitable to be solved by X-ray diffraction or NMR, so that it is common that only information of a "similar" structure in a "similar" conformation is available. For this reason, two problems are typically faced. The first one consists on the flexible modification of a similar structure to fit into a 3D-EM map, in order to get the convenient modeled protein domain and having into account that conformational changes (changes in atom positions as independent units within the atomic structure) must be biologically realistic. The second problem is the fitting of the modeled protein domain into the 3D-EM map by exploring all possible rotations and translations. This is a rigid registration problem with six degrees of freedom (three translations and three rotations) traditionally solved by maximizing the correlation of the two volumes [1].

In this article, we propose a registration algorithm that takes into account both flexible and rigid movements of the domain to fit in the 3D-EM map. The algorithm employs a biological and evolutionary method proposed by Velázquez-Muriel et al [2] (S-flexfit method) to obtain different modeled protein domains from a reference domain that belongs to a super-family. The best fitted model is found automatically by using the proposed registration algorithm whose optimizer guides the generation of models in the proper direction until convergence. This optimization method improves the traditional use of a grid of sampled models [2], since the model that best fits in the 3D-EM map could not be necessarily included in the grid.

In this work, we study the possibility of using a local gradient-based optimizer combined with Mutual Information (MI) similarity measure in order to automate the generation of models. Furthermore, we study the possibility of making an additional rigid registration between each generated model and the 3D-EM map before applying the MI similarity criteria in order to correct possible local misalignments.

* Corresponding author: mledesma@die.upm.es. Funding was provided by the European Union (FP6-502828 and EMIL LSHC-CT-2004-503569) the US National Institutes of Health (HL070472), the Spanish Ministerio de Educación y Ciencia (CSD2006-00023, BIO2007-67150-C03-1 and CDTEAM) and the Comunidad de Madrid (S-GEN-0166-2006).

978-1-4244-2003-2/08/$25.00 ©2008 IEEE

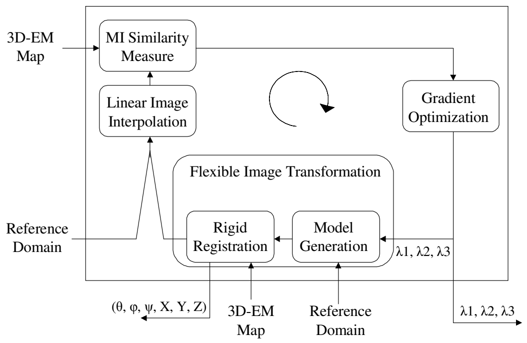

Fig. 1. Diagram of proposed registration steps**.**

2. METHODS

2.1. Flexible and rigid registration scheme.

In order to build the deformed source domain at each flexible image transformation stage, the proposed registration algorithm employs the generation of models process of the program S-flexfit [2]. S-flexfit builds stereo-chemically correct and refined deformed models from a reference protein domain belonging to a super-family. It works by studying the evolutionary structural variability that exists among the domains of the super-family as described in the structural database CATH (http://www.cathdb.info) [3]. The space of conformations of the super-family is decomposed by using singular value decomposition (SVD). The first three singular values of SVD (*λ1*, *λ2* and *λ3)* codify at least 50% of the structural variability of a super-family [2], so only these three values are used to represent the variational space due to computational power constraints. Consequently, the flexible transformation of the reference domain is defined by *λ1*, *λ2* and *λ3* and these are the parameters that the proposed registration algorithm optimizes. These parameters must always be in a range that guarantees biologically meaningful deformations. That is made by taking ranges from CDV (coordinate displacement vectors) provided by S-flexfit [2]. Generated atomic models must be converted into a density map with the resolution of the 3D-EM map before registration.

In most cases, modeled domains must be additionally rotated and translated to fit them exactly into the 3D-EM map. For this objective, an additional rigid registration algorithm is incorporated inside the flexible image transformation process. The algorithm should converge to the best fitted model defined by three flexible parameters *λ1*, *λ2* and *λ3*. There are six additional rigid parameters (three Euler rotations *θ*, *φ*, *ψ* and three translations *X*, *Y* and *Z*) that define the transformation applied over the image of the model as a result of the rigid registration algorithm. The rotation center is the center of mass of the 3D-EM images.

