Structure, Vol. 9, 779–788, September, 2001, ©2001 Elsevier Science Ltd. All rights reserved. PII S0969-2126(01)00648-7

Modeling Tricks and Fitting Techniques for Multiresolution Structures

Ways & Means

Willy Wriggers[1] **and Pablo Chacón**
Department of Molecular Biology
The Scripps Research Institute
10550 N. Torrey Pines Road
La Jolla, California 92037

Introduction

The spectacular developments of recent years toward solving atomic structures of ever-increasing complexity underscore the importance of relating 3D structures of multicomponent cellular machines to molecular mechanisms. To understand the workings of these machines [1], we face many challenges, one of which is to describe the structure and functional mechanisms at the atomic level. The atomic structure of a molecular machine in a particular state of processing can be inferred by building it from its components, i.e., by combining multiresolution data from a variety of biophysical sources. This hybrid modeling approach holds much promise, provided that the docking procedure is reproducible and incorporates the constraints of molecular interactions and architecture. In the following, we present an overview of state-of-the-art computational fitting techniques and, wherever possible, we put them to a stringent test to discuss their advantages and limitations. The assessment of the scope and validity of individual methods will hopefully serve as a “consumer guide” that allows the reader to identify the most suitable docking criterion given a specific fitting problem.

As a computational research area, multiresolution modeling is still in its infancy, but it will become increasingly attractive in the near future, when more and more low-resolution structures of large complexes and atomic structures of their components become available. This development is prompted in part by the advent of structural genomics, which promises to bring the sequences of most single-domain proteins within homology-modeling distance of a known structure [2, 3] through a substantial increase in the number of solved atomic structures. In addition, cryogenic electron microscopy (cryo-EM) has evolved to a standard technique for the study of large-scale assemblies, as it permits visualization of the structures at an intermediate level of resolution [4–7]. Unlike crystallization, cryo-EM poses few restrictions on the conformational range of multicomponent complexes and is capable of yielding low- and intermediate-resolution density maps under a wide range of biochemical conditions. Progress in automated sample preparation and image processing will produce intermediate-resolution EM structures at high-throughput pace [8]. By combining data from high-throughput cryo-EM and structural genomics, multiresolution modeling will produce approximate but reasonably accurate atomic models of macromolecular assemblies. These models of large assemblies will be created routinely, often years before a comparable crystallographic structure can be solved by the more laborious, traditional X-ray or electron crystallography methods.

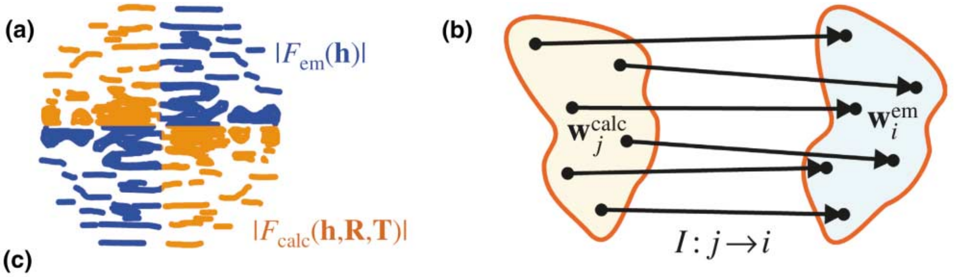
In the past decade, significant progress was made by combining cryo-EM data with high-resolution structures determined by NMR spectroscopy or X-ray crystallography. The first EM maps into which atomic structures were fitted included actomyosin filaments [9, 10] and icosahedral viruses [11–13]. The most common of these hybrid strategies involves the visual docking of atomic structures into envelopes derived from low-resolution data [14, 15]. The successful construction of such hybrid models for virus capsids and cytoskeletal motor-filament complexes constitutes a clear indication of the value of the visual approach [5, 7, 16]. More recently, several groups have recognized the need for computational tools to perform the fitting in a reliable and reproducible manner (Figure 1).

A low-resolution image reconstruction of a macromolecular assembly from electron micrographs can be viewed as a convolution of an atomic structure of the assembly with a smoothing kernel (point-spread function). In general, the point-spread function depends on the electron optics of the microscope [4]. Computationally, it is straightforward to lower the resolution of an atomic structure, e.g., by convolution with a Gaussian [17] that approximates the point-spread function. The reverse problem, then, is the deconvolution of the low-resolution density map utilizing the atomic structure of components. This multiresolution deconvolution, i.e., docking, poses a challenging computational problem that is the subject of this review.

A variety of computational docking algorithms have recently become available (Figure 1). Some algorithms have been adopted by individual laboratories for their own use, while others are openly disseminated within the EM community. It is not possible in this review to do justice to all existing algorithms, since many laboratories devise a mix of individual docking techniques for particular practical applications [18]. Also, we omit specialized methods that impose a particular symmetry on the refined data, in particular, icosahedral symmetry in the case of virus capsids [19]. Instead, we focus here on describing the central principles underlying the most commonly used fitting methods, and we refer to the original articles for the detailed implementation.

Following a discussion of differences and similarities between Fourier space and direct space fitting criteria, we sketch the advantages and limitations of data reduction techniques that reduce the complexity of direct space data for interactive and flexible fitting applications. We describe the use of crosscorrelation and convolution methods and outline how their viable resolution range can be extended by modifying the underlying correlation criterion. We conclude by outlining some outstanding problems that put forth tasks for future research such as the systematic evaluation of fitting methods and their use as database query tools.

[1] Correspondence: wriggers@scripps.edu


(a)
$|F_{\mathrm{em}}(\mathbf{h})|$
$|F_{\mathrm{calc}}(\mathbf{h},\mathbf{R},\mathbf{T})|$
(b)
$\mathbf{w}_j^{\mathrm{calc}}$
$\mathbf{w}_i^{\mathrm{em}}$
$I : j \rightarrow i$

(c)

Technique	Optimization Criterion	Pros	Cons	CPU Cost
R Value Refinement or Fourier-Space Least-Squares Fit [21,22,25,26,32]	$R_1 = \sum_{\mathbf{h}} \left\| \lvert F_{\mathrm{em}}(\mathbf{h}) \rvert - \lambda \lvert F_{\mathrm{calc}}(\mathbf{h},\mathbf{R},\mathbf{T}) \rvert \right\|^n, n = 1,2$ $R_2 = \sum_{\mathbf{h}} \lvert F_{\mathrm{em}}(\mathbf{h}) - \lambda F_{\mathrm{calc}}(\mathbf{h},\mathbf{R},\mathbf{T}) \rvert^n, n = 1,2$	Natural criterion for diffraction data.	R_1 ignores phase information.	1-6 hours
Vector Quantization [42,43,44]	$V = \sqrt{\frac{1}{k} \sum_{j=1}^{k} \lvert \mathbf{w}_{I(j)}^{\mathrm{em}} - \mathbf{w}_j^{\mathrm{calc}}(\mathbf{R},\mathbf{T}) \rvert^2}$	Instant results. Flexible fitting (in addition to rigid-body fitting) is optional.	All density must be accounted for.	1 minute
Linear Correlation / Template Convolution [46,49,52,53]	$C_1 = \int \rho_{\mathrm{em}}(\mathbf{r})\ \rho_{\mathrm{calc}}(\mathbf{r},\mathbf{R},\mathbf{T})\ \mathrm{d}^3\mathrm{r}$	Seasoned refinement tool. Reliable for resolutions <10Å. FFT accelerated.	Most of the density must be accounted for.	6 hours
Density Masking (Local Normalization) [59]	$C_2 = \frac{\int_{\mathrm{mask}} \rho_{\mathrm{em}}(\mathbf{r})\ \rho_{\mathrm{calc}}(\mathbf{r},\mathbf{R},\mathbf{T})\ \mathrm{d}^3\mathrm{r}}{\sqrt{\int_{\mathrm{mask}} \rho_{\mathrm{em}}^2(\mathbf{r})\ \mathrm{d}^3\mathrm{r}} \sqrt{\int_{\mathrm{mask}} \rho_{\mathrm{calc}}^2(\mathbf{r},\mathbf{R},\mathbf{T})\ \mathrm{d}^3\mathrm{r}}}$	Local normalization extends the reliability of correlation-based fitting (<15Å).	Not possible to accelerate by FFT.	48 hours
Density Filtering (Laplacian) [58]	$C_3 = \int \left(\nabla^2 \otimes \rho_{\mathrm{em}}\right)(\mathbf{r})\ \left(\nabla^2 \otimes \rho_{\mathrm{calc}}\right)(\mathbf{r},\mathbf{R},\mathbf{T})\ \mathrm{d}^3\mathrm{r}$	Most robust correlation method (<25Å). FFT accelerated.	Map filtering may amplify noise.	6 hours

