Protein–protein docking by fast generalized Fourier transforms on 5D rotational manifolds

Dzmitry Padhorny, Andrey Kazennov, Brandon Zerbe, Kathryn Porter, Bing Xia, Scott Mottarella, Yaroslav Kholodov, David Ritchie, Sandor Vajda, Dima Kozakov

Protein–protein docking by fast generalized Fourier transforms on 5D rotational manifolds

Dzmitry Padhorny[a,b], Andrey Kazennov[b], Brandon S. Zerbe[c], Kathryn A. Porter[c], Bing Xia[c], Scott E. Mottarella[c], Yaroslav Kholodov[b,d,e], David W. Ritchie[f], Sandor Vajda[c], and Dima Kozakov[a,g,h,1]

[a]Department of Applied Mathematics and Statistics, Stony Brook University, Stony Brook, NY 11794; [b]Moscow Institute of Physics and Technology, Moscow Region 141700, Russia; [c]Department of Biomedical Engineering, Boston University, Boston, MA 02215; [d]Innopolis University, Innopolis 420500, Russia; [e]Institute of Computer Aided Design of the Russian Academy of Sciences, Moscow 123056, Russia; [f]Inria Nancy, 54600 Villers-les-Nancy, France; [g]Laufer Center for Physical and Quantitative Biology, Stony Brook University, Stony Brook, NY 11794; and [h]Institute for Advanced Computational Sciences, Stony Brook University, Stony Brook, NY 11794

Edited by Michael Levitt, Stanford University School of Medicine, Stanford, CA, and approved May 13, 2016 (received for review March 8, 2016)

Energy evaluation using fast Fourier transforms (FFTs) enables sampling billions of putative complex structures and hence revolutionized rigid protein–protein docking. However, in current methods, efficient acceleration is achieved only in either the translational or the rotational subspace. Developing an efficient and accurate docking method that expands FFT-based sampling to five rotational coordinates is an extensively studied but still unsolved problem. The algorithm presented here retains the accuracy of earlier methods but yields at least 10-fold speedup. The improvement is due to two innovations. First, the search space is treated as the product manifold $SO(3)\times(SO(3)\backslash S^1)$, where SO(3) is the rotation group representing the space of the rotating ligand, and $(SO(3)\backslash S^1)$ is the space spanned by the two Euler angles that define the orientation of the vector from the center of the fixed receptor toward the center of the ligand. This representation enables the use of efficient FFT methods developed for SO(3). Second, we select the centers of highly populated clusters of docked structures, rather than the lowest energy conformations, as predictions of the complex, and hence there is no need for very high accuracy in energy evaluation. Therefore, it is sufficient to use a limited number of spherical basis functions in the Fourier space, which increases the efficiency of sampling while retaining the accuracy of docking results. A major advantage of the method is that, in contrast to classical approaches, increasing the number of correlation function terms is computationally inexpensive, which enables using complex energy functions for scoring.

protein docking | manifold | FFT

Determining putative protein–protein interactions using genome-wide proteomics studies is a major step toward elucidating the molecular basis of cellular functions. Understanding the atomic details of these interactions, however, requires further biochemical and structural information. Although the most complete structural characterization is provided by X-ray crystallography, solving the structures of protein–protein complexes is frequently very difficult. Thus, it is desirable to develop computational docking methods that, starting from the coordinates of two unbound component molecules defined as receptor and ligand, respectively, are capable of providing a model of acceptable accuracy for the bound receptor–ligand complex (1–4). In view of the large number of putative protein–protein interactions, the computational efficiency of docking is also a concern.

Most global docking methods start with rigid body search that assumes only moderate conformational change upon the association, accounted for by using a smooth scoring function that allows for some level of steric overlaps (3). Rigid docking was revolutionized by the fast Fourier transform (FFT) correlation approach, introduced in 1992 by Katchalski-Katzir et al. (5). The major requirement of the method is to express the interaction energy in each receptor–ligand orientation as a sum of P correlation functions, i.e., in the form

$$E(\alpha,\beta,\gamma,\lambda,\mu,\nu)=\sum_{p=1}^{P}\int \overline{R_p(x,y,z)}\hat{T}(\lambda,\mu,\nu)\hat{D}(\alpha,\beta,\gamma)L_p(x,y,z)dV, \quad [1]$$

where R_p and L_p are defined on the receptor and ligand, respectively, $\hat{T}$ and $\hat{D}$ denote translational and rotational operators, and α,β,γ and λ,μ,ν are the rotational and translational coordinates. To illustrate how such functions can be used for docking, consider the very simple case with $P=1$, $R_p=-1$ on a surface layer and $R_p=1$ on the core of the receptor, $L_p=1$ on the entire ligand, and $R_p=L_p=0$ everywhere else. It is clear that this scoring function, which is essentially the one used by Katchalski-Katzir et al. (5), reaches its minimum on a conformation in which the ligand maximally overlaps with the surface layer of the receptor, thus providing optimal shape complementarity. In later FFT-based methods, the scoring function has been expanded to include electrostatic and solvation terms (6, 7) and, more recently, structure-based interaction potentials (8, 9), substantially improving the accuracy of docked structures. As mentioned, in all scoring functions, the shape complementarity term allows for some overlaps, thereby accounting for the differences between bound and unbound (separately crystallized) structures.

Most FFT-based methods (6–8, 10–12) define R_p and L_p on grids, and use a 3D Cartesian FFT approach to accelerate the sampling of the translational space. The method is based on the idea that the energy function, given by Eq. **1**, can be expressed in

Significance

Expressing the interaction energy as sum of correlation functions, fast Fourier transform (FFT) based methods speed the calculation, enabling the sampling of billions of putative protein–protein complex conformations. However, such acceleration is currently achieved only on a 3D subspace of the full 6D rotational/translational space, and the remaining dimensions must be sampled using conventional slow calculations. Here we present an algorithm that employs FFT-based sampling on the 5D rotational space, and only the 1D translations are sampled conventionally. The accuracy of the results is the same as those of earlier methods, but the calculation is an order of magnitude faster. Also, it is inexpensive computationally to add more correlation function terms to the scoring function compared with classical approaches.

