BIOINFORMATICS ORIGINAL PAPER

Vol. 23 no. 4 2007, pages 427–433
doi:10.1093/bioinformatics/btl625

Structural bioinformatics

ADP_EM: fast exhaustive multi-resolution docking for high-throughput coverage

José Ignacio Garzón, Julio Kovacs[1], Ruben Abagyan[1] and Pablo Chacón*

Centro de Investigaciones Biológicas, CSIC, Ramiro de Maeztu 9, 28040 Madrid, Spain and
[1]Department of Molecular Biology, The Scripps Research Institute La Jolla, CA 92037, USA

Received on September 28, 2006; revised on November 28, 2006; accepted on December 4, 2006
Advance Access publication December 6, 2006
Associate Editor: Alex Bateman

ABSTRACT

Motivation: Efficient fitting tools are needed to take advantage of a fast growth of atomic models of protein domains from crystallography or comparative modeling, and low-resolution density maps of larger molecular assemblies. Here, we report a novel fitting algorithm for the exhaustive and fast overlay of partial high-resolution models into a low-resolution density map. The method incorporates a fast rotational search based on spherical harmonics (SH) combined with a simple translational scanning.

Results: This novel combination makes it possible to accurately dock atomic structures into low-resolution electron-density maps in times ranging from seconds to a few minutes. The high-efficiency achieved with simulated and experimental test cases preserves the exhaustiveness needed in these heterogeneous-resolution merging tools. The results demonstrate its efficiency, robustness and high-throughput coverage.

Availability: http://sbg.cib.csic.es/Software/ADP_EM

Contact: pablo@cib.csic.es

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

Rigid-body fitting is the standard way of interpreting the information contained in electron-microscopy (EM) maps of macromolecular structures by means of the available atomic structural components. This is a complicated jigsaw puzzle in which the low-resolution 3D EM density map of a macromolecule acts as a fuzzy frame to guide the assemblage of interlocking atomic-resolution pieces. When complete, this jigsaw puzzle produces a near-atomic-detail picture of the entire macromolecule. Thus, by solving this puzzle we can have access to a better understanding of the inner workings of the central actors in the principal cellular processes.

A number of high-performance fitting algorithms and programs have been developed over the last years. Programs like EMFIT (Rossmann, 2000), COAN (Volkmann and Hanein, 2003), DOCKEM (Roseman, 2000), FOLDHUNTER (Jiang *et al*., 2001), COLORES (Chacon and Wriggers, 2002) of SITUS (Chacon and Wriggers, 2002; Wriggers *et al*., 1999) FRM (Kovacs *et al*., 2003), URO (Navaza *et al*., 2002) and 3SOM (Ceulemans and Russell, 2004) have been employed successfully to provide relevant structural insights into macromolecular function. In essence, these tools perform an automated search of the all possible relative rotations and translations to maximize a density correlation function. This correlation is typically calculated between a target experimental EM map and a simulated probe map obtained by lowering the resolution of the atomic structure to be docked [for reviews see Wriggers and Chacon, 2001 and Fabiola and Chapman, 2005]. Despite its successful application, the exhaustive search performed by the majority of these docking tools is a very time-consuming process, and therefore they are not ready to support high-throughput fitting process.

With current structural genomics efforts, structure modeling advances and the forthcoming perspective of 3D imaging of macromolecules in their native context (Baumeister and Steven, 2000; Lucic *et al*., 2005), even faster algorithms need to be developed (Nickell *et al*., 2006; Russell *et al*., 2004; Sali *et al*., 2003).

A preferred fitting algorithm should balance efficiency and robustness, where efficiency is generally associated with reduced computational cost, and robustness with the accuracy and exhaustiveness of the docking search. A multi-resolution structural puzzle can be extremely complex, likely resulting in different docking poses due to a number of complicating factors: resolution differences, low signal-to-noise ratio of the EM map, deviations between the atomic and the EM structures, (such as missing regions, disorder and conformational changes), etc. A high speed algorithm can give a unique and critical advantage. It will allow scanning through a large number of possible alternative models for the domains fitted into a larger density map. For example, it is common that the atomic structures of individual components of the molecule imaged by EM are unknown. In this situation, one can appeal to homology modeling, which can give us an extensive set of potential atomic models. Topf and collaborators utilize MODELLER (Fiser and Sali, 2003) to produce alternative comparative models that can be placed within the target 3D EM map of the complex (Topf *et al*., 2005; Topf and Sali, 2005). These authors also demonstrated the usefulness of an inverse task in which intermediate-resolution EM maps are used for improving the comparative modeling accuracy (Topf *et al*., 2006). If related structures are not available, fold assignment and template selection procedures can be applied when the resolution of the map is better than 12 Å. SPI-EM (Velazquez-Muriel *et al*., 2005), using a combination of statistical methods and docking searches, was able to determine which CATH superfamily domains can be docked into a target EM map. Related docking programs, such as Helixhunter (Jiang *et al*., 2001) or EMatch (Dror *et al*., 2007) include template matching procedures to identify secondary structure

*To whom correspondence should be addressed.

elements in 3DEM maps. All of these approaches will benefit from new methods that can perform efficient rigid-body query searches.

Here we present a novel multi-resolution docking method with an excellent trade-off between efficiency and precision. This method is a new combination of the fast rotational matching (FRM) method (Kovacs and Wriggers, 2002) with translational scans, and can also be considered as a practical simplification of the FRM5D approach described in Kovacs *et al*., 2003. Instead of recasting the exhaustive search into a formulation involving five angles and just one translational parameter as in FRM5D, here we only speed up the three rotational degrees of freedom (DOF) using FRM, while the three translational DOF are simply scanned. By means of spherical harmonics (SH) and a convenient formulation of the 3D rotation group with an optimized code design, we are able to achieve superior efficiency and exhaustiveness for searching the rotational space. This novel approach does not suffer from the strong memory limitations of the FRM5D method (Kovacs *et al*., 2003) and constitutes a fast and robust search tool for multi-query docking searches. Results with a variety of simulated and experimental 3D EM maps confirm its efficiency and applicability. Moreover, the developed methodology is an all-purpose registration tool that can be readily applied to any 3D rigid-body registration problem.