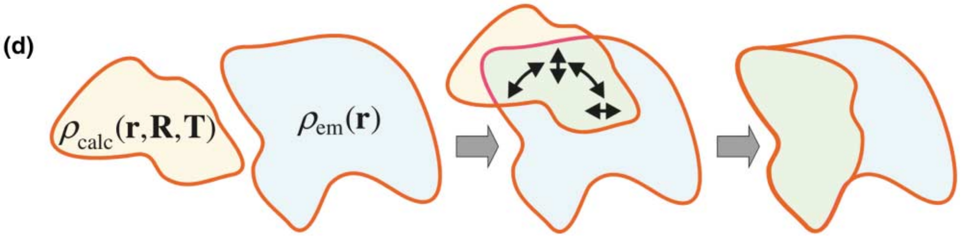

(d)
$\rho_{\mathrm{calc}}(\mathbf{r},\mathbf{R},\mathbf{T})$
$\rho_{\mathrm{em}}(\mathbf{r})$

Figure 1. Various Docking Approaches for Multiresolution Structures

Individual sketches (a, b, and d) surrounding the table (c) illustrate the various techniques. Fitting techniques include (a) Fourier space refinement, (b) direct space vector quantization, and (d) correlation-based fitting. The corresponding optimization criteria (table) are explained in the text. $\boldsymbol{F}_{\mathrm{em}}$ and $\boldsymbol{F}_{\mathrm{calc}}$ are the Fourier coefficients (structure factors) of the EM map and the probe molecule, respectively; λ is a scale factor; **h** and **r** are the coordinates in Fourier and direct space, respectively; **R** and **T** are the rotational and translational parameters of the model; $\mathbf{w}_i^{\mathrm{em}}$ and $\mathbf{w}_j^{\mathrm{calc}}$ (i,j = 1,…,k) are k codebook vectors that encode the EM map and the probe molecule, respectively, at reduced complexity; I maps probe vector indices to corresponding EM vector indices; ρ_{em} and ρ_{calc} are the direct space density distributions of EM and probe data, respectively (normalized to standard deviation 1 and centered at 0); and ($\nabla^2 \otimes \ldots$) is a convolution with a Laplacian operator:

$$\nabla^2 = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}.$$

CPU cost times are rough estimates for single processors (e.g., SGI Octane 300 MHz R12,000) using a (60 × 60 × 60)-voxel density map and a rotational sampling resolution of 9°.

Fourier Space versus Direct Space

It comes as no surprise that some of the earliest fitting tools employed by low-resolution modelers were based on well-established methods for X-ray crystallographic refinement [19–22]. The problem of rigid-body fitting can be formulated as the minimization of the discrepancy between observed and calculated structure factors in Fourier space (Figure 1c) with respect to the rotational and translational parameters of the model and the scale factor λ. If the linear discrepancy ($n = 1$ in Figure 1c) and amplitude differences (R_1, $n = 1$ in Figure 1c) are considered, this corresponds to the well-known crystallographic R factor [23]. Similarly, linear vector discrepancies (R_2, $n = 1$) correspond to the vector R factor [24]. Alternatively, it is also possible to minimize a quadratic misfit ($n = 2$) of amplitudes or vectors [25, 26].

Certain Fourier and direct space methods have equivalent formulations in either domain. For example, the minimization of the quadratic misfit (R_2, $n = 2$) is equivalent in direct space (through Rayleigh's theorem [27]) to a least-squares minimization: $R_2 = \int[\rho_{\mathrm{em}}(\mathbf{r}) - \lambda\ \rho_{\mathrm{calc}}(\mathbf{r},\mathbf{R},\mathbf{T})]^2\ \mathrm{d}^3\mathbf{r}$. After binomial expansion of the integration kernel, assuming normalized densities (Figure 1), we have $R_2 = 1 + \lambda^2 - 2\lambda C_1$, i.e., minimizing the quadratic misfit corresponds to maximizing the correlation coefficient C_1 that will be discussed in detail below.

Fourier-based methods (R_1) are the only recourse if diffraction amplitudes are the sole source of information, as in X-ray fiber diffraction [20]. In EM, the phases of the structure factors are known and, in principle, one can construct a 3D model of the density in direct space. Nevertheless, Fourier-based methods are valuable in EM if one wants to avoid the numerical complications of the transformation to direct space. For example, in helical-image analysis of EM micrographs, a standard reconstruction method relies on layer-line (helical-diffraction) data in reciprocal space [28, 29]. Figure 1a illustrates the appearance of a typical fiber diffraction or EM layer-line pattern. In such situations, one wishes to minimize the discrepancy between the observed (blue quadrants) and the calculated (orange quadrants) layer lines. In EM, the 3D reconstruction involves a Fourier-Bessel transform of the layer-line amplitudes and phases [30], which requires a manual indexing of the entire diffraction pattern [29, 31]. If the fitting is performed directly in Fourier space, problems with overlapping layer lines and the laborious accounting of the entire diffraction pattern are avoided.

The R value criterion, R_1, is not very suitable for the docking of EM densities, since the additional phase information available from EM image reconstructions is ignored. To maximize the amount of information actually used for the docking, the vector discrepancies R_2 should be minimized. Even if phase information is included [32], deviations between computed structure factors in Fourier space do not correspond to localized positional or orientational changes in direct space. In particular, it is difficult to refine the internal degrees of freedom of atomic structures against this data [32]. Overfitting is less of a problem in rigid-body docking, where one has only six degrees of freedom. However, problems with the "delocalized" Fourier space fitting are exemplified by a refinement of the actin monomer structure against low-resolution X-ray fiber diffraction data [21] that resulted in a reduced stereochemical quality of the fitted model compared to the crystal structure [33] (as judged with PROCHECK [34]). The reduced quality was perhaps an effect of overfitting due to the many degrees of freedom in the flexible model—three for each atom. We argue that in the case of flexible fitting a refinement in direct space gives the modeler better control over localized changes in the structure compared to Fourier space refinement.

