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

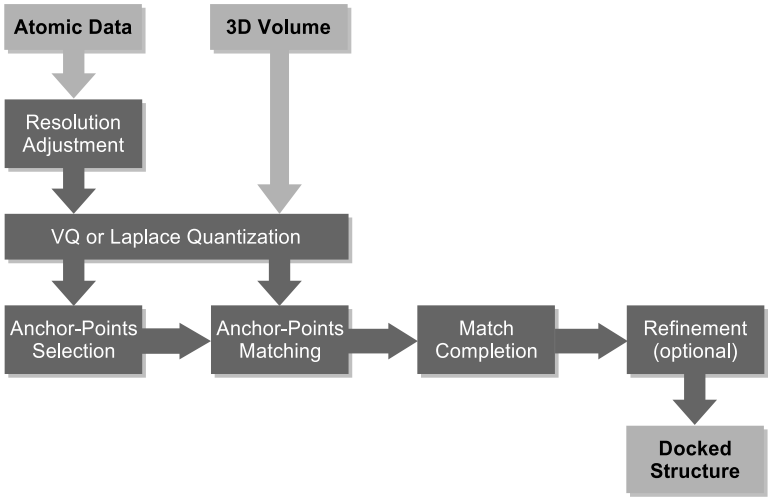

Fig. 1. Overview over the proposed multi-resolution fitting method. After the atomic structure of the probe molecule and the target density map are loaded into the program, the atomic structure is low-pass filtered to the known resolution of the 3D volume (Chacón and Wriggers, 2002). Subsequently, feature-point 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.

Hausdorff distance, which measures the deviation of the worst outlier

$$h(\mathbf{w}^{\text{calc}}, \mathbf{w}^{\text{em}}) = \max_{i\in\{0,\ldots,N\}} \min_{j\in\{0,\ldots,M\}} \left\| \mathbf{w}_i^{\text{calc}} - \mathbf{w}_j^{\text{em}} \right\|.$$

The positions of the simulated markers are determined by vector quantization, a clustering method that encodes complex multi-dimensional data using a discrete set of features, the so-called codebook. VQ is best known for its use as a 'lossy' data compression technique in speech and image processing. We briefly highlight important features of the *neural gas* algorithm, implementation details are given elsewhere (Wriggers et al., 1998).

2.1. Vector quantization

Starting from a random initial configuration, the points $\mathbf{w}_i^{\text{calc}}$ and $\mathbf{w}_j^{\text{em}}$ are determined by the neural gas network in an unsupervised training phase, in which a data point $\mathbf{v}$ is randomly picked according to a probability density function $\rho(\mathbf{v})$. For volumetric data, $\rho(\mathbf{v})$ is the (normalized) voxel intensity. In case of atomic structures, these are first low-pass filtered to the same resolution as the 3D volume, consistent with Fig. 1, i.e. $\rho(\mathbf{v})$ is proportional the low-pass-filtered mass density.

The network nodes n are updated according to

$$\Delta\mathbf{w}_n = \epsilon(t) e^{-\kappa_n(t)/\lambda(t)} (\mathbf{v} - \mathbf{w}_n),$$

where $\epsilon(t)$ and $\lambda(t)$ are iteration-dependent training parameters and $\kappa_n(t)$ describes the closeness rank of node n. In contrast to $\epsilon(t) = \epsilon_0(\epsilon_{\text{fin}}/\epsilon_0)^{t/t_{\text{fin}}}$ and $\lambda(t) = \lambda_0(\lambda_{\text{fin}}/\lambda_0)^{t/t_{\text{fin}}}$, the neighborhood ranking κ_n is comparatively expensive to compute and stands for the number of nodes $\mathbf{w}_k$ with $\|\mathbf{v} - \mathbf{w}_k\| < \|\mathbf{v} - \mathbf{w}_n\|$. The user-defined initial and final parameters λ_0 and λ_{fin} determine the plasticity of the network, whereas ϵ_0 and ϵ_{fin} adjust the level of adaptation during each iteration.

The stochastic neural gas algorithm is known to minimize the distortion error (Martinetz et al., 1993), a measure for the information loss due to the quantization. This guarantees that the calculated codebook is the best possible reduced representation of the original object. Therefore, the feature points generated by the neural gas algorithm are stable under changes of resolution and describe the overall shape and density distribution of the biological object.

In (Wriggers and Birmanns, 2001) extensive tests on the statistical stability of the feature points were performed. Ten different test systems were filtered to various resolutions and were quantized with codebooks of different

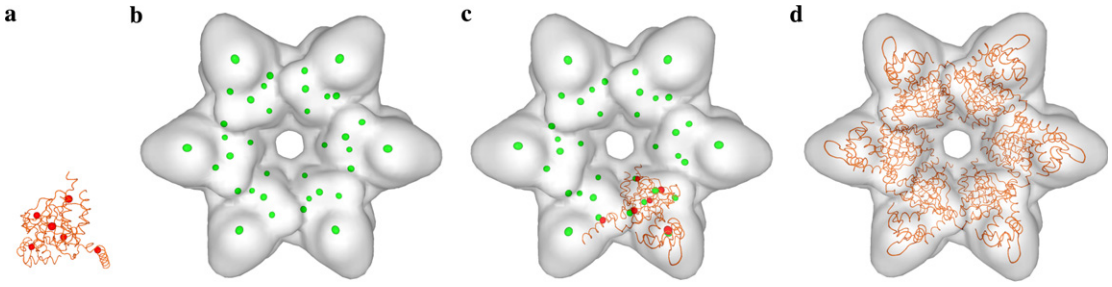

Fig. 2. Feature-based multi-resolution docking: Construction of the RecA helicase (PDB entry 2REC, simulated volumetric map at 15 Å resolution). (a) Monomeric structure with feature points; (b) vector quantized volumetric data set; (c) docked substructure; (d) constructed assembly.