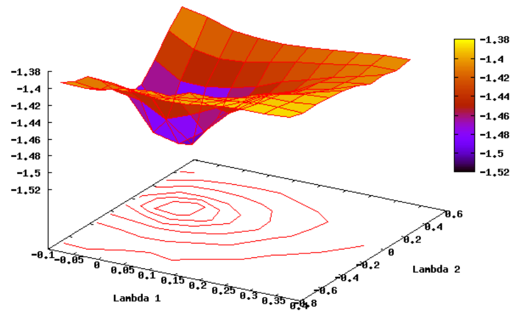

Fig. 2. Normalized Mutual Information calculated from *2rdvA0* at different (*λ1, λ2*) values for the best *λ3* value.

Figure 1 shows the main steps of the proposed registration technique. The registration algorithm has been implemented in C++ within the framework provided by Insight Segmentation and Registration Toolkit (ITK) [4].

2.2. MI Similarity measure

The similarity measure is employed to establish a level of correspondence between images by comparing pixel intensities or other image features. Many measures have been proposed in the literature [5]. In the context of molecular registration, the similarity measure most widely used is the Cross Correlation Coefficient (CCC) [6]. However, recent studies show that the use of Mutual Information (MI) is a good alternative to be employed in molecular image registration algorithms [7]. The Mutual Information similarity measure does not assume that intensities of homologous voxels are the same, as Cross Correlation does, and consequently, is more convenient for images from different modalities.

In this work, we propose to use a Normalized Mutual Information similarity measure computed within the region occupied by the 3D-EM map, in order to establish the correspondence between the 3D-EM image and every flexible transformed domain. The implementation of the Normalized MI is less sensitive to changes in overlap, and is defined by

$$NMI(A,B) = \frac{H(A) - H(B)}{H(A,B)} \qquad (1)$$

where *H(A)* and *H(B)* are the Shannon entropies of images *A* and *B*, and *H(A,B)* the joint entropy of both images. We have actually used an implementation of NMI which estimates entropies by employing joint histograms [8].

The suitability of the NMI similarity measure can be observed in Figure 2, where the NMI has been calculated between the Protein Data Bank (PDB) entry *2rdvA0* and different modeled domains generated from itself with different *λ* values. The figure shows that there is a single well-pronounced global minimum corresponding to the best alignment of the two images, that, in this ideal example, corresponds to (*λ1, λ2, λ3*)=(0,0,0).

Target Domain (PDB code)	Reference Domain (PDB code)	Super-family	*λ1* range	*λ2* range	*λ3* range	cl.	% sim.	% var.	Min. RMSD (Å)
1cll02	1wdcB2	1.10.238.10	(0.187,0.261)	(-0.559, 0.221)	(-0.304, 0.469)	α	30	76.19	3.318
1e5dB2	1mqoA0	3.60.15.10	(0.001, 0.261)	(-0.150, 0.290)	(-0.398, 0.156)	αβ	18	40.67	3.098
1ls9A0	1b7vA0	1.10.760.10	(-0.190, 0.001)	(-0.240, 0.140)	(-0.230, 0.360)	α	31	52.28	2.895
1nm7A0	1ng2A1	2.30.30.40	(0.009, 0.150)	(-0.120, 0.330)	(-0.190, 0.420)	β	29	35.84	2.840
1oyjA1	1jlvC1	3.40.30.10	(0.020, 0.160)	(-0.400, 0.060)	(-0.620, 0.410)	αβ	29	76.60	1.492

Table. 1. Entries of the Protein Data Bank used to simulate data (Target and Reference). Super-families, *λ* value intervals that guarantee biologically meaningful deformations of the reference, structural domain class (cl.), percentage of sequence identity between the target and reference domains (% sim.), percentage of super-family variability contained in the first three singular values of the SVD (% var.) and approximate minimum RMSD values (see text).

2.3. Gradient Optimization

The optimization process has a significant role in the search of the correct transformation that produces the best possible alignment between images. The flexible transformation of the reference domain is defined by *λ1*, *λ2* and *λ3*, so these are the parameters that the registration algorithm optimizes. The nature of our problem results in a similarity measure function with a global minimum (Figure 2). Finding the optimal solution can be solved by using Regular Step Gradient Descent Method (RSGD). Gradient based methods permit to find local minima with precision, and, although they are not accurate with functions with many local minima, in this case its behavior is the most convenient.