2 METHODS

The computational solution of the search problem can be reduced to finding the relative orientation and translation, which maximizes the density cross-correlation of the structures/maps to be docked. In this case, and for a given rotation and translation, the fitting criterion is typically defined as the scalar product between the EM experimental map ρ_{low}, and a low-pass filtered version of the atomic structure, ρ_{high}, mathematically:

$$C(T,R) = \int_{\mathbb{R}^3} \rho_{\text{low}} \times \Omega_T \Lambda_R \rho_{\text{high}},$$

where Ω_T and Λ_R denote the translational and rotational operators, respectively. To find the highest correlation values, previous approaches would perform a systematic rotational scan of a probe structure (usually ρ_{high}) relative to a fixed reference (ρ_{low}), combining it with a fast fourier transform (FFT)-accelerated translational search based on the convolution theorem. This well-known exhaustive search protocol is borrowed from the protein–protein docking field (Gabb *et al*., 1997; Katchalski-Katzir *et al*., 1992; Vakser *et al*., 1999), and is used, among others, by the *de facto* standard multi-resolution docking tool COLORES (Chacon and Wriggers, 2002). As an alternative to speeding up the translational DOF by means of FFTs, we accelerate the rotational search, thereby providing, as shown in this paper, a much higher efficiency. This method, termed FRM, uses a suitable parametrization of the 3D rotation group with SH to efficiently compute the rotational part of the correlation function. A detailed description of the theory underlying the FRM method was given elsewhere (Kovacs *et al*., 2003; Kovacs and Wriggers, 2002). Briefly, the density functions to be docked are first approximated by expansions in SH functions. To this end, the density volume is partitioned into concentric spherical layers (like onion shells) each of which is approximated by finite sums as:

$$\begin{aligned}\rho_{\text{low}}(r,\beta,\lambda) &= \sum_{l=0}^{B-1} \sum_{m=-l}^{l} C_{lm}^{\text{low}}(r) Y_{lm}(\beta,\lambda)\\ \rho_{\text{high}}(r,\beta,\lambda) &= \sum_{l=0}^{B-1} \sum_{m=-l}^{l} C_{lm}^{\text{high}}(r) Y_{lm}(\beta,\lambda),\end{aligned} \tag{1}$$

where:

- $C_{lm}(r)$ are coefficients associated with a specific, complex-valued spherical harmonic function $Y_{lm}(\beta,\lambda)$ defined on the unit sphere.
- $l \geq 0$ and $-l \leq m \leq l$ are the SH degree and order, and β and λ are the co-latitude and longitude, respectively.
- According to the sampling theorem, the number of sampling points (in each β and λ) used is equal to twice of the bandwidth B.

Instead of recasting the exhaustive search into a formulation involving five angles and just one translational parameter (Kovacs *et al*., 2003), here we only accelerate the three rotational DOF, while the three translational ones are simply scanned. Considering only the rotational part, the fitting function can now be expressed in terms of an inverse Fourier transform of the SH transforms [Equation (1)] of the density maps (Kovacs and Wriggers, 2002):

$$C(R) = FT^{-1}_{m,h,m'}\left[\sum_l d^l_{mh}\, d^l_{hm'} \int_0^\infty C^{\text{low}}_{lm}(r)\overline{C^{\text{high}}_{lm'}(r)}\, r^2 dr\right], \tag{2}$$

where the d^l_{mn} are real coefficients that define the matrix elements of the irreducible representations of the 3D rotation group. This expression can be computed very efficiently by precomputing the coefficients d^l_{mn} and by using as upper limit of integration the maximum shell radius for which the density has non-zero values. In this way, Equation (2) allows, for a given translation, a very fast calculation of the correlation function for all rotations, which will come out sampled at twice the bandwidth B used in the harmonic transformation of the maps [Equation (1)]. For example, $B = 16$ corresponds to scanning ∼16 000 rotations with a sampling step of 11.25°. If the rotational sampling step is set to 5.6° ($B = 32$), >130 000 rotations will be explored. Thus, this method offers an adaptable and fine rotational screening.

The exhaustive search is then performed by applying this FRM rotational scan on a list of sampled points that uniformly covers the translational search space. To prevent exploring points without physical meaning, the translational space is limited to positions on which the dimension of the probe (atomic structure) roughly fits inside the experimental EM map. To this end a mask is defined by points inside the target map and eroded by the minimum radius of the probe structure. Alternative (and more efficient) translational search strategies have been implemented, such as radial search (useful for structures with holes) or center-based search (practical for docking structures with similar dimensions) (Kovacs *et al*., 2003). These sampling schemes take advantage of geometry but their application range is not universal as the uniform sampling scheme using a mask. Therefore, here we only report the results obtained with the masking strategy. Although larger translational samplings can be used, in our tests we found that by exploring every other voxel we did not miss any significant correlation peak. This was possible because the correlation is interpolated using a simple parabolic approximation between the six 3D neighboring positions. The selection of the translational sampling was a practical solution of compromise between exhaustiveness and efficiency. Further work will be done to establish larger sampling limits in which the exhaustiveness is granted or develop new efficient coarser grid search strategies.