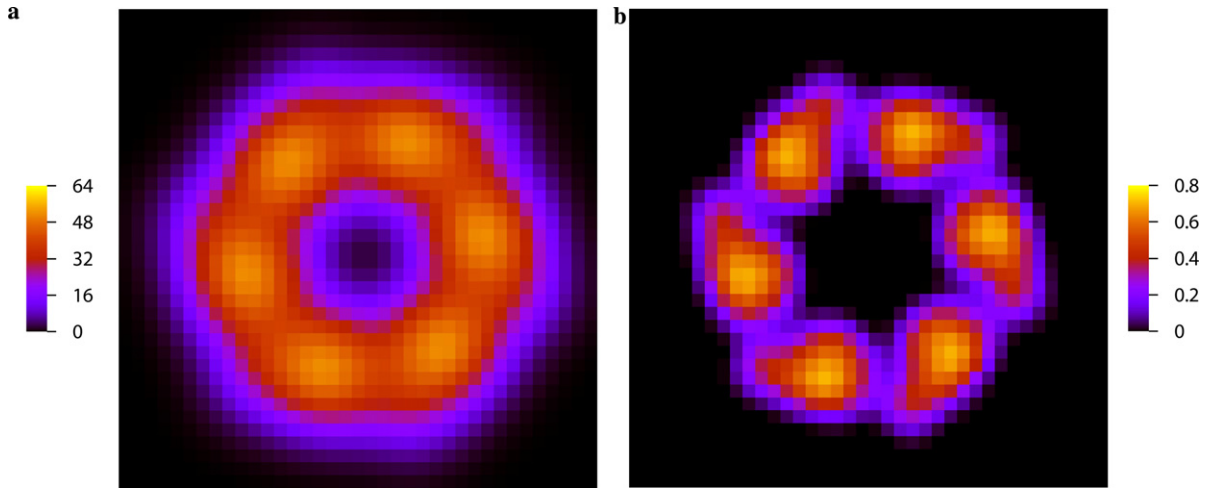

Fig. 3. Volumetric map of the RecA hexamer: Cross-section of (a) volumetric map and (b) interior voxels of Laplace-filtered map (absolute values).

cardinality. It was shown that systems with a low statistical variability of the feature points typically achieve more accurate docking results and that the variability increases with decreasing resolution of the volumetric data. For similar shaped target and probe molecules, stable features and accurate docking results were observed already for small codebooks of $3 \leqslant N \leqslant 9$, demonstrating the promise of this approach. Historically, docking by feature points of similar-sized structures was the algorithmic approach implemented in the early versions of the Situs package (Wriggers et al., 1999).

Here we have expanded the classic approach to identify subunits in larger assemblies. The new application area demands to revisit certain aspects of the feature identification process. Initial practical tests have revealed that the stability of the feature points in the case of a multi-molecule docking scenario is more problematic than in the single-molecule docking case. To further stabilize the feature points we introduce in the following an extended clustering procedure using a Laplacian filter.

2.2. Laplace quantization

When the resolution of test systems is lowered, the accuracy of algorithmic docking often breaks down early, depending on the shape of the matched biomolecular objects (Wriggers and Chacón, 2001). This is due to the loss of interior (or secondary structure) information at a resolution below 10 Å. Among several attempts to push the limits of automatic docking programs, the application of a Laplacian edge enhancement filter has proven to be very successful (Chacón and Wriggers, 2002). The Laplacian essentially boosts the contrast of volumetric maps, and thereby enhances the contour and also the interior detail.

The success of the Laplace-filter for the correlation-based docking calls for an adaptation to our reduced feature-point based matching. To conserve the theoretical properties of the original VQ algorithm described above, we impose the Laplacian by a preprocessing procedure.

As described in the previous section, our stochastic VQ approach is trained by randomly selected voxels according to a probability $\rho(\mathbf{v})$. Since the Laplace-filtered map features positive and negative intensities, the mapping of the intensities to a positive probability is no longer obvious.

The sign of the intensity values introduces a segmentation of features, negative values typically correspond to the ''interior'' of the density, whereas positive values correspond to the ''contour''. By separating the interior and the contour segments, one obtains two maps that can be clustered independently. The result are two sets of points, $\mathbf{w}_j^{\mathrm{em}_{\mathrm{int}}}$ and $\mathbf{w}_j^{\mathrm{em}_{\mathrm{cont}}}$ for the target EM map, and two sets $\mathbf{w}_i^{\mathrm{calc}_{\mathrm{int}}}$ and $\mathbf{w}_i^{\mathrm{calc}_{\mathrm{cont}}}$ for the probe molecule. A docking based on these four point sets is challenging and requires the correct weighting and interpretation of interior and contour matching. However, practical tests have revealed that the contour corresponds to a relatively thin shell that renders an independent clustering irrelevant. We have shown in (Passon et al., 2005) that the matching accuracy depends almost exclusively on the interior marker points, because the position of points in a contour is less robust if the contour becomes buried in the assembly. Therefore, we have used in this paper only the more robust interior map information. In the following Laplace quantization refers to $\mathbf{w}_i^{\mathrm{calc}_{\mathrm{int}}}$ and $\mathbf{w}_j^{\mathrm{em}_{\mathrm{int}}}$, and we drop the 'int' suffix for simplicity.

Fig. 3 shows cross-sections of a volumetric map before and after Laplacian filtering as defined above, i.e. the right image shows only 'interior' voxels. The rendering highlights the effect of the convolution with the Laplacian kernel and the effect of ignoring the contour. One can identify well-segmented densities that roughly correspond to the monomers. These segmented features are more suitable for our pattern recognition approach than the unfiltered density on the left.

The overall robustness of feature points in the assembly and the effect of the Laplacian filtering are shown in Fig. 5. The plot shows how feature points drift from the monomer positions when they become embedded into an assembly, depending on the resolution of the matching setup. Both the RMSD and the Hausdorff distance were computed. The measures increase with lower resolution, indicating a resolution-dependent drift of the feature points. However, the Laplace quantization outperforms the standard VQ significantly. Laplace quantization stabilizes the point RMSD and the deviation of the worst outlier at about 20%, and 30% of the nominal resolution, respectively.