2.4 Additional rigid registration conformation

The additional rigid registration algorithm uses a Mutual Information similarity measure calculated within the region occupied by the 3D-EM map and optimized by a Regular Step Gradient Descent Method. However, this rigid registration process just corrects small misalignments between images. Therefore, when working with real data, initial biology knowledge would be required to roughly align the 3D-EM images with the reference prior to applying the proposed registration algorithm.

3. EXPERIMENTS

In order to evaluate the usefulness of the proposed registration algorithm, simulated data from the Protein Data Bank (PDB, http://www.rcsb.org) have been used. The experiment consists on the flexible registration of a given super-family reference domain into a simulated 8 Å resolution map (called Target) generated by another domain of the same super-family (sampling rate 2 Å/voxel). The experiment has been repeated for 5 cases whose properties are represented in Table 1. These domains from each CATH Class level α, β, αβ [3] are selected as a subset of the superfamilies studied in [9]. These domains describe the relative spatial layout of the α helices and β sheets, and are representative enough to test the algorithm. Targets and references are initially aligned by MAMMOTH [10]. This test simulates the case that occurs when the actual structure at atomic level in a 3D-EM map is not solved, but its super-family and a structural relative are known. Since it is simulated data, the atomic resolution image belonging to the target is also available; and this allows calculating the Root Mean Squared Distance (RMSD) between the backbones of the target domain and the resulting best fitted model in order to verify the success of our registration algorithm. The RMSD is calculated identifying the common backbone atoms by performing a structural alignment with MAMMOTH [10], and then computing their RMSD. We assume that the best RMSD obtained by calculating distances between atomic targets and a set of 125 models can be considered as the minimum possible in order to evaluate the values of the obtained best fitted models.

4. RESULTS

Conceptually, RMSD is not directly correlated with any image medium resolution metric, so RMSD values for resulting best fitted models will be always bigger than those obtained with MAMMOTH alignments. We consider that a registration has been convenient when the difference in RMSD with the minimum is less than 2 Å (voxel size) [2]. Table 2 shows the described experiment results. In all cases, except for 1e5dB2, obtained RMSD values are below the minimum value (Table 1) plus 2 Å. For two of the five cases (1cll02, 1ls9A0), RMSD values improve those obtained by applying a grid of a finite number of models (125 models) [2] and the alternative program Colores [1] (CCC Similarity Measure), and for one of the other cases, maintains a similar value (1nm7A0). Only for 1e5dB2, that is the case with significantly less sequence similarity between reference and target, the result worsens. Figure 3 shows the fitting into the target domain 1ls9A0 of two models obtained from the reference domain 1b7vA0 without any registration (left) and after applying the proposed registration algorithm (right). Results have been obtained for the following registration options: *λi* minimum and maximum step sizes 0.001/0.05, rigid parameters minimum and maximum step sizes 0.001/0.4 and relaxation factor for Gradient optimizer 0.5.

Target Domain (PDB code)	Reference Domain (PDB code)	Resulting model parameters			RMSD (Å) (1)	RMSD (Å) (2)
		λ1 (Å)	*λ2* (Å)	*λ3* (Å)		
1cll02	1wdcB2	0.192	-0.112	-0.303	3.601	4.071
1e5dB2	1mqoA0	0.261	-0.147	0.126	8.351	4.341
1ls9A0	1b7vA0	0.001	-0.032	-0.227	3.359	3.410
1nm7A0	1ng2A1	0.150	0.006	0.176	4.083	3.942
1oyjA1	1jlvC1	0.020	-0.182	-0.057	1.950	1.611

Table. 2. Best fitted models (*λ* values) for the five simulated cases after the application of the proposed flexible registration algorithm with additional rigid registration. Comparative of RMSD values calculated for the best fitted model with the proposed algorithm (1) and with Cross Correlation Similarity Measure for a grid of 125 models (Colores [1], [2])(2).

