Available online at www.sciencedirect.com
ScienceDirect
Journal of Structural Biology 157 (2007) 271–280
Journal of Structural Biology
www.elsevier.com/locate/yjsbi

Multi-resolution anchor-point registration of biomolecular assemblies and their components

Stefan Birmanns, Willy Wriggers *

School of Health Information Sciences, University of Texas Health Science Center at Houston, Houston, TX 77030, USA

Received 8 May 2006; received in revised form 1 August 2006; accepted 4 August 2006
Available online 25 August 2006

Abstract

An atomic scale interpretation facilitates the assignment of functional properties to 3D reconstructions of macromolecular assemblies in electron microscopy (EM). Such a high-resolution interpretation is typically achieved by docking the known atomic structures of components into the volumetric EM maps. Docking locations are often determined by maximizing the cross-correlation coefficient of the two objects in a slow, exhaustive search. If time is of essence, such as in related visualization and image processing fields, the matching of data is accelerated by incorporating feature points that form a compact description of 3D objects. The complexity reduction afforded by the feature point representation enables a near-instantaneous matching. We show that such reduced matching can also deliver robust and accurate results in the presence of noise or artifacts. We therefore propose a novel multi-resolution registration technique employing feature-based shape descriptions of the volumetric and structural data. The pattern-matching algorithm carries out a hierarchical alignment of the point sets generated by vector quantization. The search-space complexity is reduced by an integrated tree-pruning technique, which permits the detection of subunits in large macromolecular assemblies in real-time. The efficiency and accuracy of the novel algorithm are validated on a standard test system of homo-oligomeric assemblies.

Keywords: Feature points; Vector quantization; Laplace filter; Docking; Interactive modeling

1. Introduction

By combining data from multiple biophysical sources at multiple levels of detail one can take advantage of the complementary strengths of various 3D structure determination methods in biology. This multi-resolution modeling approach often yields new insights into the architecture of biomolecular assemblies. Clearly, the model as a whole is then greater than the sum of its biophysical parts, in a spatial sense (considering the buildup of large functional biological 'machines' from their 'machine parts'), resolution sense (considering the interpretation of volume data in terms of atomic structures), and functional sense (considering the possible conformational polymorphism of underlying structures).

Over the last years multi-resolution modeling tools (Wriggers and Chacón, 2001; Rossmann et al., 2005) have gained in popularity among structural biologists and a large number of software packages were developed. For rigid-body fitting there are two classes of programs, interactive tools (Jones et al., 1991; Birmanns and Wriggers, 2003) that assist a manual fit 'by eye' of the user, and 'algorithmic' tools like Situs (Wriggers et al., 1999; Chacón and Wriggers, 2002), COAN (Volkmann and Hanein, 1999), DockEM (Roseman, 2000) and EMFit (Rossman, 2000) that use a quantitative scoring function to generate the results automatically. Many of the routines are able to obtain a fit even if the structure represents only a subunit of a larger assembly. At a conceptual level, the available algorithms typically aim to maximize the cross-correlation

* Corresponding author. Fax: +1 713 500 3907.
E-mail address: wriggers@biomachina.org (W. Wriggers).

1047-8477/$ - see front matter © 2006 Elsevier Inc. All rights reserved.
doi:10.1016/j.jsb.2006.08.008

coefficient C by employing an exhaustive six-dimensional search. Various definitions of C were suggested (Wriggers and Chacón, 2001) and in some cases the exhaustive search can be accelerated by Fourier-space methods.

Our goal is to combine the advantages of both classes of algorithms. We wish to compute a quantitative measure of the fit, but in real time such that the scoring function may be used for interactive exploration of the model. Therefore we are using here ideas from computer vision, pattern and speech recognition where unsupervised clustering techniques are often employed to characterize the data in a compact or compressed state. Such reduced representations can improve the robustness of manipulation and interpretation methods and simplify data analysis. Clustering techniques such as vector quantization (Gersho and Gray, 1992) (VQ), provide flexible, general purpose tools for the feature-point determination. In electron microscopy such reduced models were already successfully applied to rigid-body docking (Wriggers et al., 1998, 1999), and in the modeling of structural flexibility (Wriggers and Chacón, 2001). Vector quantization was also utilized in normal mode analysis of EM data (Tama et al., 2002; Ming et al., 2002; Chacón et al., 2003) as basis for an elastic network of mass elements.