3. Anchor-point fitting

By quantizing both data sets—the probe molecule and the volume from electron microscopy—one obtains two codebooks, $\mathbf{w}_i^{\mathrm{calc}}$ and $\mathbf{w}_j^{\mathrm{em}}$ with $i \in \{0, \ldots, N\}$ and $j \in \{0, \ldots, M\}$. Performing a multi-resolution docking in this context means to find an optimal index map $I : i \to j$ and a transformation that minimizes the RMSD.

For a given I the optimal rigid-body transformation, defined by the rotation $\mathbf{R}(I)$ and the translation $\mathbf{t}(I)$, is determined by a least-squares fit (Kearsley, 1989; Kabsch, 1976, 1978. However, there are $\frac{M!}{(M-N)!}$ possible index maps I. Even for the small point clouds of interest, the size of the search-space becomes prohibitively large.

In principle three well chosen point pairs would suffice to determine the six rigid-body degrees of freedom. Based on this observation we propose an iterative search procedure in which one first only matches three appropriately chosen point pairs. Although the resulting transformation is not necessarily optimal for the entire point set, it forms a basis for a subsequent match completion and refinement routine. The initial three points are often referred to as ''anchor-points'' and are a common theme in motif search and structure alignment algorithms. The algorithm proposed here is similar for example to Geometric Hashing (Nussinov and Wolfson, 1991) which is also based on the idea of anchor points. We will exploit the fact that the

feature point sets are not as densely packed as atomic data. This allows us to omit the complex clustering step of the Geometric Hashing algorithm and to directly refine the matching in a very efficient, iterative manner.

3.1. Anchor-point selection

The task at hand is to select three suitable features in the probe molecule as anchor points. We have already seen in Fig. 5 that there are small deviations between the points corresponding to probe and target. To minimize the effect of such deviations on the matching accuracy, the anchors should be sufficiently separated. Here we chose as first anchor the point which exhibits the largest radial distance from the center of mass of the probe molecule, as second anchor the one which exhibits the largest distance from the first one and as third the one which exhibits the largest distance from the first two.

As an alternative anchor point selection scheme one may also pick feature points that exhibit the smallest statistical VQ variability. This variability arises due to the stochastic nature of the neural gas algorithm, and gives an indication of the convergence and robustness of an individual feature point (Wriggers et al., 1999). Typically, a small number of VQ calculations (here: eight) are repeated with a different seed of the random number generator. One may then pick the three points with the smallest mean-square variation as anchors.

In principle, any three points in $\mathbf{w}^{\text{calc}}$ can serve as anchor points in the matching procedure. In the following we have implemented and considered results from the above two anchor point selection schemes. The run time of the matching (on order of a second) is sufficiently short to allow the implementation of multiple strategies. This extensible approach allows us to add selection schemes in the future that reflect empirical observations about the stability and suitability of feature points.

To reduce the complexity of the point matching problem, the selected anchors $\mathbf{w}_A^{\text{calc}}, \mathbf{w}_B^{\text{calc}}, \mathbf{w}_C^{\text{calc}}$ of the probe structure are only tested with points of the target map that are part of a similar shaped triangle. Therefore the target point set is pre-filtered in a two-step procedure:

- First we screen out all features i,j with $\|\mathbf{w}_i^{\text{em}} - \mathbf{w}_j^{\text{em}}\| < \|\mathbf{w}_A^{\text{calc}} - \mathbf{w}_B^{\text{calc}}\| + \delta$, where δ is an edge length tolerance parameter, and $\|.\|$ denotes the Euclidean distance.
- Subsequently, within this set we screen out a subset for which a k exists such that $\|\mathbf{w}_i^{\text{em}} - \mathbf{w}_k^{\text{em}}\| < \|\mathbf{w}_A^{\text{calc}} - \mathbf{w}_C^{\text{calc}}\| + \delta$ and $\|\mathbf{w}_k^{\text{em}} - \mathbf{w}_j^{\text{em}}\| < \|\mathbf{w}_C^{\text{calc}} - \mathbf{w}_B^{\text{calc}}\| + \delta$ are satisfied.

The resulting set of triangles forms the search space for the next stage of the algorithm. The choice of tolerance parameter δ depends on the robustness and density of the feature points, practical values are 5–15 Å.

Although it was not important here, it is straightforward to implement an optimization of the triangle matching routine using advanced geometric data structures (Langetepe and Zachmann, 2006). This would enable a comparison of very large point clouds.

3.2. Match completion

The initial rigid body transformation is not necessarily optimal for the entire object as it only relies on three point pairs instead of the entire feature sets. However, the initial transformation can be improved by augmenting the index map I with more point pairs in addition to the anchor points. In the match completion stage of the algorithm, the features of the transformed probe point set are therefore matched with their neighbors in the target map. Each matched point pair is investigated in a recursive manner, forming a search tree that is visited using a depth first strategy.

The match completion algorithm starts with the initial index map I_0 (and corresponding least-squares fit) of the anchor points, providing a transformation $\mathbf{R}(I_0)$, $\mathbf{t}(I_0)$. Subsequently, for each unmatched point $\mathbf{w}_k^{\text{calc}}$ the nearest $\{\mathbf{w}_l^{\text{em}} | \|\mathbf{w}_l^{\text{em}} - \mathbf{R}\mathbf{w}_k^{\text{calc}} - \mathbf{t}\| < \gamma\}$ are considered as possible matches, where γ is an adjacency tolerance that limits the radius of the zone that is searched for unmatched target points $\mathbf{w}_l^{\text{em}}$. By adding the corresponding point pairs ($\mathbf{w}_l^{\text{em}}$, $\mathbf{w}_k^{\text{calc}}$) one at a time to I_0, a group of new possible index matches $I'_0, I'_1, \ldots, I'_K$ is generated, where K is the total number of pairs satisfying the adjacency criterion. Each index map generates a new transformation and so forth. Since all the potential matchings I'_k are investigated, the search tree grows K new branches. The search is terminated for $K = 0$ (tree pruning) or when all N point pairs are matched.