The density cross-correlation works reasonably well, although in particular cases its use as docking criterion may lead to ambiguous matches or false positives. This can be critical at low resolutions (worse than 15 Å) when small components are to be placed in a large density map. Several alternatives can be adopted to improve the fitting contrast. For example, the fitting can be performed by a local correlation criterion (Rath *et al*., 2003; Roseman, 2000), or the maps can be pre-filtered with a Laplacian kernel (Chacon and Wriggers, 2002). Since its implementation does not need any change in the registration scheme, here the partial docking is performed with Laplacian-filtered maps instead of the original density maps. The strategy of convolving the maps with a Laplacian kernel improves the numerical contrast among potential solutions, by including both density and contour overlap. Despite its known limitations, such as sensitivity to high-frequency noise and cases where the surface exposure of the probe structure is relatively limited, many successful applications have been reported; see for example (Golas *et al*., 2003; Laurinmäki *et al*., 2005; Leiman *et al*., 2004; Opalka *et al*., 2003; Petosa *et al*., 2004; Samso *et al*., 2006; Sandin *et al*., 2004; Sewell *et al*., 2003).

To extend multi-resolution docking techniques to higher-throughput coverage, the implementation of the new combination of FRM and the translational scan was carefully designed and optimized to achieve maximal runtime savings. This new algorithm, called ADP_EM (Another Docking Platform for EM) was coded in C++ to gain the flexibility and reusability of an object-oriented approach.

3 RESULTS

3.1 Docking benchmark

The performance of our novel docking algorithm was firstly tested on 28 simulated docking cases, comprising a wide-variety of macromolecular shapes (see Fig. 1 for a detailed list). Each test case consists of five simulated 3D-EM maps at experimental resolutions of 10, 15, 20, 25 and 30 Å, and an atomic subunit or component of the macromolecular structure used to generate such density maps. By carrying out the docking procedure, the atomic subunit for each test case should be correctly positioned into the corresponding complete EM map. Thus, using this benchmark, we evaluated the performance of our method in the most challenging situation when the atomic structure to be docked represents only a portion of the low-resolution density map. To have statistical significance and avoid pre-alignment situations, for each atomic component the registration search has been repeated 50 times starting from different relative positions. In addition, three different rotational samplings have been used: ~11°, 8° and 6°, which correspond to harmonic bandwidths of 16, 24 and 32, respectively. (See Fig. 1 for a full description of the parameters used.)

The results obtained in this thorough validation test showed that even at low-resolution the algorithm was able to find the correct position with reasonable precision for the vast majority of the 21 000 docking searches performed. In Figure 1, the rmsd between the best docking results and the original target structure is shown as a function of both the resolution and the bandwidth used. As expected, the docking accuracy is gradually degraded as the resolution of the maps is lowered. The rmsd at 10 Å resolution is <1 Å, which can be considered a perfect match. At 20 Å the rmsd is under 2 Å, and for 30 Å it is still close to 3 Å. Only at very low-resolution was it not possible to get a unique and reliable docking result for all the test cases. The docking failed at 30 Å with 5-aminolaevulinate dehydratase, pilin, methyl-coenzyme M reductase and the ribosomal protein S2. Another case for which it was not possible to find the correct pose among the top-scoring solutions was GroES at resolutions below 15 Å. GroES was already identified by other authors as a very difficult docking case (Ceulemans and Russell, 2004; Rossmann, 2000). In cases like this, the relatively small size and the low-resolution prevented a reliable docking.

Empirically, the maximum accuracy that can be obtained with simulated and experimental maps is always at most 1/10 of the nominal resolution of the map (Chacon and Wriggers, 2002). Other authors extend the maximum accuracy achievable with EM experimental maps to 4/10 of the map resolution (Fabiola and Chapman, 2005).

As it can be seen in Figure 1, the accuracy of ADP_EM is below such limits, even with the smallest bandwidth used. There is also a gain in the docking precision when the bandwidth is increased from 16 to 24 and less pronounced from 24 to 32. These improvements are logically due to both the finer rotational sampling and the more accurate harmonic description. Nevertheless, the maximum gain in the best case is only 0.4 Å, which is almost negligible taking into account the maximum accuracy than can be expected of this hybrid multi-resolution docking approach. Thus, we think a bandwidth of 16 suffices to identify the correct pose in the majority of docking scenarios. Only if extra-rotational accuracy is needed should higher harmonic order be considered.