Author contributions: D.P., A.K., Y.K., D.W.R., S.V., and D.K. designed research; D.P., A.K., B.S.Z., K.A.P., B.X., S.E.M., and D.K. performed research; D.P. and A.K. analyzed data; and D.P., A.K., D.W.R., S.V., and D.K. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence should be addressed. Email: midas@laufercenter.org.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1603929113/-/DCSupplemental.

terms of the Fourier transforms r_p of R_p and l_p of L_p. Because the translational operator applied to l_p in the Fourier space is given by

$$T(\lambda,\mu,\nu)l_p(n,m,l) = e^{-2\pi i/N(n\lambda+l\mu+m\nu)}l_p(x,y,z), \quad [2]$$

where $i=\sqrt{-1}$, accounting for the orthonormality of Fourier basis functions and interchanging the order of integration and summation yield

$$E(\alpha,\beta,\gamma,\lambda,\mu,\nu) = \sum_{p=1}^{P}\sum_{nlm}\overline{r_p(n,l,m)}l_p(\alpha,\beta,\gamma,n,l,m)e^{-\frac{2\pi i}{N}(n\lambda+l\mu+m\nu)}, \quad [3]$$

which is the expression for the inverse Fourier transform of the product of the Fourier images $r_p(n,m,l)$ and $l_p(\alpha,\beta,\gamma,n,l,m)$ as stated by the convolution theorem. Thus, for a given rotation, E can be calculated over the entire translational space using P forward and one inverse FFT. If N denotes the size of the grid in each direction, then the efficiency of this approach is $O(N^3 log N^3)$ compared with $O(N^6)$ when energy evaluations are performed directly. Owing to the high numerical efficiency of the FFT-based algorithm, it became computationally feasible, for the first time, to systematically explore the conformational space of protein–protein complexes evaluating the energies for billions of conformations, and thus to dock proteins without any a priori information on the expected structure of their complex.

Despite the usefulness of the above algorithm, using FFTs only in translational space has three major limitations. First, FFTs on a new grid must be computed for each rotational increment of the rotating molecule; thus acceleration applies only to half of the degrees of freedom (Fig. 1). Second, each term in the scoring function requires a separate FFT calculation. Thus, accounting for electrostatics, desolvation, and, particularly, pairwise interactions substantially increases the required computational efforts. Third, experimental techniques such as NMR Nuclear Overhauser effect measurements and chemical cross-linking yield information on approximate distances between interacting residues across the interface, and this information can be used to perform the docking subject to pairwise distance restraints. Unfortunately, each pairwise distance restraint requires a new correlation function term. Because the required computational effort is proportional to P, the number of correlation functions in the energy expression, the increasing complexity reduces the numerical advantage of the FFT approach.

In principle, the above problems can be avoided by applying the transforms first, and then moving the proteins in the Fourier space without the need for recomputing the transforms. However, it is difficult to carry out rotations in the translational Fourier space, and, thus, to perform rotations efficiently, it is natural to use spherical coordinates. This approach was applied to crystallography in the early 1970s by Tony Crowther, who realized that the rotation function can be computed more quickly using the FFT, expressing the Patterson maps as spherical harmonics (13). A few groups also used this idea for the development of docking algorithms (14, 15). Most notable is the Hex method of Ritchie and Kemp (14), which represents protein shapes using Fourier series expansions of spherical harmonic and Gauss–Laguerre polynomials. This representation allows rotational searches to be accelerated by angular FFTs, and it enables translations to be calculated analytically in the Fourier basis (15). A similar approach has been developed by Chacon's group (16, 17), in which translations are calculated numerically. However, both approaches were found to have lower accuracy than traditional Cartesian FFT sampling (15). This may be attributed to three main factors. Firstly, the energy functions used were less detailed than in some of the Cartesian approaches. In particular, we used only van der Waals and electrostatic terms (15). Secondly, because the computational cost of the polar Fourier translation matrices grows as $O(N^5)$, the polar Fourier representation is limited to using relatively low order expansions, which limits the achievable accuracy. Finally, the manifold structure of the 5D rotational space was not fully considered, and this resulted in a memory-intensive algorithm that mapped less efficiently onto modern multiprocessor computer architectures than simple 1D FFTs (18). Although we showed previously that the polar representation allows an elegant 5D factorization of multiterm potentials (15), previous efforts to exploit this property have, until now, had limited success.

In this paper, we describe a fast manifold Fourier transform (FMFT) algorithm that eliminates the above shortcomings and, on

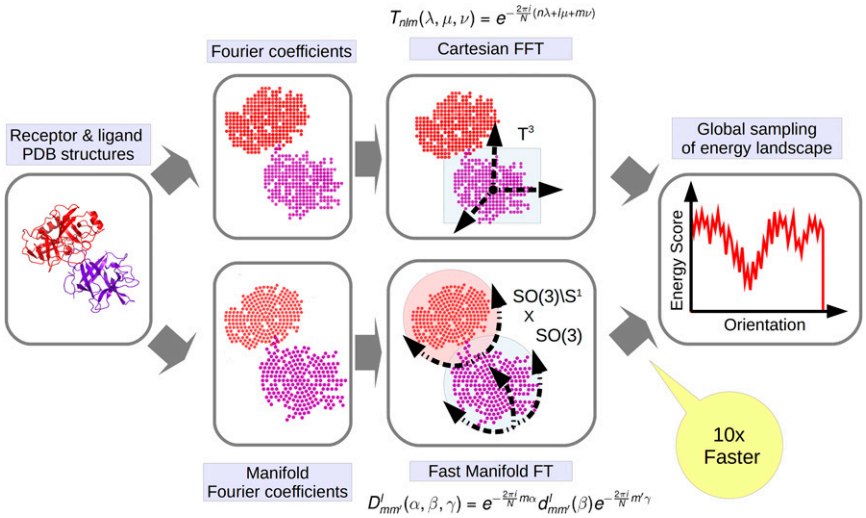

Fig. 1. Schematic representation of FFT-based docking methods. In Cartesian FFT sampling (upper path), the ligand protein is translated along three Cartesian coordinates in Fourier space using the translational operator T. The translation must be repeated for each rotation of the ligand. In 5D FMFT docking (lower path), the direction of the vector from the center of the receptor to the center of the ligand is defined by two Euler angles, and the ligand is rotated around its center, resulting in the search space $(SO(3)\backslash S)\times SO(3)$. All rotations are performed in generalized Fourier space, where D denotes the rotational operator. The only traditional search is the 1D translation along the vector between the centers of the two proteins.