The size of the search tree depends on the number of potential matching partners K, which in turn depend on the density of the point distribution and on the tolerance γ. As with δ, the tolerance γ is typically set to 5–15 Å in practical applications. This results in only a few $K = 1$ or $K = 2$ matching partners. We limit the maximum number of investigated neighbors to a user definable cap (here: $K = 5$) to ensure a convergence of the algorithm even in worst-case scenarios. The scheme in Fig. 4 provides an overview of the initial anchor match and the match completion phase.

The order in which the unmatched points $\mathbf{w}_k^{\text{calc}}$ are analyzed becomes a factor if noise is present in the experimental data. Initially the number of matched point pairs is small, so if an outlier is matched early, the probe molecule will be significantly displaced from the ideal docking position and the rest of the feature points will be missed or mismatched. To avoid this, we sort the unmatched points $\mathbf{w}_k^{\text{calc}}$ by the distance to the nearest $\mathbf{w}_l^{\text{em}}$.

3.3. Complexity and efficiency

The complexity of the first part of the algorithm—the anchor point matching—is bound by the potential number of feature points with compatible distances to the anchor points. Theoretically, the maximal number of potential

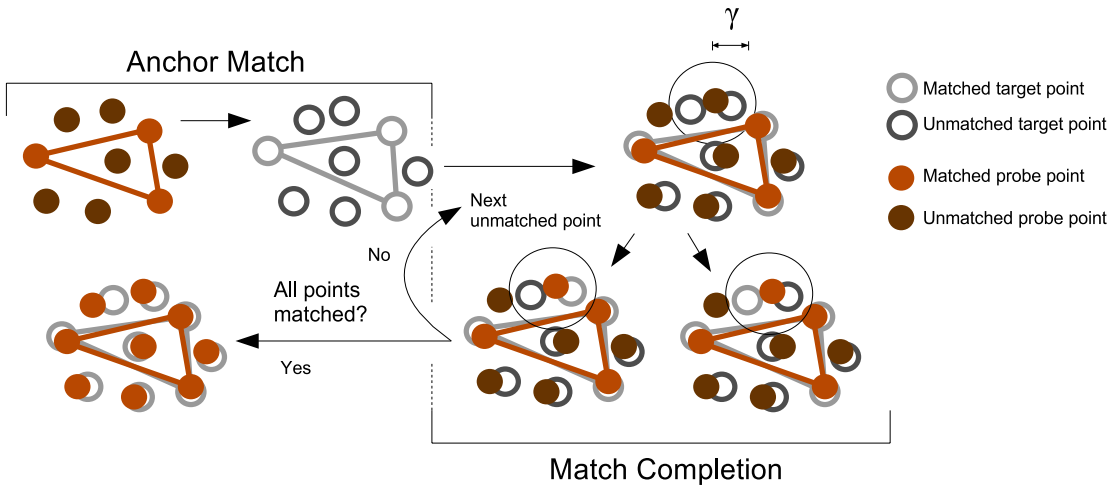

Fig. 4. Schematic rendering of the hierarchical matching algorithm. In the initial anchor matching phase three point pairs are matched and the resulting transformation is applied to the entire probe molecule. In the match completion phase, this initial match is completed by adding unmatched target points in the vicinity of already matched probe points. With each point pair added, the transformation is refined, i.e. a new least-squares fit is performed. Multiple potential matching partners $K > 1$ are investigated independently, leading to a search tree.

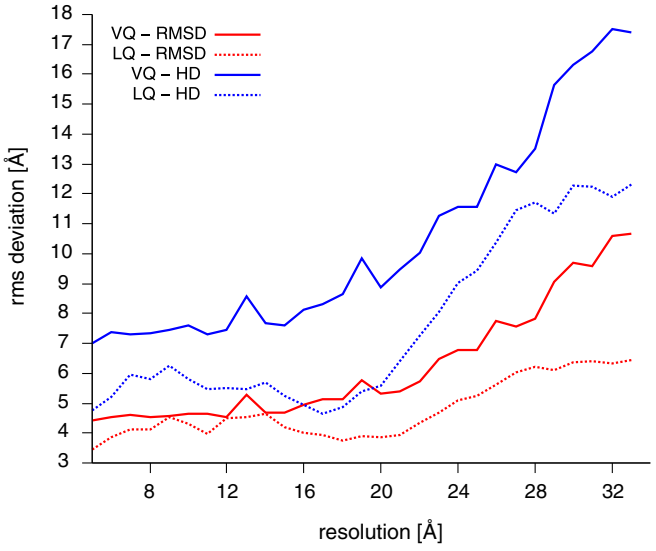

Fig. 5. Robustness of feature points in a multi-component fitting scenario. A catalase tetramer (PDB entry 7CAT) and one of its monomers were low-pass filtered (Chacón and Wriggers, 2002) and vector quantized to create an idealized matching setup. The RMSD and Hausdorff distance between the two point clouds were computed (see text) as a function of resolution, to provide a measure of the invariance of the feature points under polymerization. The results for the standard VQ algorithm and the proposed Laplacian quantization (LQ) are shown.

matching triangles in the target point set is $\binom{N}{3} = O(N^3)$, in the unrealistic case that all edges have the same length. In practice the features do not have such homogeneous distances and one typically experiences a complexity of around $O(N^2)$. The second part of the algorithm is bound by the number of potential matching partners K in the vicinity of an unmatched probe feature. As discussed above this branching number K is commonly very small. An upper bound of the complexity is $O(K^{N-3})$ since every potential matching partner has to be considered in a recursive manner, but some of the branches are pruned. Table 1 shows some actual docking times as a function of cardinality (N, M) on a standard PC-Linux computer. The exact run times depend also on the choice of tolerances δ and γ, as well as the feature robustness, resolution and number of anchor point matches, and are therefore system dependent.