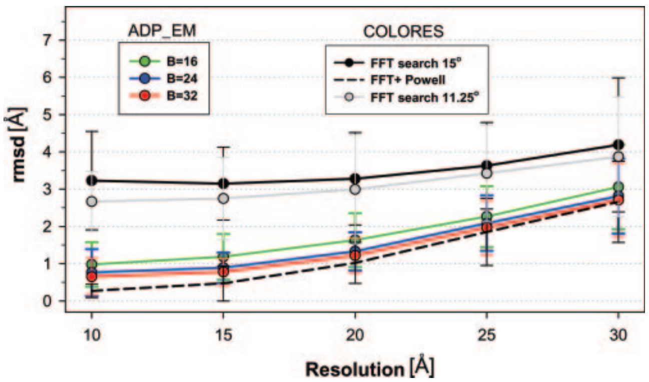

Fig. 1. Registration accuracy. The docking tests were performed on simulated EM maps calculated from 28 atomic macromolecular structures by lowering their resolution using the PDB2MRC program of the EMAN package (Ludtke *et al.*, 1999). The chosen resolutions (10, 15, 20, 25 and 30 Å) correspond to typical experimental ranges of EM measurements. The test structures used are: 5-aminolaevulinate dehydratase (PDB: 1aw5,8× symmetry related copies); chains A(2×), B(2×) and BC(2×) of methyl-coenzyme M reductases (1e6v); glutamine synthetase (1fpy,10×); ATP sulfurylase (1g8g,6×); tricorn protease as homohexamer (1k32,6×); chain A(2×) and CD(2×) of quinol-fumarate reductase (1kf6); ATP sulfurylase (1g8g,6×), holo-glyceraldehyde-3-phosphate dehydrogenase (1gd1,4×); glutamate dehydrogenase (1l1f,6×); tricorn protease as homotrimer (1n6d,3×); Cu-nitrite reductase (1nic,3×); proteasome α-ring(1j2p,7×); cadherin (1q5b,3×); lectin (1w3a,6×); hemagglutinin (1ruz,3×) RecA (1xmv,6×); chain A of voltage-gated potassium channel $\beta 2$-subunit (2a79,4×); pilin (2pil,5×); catalase (7cat,4×); GroEL ATP (7×) and ADP subunits (7×), and GroES (1aon,7×); thermosome (1a6d,14×); large and small subunit of ribosome (1ffk + 1fjf); and ribosomal protein S2 (1ffk − 1fjf). To enhance statistics and prevent pre-alignment situations for each test structure, the registration search was repeated 50 times starting from different translated and rotated replicas of the atomic structure. The registration procedure with all the test cases was performed with three different rotational samplings steps: 11°, 8° and ~6°, which correspond to harmonic bandwidths of 16, 24 and 32, respectively. In all cases, the translational sampling was chosen to be 6 Å, which is twice the grid size of the maps. For validation purposes, the full-atom rmsd has been computed between the highest-correlation fitted subunits and the equivalent structures included in the original macromolecule (used to generate the simulated map). To measure the registration accuracy all the cases have been considered except those in which the method fails. The few failed cases are at 30 Å resolution with 1aw5, 2pil, 1e6v and ribosomal protein S2, and at resolutions below 15 Å with GroES. To perform the FFT translational searches, COLORES was used with the '–nopowell' option and with rotational sampling steps of 15° and 11.25° (the latter corresponding to $B = 16$) (solid black and gray lines, respectively), using default values for all other parameters. The results shown represent the average of 10 runs starting from different relative positions. Subvoxel precision has been obtained by Powell minimization (dashed black line) of the best-fitting results obtained in the Fourier-based COLORES search at 15°.

In Table 1, timings of the docking searches are shown as function of resolution and bandwidth. As can be seen, the dependence on the resolution is marginal, but the increase in the harmonic order considerably increases the docking times. For B = 32, the average time to perform a docking exhaustive search is nearly 222 s, and drops to 113 and 34 s using B = 24 and B = 16, respectively. The search times range from 8 s for a small-size map (e.g., 1nic, with 40^3 voxels) to 3 min for the large ribosome map (100^3 voxels). These timing results show the high-efficiency achieved with this new docking approach.

To our knowledge this is the most complete validation test that has been performed over any multi-resolution docking tool. In addition, for future developments of this and other methods, the docking benchmark has been made available on-line.

3.2 Comparative results

One of the most popular docking program is Situs (Wriggers *et al*., 1999) and its correlation-based exhaustive search tool, COLORES (Chacon and Wriggers, 2002). By means of FFT, this is the fastest cross-correlation maximization method. Even though it is difficult to make a truly fair comparison with ADP_EM because of the opposite philosophies of speeding up either the translational or the rotational search, the approach proposed here clearly offers a great advantage in efficiency. We applied COLORES to the same docking benchmark described in the previous section using a rotational sampling step of 15°. Note that employing the finer rotational sampling used in ADP_EM (11.25°) would increase substantially the number of rotations to be explored (from 4416 to 10 496) and consequently also the computational time. However, a 15° sampling was enough to obtain an overall fitting accuracy similar to our approach. In fact, the fits are correct for the majority of the benchmark cases, whereas the failing cases are the same as those obtained with ADP_EM.

The average time required to perform an FFT translational search using COLORES (without Powell minimization for subvoxel refinement) for all the benchmark test cases was 25 min, whereas with our approach it was only 34 s with B = 16 ($\sim$11°), and <4 min for B = 32 ($\sim$6°). Moreover, our approach scales better with size. As the resolution is gradually lowered, the density spreads out over a larger volume. Thus, in our benchmark we have larger maps as resolution decreases. In Table 1, we can observe how timings of the FFT search progressively increased with resolution due to this effect. This behavior is not observed with APD_EM, thus demonstrating its significantly better scaling performance.

Most importantly, our approach yields better rmsd values. The FFT translational search is limited by the voxel size, and the best fits are >3 Å away from the correct solutions (Fig. 1, solid black line). In contrast, our FRM search achieves higher precision, especially at higher resolutions (Fig. 1, colored lines). Extra increase of accuracy can be obtained by refinement of the best hits. By default COLORES uses a Powell minimization step to achieve the highest accuracy (Fig. 1, black dashed line), but with significant extra computational cost (Table 1). These very low rmsd values are probably meaningless in most real problems where inconsistencies (small missing or disorder regions, minor conformational changes, etc.) between the EM map and the probe structure will always prevent a perfect match. The refinement is a necessary step in COLORES to overcome the translational limit imposed by the voxel size. On the contrary, ADP_EM does not need further refinement because it achieves, even at low harmonic order B, reasonable rmsd values, which are below the empirical limits of multi-resolution docking. Only in particular cases at high-resolutions, where we have an excellent correspondence between the map and the underlying atomic structure, should a local minimization be considered.