the average, results in a 10-fold decrease in computing time while retaining the accuracy of the traditional Cartesian FFT-based docking. As will be further emphasized, even more important is that, using FMFT, the computational efforts required are essentially independent of the number of correlation function terms in the scoring function, thus enabling the efficient use of more accurate but also more complex energy expressions, as well as accounting for any number of pairwise distance restraints. Developing the method, we took advantage of the generalization of the Cartesian FFT approach to the rotational group manifold $SO(3)$ by Kostelec and Rockmore (19). The basis for using this algorithm was recognizing that the 5D rotational search space can be regarded as the product manifold $SO(3)\times(SO(3)\backslash S^1)$, where the rotation group $SO(3)$ represents the space of the rotating ligand and $(SO(3)\backslash S^1)$ is the space spanned by the two Euler angles that define the orientation of the vector from the center of the fixed receptor to the center of the ligand (Fig. 1 and Fig. S1). This is important, because the algorithm by Kostelec and Rockmore (19) can be easily extended to the $SO(3)\times(SO(3)\backslash S^1)$ manifold.

As already mentioned, a general shortcoming of using Fourier decomposition in spherical spaces is the relatively slow convergence of the series of spherical basis functions. Thus, using a large number of terms reduces computational efficiency, whereas truncating the series limits the accuracy of the energy values calculated by the method. Therefore, a key factor explaining the success of our manifold FFT docking method is that we select the centers of highly populated clusters of low-energy docked structures rather than simply low-energy conformations as predictions of the complex. Such clusters occur in low-energy regions around the local minima in the conformational space. The size of each cluster represents the width of the corresponding energy well, and hence provides some information on entropic contributions to the free energy. Model selection based on cluster size has been used in our very successful docking server ClusPro and, in a substantial fraction of docking problems, enabled the identification of the docked structure closest to the native complex (20). We note that a similar clustering step is implemented in the protein structure prediction program Rosetta (21). For a somewhat more formal justification of the cluster-based approach to model selection, we argue that, using FFT, we globally and systematically sample the energy landscape of the interacting protein pair on a grid, and hence we can calculate an approximate partition function of the form $Z=\sum_j exp(-E_j/RT)$, where E_j is the energy of the jth pose, and we sum over all poses. For the kth low-energy cluster, the partition function is given by $Z_k=\sum_j exp(-E_j/RT)$, where the sum is restricted to poses within the cluster. Based on these values, the probability of the kth cluster is given by $P_k=Z_k/Z$. However, because the low-energy structures are selected from a relatively narrow energy range, and the energy values are calculated with considerable error, it is reasonable to assume that these energies do not differ from each other, i.e., $E_j=E$ for all j in the low-energy clusters. This simplification implies that $P_k=exp(-E/RT)\times(N_k/Z)$, and thus the probability P_k is proportional to N_k, where N_k is the number of structures in the kth cluster. Therefore, we select the centers of highly populated clusters of docked structures, rather than low-energy conformations, as predictions of the complex. Although neglecting the energy differences within the low-energy clusters seems to be arbitrary, the success of the ClusPro server demonstrates that the approximation is valid in a large fraction of cases. The significance of model selection based on cluster size rather than energy values is that it does not require very accurate energy evaluation, and hence, in FMFT, it is sufficient to use a limited number of spherical basis functions in the Fourier space, increasing numerical efficiency without noticeable loss of docking accuracy.

The high efficiency of the FMFT algorithm enables solving very demanding docking problems, way beyond what was considered feasible in the past. After demonstrating that the accuracy of FMFT is comparable to that of the traditional Cartesian FFT-based docking, we present here a few applications that require a large number of docking calculations. Such problems include docking ensembles of models obtained by NMR or homology modeling, and exploring a large number of putative peptide conformations in peptide–protein docking. As will be described, an additional and very favorable property of the FMFT algorithm is that the required computational efforts are almost completely independent of the number P of the correlation function terms in the energy expression given by Eq. **1**, and hence the method can be efficiently used with scoring functions of arbitrary complexity. In contrast, in the traditional FFT approach, the efforts are proportional to P, and hence it is difficult to perform docking subject to pairwise distance restraints, as each restraint gives rise to an additional term in the scoring function. Using FMFT, we demonstrate that this problem can be solved effectively without significant increase in running times (Fig. S2).

Results and Discussion

FFT-Based Docking on 5D Rotational Manifolds. Here we demonstrate that, by taking advantage of the special geometry of the space characterizing molecular movement upon protein–protein association, it is possible to construct an extremely efficient FFT-based docking algorithm. We present the basic idea of this algorithm as the generalization of the translational FFT method described in the Introduction. Because we plan to work in the rotational space, we change the Cartesian coordinates to polar coordinates $(x,y,z)\rightarrow(r,\theta,\phi)$, and consider the generalization of the Fourier transform on the sphere

$$R(r,\theta,\phi)=\sum_{nlm}^{N} r(n,l,m)R_{nl}(r)d_{lm}(cos\theta)e^{-im\phi} \qquad [4]$$

where $R_{nl}(r)$ are radial basis functions, $r(n,l,m)$ are generalized Fourier coefficients, $d_{lm}(cos\theta)$ are Legendre polynomials (22), and N is the order of expansion used. Eq. **4** looks like a Fourier transform, but $e^{-im\phi}$ is replaced by $d_{lm}(cos\theta)e^{-im\phi}$, which shows the non-Cartesian properties of the sphere (23).