We tested the scalability of the algorithm with respect to the size of the target map. Actin is an ideal test-system as it allows the construction of polymeric filaments (F-actin) of variable length which can be utilized for efficiency validation with a constant point density (Fig. 6). Oligomeric F-actin structures of variable length were constructed starting from the atomic structure of a single G-actin monomer (Wriggers and Schulten, 1999) following the helical symmetry of the filament (Lorenz et al., 1993). Fig. 6 shows the observed run time of the algorithm as a function of monomeric subunits. The moderate increase in run time

Table 1
Efficiency of the feature point matching algorithm for various cardinality (number of points) N and M

System	N	M	δ (Å)	γ (Å)	Run time (s)
RecA	6	36	15	12	0.32
RecA	10	60	12	9	1.69
RecA	18	108	8	5	3.92
GroEL	6	84	15	12	2.96
GroEL	10	140	12	9	8.52
GroEL	18	252	8	5	26.41

The input data were a RecA hexamer (Yu and Egelman, 1997) low-pass filtered to 15 Å resolution, and the GroEL 6 Å 3 D reconstruction available at the EBI EM data base. The atomic structures of the monomers were obtained from the PDB entries 2REC and 1OEL. The run times were measured on an Intel Pentium 4, 3.0 GHz Linux PC. Overall, the high performance of the proposed algorithm is due to the branch limit that prevents the combinatorial explosion of a full search.

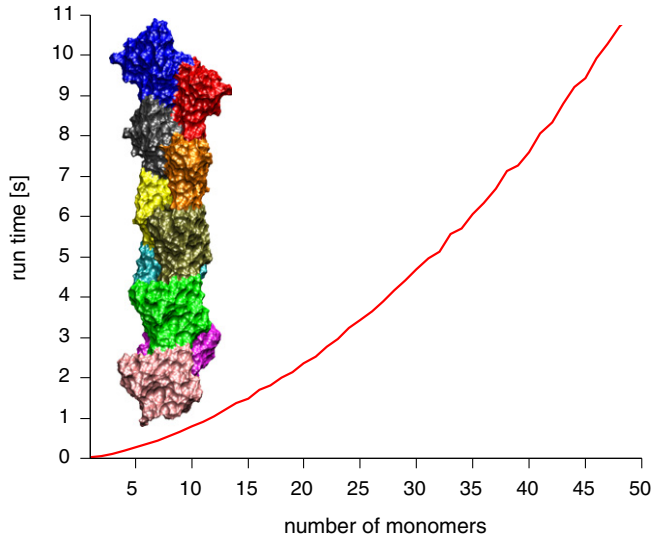

Fig. 6. Run-time of the algorithm as function of the size of the target system. A single actin monomer structure (Wriggers and Schulten, 1999) was docked into simulated EM maps created from oligomers. The insert shows the arrangement of multi-colored monomers in the assembly. Oligomers of variable length (1–50 monomers) were created following the helical symmetry of the actin filament Lorenz et al., 1993. Each monomer was represented by $N = 5$ feature points, and the simulated EM maps by $M = 5$ to $M = 250$ points dependent on the number of monomers. The parameters $\delta = 15$ Å and $\gamma = 12$ Å were used for the matching algorithm.

demonstrates that the algorithm scales well towards large assemblies ($N \ll M$), enabling fast matching in real compute time.

3.4. Outlier suppression

The practical tests with EM maps reveal that even with Laplace quantization there are certain shifts in the position of feature points that increase with lower resolution (Fig. 5). These shifts may be due to a lack of interior detail in low-resolution maps, or due to the effect of polymerization on the density and on the corresponding point distribution. To further reduce the effect of these discrepancies we implemented an optional approximative matching with outlier suppression that is described in the following.

We define a wildcard index map $I(j) = -1$ that enables the algorithm to skip outliers in the special case $K = 0$. The optimization criterion is then defined as

$$\mathrm{RMSD}(I) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\mathrm{dist}^2(i)}, \quad \text{where}$$

$$\mathrm{dist}(i) = \begin{cases} \left\| (\mathbf{R}\mathbf{w}_i^{\mathrm{calc}} + \mathbf{t}) - \mathbf{w}_{I(i)}^{\mathrm{em}} \right\| & \text{if } I(i) > 0 \\ 0 & \text{else} \end{cases}$$

Note that we have defined an alternative way of handling the case $K = 0$. With outlier suppression, dead branches are no longer automatically pruned, instead we are free to set the adjacency threshold γ to a smaller allowed point deviation which results in more dead branches. In practice, the number of these wildcards should be limited since a too liberal use may lead to false positives if significant parts of the probe molecule are no longer represented in the matching. Practical tests suggest that not more than 0.1 N wildcards should be introduced.

Also we note that the above RMSD criterion tends to favor wildcard matches as they do not contribute to the RMSD. To encourage complete point assignments, wildcard matches should be penalized. If N' is the number of matches i with $I(i) > 0$, the point cloud similarity measure we actually implemented is given by:

$$\mathrm{RMSD}(I) = \sqrt{\frac{1}{N}\left((N - N')p_{\mathrm{wc}}^2 + \sum_{i=1}^{n}\mathrm{dist}^2(i)\right)},$$

where p_{wc} is the wildcard penalty distance. In practice p_{wc} should be of the order of the feature point separation to ensure proper weighting relative to the standard deviations. The outlier suppression scheme is an approximative point matching technique since it considers only $N' < N$ points. We refer to Alt and Guibas (1996) for a general review of approximative matching algorithms.

4. Post-matching refinement

The robustness and short run time of the proposed feature-point docking approach is in part based on the compactness of the description of the biological objects. Instead of considering all the voxel intensity values, here only a small number of feature points is used to register the probe molecule in the polymeric assembly. The downside of such a compact description is that a small error is unavoidable in practice. To further reduce the fitting error, we explored two additional refinement techniques.

Most algorithmic correlation-based fitting methods rely on a grid search with translational and angular steps and therefore typically benefit from a real-space refinement of the found solution, leading to a local optimization of a found fit. In (Chacón and Wriggers, 2002) an off-lattice refinement step was developed, based on the standard cross-correlation coefficient

$$C(\mathbf{R},\mathbf{t}) = \int \rho^{\mathrm{em}}(\mathbf{v})\rho^{\mathrm{calc}}(\mathbf{v},\mathbf{R},\mathbf{t})\,\mathrm{d}^3\mathbf{v}. \tag{1}$$

This refinement step has been implemented as a standalone tool *colacor* in version 2.2 of the Situs docking package (http://situs.biomachina.org). However, the full computation of the cross-correlation does not benefit from the reduced quantization of the data we propose in this paper, and requires significant extra time.