Table 1. Timing results, in seconds, obtained with the benchmark described in Figure 1

	Sampling	Resolution				
	B/°	10Å	15Å	20Å	25Å	30Å
ADP_EM	16/11°	28	31	35	34	38
	24/8°	100	108	119	118	123
	32/6°	226	220	225	216	221
FFT search	–/15°	1697	1926	2341	5028	6681
Powell minim	–/15°	375	918	1747	3739	6597

All runs have been performed on a PC Linux box with a Xeon processor at 2.8 GHz. The COLORES (version 2.2.1) program of the Situs package (http://situs.biomachina.org) was used with 15° of accuracy. To perform only FFT translational searches (solid dark line in Fig. 1), the '–nopowell' option has been used. The Powell minimization timings (dashed dark line) consider only the time spent in this off-lattice refinement step. Thus, to compare ADP_EM with the standard running time spent by COLORES, the timings in the last two rows must be added together. In all cases the standard deviation was of the order of the magnitude of the average. In both programs the Laplacian filter was used to improve fitting contrast.

To our knowledge, 3DSOM is the fastest alternative for fitting atomic structures into low-resolution EM maps (Ceulemans and Russell, 2004). This method, based on surface overlap maximization, gives many solutions and therefore it is usually difficult to distinguish the correct solutions from the incorrect ones. In fact, it is necessary to visually inspect a large number of best-scoring solutions and their variations to find fits relatively close to the correct ones. Therefore, we were not able to perform a systematic test with our benchmark. The authors of 3DSOM have already pointed out this limitation, as well as the fact that the rmsd from the correct pose might be slightly higher than those obtained by other methods (Ceulemans and Russell, 2004). In any case, although this approach is faster for small maps (5 s for the smallest map), ADP_EM scales much better with the map size. ADP_EM takes 3 min to dock the large subunit into the whole ribosome map, where 3DSOM takes almost an hour.

We have focused our comparison with the *de facto* standard docking program COLORES and the fastest alternative 3DSOM, since others would be less efficient. Likewise, we do not consider the FRM5D approach (Kovacs *et al*., 2003) because of its strong memory limitations. For example, a bandwidth of B = 32 requires 4.5 GB of memory, which clearly complicates its use in current workstations. Nevertheless, the expected performance of FRM5D ($\sim$12 min in average on the whole benchmark using B = 16) is clearly worse than ADP_EM's.

3.3 Testing with experimental maps

In contrast to a simulated benchmark, there are very few 'gold-standard' X-ray/EM experimental test cases against which new methods can be validated. Here we show results obtained with five EM maps previously used for testing other methods or kindly provided by other labs. The first case is an *Escherichia coli* GroEL–GroES complex at 23 Å [Macromolecule Structure Database (EMD) ID 1046]. The GroEL atomic subunits extracted from the

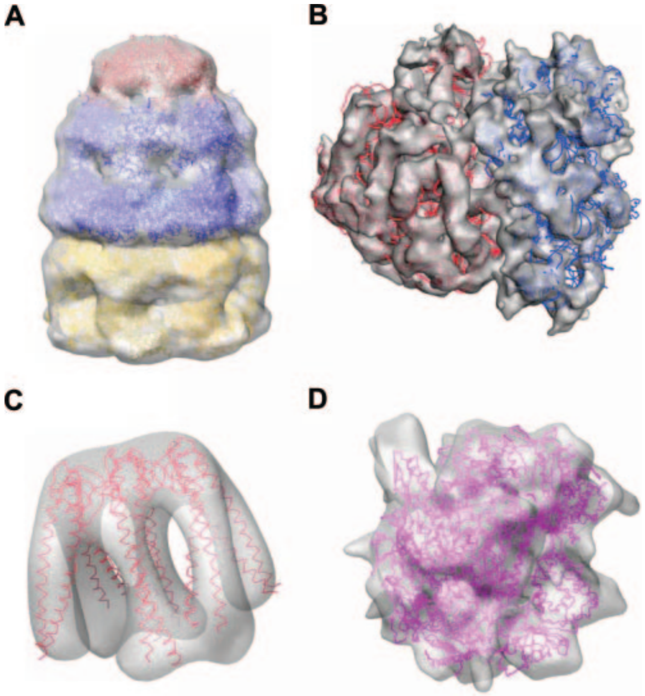

Fig. 2. Docking results with experimental EM data. (**A**) *E.coli* GroES-ADP7-GroEL-ATP7 from *E.coli* at 23.5 Å (EMD ID 1046, PDB: 1ml5); ADP and ATP GroEL subunits have been docked independently to reconstruct the cis and trans heptameric rings of the complex. For GroES the whole heptamer was used. (**B**) Docking of 30S and 50S subunits into *E.coli* ribosome map at 14 Å (EMD ID 1046, PDB: 1gix/1giy). Single-molecule docking of prefoldin (**C**) at 23 Å (Martin-Benito *et al*., 2002), PDB: 1l6h, and of yeast RNA polymerase II (**D**) at 15 Å (Craighead *et al*., 2002), PDB: 1fxk.

PDB entry 1AON were correctly fitted into the corresponding heptameric double-ring of the whole chaperonin system (Fig. 2A). The rmsd differences between the X-ray structures of cis/trans rings and the corresponding reconstructed structure are <1.3 Å. However, as was the case with simulated maps, the docking of the GroES subunits failed. The low-resolution and the relatively very small size of this subunit are likely to be the reasons for this mismatch. Another GroEL-ATP map (EMD ID 1047) at 14.9 Å was successfully reconstructed from its subunits (data not shown). In this case, the rmsd difference between the docked and original structure of the heptameric rings was 1.1 Å.

We obtained excellent fits even if there is not an exact correspondence between the atomic structures and the EM reconstructions. For example, the large and small ribosomal subunits were separately and correctly docked into a 14 Å map (EMD entry 1005, Fig. 2B). The overall 'jellyfish' atomic structure of prefoldin (blue ribbons) fits well into the EM density, with the exception of small differences in the positions of known flexible tentacles (Fig. 2C). We also performed a docking of the eukaryotic RNAPII map with a crystal structure of its *Methanococcus jannaschii* homolog (Fig. 2D). The result obtained reproduces the tedious manual fitting that furnished a model of the nascent RNA and thereby a hypothesis of how RNAPII interacts with promoter DNA (Asturias, 2004).