A feature-based shape description recasts the multi-resolution fitting problem into a point-cloud matching task. An exhaustive search as described in Wriggers et al. (1999) enables a matching of similar shapes for a small number of feature points. Here we have extended the earlier approach to cases where a smaller probe structure is to be matched with a much larger oligomeric assembly. This leads to a pattern recognition task where one has to find a similar subset of points in a larger point cloud. The complexity of such a scenario renders an exhaustive search unfeasible in practical applications, but a tree-pruning algorithm keeps the number of plausible combinations of matched points reasonably limited. Our novel anchor-point matching uses a hierarchical search strategy that exploits the point density properties of the VQ data sets. The accuracy of the new algorithm enables the detection of subcomponents in large assemblies, and its efficiency enables data-mining in collections of volumetric or atomic structures. Fig. 1 describes the work flow of the novel docking approach. Fig. 2

In the following three sections we will describe the computation of feature points, the anchor-point fitting (tree pruning), and the refinement of roughly aligned structures. Subsequently, we describe the results of our performance tests and validations as well as implementation details.

2. Feature-based shape description

The registration proposed in this paper does not directly correlate the probe and target structures, instead it depends on the comparison of intermediate feature vectors. The similarity of two sets of points $\mathbf{w}_i^{\text{calc}}$ (corresponding to high-resolution data) and $\mathbf{w}_j^{\text{em}}$ (corresponding to low-resolution data) with $i \in \{0, \ldots, N\}$ and $j \in \{0, \ldots, M\}$ $(N \leqslant M)$ determines the optimal position and orientation of the probe molecule within the target map. The similarity of point sets and its quantification are key challenges in related disciplines like computer vision or pattern matching (we refer to (Alt and Guibas, 1996) for a review). In this paper we rely on the root-mean-square deviation (RMSD) as error metric. The RMSD is defined as

$$\mathrm{RMSD}(I) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\left\|\mathbf{w}_i^{\text{calc}} - \mathbf{w}_{I(i)}^{\text{em}}\right\|\right)^2},$$

where the index map $I: i \to j$ identifies corresponding feature points in the atomic data and the low-resolution map. We also implemented an alternative metric, the

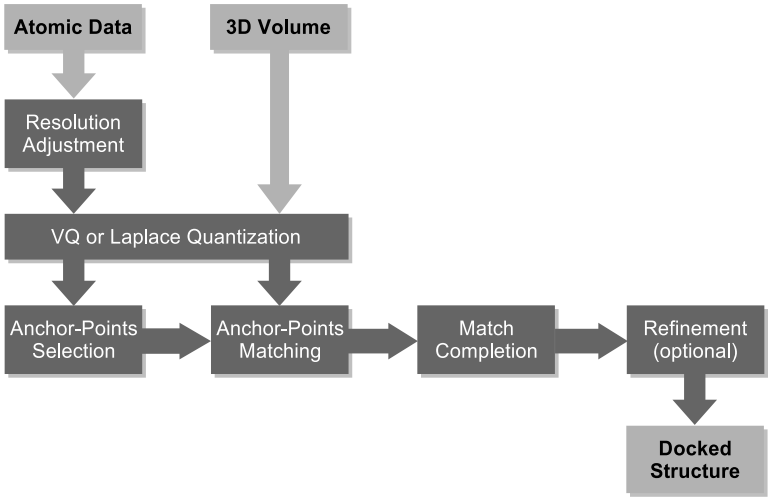

Fig. 1. Overview over the proposed multi-resolution fitting method. After the atomic structure of the probe molecule and the target density map are loaded into the program, the atomic structure is low-pass filtered to the known resolution of the 3D volume (Chacón and Wriggers, 2002). Subsequently, feature-point sets for both objects are calculated. In the case of the probe molecule three anchor points are selected and matched with the feature points of the target map. The match is completed in a next step and the result is refined in an optional post-processing step. Finally, the results are presented to the user.