In (Birmanns and Wriggers, 2003) we proposed an alternative, more efficient, cross-correlation algorithm for the high-force update rates required in interactive modeling applications. This algorithm is based on a vector quantization of the probe molecule $\mathbf{w}_i^{\mathrm{calc}}$ which leads to the following approximation:

$$\rho^{\text{calc}}(\mathbf{v},\mathbf{R},\mathbf{t}) \approx \sum_{i=1}^{N} \delta(\mathbf{v} - \mathbf{w}_i^{\text{calc}}(\mathbf{R},\mathbf{t}))$$

After introducing the approximation into Eq. (1) the coefficient then assumes the following reduced form

$$C(\mathbf{R},\mathbf{t}) = \sum_{i=1}^{N} \rho^{\text{em}}(\mathbf{w}_i^{\text{calc}}(\mathbf{R},\mathbf{t})) \quad (2)$$

We have shown in (Birmanns and Wriggers, 2003) that this reduced coefficient can be evaluated on the microsecond time scale on a standard PC. In addition, we have shown that the precision of the fast correlation measure in an algorithmic docking application is nearly identical to the full correlation coefficient if a sufficient number of feature points (about 1% of the number of atoms) is used. We have therefore combined Eq. (2) with Powell's optimization method (Press et al., 1992). The resulting local refinement procedure typically converges after less than 50 Powell iterations, leading to a run-time cost of less than one second on a typical PC.

5. Results

The performance of the described multi-resolution docking method was tested on our standard set of simulated oligomeric density maps (Chacón and Wriggers, 2002), employing a low-pass filter to generate simulated EM maps from known atomic structures at different levels of resolution. In a series of tests a single monomer of each system was docked into a low-resolution map of the corresponding oligomer. The docking was validated with four different homo-oligomeric test systems. A dimer (1AFW Mathieu et al., 1997), a tetramer (7CAT Fita and Rossmann, 1985), a trimer (1NIC Adman et al., 1995) and a hexamer (2REC Yu and Egelman, 1997) each exhibit different shape and size properties.

Fig. 7 presents the docking precision, i.e. the RMSD between the docked subunit and the known oligomeric structure, as a function of the resolution. The feature point cardinality was set in this example to $N = 6$ and $M = SN$, with S being the number of symmetry-related subunits in the particular system. The size N of the point set representing the monomer is not critical (see below). For hetero-oligomeric systems the numbers N and M should be adjusted based on the relative volumes of probe molecule and target map, i.e. the multiplier S may be non-integer in such cases.

The docking precision results are very similar to the published validation (Chacón and Wriggers, 2002) of the *colores* docking tool that maximizes Eq. (1) in an exhaustive search. In a number of cases (for example 1NIC and 1AFW) the new algorithm is able to push the resolution limit of acceptable docking accuracy below that of the conventional maximization of C. Although at lower resolutions typically a higher RMSD is observed compared to conventional methods (Chacón and Wriggers, 2002), a cat-

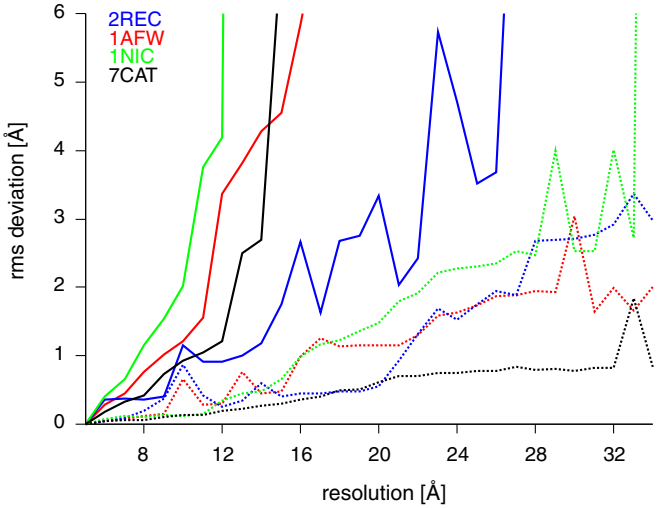

Fig. 7. Docking precision tested on four systems: thiolase, PDB entry 1AFW; catalase, PDB entry 7CAT; oxidoreductase, PDB entry 1NIC and helicase, PDB entry 2REC. The accuracy of the fitting was measured as the RMSD of the docked monomer to the known l oligomeric structure that was used to generate the simulated EM maps (see text). The docking accuracy of the novel fitting algorithm was recorded for both VQ (continuous lines) and Laplace quantization (dotted lines). The synthetic maps were generated with the following voxel sizes: 4 Å for resolution values $r > 12$ Å, 3 Åfor resolutions $8 < r \leqslant 12$ Å and 2 Å for $r \leqslant 8$ Å. The anchor matching algorithm was parameterized with $\delta = 15$ Å, $\gamma = 12$ Å, no wildcard matches were allowed in this example.

astrophic mismatch is often seen at a later stage (e.g. the docking position for RecA helicase can be identified for resolutions as low as 22 Å even without the Laplacian filter). The results are in agreement with earlier findings that suggest that the use of fiducials combined with pattern

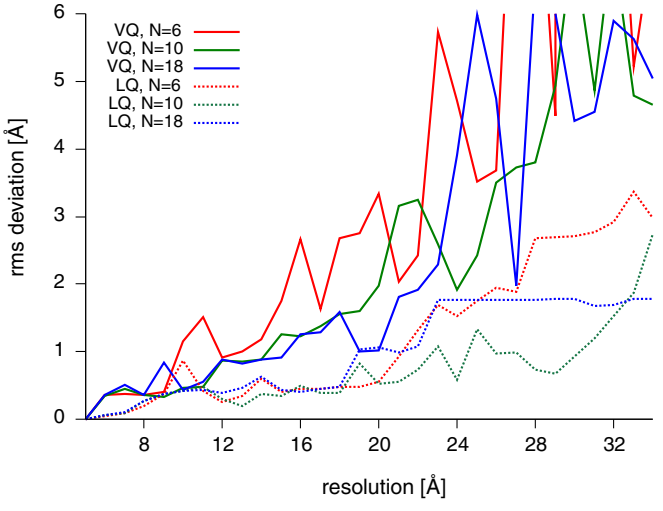

Fig. 8. Docking precision as a function of codebook-size: RecA helicase monomer and hexamer were represented by $N = 6$, 10, 18 and $M = 36$, 60, 108 feature points, respectively. The docking precision was measured as in Fig. 7. For $N = 6$ no wildcards were used, whereas in cases $N = 10$ and $N = 18$, 3 and 4 wildcards, respectively, were allowed ($p_{wc} = 1.0$ Å). The tolerances $\delta = 15$ Å, 12 Å, 10 Å and $\gamma = 12$ Å, 7 Å, 5 Å for $N = 6$, 10, 18 were used.