Direct Space Data Reduction

Direct space refinement is a "WYSIWYG" (what you see is what you get) modeling approach. It is straightforward, in direct space, to combine EM-based refinement with geometric constraints from biochemical experiments and with molecular force fields that govern the physical interactions of the atoms [35]. One can count the number of independent pieces of information available for the fitting of a model in direct space by dividing the volume of the structure by the volume of a resolution element, i.e., a cube whose length corresponds to the spatial resolution. For medium-resolution (~10–30 Å) EM maps of single molecules, this number is surprisingly small, ranging from the lower single digits in the cases of actin [36], tubulin [37], and kinesin [38] to a few dozen in the case of the ribosomal elongation factor EF-G [39]. Clearly, it would be beneficial for direct space fitting and modeling if we could represent this small number of shape-defining fiducials in a reliable and reproducible fashion.

Clustering techniques have been used since the 1950s for digital signal compression in engineering applications such as digital speech and image processing [40]. Such "lossy" data compression methods seek to represent a complex signal by a reduced number of vectors that identify the signal cluster centers. Certain clustering methods are rooted also in neural computing, where the goal is to find "faithful" neighborhood-preserving maps from an input space of sensory signals to a discrete network of neurons in the cortex [41]. From both fields, algorithms emerged that essentially perform density estimation, yielding discrete estimates of data manifolds. Electron microscopists may be familiar with such methods in the context of classification of images during the data acquisition and management phase to reduce signal loss and to achieve improvements in resolution [4]. Recently, it was proposed to utilize a clustering technique, termed vector quantization, for a reduced representation of 3D data that allows EM data to be matched with atomic structures [42–44].

In vector quantization, a single-molecule data set is represented by k so-called codebook vectors (Figure 1b): $\mathbf{w}_j^{calc}$ (corresponding to high-resolution data) or w_i^{em} (corresponding to low-resolution data; $i,j = 1,\ldots,k$). An index map $I : j \rightarrow i$ defines the k pairs of corresponding vectors. There are two applications of this technology; in rigid-body fitting, I is not known a priori, and an exhaustive search of the k! possible permutations $(I(1),\ldots,I(k))$ is carried out. The fits are then ranked by the residual rms deviation V (Figure 1c) after least-squares fitting of the vectors $\mathbf{w}_j^{calc}$ to the $\mathbf{w}_{I(j)}^{em}$. In flexible fitting, rigid-body docking alone gives poor alignment of the crystal structure and the low-resolution data set. For example, the structure of the ribosomal protein EF-G exhibits a striking "induced fit" conformational change on the 70S ribosome involving three protruding domains [14, 39]. In such situations I is known, and the deviating atomic structure can be brought into register with the EM density, effectively by forcing V (Figure 1c) to vanish. This is done in a molecular dynamics refinement of the atomic structure where a quantity equivalent to V forms a penalty that is imposed by distance constraints (see [44] for details).

The major advantage of vector quantization, apart from its obvious value for flexible fitting, is computational speed. Both quantization and docking by the reduced representation can be carried out within seconds of compute time. In contrast, the exhaustive search of all rigid-body degrees of freedom for the full data sets can take many hours (Figure 1c). Despite the fact that multiresolution structural data is docked indirectly (by means of the vector quantization), the accuracy that can be achieved in flexible and rigid-body docking with simulated (noise-free) data is one order of magnitude above the nominal resolution of the EM map, or better [44].

The identification of spatial features by vector quanti-

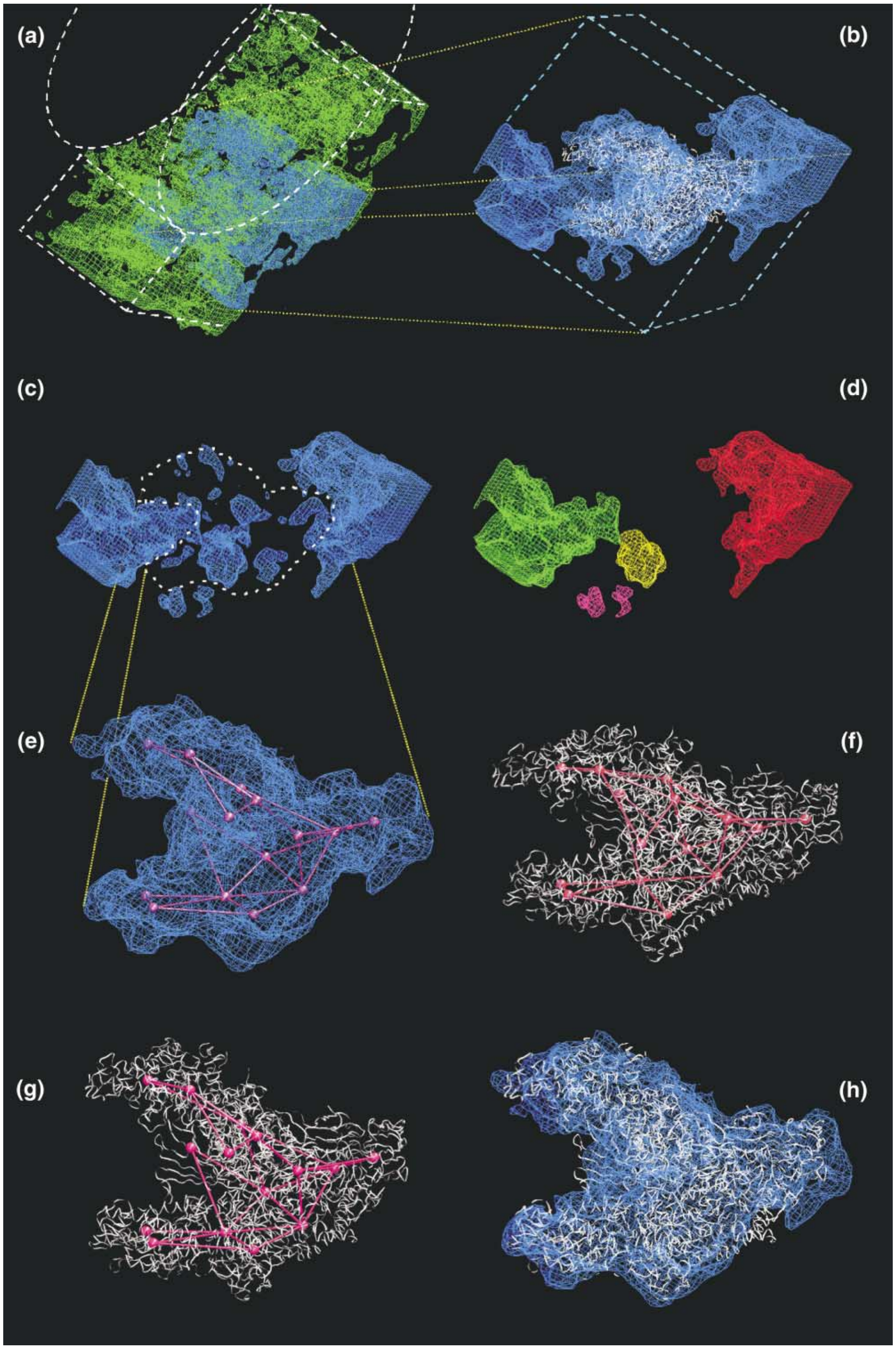


Figure 2. Direct Space Flexible Fitting with Skeletons

This overview illustrates the map editing and skeleton-based modeling steps that were employed in the recent flexible refinement of the *T. aquaticus* (Taq) crystal structure of RNA polymerase (RNAP) against *E. coli* EM data from the laboratory of Seth Darst.