Consider again the derivation of the convolution theorem (Eq. **1**) but, this time, on the manifold $(SO(3)\backslash S)\times SO(3)$ shown in the lower path of Fig. 1. The translation of the ligand can be represented as the rotation of the receptor, followed by the translation of the ligand along the z axis,

$$E(z,\beta,\gamma,\alpha',\beta',\gamma')=\sum_{p=1}^{P}\int \hat{T}(-z)\hat{D}(0,\beta,\gamma)R_p(\rho,\theta,\phi) \times\hat{D}(\alpha',\beta',\gamma')L_p(\rho,\theta,\phi)dV. \qquad [5]$$

Rotations of the receptor can be expressed as follows:

$$D(\alpha,\beta,\gamma)R(r,\theta,\phi)=\sum_{nlm}R_{nl}(r)Y_{lm}(\theta,\phi)\sum_{m_1}D^{l}_{mm_1}(\alpha,\beta,\gamma)r(n,l,m_1), \qquad [6]$$

where $Y_{lm}(\theta,\phi)$ denotes spherical harmonics, and

$$D^{l}_{mm'}(\alpha,\beta,\gamma)=e^{-im\alpha}d^{l}_{mm'}(\beta)e^{-im'\gamma} \qquad [7]$$

are Wigner rotation matrices with $d^{l}_{mm1}(\beta)$ denoting Wigner d functions, related to Jacobi polynomials (19). Eqs. **6** and **7** show that the rotational operator in the rotational group $SO(3)$ acts on generalized Fourier coefficients the same way as the translation operator acts on Fourier coefficients in the Cartesian space (Eq. **2**), apart from the asymmetry of the middle angle β, which requires special treatment. Describing the translation of the ligand along

the z axis in the Fourier space is far from simple, and requires updating a set of coefficients. However, it is only 1 degree of freedom (as opposed to 3 degrees in the Cartesian space), and hence it can be accomplished relatively efficiently (24). Now we apply the translation operator and the rotation operator (Eq. **7**) to the integral in Eq. **5**. Based on the orthonormality of the generalized Fourier basis functions, interchanging the order of integration and summation yields

$$E(z,\beta,-\gamma,\alpha',\beta',\gamma') = \sum_{\substack{ll_1\\ mm_1m_2}} \left(\sum_{nn_1} \sum_{p} \overline{r_p(n_1,l_1,m_1)} l_p(n,l,m_2) T^{|m|}_{nln_1l_1}(z) \right) \times d^{l}_{mm_1}(\beta) d^{l}_{mm_2}(\beta') e^{-i(m\alpha'+m_1\gamma+m_2\gamma')}. \quad [8]$$

Note that Eq. **8** is similar to Eq. **3** in Cartesian coordinates, with the difference that, instead of a 3D inverse Fourier transform, we have a generalized FMFT, which involves the Wigner d functions $d^{l}_{mm1}(\beta)$. However, the really important difference is in the order of the transforms and the summation of correlation functions. In Eq. **3**, for each rotation of the ligand, we have to calculate the Fourier transforms $l_p(\alpha,\beta,\gamma,n,l,m)$ for each of the P components of the ligand energy function separately, form the product with the transform $r_p(n,m,l)$ of the pth component of the receptor energy function, sum all terms, and take the inverse transform. In contrast, according to Eq. **8**, we calculate the sum of initial precalculated generalized Fourier coefficients in the internal loop only once, and perform all rotations in Fourier space rather than calculating an FFT for each rotation. This allows us to calculate multiple energy terms using a single FMFT for each translation. Thus, as already emphasized, the computational efforts are essentially independent of the number P of the correlation function terms in the energy expression. Because inverse manifold Fourier transforms can be efficiently calculated by methods due to ref. 19, this approach provides substantial computational advantage, particularly if P is high.

Execution Times. Execution times of the FMFT sampling algorithm were measured by docking unbound structures of component proteins in 51 enzyme–inhibitor pairs from the established Protein Docking Benchmark (25) (Table S1). The times were compared with those required for docking the same proteins using PIPER, a protein docking program based on the Cartesian FFT approach (8). The FFTW (Fastest Fourier Transform in the West) library (26) was used for FFT calculations. All runs were performed using the standard PIPER scoring function, consisting of eight correlation function terms. Execution times were measured on one or several Intel Xeon E5-2680 processors. Using the FMFT algorithm, the average execution time was 15.39 min. In comparison, the average execution time for the same set of proteins using PIPER was 232.15 min, indicating that FMFT speeds up the calculations ~15-fold. Using parallel versions of the algorithms on 16 CPU cores, the average execution times measured were 2.67 min and 20.19 min for FMFT and PIPER, respectively, which shows about a 7.5-fold speedup.

Application 1: Constructing Enzyme–Inhibitor Complexes. The quality of FMFT and PIPER results was determined by docking the same 51 enzyme–inhibitor pairs that we have used for comparing execution times (Table S2). In both cases, the scoring function was the same one normally used in PIPER for docking enzyme–inhibitor pairs, and it consisted of attractive and repulsive van der Waals, Coulombic electrostatics, generalized Born, and knowledge-based Decoys As the Reference State terms, the latter representing nonpolar solvation (27). The docking procedure for these cases was the one normally used by PIPER (20). First, the conformational space was sampled using either the FMFT or the PIPER protocol. After docking, the 1,000 lowest energy poses were retained and clustered using interface C_α rms deviation (RMSD) as the distance metrics with a fixed 9 Å clustering radius. The clusters were ranked according to cluster populations (i.e., number of poses in the cluster), and the centers of up to 30 largest clusters were reported as putative models of the complex (Table S2).