recognition can give meaningful results even in cases where the conventional cross-correlation breaks down due to lack of interior detail in the matched objects (Wriggers et al., 1999).

The complexity of the fitting algorithm is dependent the number of feature points N and M. We have therefore investigated whether the docking accuracy is sensitive to the number of feature points employed. In Fig. 8 the docking precision is shown for three codebook sizes: $N = 6$, 10, 18 and $M = 36$, 60, 108, respectively. For all of these we employed both standard VQ and the Laplace quantization. The test shows that the best overall results were achieved with ($N = 10$, $M = 60$) with outlier suppression, but additional points did not significantly improve the performance.

The current algorithm was designed for identifying components of large assemblies. It is possible to improve both efficiency and accuracy of the docking if the components can be isolated directly in the EM density. Two possible approaches are segmentation (Yu and Bajaj, 2005) and successive subtraction (discrepancy mapping) of known structures (Volkmann et al., 2000). We conducted a test separating a single monomer of the thiolase dimer (1AFW), then docking it into a synthetic map of variable resolution of the subunit ($N = M = 6$). The algorithm finds an accurate docking position up to 21Å, whereas docking the monomer into a complete map of the dimer already produces a mismatch at 12 Å (as shown in the previous test, see Fig. 7.) However, in practice it will be difficult to isolate components of larger assemblies without error in the artificial segmentation boundaries. Therefore, the general case ($N \leqslant M$) was considered here to allow for situations where not all EM density is accounted for.

6. Conclusions

The proposed point-set matching technique offers an efficient solution to the multi-resolution docking problem. Because of the speed and accuracy of the algorithm, subcomponents can be placed reliably into large macromolecular assemblies. The efficiency of the algorithm also enables a more interactive workflow, making it feasible to embed the method into a visualization tool in our quest to provide more user-friendly software to the scientific community.

Compared to conventional correlation-based docking programs our approach is not only more time-efficient, but also has other fundamental advantages. Since feature points describe the shape of the biological objects at a higher, more abstract level, they offer a very stringent criterion for matching, avoiding shifts and mismatches often observed with density-based criteria. The method also tends to yield fewer false positives, leading to a more compact and meaningful ranking of results. Although individual point deviations may reach values of 20–30% of the nominal resolution, the overall docking precision that can be achieved by the point cloud matching is much higher, 10% of the nominal resolution, which is on a par with other algorithmic approaches (Wriggers and Chacón, 2001).

On the other hand the method also has some intrinsic limitations. Firstly, the detection of smaller entities like secondary structure elements in "swiss cheese" like volumetric maps would require to cluster the objects at a much finer level of detail. This would lead to a significant increase in the number of feature points and in the complexity of the matching algorithm. This could be addressed in the future by an adaptive clustering technique that enables a hierarchical matching. However, in its present form the intended application is limited to intermediate resolution docking scenarios in the absence of secondary structure detail. Secondly, the new implementation introduces a number of parameters that need to be adjusted by the user.

This work was designed as a feasibility study so we have not fully explored the parameter space. However, we are able to provide the following brief "road map" to users of the method.

- The results are not sensitive to the cardinality (N, M). An upper bound for the number M of features describing the volumetric data can be found by dividing the volume of the target map by the volume of a resolution element. The number N for the probe molecule should be proportionally reduced relative to M based on the relative volume differences between target and probe.
- The speed of the algorithm depends on the tolerances γ and δ which should both be of the order of the nearest-neighbor separation of feature points. If the two point-clouds are similar, the variables are not critical for the docking precision. The expected run time of the reduced search increases significantly with larger tolerances.
- If there are discrepancies between matched data sets and/or their point clouds, one should use outlier suppression by choosing a small number of wild cards (no more than 0.1 N) and a distance penalty p_{wc} smaller or equal to the nearest-neighbor separation of feature points. One may also set γ to a lower value to further enforce that no outliers corrupt the matching.
- Any suboptimal results will still be refined in the subsequent fast Powell optimization that finds the nearest maximum of the (reduced) cross-correlation coefficient. This subsequent refinement further limits the effect of parameter choices on the final docking precision.

The described fitting method was implemented in our novel modeling program Sculptor (executables available at http://sculptor.biomachina.org; Birmanns and Wriggers, to be published). Because of the efficient nature of the matching algorithm, the program allows a higher level of interactivity compared to earlier "black-box"-style fitting tools. The program supports the clustering and filtering techniques described in this paper. After docking, the user can immediately explore the found solutions within an intuitive graphical user interface. The feature-point matchings are ranked and listed in a dialog box, where one can

select and organize solutions for further rendering. The final solutions are visualized in standard molecular graphics modes, or exported to a high-quality raytracing program.

Acknowledgments

We thank Oliver Passon, Brian Chen, and Lydia Kavraki for discussions, and Maik Boltes for the help with the implementation. This work was supported by NIH Grant R01-GM62968, by Human Frontier Science Program Grant RGP0026/2003, by Alfred P. Sloan Foundation Grant BR-4297 (to W.W.), as well as by a training fellowship from the W.M. Keck Foundation to the Gulf Coast Consortia through the Keck Center for Computational and Structural Biology (to S.B.).