(a) Original 3D reconstruction. Bacterial RNAP and lipid molecules are arranged in tubular crystals. The extent of the lipid-RNAP tube is indicated by the dashed lines.

(b) Single RNAP strand extracted. The map was cropped (blue box), and the central contiguous density corresponding to RNAP was extracted from the lipid background by using segmentation at the surface density level of RNAP with the Situs [43] *floodfill* tool. Subsequently, a single Taq RNAP structure (white) was docked to the density using the Laplacian correlation coefficient C_3 (Figure 1c).

(c) Discrepancy mapping. The resolution of the fitted molecule was lowered to 15 Å with the Situs *pdblur* tool, and the resulting map was subtracted from the map in (b) after a rescaling of the density to match the isocontour level (dashed line).

zation is sensitive to noise that might originate from experimental limitations. One of the open questions in flexible docking is how to maintain the stereochemical quality of a fitted structure [45], since any overfitting to noise-induced vector displacements would compromise the quality of the atomic model. In a recent approach, intervector distances along the connected polypetide chain are constrained (Figure 2). The resulting vector skeletons (distance-constrained vectors) eliminate the longitudinal degrees of freedom that are deemed inessential for the flexible docking while permitting lateral flexibility. Thereby, the skeleton-based fitting approach provides additional robustness against the effects of noise and experimental uncertainty [44].

Figures 2a–2e illustrate one limitation of vector quantization and a possible remedy: all density should be accounted for by the atomic structure. If there are extraneous densities, e.g., due to sequence insertions or neighboring structures as in the case of RNA polymerase, they should be identified and subtracted by discrepancy mapping [46] prior to any docking. Discrepancy mapping can involve an iterative strategy in which the structure is first docked in a course manner, and then the refinement is done at a later stage once the extraneous densities have converged [S.A. Darst, N. Opalka, P.C., A. Polyakov, C. Richter, G. Zhang, and W.W., submitted].

Crosscorrelation and Convolution

The quantitative docking methods discussed so far involve symmetrical systems or systems where the subunit to be docked can be isolated. The methods presented in this section are capable—to varying degrees—of docking components into larger densities present in biomolecular assemblies. Naturally, these exhaustive search methods are computationally demanding and, at present, are limited to rigid-body docking. For the first time, we have evaluated the docking performance of three state-of-the-art criteria on simulated low-resolution data generated from known atomic structures.

Perhaps the oldest fitting criterion is the (globally normalized) crosscorrelation coefficient C_1 (Figure 1c). The method has been adopted by a large number of authors [13, 46–52], and a number of computer programs are readily available [46, 49, 52, 53]. The idea is to maximize C_1 with respect to the translational and rotational degrees of freedom and thereby minimize the mean-square density discrepancy of two structural data sets (Figure 1d). In a crystallographic context, the method has been termed template convolution [49] because it enhances features of an electron density map in direct space by matching it with a template structure. To this end, one considers C_1 to be a function of the translations (**T**) only, by projecting out—for a given **T**—the maximum C_1 for all rotations (**R**). Although convolution and correlation are related and frequently used interchangeably, we note that the mathematical definitions differ [27]. In this paper we refer to convolution only in the context of the "smoothing" of a density with a point-spread function (Figure 3).

One important aspect of globally normalized correlation-based fitting is that the performance of the algorithms can be significantly enhanced by the use of fast Fourier transform (FFT) techniques. The Fourier-based shape complementarity algorithm of Katchalski-Katzir et al. [54] is widely used in external ligand docking programs [55–57]. This molecular surface recognition method takes advantage of Fourier correlation theory and FFT to scan rapidly the translation of a probe molecule relative to a (fixed) reference molecule. It is straightforward to adapt this technique for the calculation of globally normalized correlations (Figure 3). The underlying idea is that the correlation coefficient can be expressed in Fourier space as the product of the structure factors. The calculation of the correlation in direct space (Figure 3) is very expensive, since it requires M^2 multiplications for each translation **T**, where M is the total number of voxels. However, the reverse FFT of the structure factors is significantly faster than the direct space calculation, because we need to perform only two transformations (each of them scale as $M \log M$). Note that the FFT corresponding to the fixed EM map can be omitted from all but the first calculation. In addition, the use of parallelization may further reduce the cost. This is particularly advantageous for the orientations since sets of Euler angles can be farmed out to various processors. The fast, FFT-based computation of C_1 and C_3 (Figure 1c) has recently been adopted also for density-based docking [53, 58].

The standard correlation coefficient C_1 is nonspecific in terms of deviations of corresponding features; i.e., unlike the rms deviation that is commonly used for comparing atomic structures, the correlation measure is not very sensitive to positional or orientational changes [43] and exhibits relatively broad distributions of possible solutions sets [46]. A second limitation has been pointed out [59] concerning the onset of false solutions if the probe object represents only a portion of the EM map. In such situations, the density in the EM map that does not correspond to the component being docked acts as noise, and the global optimization of the correlation coefficient may actually worsen the fit, as judged by the

(d) Segmentation of foreign densities. Densities corresponding to the neighboring RNAP subunits (red and green) and the *E. coli* dispensable regions DR1 (pink) and DR2 (yellow) were tagged and segmented with *floodfill*.

(e) Single-molecule skeleton. After subtracting the foreign densities in (d) from the whole map in (b), one obtains a single-molecule "Taq-like" map that retains the information about the *E. coli* conformational change. A skeleton was fitted to this map using vector quantization and distance constraints.

(f) Parametrization of the skeleton. The connectivities and codebook vector distances in (e) were based on a vector quantization of the Taq atomic structure. Codebook vector connectivities follow the polypetide chain.

(g) Flexible fitting. The flexible refinement was carried out with X-PLOR [24] as will be described in detail (S.A. Darst, N. Opalka, P.C., A. Polyakov, C. Richter, G. Zhang, and W.W., submitted).

(h) Comparison of flexibly fitted model with the single-molecule map from (e). The images (a)–(h) were created with Situs [43] and VMD [68].