Fig. 2 *A–C* shows the results of docking. The number of hits shown in Fig. 2*A* is the number of near-native poses, defined as having less than 10 Å C_α interface RMSD (IRMSD) from the native complex, generated by each of the two algorithms. Note that IRMSD is calculated for the backbone atoms of the ligand that are within 10 Å of any receptor atom after superimposing the receptors in the X-ray and docked complex structures. We found that the number of poses with less than 10 Å IRMSD is a good measure of the quality of sampling of the energy landscape in the vicinity of the native structure. Fig. 2 *B* and *C* shows the properties of models obtained by clustering low-energy poses using pairwise IRMSD as a distance metric. A large number of low-energy poses typically yields a well-populated and thus highly ranked near-native cluster, reported as one of the final models. Based on all these results, FMFT and PIPER show comparable docking performance, both in terms of the number of near-native structures (Fig. 2*A*), the ranks of the clusters that define the final near-native models (Fig. 2*B*), and the IRMSD (Fig. 2*C*) of these models.

Application 2: Docking Interacting Protein Domains. We further compared FMFT and PIPER by docking interacting domains extracted from proteins that are defined as "Other" type in the Protein Docking Benchmark (25) (Tables S3 and S4). This problem is generally more challenging than docking inhibitors to enzymes because the Other category includes complexes with highly variable properties. Restricting consideration to individual domains eliminates the additional problem that the domains in multidomain proteins may shift relative to each other, affecting the docking results. Thirty cases representing domain–domain binding were selected from the Others section of the Protein Docking Benchmark (Table S4). Nineteen cases from this set represent binding of single-domain proteins (or single domains taken from larger proteins), and thus full protein structures were used for docking. In another 11 cases, receptor and/or ligand are composed of several domains, so reduced representations of protein structures were prepared: Only the binding domains were retained for docking, and the rest of the structure was cleaved. Residue ranges for binding domains were assigned according to structural classification of proteins (SCOP) domain classification (28). To prevent possible association at intra-protein domain–domain binding interfaces exposed by the cleavage, additional repulsion grids were used in the docking procedure. These were constructed by taking the backbone atoms of the original structure lying within 10 Å (but not closer than 5 Å) of the binding domain and placing repulsive spheres with 0.5-Å radius at the positions of those atoms. The 5-Å lower bound to the distance range specifying the thickness of this "repulsive padding" was introduced to ensure that additional repulsion doesn't affect binding to the relevant portion of protein surface. During the docking process, such repulsive padding grid was correlated with the standard repulsive van der Waals grid of the binding partner. The docking procedure overall was the same as that used for enzyme–inhibitor targets, except that 1,500 low-energy poses were used for clustering, generated from three docking runs (500 poses from each) performed with differently weighted components of the scoring function (20). Similarly to the results obtained for enzyme–inhibitor complexes, FMFT and PIPER show comparable performance (Fig. 2 *D–F* and Table S3). Although PIPER generates large numbers (>200) of near-native structures for more complexes than FMFT, the number of complexes with very few (<10) such near-native structures is substantially smaller using FMFT than using PIPER. Thus, FMFT shows better performance for the more difficult-to-dock complexes (Fig. 2*D*). In addition, using PIPER, the number of models that are not

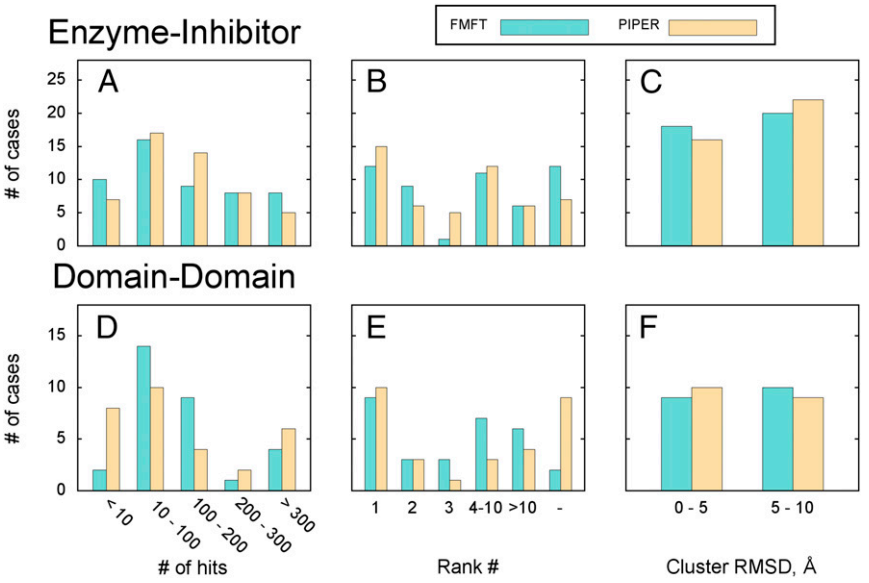

Fig. 2. Results of docking enzyme–inhibitor and domain-domain pairs. Bar heights represent the number of docking cases that fall into an appropriate category. (*A*) The number of hits among the 1,000 low-energy poses generated for enzyme–inhibitor complexes. (*B*) Ranking of final near-native models for enzyme–inhibitor complexes. (*C*) C_α IRMSD of the final model for enzyme–inhibitor complexes (here only cases with both FMFT and PIPER producing a near-native model were taken into account). (*D*) The number of hits among the 1,500 low-energy poses generated for domain–domain complexes. (*E*) Ranking of final near-native models for domain–domain complexes. (*F*) C_α IRMSD of the final model for domain–domain complexes. As in *C*, only cases with both FMFT and PIPER producing a near-native model were taken into account.

ranked in the top 10 is much higher than using FMFT (Fig. 2*E*). Based on these results, FMFT performs as well as PIPER.