References

Adman, E.T., Godden, J.W., Turley, S., 1995. The structure of copper-nitrite reductase from achromobacter cycloclastes at five pH values, with NO_2-bound and with type II copper depleted. J. Biol. Chem. 270, 27458–27474.

Alt, H., Guibas, L.J., 1996. Discrete Geometric Shapes: Matching, Interpolation, and Approximation—A Survey. Tech. rept. B 96-11. Freie Universität Berlin. Department of Mathematics and Computer Science, Germany.

Birmanns, S., Wriggers, W., 2003. Interactive fitting augmented by force-feedback and virtual reality. J. Struct. Biol., 123–131.

Chacón, P., Tama, F., Wriggers, W., 2003. Mega-dalton biomolecular motion captured from electron microscopy reconstructions. J. Mol. Biol. 326, 485–492.

Chacón, P., Wriggers, W., 2002. Multi-resolution contour-based fitting of macromolecular structures. J. Mol. Biol. 317, 375–384.

Fita, I., Rossmann, M.G., 1985. The NADPH binding site on beef liver catalase. Proc. Natl. Acad. Sci. USA 82, 1604–1608.

Gersho, A., Gray, R.M., 1992. Vector Quantization and Signal CompressionCommunications and Information Theory. Kluwer Academic Publishers, Norwell, MA, USA.

Jones, T.A., Zou, J.-Y., Cowan, S.W., 1991. Improved methods for building protein models in electron density maps and the location of errors in these models. Acta Cryst. A 47, 110–119.

Kabsch, W., 1976. A solution for the best rotation to relate two sets of vectors. Acta Cryst. A 32, 922.

Kabsch, W., 1978. A discussion of the solution for the best rotation to relate two sets of vectors. Acta Cryst. A 34, 827–828.

Kearsley, S.K., 1989. On the orthogonal transformation used for structural comparisons. Acta Cryst. A45, 208–210.

Langetepe, E., Zachmann, G., 2006. Geometric Data Structures for Computer Graphics. ISBN 1-56881-235-3.

Lorenz, M., Popp, D., Holmes, K.C., 1993. Refinement of the F-actin model against X-ray fiber diffraction data by the use of a directed mutation algorithm. J. Mol. Biol. 234, 826–836.

Martinetz, T.M., Berkovich, S.G., Schulten, K., 1993. Neural gas for vector quantization and its application to time-series prediction. IEEE Trans. Neural Networks 4, 558–569.

Mathieu, M., Modis, Y., Zeelen, J., Engel, C.K., Abagyan, R.A., Ahlberg, A.e., 1997. The 1.8 Å crystal structure of the dimeric peroxisomal 3-ketoacyl-CoA thiolase of *Saccharomyces cerevisiae*: implications for substrate binding and reaction mechanism. J. Mol. Biol. 273, 714–728.

Ming, D., Kong, Y., Lambert, M., Huang, Z., Ma, J., 2002. How to describe protein motion without amino acid sequence and atomic coordinates. Proc. Natl. Acad. Sci. USA 99, 8620–8625.

Nussinov, R., Wolfson, H.J., 1991. Efficient detection of three-dimensional structural motifs in biological macromolecules by computer vision techniques. Proc. Natl. Acad. Sci. USA, 10495–10499.

Passon, O., Boltes, M., Birmanns, S., Zilken, H., Wriggers, W., 2005. Laplace-filter enhanced haptic rendering of biomolecules. In: Greiner, G., Hornegger, J., Niemann, H., Stamminger, M. (Eds.), Proceedings Vision Modeling and Visualization, pp. 311–318, 516, IOS Press, Netherlands, ISBN 1-58603-569-X.

Press, W.H., Teukolsky, S.A., Vetterling, W.T., Flannery, B.P., 1992. Numerical Recipes in C: The Art of Scientific Computing. Cambridge University Press, New York, NY, USA.

Roseman, A.M., 2000. Docking structures of domains into maps from cryo-electron microscopy using local correlation. Acta Cryst. D 56, 1332–1340.

Rossman, M.G., 2000. Fitting atomic models into electron-microscopy maps. Acta Cryst. D 56, 1341–1349.

Rossmann, M.G., Morais, M.C., Leiman, P.G., Zhang, W., 2005. Combining X-ray crystallography and electron microscopy. Structure 13, 355–362.

Tama, F., Wriggers, W., Brooks, C.L., 2002. Exploring global distortions of biological macromolecules and assemblies from low-resolution structural information and elastic network theory. J. Mol. Biol. 321, 297–305.

Volkmann, N., Hanein, D., Ouyang, G., Trybus, K.M., DeRosier, D.J., Lowey, S., 2000. Evidence for cleft closure in actomyosin upon ADP release. Nat. Struct. Biol. 7, 1147–1155.

Volkmann, N., Hanein, D., 1999. Quantitative fitting of atomic models into observed densities by electron microscopy. J. Struct. Biol. 125, 176–184.

Wriggers, W., Birmanns, S., 2001. Using Situs for flexible and rigid-body fitting of multiresolution single-molecule data. J. Struct. Biol. 133, 193–202.

Wriggers, W., Chacón, P., 2001. Modeling tricks and fitting techniques for multiresolution structures. Structure 9, 779–788.

Wriggers, W., Schulten, K., 1999. Investigating a back door mechanism of actin phosphate release by steered molecular dynamics. Proteins: Struct. Funct. Genet. 35, 262–273.

Wriggers, W., Milligan, R.A., Schulten, K., McCammon, J.A., 1998. Self-organizing neural networks bridge the biomolecular resolution gap. J. Mol. Biol. 284, 1247–1254.

Wriggers, W., Milligan, R.A., McCammon, J.A., 1999. Situs: A package for docking crystal structures into low-resolution maps from electron microscopy. J. Struct. Biol. 125, 185–195.

Yu, X., Egelman, E.H., 1997. The RecA hexamer is a structural homologue of ring helicases. Nat. Struct. Biol. 4, 101–104.

Yu, Z., Bajaj, C., 2005. Automatic ultrastructure segmentation of reconstructed cryoEM maps of icosahedral viruses. IEEE Trans. Image Process. 14 (9), 1324–1337.