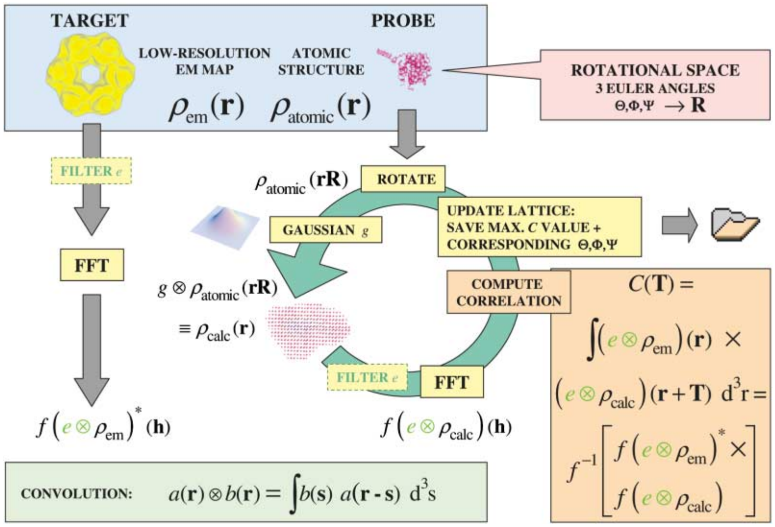

TARGET
PROBE
LOW-RESOLUTION EM MAP
ATOMIC STRUCTURE
$\rho_{\mathrm{em}}(\mathbf{r})$ $\rho_{\mathrm{atomic}}(\mathbf{r})$
ROTATIONAL SPACE
3 EULER ANGLES
$\Theta,\Phi,\Psi \rightarrow \mathbf{R}$
FILTER e
ROTATE
$\rho_{\mathrm{atomic}}(\mathbf{rR})$
GAUSSIAN g
UPDATE LATTICE:
SAVE MAX. C VALUE +
CORRESPONDING Θ,Φ,Ψ
FFT
$g \otimes \rho_{\mathrm{atomic}}(\mathbf{rR})$
$\equiv \rho_{\mathrm{calc}}(\mathbf{r})$
COMPUTE CORRELATION
$C(\mathbf{T}) =$
$\int (e \otimes \rho_{\mathrm{em}})(\mathbf{r}) \times (e \otimes \rho_{\mathrm{calc}})(\mathbf{r}+\mathbf{T})\, d^3r =$
$f^{-1}\left[f(e \otimes \rho_{\mathrm{em}})^* \times f(e \otimes \rho_{\mathrm{calc}}) \right]$
FILTER e
FFT
$f(e \otimes \rho_{\mathrm{em}})^*(\mathbf{h})$
$f(e \otimes \rho_{\mathrm{calc}})(\mathbf{h})$
CONVOLUTION: $a(\mathbf{r}) \otimes b(\mathbf{r}) = \int b(\mathbf{s})\, a(\mathbf{r}-\mathbf{s})\, d^3s$

Figure 3. Accelerated Correlation-Based Fitting with the Fast Fourier Transform (FFT)

The initial data sets are a low-resolution map (target) and an atomic structure (probe), corresponding to direct space densities ρ_{em} (**r**) and ρ_{atomic} (**r**), respectively (blue box). The probe molecule is subject to a rotation matrix **R** (red box) that can be constructed from the three Euler angles [69]. After lowering the resolution of the atomic structure (by direct space convolution with a Gaussian *g*) to that of the target map, the rotated probe molecule corresponds to the simulated density ρ_{calc} (**r**). An optional filter *e* (e.g., a Laplacian, c.f. Figure 1) can be applied to both ρ_{em} (**r**) and ρ_{calc} (**r**) before the structure factors are computed (*f* denotes the FFT and the asterisk denotes the complex conjugate). The definition of a direct space convolution of a density function *b*(**r**) with a kernel *a*(**r**) is given in the green box. The definition of the direct space correlation *C* as a function of a translational displacement **T** is given in the orange box. By virtue of the Fourier correlation theorem [27], *C* can be computed for all **T** from the inverse Fourier transform of the previously calculated structure factors. In practical situations, the coarseness of the translational sampling of *C*(**T**) is given by the lattice spacing of ρ_{em} (**r**) and ρ_{calc} (**r**). For each **T**, the *C* values (and the corresponding Euler angles) are saved to a file if they exceed values calculated for all previously sampled rotations. Subsequently, the next iteration proceeds with a new set of Euler angles.

visual agreement of discernible features [59]. This effect is particularly pronounced at resolutions below 15 Å, where there is little internal structure. At such low resolution, the probe molecule would simply drift to the highest density region in the EM map where C_1 is maximized. It is possible to improve the performance of the correlation measure by successive application of discrepancy mapping to simulate the molecular boundaries between individually docked segments [60]. Other software developers argue that the problems are intrinsic to the correlation measure itself and have sought to modify it conceptually [58, 59].

To reduce the effect of densities that are not accounted for (blue in Figure 1d), it was proposed to renormalize the correlation locally [59]. A mask corresponding to the local overlap (green in Figure 1d) is applied to the densities, and only the region under the mask contributes to the correlation measure C_2 (Figure 1c). The effect is a leveling of the distribution of coefficients across the explored space that reflect only local similarity and not the effect of any extraneous density. This eliminates the problem of drifting toward higher density regions in the interior of low-resolution maps. Due to the local renormalization, the correlation C_2 must be computed in direct space.

Alternatively, it was proposed to alter the functional form of the compared densities by applying a filter that enhances the contours in the data sets (red in Figure 1d). Correlation-based fitting is mainly used in crystallography [49], where the high resolution of the electron density does not require any special form of the density functions. In EM docking, however, the densities show little variation inside of the molecules, and after maximizing C_1, "surface" features may no longer be in register. The proposed solution [58; P.C. and W.W., submitted] is to include surface information in the fitting. A Laplacian operator, well-known in image processing [61], assigns positive densities to the contours (red in Figure 1d) and negative densities to the interior volumes (yellow and blue in Figure 1d). This approach effectively maximizes both surface and volumetric overlap when C_3 (Figure 1c) is maximized [58].

We have put the three correlation coefficients C_1, C_2, and C_3 to a stringent test using simulated (noise-free) low-resolution data derived by Gaussian real-space convolution of known oligomeric structures. The task was to place monomeric components accurately into the larger, simulated maps of the complexes. In general, this test is much more challenging than the docking of isolated molecules. Figure 4a shows the simulated 15 Å resolution map of the RecA hexamer, including a correctly and incorrectly docked monomer. The resulting crosssections of the maximal correlation coefficients in the center plane of the hexamer are presented in Figures 4b–4d. Clearly, the maximum of the standard coefficient C_1 is degenerate along a ring within the hexamer. The docking with C_1 is ambiguous, and the highest scoring fit is a false positive (shown in red in Figure 4a). In comparison, both coefficients C_2 and C_3 exhibit six narrow peaks (C_3 with highest contrast) that unambiguously correspond to the correct solution (shown in green in Figure 4a). We have also tested the performance of the coefficients as a function of resolution on four oligomeric systems (a dimer, trimer, tetramer, and hexamer) shown in Figure 4d. The task was again to reproduce the correct position of individual monomeric components in the simulated oligomeric maps. Initially, each of the coefficients performed very well (docking precision <1 Å) but eventually the fitting breaks down (docking precision >10 Å) when the resolution becomes too low. C_1 reaches the breaking point the earliest, at 8–15 Å resolution; C_2 follows at 12–24 Å resolution; and C_3 performs well up to 26–36 Å resolution.