As occurred with simulated cases, we were able to obtain the same correct fitting results (rmsd divergences below 1/10 of nominal resolution) as with the FFT-based search tool COLORES, but in a much more efficient way. For example, with ADP_EM, the reconstruction of any of the heptameric rings of the GROEL subunit into the GroEL-GroES map takes 40 s, whereas COLORES spent almost 1 h (30 min for FFT + 20 min for Powell minimization) with a rotational sampling of 15°. This difference grows with larger maps: our approach needs 296 s to dock the large ribosomal subunit, while the FFT approach required almost 11 h. As to the 3SOM approach, the registration of GroEL-GroES, GroEL-ATP, 30S, 50S, RNAP and prefoldin took, respectively, 39s, 1m37s, 22m, 64m, 7m and 13m. Our method is comparatively fast, giving acceleration ratios of 1.0, 1.0, 6, 13, 21, 130. The gain in speed is due to the much better scaling performance of ADP_EM, except in the prefoldin case where the acceleration was due mainly to its hollow structure, which simplifies the translational search masking strategy. It was also problematic to identify the correct solution among the large set of possible docking poses that 3SOM produced. It was difficult to locate some of the different heptameric GroEL correct positions, and the correct poses of large subunit of ribosome or the RNAP. In these cases, either the correct pose was hidden in secondary minima or its rmsd with respect to COLORES and ADP-EM solutions was high.

3.4 Homology modeling application

It frequently happens that the original atomic structures of the components to be docked are unknown. In this case, homology-modeling bioinformatics tools provide an extensive set of potentially useful atomic models. Selecting those models with the highest density correlation will most likely lead to the atomic characterization of the target macromolecule imaged by electron microscopy. It has been shown that comparative modeling provides structures that are more useful for fitting into EM maps than the homolog's experimentally determined structures (Topf *et al*., 2005). This docking procedure has proved useful also as a model assessment score in comparative modeling (Topf *et al*., 2006).

Here we apply our ADP_EM docking tool over a benchmark provided by these authors (Topf *et al*., 2005, http://salilab.org/modem). The benchmark is formed by eight pairs of proteins of known structures (each pair consisting of a target structure and its corresponding remote homolog, which is used as modeling template) sharing between 12 and 32% sequence identity. For each pair, 300 alternative comparative models have been built using MODELLER (Fiser and Sali, 2003). The benchmark includes several simulated maps created from the native target structures at different resolutions. The test consists in identifying the most accurate models by fitting all the alternative homology models into the corresponding density maps. To assess the geometrical accuracy of the models, we carried out the structural alignment of each target with its homolog and their corresponding comparative models using the MAMMOTH program (Ortiz *et al*., 2002).

All the models along with the target and template atomic structures have been docked into the target density maps using ADP_EM. As an example in Figure 3, the fitting correlation values are plotted versus the model alignment score using a map of 12 Å resolution of the protein 1MUP. As can be observed, there is a clear correlation between fitting and model accuracy values. The best fitting models always correspond to models structurally close to the target structure. It is important to note that models that are closer to the target structure than the template homolog generally dock better into the map than the homolog itself. As expected, the 1MUP target structure has the highest score, which logically corresponds to a perfect match (See Supplementary Table 1). The template, the best model and the best fitting model have quite similar shape (Fig. 3).

However the docking procedure is able to discriminate among models, specially the template model relative to the other two which have better structural overlap with respect to 1MUP. The best model ranks on the top of the best-correlation list (6–9th). On the contrary, the homolog template structure has significantly lower alignment scores and correlation values (148–167th). The method is robust and, independently of resolution, the same best-fitting model is obtained. The good correspondence between fitting and modeling scores, and the fact of obtaining better fits for the more accurate models rather than for templates, was also observed in all of the other benchmark cases (see Supplementary materials). This fact confirms the potential use of comparative modeling as a docking protocol, as already pointed out by Toft and collaborators regarding its Mod_EM protocol. Here we reproduce their results with a much more efficient and robust protocol. In fact, the time to perform all the 302 fittings was ~50 min, i.e. 10 s per fit. Toft *et al*. compare an improved FOLDHUNTER approach (Jiang *et al*., 2001), which takes ~10–15 min per fit, with a optimized scanning Monte Carlo protocol, Mod_EM, which takes 1–2 min per fit. In terms of fitting, all the approaches yield quite analogous results. But it terms of efficiency, ADP_EM is at least 60 times faster than the most comparable of these protocols, the Fourier-based exhaustive search protocol, FOLDHUNTER. Even though in this particular benchmark the stochastic Monte Carlo approach is still competitive, in the real world, with maps larger than the probe, it is expected that MC be even less efficient than FOLDHUNTER (Topf *et al*., 2005).