Application 3: Accounting for Pairwise Distance Restraints. An important consideration for selecting a docking method is the maximum complexity of the scoring function that still allows for solving problems with reasonable execution times. As mentioned, all FFT-based approaches require the use of scoring functions that can be written as sums of correlation functions. This is not a major limitation, because such functions may include many commonly used physics-based energy terms, such as steric repulsion, van der Waals interaction, and Coulombic electrostatics. It has also been shown that some energy terms that are not inherently correlation-based, such as the widely used pairwise interaction potentials, can be efficiently approximated by a sum of several correlation functions (27). Altogether, this makes the number of correlations a crucial parameter, because this number effectively defines the complexity of the scoring function in the particular sampling run.

One important task, especially demanding in terms of scoring function complexity, is incorporating pairwise distance restraints, based on known interactions between residue pairs, into the docking procedure. Such restraints can be derived in a variety of experiments, including NMR, cross-linking, and mutagenesis assays (29). The restraints can be implemented as short-distance attractive terms in the scoring function, but each will add a correlation function term. As emphasized, in Cartesian FFT, the number of transforms required is proportional to the number P of correlation functions (Eq. **3**), whereas, in FMFT, the number of transforms is independent of P. To demonstrate this difference, we determined the structure of the glucose-specific enzyme IIA (E2A)-histidine-containing phosphocarrier protein (HPr) complex [Protein Data Bank (PDB) entry 1GGR] (30) from the structures of its constituents in their unbound form (PDB entries 1F3G and 1POH) and 20 ambiguous interaction restraints (AIRs) based on NMR titration data (29). The docking procedure was the one used for docking enzyme–inhibitor pairs, but with 20 additional correlations terms in the scoring function due to the restraints (29). Each restraint is specified as a residue in one of the proteins, and a set of residues on the partner protein that are in contact with the first residue, where "contact" means ≤3 Å distance between any two atoms of the residue pair. To represent these restraints, receptor and ligand correlation components were constructed by placing 3-Å radius attractive spheres on the atoms of the particular residue on the first protein and the attracted point "charges" on the atoms of the interacting residues on the partner protein. Docking was performed using both FMFT and PIPER. Incorporation of restraints increased the population of the near-native cluster from 201 to 410, which became the most populated cluster and thus provided the putative model of the complex (Fig. 3 and Table S5) without any significant change in the IRMSD of the cluster center (5.25 Å for the unrestrained case versus 5.15 Å for the restrained). Adding the restraints increased the number of correlation function terms in the scoring function from 8 to 28. For PIPER, this resulted in a proportional increase in execution time (from 96.15 min to 373.80 min). In contrast, running FMFT, the execution time barely changed, from 12.32 min to 15.30 min. This result demonstrates that FMFT can be used with very complex scoring functions (Fig. S2).

Application 4: Docking Ensembles of NMR Models. Multiple docking runs may be required when one or both component proteins are given as ensembles of structures, obtained by NMR experiments or by extracting snapshots from molecular dynamics simulations. Because accounting for multiple structures may substantially improve docking results, the high efficiency of the FMFT method is particularly useful. As an example, we considered calculating the complex formed by the *Escherichia coli* Colicin E9 DNase domain and its cognate immunity protein IM9. Four different X-ray structures of the unbound E9 DNase domain (chains B, C, D, and E of PDB entry 1FSJ) were docked in a pairwise manner to 20 NMR models of the IM9 protein (PDB entry 1E0H), thus performing 80 docking calculations. Unstructured termini of the receptor were masked and didn't contribute to the calculated energy scores. The 50 lowest energy poses were extracted from each of the 80 docking runs and merged, yielding a total of 4,000 poses that were then clustered as usual in PIPER. Fig. 4 *A*–*C* shows the docking results. In short, merging the 50 lowest energy poses from each docking run,

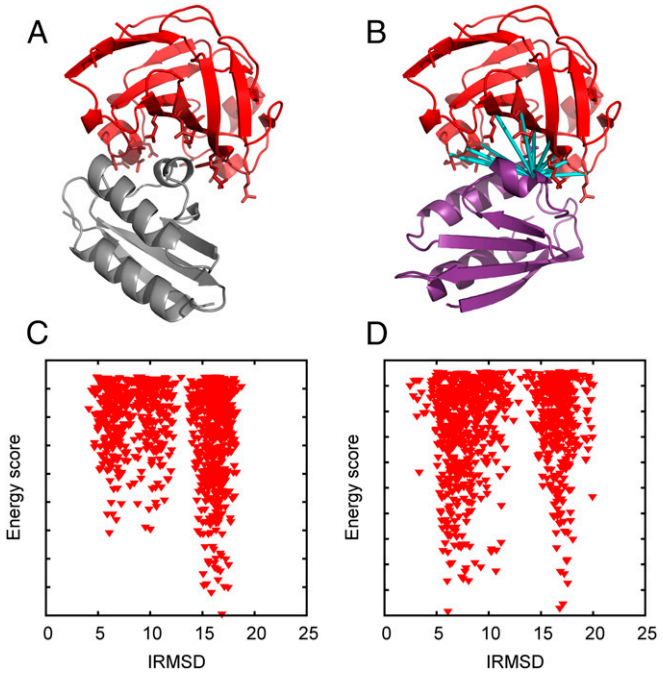

Fig. 3. Docking of E2A and HPr proteins. (*A*) Model defined by the most populated cluster obtained without restraints. (*B*) Model defined by the most populated cluster obtained with restraints. A set of cyan cylinders represents one of the 20 restraints. (*C*) IRMSD versus energy score for docking without restraints. (*D*) IRMSD versus energy score for docking with restraints. Incorporation of experimental restraints substantially increased the population of the near-native cluster.

followed by clustering, provided a 2.94-Å IRMSD model of the complex ranked fifth, where the 1IBX structure of the native complex was used for comparison when evaluating the accuracy of results. To emphasize the advantage of ensemble docking, we also docked a single pair of structures, chain B of 1FSJ and the first NMR model of the ligand from the PDB file 1E0H. The standard docking protocol was used, and we retained the 1,000 lowest energy poses for clustering. Docking the single pair, the best near-native model obtained was ranked 13, and had the IRMSD value of 3.45 Å. Thus, in the case of structural uncertainty of the component proteins, ensemble docking can substantially improve the results, and, in this type of application, the higher speed of FMFT is a major advantage. Computational efficiency will be particularly important for genome-wide analyses of protein–protein interactions, but we think that, for such applications, it will be necessary to better understand the docking of homology models (see *Application 5: Identification of Binding Sites by Docking Homology Models*), because, generally, structures are not available for a substantial fraction of proteins.