Perspective: Scope and Validity

We have witnessed considerable advancements over the past years in the development of quantitative fitting

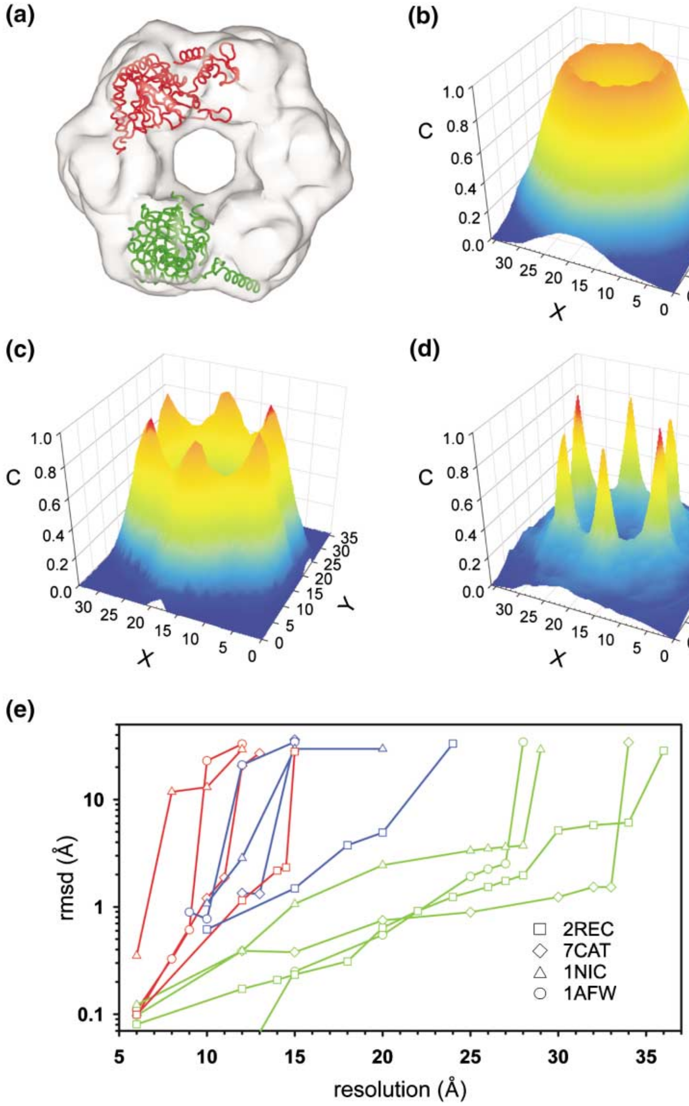


Figure 4. Performance of Various Correlation-Based Fitting Criteria

The criteria were validated on simulated low-resolution data derived from oligomeric structures.

(a) Simulated 15 Å map of the RecA hexamer, PDB code 2REC, including correctly (green) and incorrectly (red) docked subunits. (b–d) Maximum correlation coefficients (Figure 1c) as a function of (x, y) coordinates (central crosssections). For each (x, y) position, the coefficients were maximized with respect to all rotations. The correlation values are normalized to the unit interval and represented with a color spectrum from blue ($C = 0$) to red ($C = 1$).

(b) Linear correlation coefficient C_1.

(c) Locally normalized correlation coefficient C_2.

(d) Laplacian correlation coefficient C_3.

(e) Docking precision (rms deviation from target) as a function of resolution, using C_1 (red), C_2 (blue), and C_3 (green). The correlation criteria were tested on a protein dimer (peroxisomal thiolase, PDB code 1AFW), trimer (copper-nitrite reductase, 1NIC), tetramer (catalase, 7CAT), and hexamer (RecA, 2REC). The on-lattice rigid-body displacements were sampled in 5 Å and 9° steps, followed by off-lattice Powell (gradient ascent) maximization [70]. The average rms deviations of the six highest scoring results are 28 Å, 1.5 Å, and 0.24 Å for C_1, C_2, and C_3, respectively. The figure was created with Situs [43] and VMD [68] (a), as well as SPSS SigmaPlot (b–e).

algorithms. Each of the presented algorithms have found their niche in the modeler's tool chest, and methods may be mixed and combined to form more complex fitting strategies [18]. To conclude, we list five quality standards that we deem highly desirable for achieving satisfying results in a variety of practical situations. We argue that these standards should direct the current and ongoing inception of robust methods for combining multiresolution data. They are (1) computational speed, (2) exhaustive search, (3) discriminative scoring, (4) scope of applicability, and (5) robustness under experimental limitations.

Computational Speed

Vector quantization is the most efficient fitting algorithm due to the reduced representation of the data that drastically reduces the complexity of the conformational search. The main application areas of this data-reduction technique are flexible fitting, for which it is uniquely suited, as well as interactive fitting and database query, which both require sufficient computational efficiency. By integrating vector quantization with molecular visualization routines, scientists could, in principle, build models and perform flexible and rigid-body docking interactively within a single computational environment. Fitting by data reduction also has the functionality of a database query tool. The reduced representation and fast recognition of templates in EM maps will be of increasing importance in the future, as a number of microscopists have proposed to build a standardized database of 3D volumetric data in analogy to the Protein Data Bank in crystallography. A prototype microscopy database is already operational [62]. We expect that the method could be easily adapted toward a shape-based query of such databases.

Exhaustive Search

Algorithms should ideally perform a full rigid-body search to systematically explore all possible solutions. A six-dimensional exhaustive search as offered by the correlation-based methods will be the likely choice of more conservative investigators. The computational time requirements (Figure 1) are not prohibitive considering that much more time is spent on the experimental discovery of high- and low-resolution structures.

Discriminative Scoring

An exhaustive search may lead to ambiguous matches or false positives. The ranking criterion should be discriminative enough to eliminate spurious matches and to find unique solutions if warranted by the underlying data. Our test calculations (Figure 4) demonstrate that density masking (local renormalization of the correlation) and filtering (Laplacian operator) are useful techniques that eliminate spurious fits and improve the discriminative contrast of the correlation-based fitting criterion. To resolve ambiguities of the standard correlation criterion, it is also possible to combine multiresolution docking with additional information from a variety of biological or biochemical sources. Such constraints are typically derived from biochemical footprinting of contact interfaces between proteins, covalent crosslinks such as disulfide bridges, or in the form of orientational markers such as heavy-metal clusters that can be observed with the microscope. In one study, the EM-based fitting of fimbrin to actin was improved significantly by using information from peptide and deletion studies and from mutagenesis [63]. In another study, the use of gold labels favored clearly one of the three published orientations of kinesin docked to the microtubule [64], although at present only one such model can be ruled out, and the debate about the remaining two solutions [65] argues for the need to develop additional computational tools that objectively measure the agreement of models with constraints from metal labeling.

Scope of Applicability

The docking should not be limited to isolated components alone and should enable, for example, the fitting of monomeric subunits to large complexes or the localization of small components in macromolecular low-resolution structures. Standard linear correlation works well for high resolution (<8 Å) where the internal structure of the EM data is sufficient for the recognition of structural templates. In our tests on idealized systems, density masking and filtering outperform the standard correlation criterion and significantly extend the viable resolution range. Laplacian filtering is the best-performing docking criterion and works well for resolutions as low as 26 Å. We note that Laplacian filtering corresponds to a multiplication of the structure factors with a harmonic function in Fourier space [27]. Hence, the Laplacian acts as a high-pass filter that suppresses low frequencies. On the one hand, this effect is desirable, because it reduces low-resolution terms beyond the spatial frequencies present in the probe molecule that contribute to the localized noise from extraneous densities. On the other hand, the Laplacian also amplifies the random high-resolution noise present in experimental EM maps. We have tested the satisfactory performance of the Laplacian on experimental maps [58], and we expect that future validations will continue to demonstrate the merit of this novel criterion.