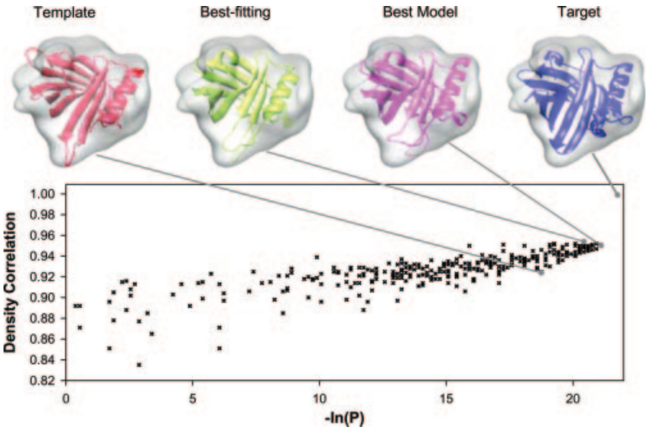

Fig. 3. ADP_EM fitting results obtained with the 1MUP test case of comparative homology-modeling docking benchmark described in (Topf *et al*., 2005) at 12 Å resolution. The target and template structures together with the 300 atomic models have been docked using ADP_EM with a rotational accuracy of 11°. The density correlation of the best-fitting pose obtained in the docking of each model is plotted against the MAMMOTH alignment score between the native target structure and the 300 comparative models used. The alignment score is defined as −ln(P), where P is the probability of obtaining the given proportion of aligned residues with respect to the shortest model by chance (Ortiz *et al*., 2002). Pointing at their corresponding values, the best fits of the template structure (1rbp, red ribbons), of the best-fitting structure (green), and of the most accurate model (pink) are also shown. The template structure fitted with lower density correlation than the other two more accurate models. The fitting difference between the latter two is almost insignificant. In fact, the best-fitting and the best-model structures are very similar to each other, and only small differences in the loops and at the ends of the helix can be observed. Note that this is a stringent test since all of these structures have quite similar shapes.

4 DISCUSSION

ADP_EM offers a practical advancement over existing methods, since it provides a faster and reliable tool for fitting X-ray crystal structures into low-resolution density maps. This new approach reduces docking timings to only a few minutes or even seconds on a standard PC. The high-efficiency achieved with simulated and experimental test cases preserves the exhaustiveness needed in these heterogeneous-resolution merging tools. In addition to time savings, the major advantage of our approach is the fine rotational sampling step (between 11 and 6 degrees) that can be used in the docking search while still keeping superior efficiency. This ensures a thorough 6D exploration avoiding overlooking possible valid docking alternatives.

The level of performance reached, which overcomes previous approximations, opens up a new application window, where fast and robust 6D exhaustive searches are needed. For a given low-resolution structure, the usual practice is to perform multiple dockings, either with a number of different probes, or to resolve scaling uncertainties. This will effectively contribute to obtain accurate near-atomic interpretations of large macromolecular complexes. Moreover, this new approach greatly simplifies the large-scale merging of 3D information data coming from diverse structural sources including bioinformatics modeling. In this context, here we report a homology-modeling test case as an illustrative example of an improved optimization protocol for fitting comparative modeling structures into EM reconstructions. This example can be easily scaled to support a larger number of comparative models from different methods (Ginalski, 2006; Tress *et al*., 2005), including the use of automated web servers (Fischer, 2006). There are additional scenarios where high-throughput coverage is needed. The most attractive are template matching approaches used in cryo-electron tomography (Nickell *et al*., 2006) or in hybrid approaches used for locating CATH superfamilies into 3D-EM reconstructions (Velazquez-Muriel *et al*., 2005). The latter has been recently extended for flexible fitting by exploiting the superfamily's structural variability (Velazquez-Muriel *et al*., 2006). Moreover, ADP_EM will be a very interesting tool to scan multiple flexibility based variants of a model, e.g. generated by using the relevant normal modes of a low-resolution model. All these promising strategies are based on extensive model fitting steps that could strongly profit from our ultra-fast and reliable docking tool.

Since the method constitutes an efficient general 3D registration algorithm, its application range could be extended to other fields. We are currently pursuing the application of the proposed algorithm to protein–protein docking.

ACKNOWLEDGEMENTS

The authors would like to thank J. M. Valpuesta [CNB(CSIC), Spain] for Prefoldin and F. Asturias (The Scripps Research Institute, USA) for RNAPII reconstruction. This work was supported by funds from MEC BFU2004-01282/BMC and Fundación BBVA (P.C.) and NIH 1-R01-GM071872-01 (R.A.).

Conflict of Interest: none declared.

REFERENCES

Asturias,F.J. (2004) RNA polymerase II structure, and organization of the preinitiation complex. *Curr. Opin. Struct. Biol.*, **14**, 121–129.

Baumeister,W. and Steven,A.C. (2000) Macromolecular electron microscopy in the era of structural genomics. *Trends Biochem. Sci.*, **25**, 624–631.

Ceulemans,H. and Russell,R.B. (2004) Fast fitting of atomic structures to low-resolution electron density maps by surface overlap maximization. *J. Mol. Biol.*, **338**, 783–793.

Chacon,P. and Wriggers,W. (2002) Multi-resolution contour-based fitting of macromolecular structures. *J. Mol. Biol.*, **317**, 375–384.

Craighead,J.L. *et al.* (2002) Structure of yeast RNA polymerase II in solution: implications for enzyme regulation and interaction with promoter DNA. *Structure (Camb)*, **10**, 1117–1125.

Dror,O. *et al.* (2007) EMatch: an efficient method for aligning atomic resolution subunits into intermediate-resolution cryo-EM maps of large. *Acta Crystallogr. D. Biol. Crystallogr.*, **63**, 42–49.

Fabiola,F. and Chapman,M.S. (2005) Fitting of high-resolution structures into electron microscopy reconstruction images. *Structure (Camb)*, **13**, 389–400.

Fischer,D. (2006) Servers for protein structure prediction. *Curr. Opin. Struct. Biol.*, **16**, 178–182.

Fiser,A. and Sali,A. (2003) Modeller: generation and refinement of homology-based protein structure models. *Meth. Enzymol.*, **374**, 461–491.

Gabb,H.A. *et al.* (1997) Modelling protein docking using shape complementarity, electrostatics and biochemical information. *J. Mol. Biol.*, **272**, 106–120.