Application 5: Identification of Binding Sites by Docking Homology Models. It has been shown that protein–protein interaction sites can be found by determining the highly populated interfaces in the ensemble of structures generated by global docking (31, 32). We implemented this approach by clustering the "interfacial" atoms in the low-energy docked poses. Although this method usually requires structures of the component proteins, we extended the approach to proteins with yet undetermined structures by docking multiple homology models. The extended method was applied to determining the interface in the Nef–Fyn(R96I)SH3 complex (PDB entry 1EFN). Ten models of the receptor (SH3 domain) and 2 models of the ligand (HIV-1 Nef protein) were constructed using the MODELLER program (33) and based on homologous templates with 30–60% sequence identity (see Table S6 for the list of templates used). All possible receptor–ligand model pairs were docked using the approach developed for Other type of complexes (20). From each of the 20 docking runs, we selected the 1,500/20 = 75 lowest energy poses that were merged and clustered using RMSD as the distance metrics. The structures at the centers of these clusters were used to define interface atoms as atoms located within 5 Å of any atom of the partner protein. These interfacial atoms were then subjected to bottom-up hierarchial clustering using the Euclidian distance as the metrics. Clustering was terminated, i.e., neighboring clusters were not merged, if the minimal distance between a pair of their atoms was larger than the value of a separation parameter. The resulting clusters were ranked according to cluster population (i.e., the number of atoms in each cluster), and the largest cluster was considered to be the most probable prediction of the protein–protein interaction site. For comparison, we also predicted the interaction site by docking a single pair of homology models based on the templates with the highest sequence identity. In this case, a slightly larger value of the clustering separation parameter was used (1.35 Å rather than 1.30 Å). This change was due to the fact that a single docking run provided fewer interfacial atoms for hierarchical clustering, resulting in clusters that were too small. Therefore, the value of the cutoff parameter was increased to ensure that the relative population of the largest cluster was comparable to that obtained by merging the results from 20 docking runs. As shown in Fig. 4 *D* and *E*, docking of multiple homology models of the component proteins increased the accuracy of binding site prediction, compared with the result of using the maximum sequence identity models alone.

Application 6: Docking Flexible Peptides. The difficulty in docking short linear peptides is that their structure in solution is generally unknown and may be ill-defined. One possible solution is to dock a variety of peptide conformations, thus requiring multiple docking runs. We have recently developed an algorithm based on the use of structural templates extracted from the PDB with sequences that matched the known sequence motif in the peptide. These templates were docked individually using the FMFT algorithm. From each run, a number of low-energy poses were retained, the pooled peptide structures were clustered, and the highly populated cluster centers were reported as final models as in all applications of our docking algorithm.

Here we demonstrate this algorithm by docking the ace-PQQATDD peptide to the tumor necrosis factor receptor-associated factor 2 (TRAF2). For this peptide, the PXQ motif sequence known from the literature was extended to length 7 (PXQXXDD) and used to extract 316 structural templates from the PDB database. These templates were then used to model the target peptide. The models were aligned and clustered using the backbone RMSD as the distance measure, with 0.5 Å as the fixed clustering radius. Peptide structures corresponding to the centers of the 25 most-populated clusters were docked to the unbound receptor structure (chain A of PDB entry 1CA4).

The 250 lowest energy poses were retained from each docking run. The poses were merged and clustered using backbone RMSD as a distance measure with 3.5 Å as the fixed clustering radius. Cluster centers were ranked according to cluster populations and reported as final models. Docking results were evaluated using the backbone RMSD from the structure of the peptide in the native complex (chain A of PDB entry 1CZY). A near-native model of the protein–peptide complex was ranked fourth and had the backbone RMSD of 3.3 Å from the conformation in the X-ray structure (Table S7 and Fig. 5*A*). Note that docking only the most frequently occurring structural template provides less accurate models, as demonstrated in Fig. 5 *B* and *C*.

Conclusions

Extending the classical 3D Cartesian implementation of the FFT correlation approach to perform rotations in Fourier space without

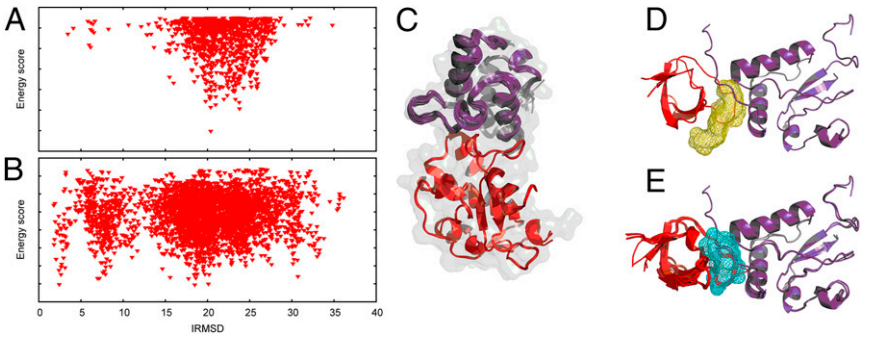

Fig. 4. Docking of structural ensembles. (*A*) Sampling the interaction energy landscape using a single E9 DNase domain structure and the first NMR model of IM9. The docking does not capture any near-native energy minimum. (*B*) Consensus energy values from the 80 pairwise dockings of four different X-ray structures of the E9 DNase domain to 20 NMR models of the IM9 protein. (*C*) Cartoon representation of the four E9 DNase domain and 20 IM9 structures used for docking, superimposed on the structure of the native complex (gray shade). (*D*) Binding site identification for the Nef–Fyn(R96I)SH3 complex obtained by docking the highest sequence identity models alone. (*E*) Using multiple homology models of the receptor and the ligand to identify the binding site for the Nef–Fyn(R96I) SH3 complex results in a more specific prediction.