Robustness under Experimental Limitations

Even when taking into account the resolution differences, there is often only an approximate correspondence between crystal structures and EM data. EM densities may be missing certain regions that are accounted for in the atomic structure (and vice versa) due to disorder [4]. Also, the relationship between corresponding amplitudes from crystallographic electron densities and EM maps is often frequency dependent [66, 67]. Several authors have arrived at a consensus estimate that a docking precision one order of magnitude (or better) above the nominal spatial resolution of an EM map can be achieved [18, 44, 60] in noise-free situations. But how tolerant are these fitting strategies of the differences between EM and crystallographic densities? How can one estimate the docking precision?

Our results demonstrate that the contrast and resolvability of various correlation-based criteria is quite variable (Figures 4b–4d). A statistical analysis that estimates the positional and orientational variabilities of solution sets based on the spatial distribution of underlying scores [44, 46] estimates only the intrinsic uncertainty of a method, and therefore provides only a lower bound estimate of the total fitting error. If one is concerned with the systematic error of a scoring function, what matters is not the distribution of the scores around the maximum peak but the deviation of the top scoring model (Figures 4b–4d) from the true structure. Therefore, we argue that method validation should proceed with experimental maps, if corresponding atomic structures exist [59]. In the absence of atomic structures, simulated EM maps could be calculated from a docking model by resolution lowering [17], and the fitting strategy could be applied a second time to test how well a particular model can be reproduced. Undoubtedly, such self-imposed quality standards will help multiresolution modeling efforts gain acceptance among traditional structural biologists and will reach ubiquity in the near future.

Acknowledgments

Support by the National Institutes of Health (R01 GM62968 and P41 RR 12255) and by the La Jolla Interfaces in Science Program/ Burroughs Wellcome Fund is gratefully acknowledged.

References

1. Alberts, B. (1998). The cell as a collection of protein machines: preparing the next generation of molecular biologists. Cell *92*, 291–294.
2. Šali, A., and Kuriyan, J. (1999). Challenges at the frontiers of structural biology. Trends Cell Biol., *9*, M20–M24.
3. Sánchez, R., et al., and Šali, A. (2000) Protein structure modeling for structural genomics. Nat. Struct. Biol. *7*, 986–990.
4. Frank, J. (1996). Three-Dimensional Electron Microscopy of Macromolecular Assemblies (San Diego: Academic Press).
5. Baker, T.S., and Johnson, J.E. (1996). Low resolution meets high: towards a resolution continuum from cells to atoms. Curr. Opin. Struct. Biol. *6*, 585–594.
6. DeRosier, D.J., and Harrison, S.C. (1997). Macromolecular assemblages: sizing things up. Curr. Opin. Struct. Biol. *7*, 237–238.
7. Baumeister, W., and Steven, A.C. (2000). Macromolecular electron microscopy in the era of structural genomics. Trends Biochem. Sci. *25*, 624–631.
8. Carragher, B., et al., and Reilein, A. (2000). Leginon: an automated system for acquisition of images from vitreous ice specimens. J. Struct. Biol. *132*, 33–45.
9. Schröder, R.R., et al., and Spudich, J.A. (1993). Three-dimen-

sional atomic model of F-actin decorated with *Dictyostelium* myosin S1. Nature *364*, 171–174.