Ginalski,K. (2006) Comparative modeling for protein structure prediction. *Curr. Opin. Struct. Biol.*, **16**, 172–177.

Golas,M.M. *et al.* (2003) Molecular architecture of the multiprotein splicing factor SF3b. *Science*, **300**, 980–984.

Jiang,W. *et al.* (2001) Bridging the information gap: computational tools for intermediate resolution structure interpretation. *J. Mol. Biol.*, **308**, 1033–1044.

Katchalski-Katzir,E. *et al.* (1992) Molecular surface recognition: determination of geometric fit between proteins and their ligands by correlation techniques. *Proc. Natl Acad. Sci. USA*, **89**, 2195–2199.

Kovacs,J.A. *et al.* (2003) Fast rotational matching of rigid bodies by fast Fourier transform acceleration of five degrees of freedom. *Acta Crystallogr. D. Biol. Crystallogr.*, **59**, 1371–1376.

Kovacs,J.A. and Wriggers,W. (2002) Fast rotational matching. *Acta Crystallogr. D. Biol. Crystallogr.*, **58**, 1282–1286.

Laurinmäki,P.A. *et al.* (2005) Membrane proteins modulate the bilayer curvature in the bacterial virus Bam35. *Structure*, **13**, 1819–1828.

Leiman,P.G. *et al.* (2004) Three-dimensional rearrangement of proteins in the tail of bacteriophage T4 on infection of its host. *Cell*, **118**, 419–429.

Lucic,V. *et al.* (2005) Structural studies by electron tomography: from cells to molecules. *Annu. Rev. Biochem.*, **74**, 833–865.

Ludtke,S.J. *et al.* (1999) EMAN: semiautomated software for high-resolution single-particle reconstructions. *J. Struct. Biol.*, **128**, 82–97.

Martin-Benito,J. *et al.* (2002) Structure of eukaryotic prefoldin and of its complexes with unfolded actin and the cytosolic chaperonin CCT. *EMBO J.*, **21**, 6377–6386.

Navaza,J. *et al.* (2002) On the fitting of model electron densities into EM reconstructions: a reciprocal-space formulation. *Acta Crystallogr. D. Biol. Crystallogr.*, **58**, 1820–1825.

Nickell,S. *et al.* (2006) A visual approach to proteomics. *Nat. Rev. Mol. Cell. Biol.*, **7**, 225–230.

Opalka,N. *et al.* (2003) Structure and function of the transcription elongation factor GreB bound to bacterial RNA polymerase. *Cell*, **114**, 335–345.

Ortiz,A.R. *et al.* (2002) MAMMOTH (matching molecular models obtained from theory): an automated method for model comparison. *Protein Sci.*, **11**, 2606–2621.

Petosa,C. *et al.* (2004) Architecture of CRM1/Exportin1 suggests how cooperativity is achieved during formation of a nuclear export complex. *Mol. Cell*, **16**, 761–775.

Rath,B.K. *et al.* (2003) Fast 3D motif search of EM density maps using a locally normalized cross-correlation function. *J. Struct. Biol.*, **144**, 95–103.

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

Rossmann,M.G. (2000) Fitting atomic models into electron-microscopy maps. *Acta Crystallogr. D.*, **56**, 1341–1349.

Russell,R.B. *et al.* (2004) A structural perspective on protein-protein interactions. *Curr. Opin. Struct. Biol.*, **14**, 313–324.

Sali,A. *et al.* (2003) From words to literature in structural proteomics. *Nature*, **422**, 216–225.

Samso,M. *et al.* (2006) Structural characterization of the RyR1-FKBP12 interaction. *J. Mol. Biol.*, **356**, 917–927.

Sandin,S. *et al.* (2004) Structure and flexibility of individual immunoglobulin G molecules in solution. *Structure (Camb)*, **12**, 409–415.

Sewell,B.T. *et al.* (2003) The cyanide degrading nitrilase from Pseudomonas stutzeri AK61 is a 2-fold symmetric, 14-subunit spiral. *Structure*, **11**, 1413–1422.

Topf,M. *et al.* (2005) Structural characterization of components of protein assemblies by comparative modeling and electron cryo-microscopy. *J. Struct. Biol.*, **149**, 191–203.

Topf,M. *et al.* (2006) Refinement of protein structures by iterative comparative modeling and CryoEM density fitting. *J. Mol. Biol.*, **357**, 1655–1668.

Topf,M. and Sali,A. (2005) Combining electron microscopy and comparative protein structure modeling. *Curr. Opin. Struct. Biol.*, **15**, 578–585.

Tress,M. *et al.* (2005) Assessment of predictions submitted for the CASP6 comparative modeling category. *Proteins*, **61** (Suppl. 7), 27–45.

Vakser,I.A. *et al.* (1999) A systematic study of low-resolution recognition in protein–protein complexes. *Proc. Natl Acad. Sci. USA*, **96**, 8477–8482.

Velazquez-Muriel,J.A. *et al.* (2005) SPI-EM: towards a tool for predicting CATH superfamilies in 3D-EM maps. *J. Mol. Biol.*, **345**, 759–771.

Velazquez-Muriel,J.A. *et al.* (2006) Flexible fitting in 3D-EM guided by the structural variability of protein superfamilies. *Structure*, **14**, 1115–1126.

Volkmann,N. and Hanein,D. (2003) Docking of atomic models into reconstructions from electron microscopy. *Meth. Enzymol.*, **374**, 204–225.

Wriggers,W. and Chacon,P. (2001) Modeling tricks and fitting techniques for multiresolution structures. *Structure (Camb)*, **9**, 779–788.

Wriggers,W. *et al.* (1999) Situs: a package for docking crystal structures into low-resolution maps from electron microscopy. *J. Struct. Biol.*, **125**, 185–195.