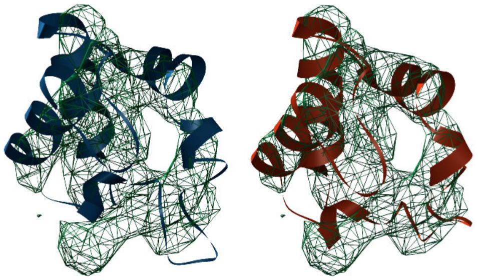

Fig. 3. Left: model from reference domain 1b7vA0 (blue) and target domain 1ls9A0 (mesh) without registration. Right: resulting best fitted model after registration (red) overlaid with target (mesh)(visualized with UCSF Chimera).

5. CONCLUSION

In this work we propose an image registration algorithm that introduces three aspects. Firstly, it permits to automate the generation of protein domain models in an optimized direction, generating a number of models that depends on the convergence of the algorithm for each particular case. Secondly, the approach employs Mutual Information as an alternative to Cross Correlation Coefficient that is traditionally the most widely used in the context of molecular image registration. We have seen that MI improves results in terms of accuracy in most of the cases. Finally, the proposed approach incorporates an additional rigid registration algorithm that is able to compensate small misalignments between images.

In our study, the proposed registration algorithm behaves properly, leading to convenient results in most of the cases. Only in one case (1e5dB2), the proposed algorithm results are not satisfactory, indicating the possible need of a more global optimization strategy to avoid local minima. Further studies are guaranteed in this direction.

The proposed registration algorithm can be used as an alternative to other programs, improving outcomes for some of the cases. Results seem promising and should be confirmed on experimental data. Experimental data has not been used yet due to predicted possible problems. First of all, the number of members of the super-family of the atomic structure could not be enough to compute a valid super-family variability. Secondly, calculating RMSD between the backbones of the target domain and the resulting best fitted model in order to verify the success of the registration algorithm would not be possible. At this point, it is necessary also to consider, for future research, extending the number of singular values of SVD in order to increase the percentage of super-family variability contained in *λ* parameters.

6. REFERENCES

[1] P. Chacón and W. Wriggers, "Multi-resolution contour-based fitting of macromolecular structures," *J Mol Biol*, vol. 317, no. 3, pp. 375-384, Mar. 2002.

[2] J. Á. V. Muriel, M. Valle, A. S. Pang, I. A. Kakadiaris, and J. M. Carazo, "Flexible Fitting in 3D-EM Guided by the Structural Variability of Protein Superfamilies," *Structure*, vol. 14, pp. 1115-1126, July 2006.

[3] C. A. Orengo, F. M. G. Pearl, and J. M. Thornton, "The Cath Domain Structure Database," *Methods Biochem.*, vol. 44, pp. 249-271, 2003.

[4] L. Ibañez, W. Schroeder, L. Ng, and J. Cates, *The ITK Software Guide*. Insight Software Consortium, Nov. 2005.

[5] B. Zitová and J. Flusser, "Image registration methods: a survey,"*Image Vis Comput* vol. 21, pp.977–1000, 2003.

[6] J. Frank, *Three-Dimensional Electron Microscopy of Macromolecular Assemblies. Visualization of Biological Molecules in Their Native State.* Oxford Univ. Press, 2006.

[7] B. Telenczuk, M. J. Ledesma-Carbayo, J. A. Velazquez-Muriel, C. O. S. Sorzano, J.-M. Carazo, and A. Santos, "Molecular Image Registration using Mutual Information and Differential Evolution Optimization," in *3rd IEEE Int Symp. on Biomedical Imaging*, Washington DC, 2006.

[8] F. Maes, A. Collignon, D. Vandermeulen, G. Marchal, and P. Suetens, "Multimodality Image Registration by maximization of Mutual Information," *IEEE Trans Med Imaging*, vol. 16, no. 2, pp. 187-198, April 1997.

[9] A. Leo-Macias, P. Lopez-Romero, D. Lupyan, D. Zerbino, and A. R. Ortiz, "An Analysis of Core Deformations in Protein Superfamilies," *Biophys*, vol. 88 pp. 1291-1299, 2005.

[10] A. Ortiz, C. Strauss, and O. Olmea, "MAMMOTH (matching molecular models obtained from theory): an automated method for model comparison," *Protein Sci*, vol. 11, pp. 2606-2621, 2002.