10. Rayment, I., et al., and Milligan, R.A. (1993). Structure of the actin-myosin complex and its implications for muscle contraction. Science *261*, 58–65.
11. Olson, N.H., et al., and Rossmann, M.G. (1993). Structure of a human rhinovirus complexed with its receptor molecule. Proc. Natl. Acad. Sci. USA *90*, 507–511.
12. Smith, T.J., et al., and Baker, T.S. (1993). Structure of human rhinovirus complexed with Fab fragments from a neutralizing antibody. J. Virol. *67*, 1148–1158.
13. Stewart, P.L., Fuller, S.D., and Burnett, R.M. (1993). Difference imaging of adenovirus: bridging the resolution gap between X-ray crystallography and electron microscopy. EMBO J. *12*, 2589–2599.
14. Agrawal, R.K., Heagle, A.B., Penczek, P., Grassucci, R.A., and Frank, J. (1999). EF-G-dependent GTP hydrolysis induces translocation accompanied by large conformational changes in the 70S ribosome. Nature Struct. Biol. *6*, 643–647.
15. Moores, C.A., Keep, N.H., and Kendrick-Jones, J. (2000). Structure of the utrophin actin-binding domain bound to F-actin reveals binding by an induced fit mechanism. J. Mol. Biol. *297*, 465–480.
16. Stowell, M.H.B., Miyazawa, A., and Unwin, N. (1998). Macromolecular structure determination by electron microscopy: new advances and recent results. Curr. Opinion Struct. Biol. *8*, 595–600.
17. Belnap, D.M., Kumar, A., Folk, J.T., Smith, T.J., and Baker, T.S. (1999). Low-resolution density maps from atomic models: how stepping "back" can be a step "forward." J. Struct. Biol. *125*, 166–175.
18. Rossmann, M.G. (2000). Fitting atomic models into electron-microscopy maps. Acta Crystallogr. D *56*, 1341–1349.
19. Baker, T.S., Olson, N.H., and Fuller, S.D. (1999). Adding the third dimension to virus life cycles: three-dimensional reconstruction of icosahedral viruses from cryo-electron micrographs. Microbiol. Mol. Biol. Rev. *63*, 862–922.
20. Holmes, K.C., Popp, D., Gebhard, W., and Kabsch, W. (1990). Atomic model of the actin filament. Nature *347*, 44–49.
21. Lorenz, M., Popp, D., and 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.
22. Mendelson, R.A., and Morris, E. (1994). The structure of F-actin. Results of global searches using data from electron microscopy and X-ray crystallography. J. Mol. Biol. *240*, 138–154.
23. Drenth, J. (1999). Principles of Protein X-Ray Crystallography, Second Edition (New York: Springer Verlag).
24. Brünger, A.T. (1992). X-PLOR, Version 3.1: A System for X-Ray Crystallography and NMR (New Haven, CT: Yale University Press).
25. Huber, R., and Schneider, M. (1985). A group refinement procedure in protein crystallography using Fourier transforms. J. Appl. Cryst. *18*, 165–169.
26. Mathieu, M., et al., and Rey, F.A. (2001). Atomic structure of the major capsid protein of rotavirus: implications for the architecture of the virion. EMBO J. *20*, 1485–1497.
27. Bracewell, R.N. (1986). The Fourier Transform and Its Applications, Second Edition (New York: McGraw-Hill).
28. DeRosier, D.J. and Klug, A. (1968). Reconstruction of three dimensional structures from electron micrographs. Nature, *217*, 130–134.
29. Stewart, M. (1988). Computer image processing of electron micrographs of biological structures with helical symmetry. J. Electron Microsc. Tech. *9*, 325–358.
30. Cochran, W., Crick, F.H., and Vand, V. (1952). The structure of synthetic polypeptides. I. The transform of atoms on a helix. Acta Crystallogr. *5*, 581–586.
31. Hawkes, P.W., and Valdrè, U. (1990). Biophysical Electron Microscopy (London: Academic Press).
32. Mendelson, R., and Morris, E.P. (1997). The structure of acto-myosin subfragment 1 complex: Results of searches using data from electron microscopy and x-ray crystallography. Proc. Natl. Acad. Sci. USA *94*, 8533–8538.
33. Kabsch, W., Mannherz, H.G., Suck, D., Pai, E.F., and Holmes, K.C. (1990). Atomic structure of the actin:DNase I complex. Nature 347, 37–44.
34. Laskowski, R.A., MacArthur, M.W., Moss, D.S., and Thornton, J.M. (1993). PROCHECK: a program to check the stereochemical quality of protein structures. J. Appl. Crystallogr. *26*, 283–291.
35. Mueller, F., et al., and Brimacombe, R. (2000). The 3D arrangement of the 23 S and 5 S rRNA in the *Escherichia coli* 50S ribosomal subunit based on a cryo-electron microscopic reconstruction at 7.5 Å resolution. J. Mol. Biol. *298*, 35–59.
36. Galkin, V.E., Orlova, A., Lukoyanova, N., Wriggers, W., and Egelman, E.H. (2001). ADF stabilizes an existing state of F-actin and can change the tilt of F-actin subunits. J. Cell Biol. *153*, 75–86.
37. Llorca, O., et al., and Valpuesta, J.M. (2000). Eukariotic chaperonin CCT stabilizes actin and tubulin folding intermediates in open quasi-native conformations. EMBO J. *19*, 5971–5979.
38. Kikkawa, M., Okada, Y., and Hirokawa, N. (2000). 15 Å resolution model of the monomeric kinesin motor, KIF1A. Cell *100*, 241–252.
39. Wriggers, W., Agrawal, R.K., Drew, D.L., McCammon, J.A., and Frank, J. (2000). Domain motions of EF-G bound to the 70S ribosome: insights from a hand-shaking between multi-resolution structures. Biophys. J. *79*, 1670–1678.
40. Makhoul, J., Roucos, S., and Gish, H. (1985). Vector quantization in speech coding. Proc. IEEE *73*, 1551–1588.
41. van Hulle, M.M. (2000). Faithful Representations and Topographic Maps: From Distortion- to Information-Based Self-Organisation (New York: John Wiley and Sons).
42. Wriggers, W., Milligan, R.A., Schulten, K., and McCammon, J.A. (1998). Self-organizing neural networks bridge the biomolecular resolution gap. J. Mol. Biol. *284*, 1247–1254.
43. Wriggers, W., Milligan, R.A., and McCammon, J.A. (1999). Situs: a package for docking crystal structures into low-resolution maps from electron microscopy. J. Struct. Biol. *125*, 185–195.
44. Wriggers, W., and Birmanns, S. (2001). Using Situs for flexible and rigid-body fitting of multi-resolution single molecule data. J. Struct. Biol. *133*, 193–202.
45. Rice, W.J., Young, H.S., Martin, D.W., Sachs, J.R., and Stokes, D.L. (2001). Structure of Na^+, K^+-ATPase at 11-Å resolution: comparison with Ca^{2+}-ATPase in E^1 and E^2 states. Biophys. J. *80*, 2187–2197.
46. Volkmann, N., and Hanein, D. (1999). Quantitive fitting of atomic models into observed densities derived by electron microscopy. J. Struct. Biol. *125*, 176–184.
47. Zhang, X., et al., and Freemont, P.S. (2000). Structure of the AAA ATPase p97. Mol. Cell *6*, 1473–1484.
48. Nogales, E., Whittaker, M., Milligan, R.A., and Downing, K.H. (1999). High-resolution model of the microtubule. Cell *96*, 79–88.
49. Kleywegt, G.J., and Jones, T.A. (1997). Template convolution to enhance or detect structural features in macromolecular electron-density maps. Acta Crystallogr. D *53*, 179–185.
50. de Groot, B.L., Heymann, J.B., Engel, A., Mitsouka, K., Fujiyoshi, Y., and Grubmüller, H. (2000). The fold of human aquaporin 1. J. Mol. Biol. *300*, 987–994.
51. Meurer-Grob, P., Kasparian, J., and Wade, R.H.. Microtubule structure at improved resolution. Biochemistry *40*, 8000–8008, 2001.
52. Jiang, W., Baker, M.L., Ludtke, S.J., and Chiu, W. (2001). Bridging the information gap: computational tools for intermediate resolution structure interpretation. J. Mol. Biol. *308*, 1033–1044.
53. Cowtan, K. (1998). Modified phase translation functions and their application to molecular fragment location. Acta Crystallogr. D *54*, 750–756,.
54. Katchalski-Katzir, E., Shariv, I., Eisenstein, M., Friesem, A.A., Aflalo, C., and Vakser, I.A. (1992). Molecular surface recognition: determination of geometric fit between proteins and their ligands by correlation techniques. Proc. Natl. Acad. Sci. USA *89*, 2195–2199.
55. Gabb, H.A., Jackson, R.M., and Sternberg, M.J. (1997). Modelling protein docking using shape complementarity, electrostatics and biochemical information. J. Mol. Biol. *272*, 106–120.
56. Mandell, J.G., et al., and Ten Eyck, L.F. (2001). Protein docking using continuum electrostatics and geometric fit. Protein Eng. *14*, 105–113.

57. Vakser, I.A., Matar, O.G., and Lam, C.F.A. (1999). Systematic study of low-resolution recognition in protein-protein complexes. Proc. Natl. Acad. Sci. USA *96*, 8477–8482.
58. Chacón, P., and Wriggers, W. (2001). Fitting multi-resolution structures with Fourier template convolution. (Conference abstract) Biophys. J. *80* (no. 1, pt. 2), 414A.
59. Roseman, A.M. (2000). Docking structures of domains into maps from cryo-electron microscopy using local correlation. Acta Crystallogr. D *56*, 1332–1340.
60. Volkmann, N., Hanein, D., Ouyang, G., Trybus, K.M., DeRosier, D.J., and Lowey, S. (2000). Evidence for cleft closure in actomyosin upon ADP release. Nat. Struct. Biol. *7*, 1147–1155.
61. Russ, J.C. (1998). The Image Processing Handbook, Third Edition (Boca Raton, FL: CRC Press).
62. Marabini, R., Vaquerizo, C., Fernandez, J.J., Carazo, J.M., Engel, A., and Frank, J. (1996). Proposal for a new distributed database of macromolecular and subcellular structures from different areas of microscopy. J. Struct. Biol. *116*, 161–166.
63. Hanein, D., et al., and Matsudaira, P. (1998). An atomic model of fimbrin binding to F-actin and its implications for filament crosslinking and regulation. Nat. Struct. Biol. *5*, 787–792.
64. Rice, S., et al., and Vale, R.D. (1999). A structural change in the kinesin motor protein that drives motility. Nature *402*, 778–784.
65. Amos, L.A. (2000). Focusing-in on microtubules. Curr. Opin. Struct. Biol. *10*, 236–241.
66. Fuller, S.D. (1987). The T=4 envelope of sindbis virus is organized by interactions with a complementary T=3 capsid. *Cell* 48, 923–934.
67. Penczek, P., Ban, N., Grassucci, R.A., Agrawal, R.K., and Frank, J. (1999). *Haloarcula marismortui* 50S subunit—complementarity of electron microscope and x-ray crystallographic information. J. Struct. Biol. *128*, 44–50.
68. Humphrey, W.F., Dalke, A., and Schulten, K. (1996). VMD—visual molecular dynamics. J. Mol. Graphics *14*, 33–38.
69. Goldstein, H. (1980). Classical Mechanics (Reading, MA: Addison-Wesley Publishing Co.).
70. Press, W.H., Teukolsky, S.A., Vetterling, W.T., and Flannery, B.P. (1992). Numerical Recipes in C, Second Edition (New York: Cambridge University Press).