the need for recalculating the transforms has been a long-outstanding and extensively studied problem. The main difficulty in developing such methods is that, to achieve numerical efficiency, one can use only a moderate number of spherical basis functions to span the search space, and this may reduce the accuracy of energy evaluation. However, because we base model selection on the population of low-energy clusters rather than on energy values, minor deviations in energy generally do not affect the accuracy of final models. Here we present an elegant manifold FFT implementation of 5D search that is more than 10-fold faster than the traditional 3D approach. A major advantage of the method is that adding correlation function terms in the scoring function is computationally inexpensive, and hence the method works efficiently with very complex energy evaluation models, possibly including pairwise distance restraints that are difficult to deal with in traditional FFT-based docking. The improved efficiency implies that we can solve new classes of docking problems, including the docking of large ensembles of proteins rather than just a single protein pair, docking homology models, and flexible peptides that may have a large number of potential conformations. We note that the beta version of a code implementing the FMFT algorithm can be downloaded from https://bitbucket.org/abcgroup_midas/fmft_dock/, thus providing an opportunity for testing and using the method. In addition, we are in the process of adding FMFT as a new option to the server.

Materials and Methods

This section summarizes the implementation of the FMFT approach. For the mathematical details of the algorithm, see *SI Materials and Methods*.

The procedure starts with receptor- and ligand-associated components of each correlation term of the energy function being represented as sets of coefficients $r(n,l,m)$, $l(n,l,m)$ that appear in the expansion shown as Eq. **4**. Here $1 \le n \le N$, $0 \le l \le n-1$, and $-l \le m \le +l$, where N governs the order at which the series is truncated. These coefficients, together with the translation range to be sampled (i.e., minimal and maximal distances between protein centers, calculated from the geometrical properties of the proteins), are submitted as input parameters to the program performing the FMFT-based sampling. To improve efficiency, two stages of FMFT sampling are being executed: The first one, performed with a maximal coefficient order $N=20$ on a small FFT grid, is computationally inexpensive and provides a crude approximation of the energy landscape, which is then used to focus the search to the translation range potentially containing the energy minima, whereas the second one is executed with $N=30$ on a full-sized FFT grid but performs the sampling only in the refined translation range, thus saving computational resources.

The actual sampling stage can be described as follows: After loading the input parameters, the program starts to iterate the allowed translation range in steps of 1 Å. For each translation step, the $\sum_{nn_1}\sum_{p} r_p(n_1,l_1,m_1) l_p(n,l,m_2) T^{|m|}_{nln_1l_1}(z)$ product of coefficients and translation matrix elements is calculated, followed by a manifold FFT, which provides the values of energy score for all receptor–ligand orientations corresponding to a fixed distance between the centers of the two proteins. The resulting samples are located on the $(\beta,\gamma,\alpha',\beta',\gamma')$ Euler angle grid with dimensions of $30\times59\times59\times30\times59$ (or $16\times30\times30\times16\times30$ for the low-order scan). K (on the order of 1,000 for a typical sampling run) lowest energy

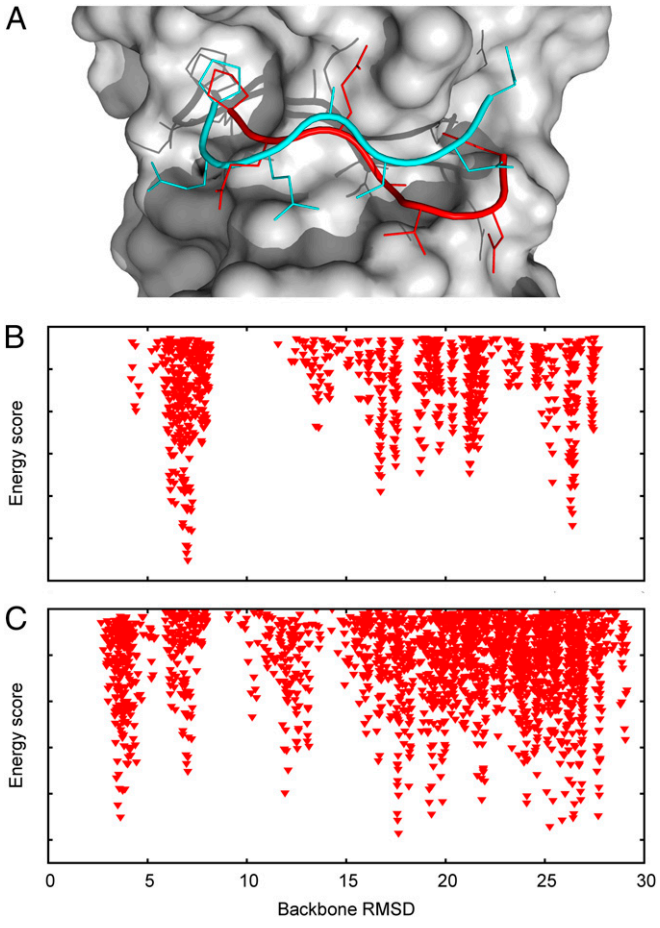

Fig. 5. Docking of the ace-PQQATDD peptide to TRAF2. (*A*) Bound structure of the peptide (red) and the 3.3-Å model, ranked fourth (cyan). (*B*) Peptide backbone RMSD versus scoring function when docking the most common structural template alone. (*C*) Peptide backbone RMSD versus scoring function when using all 25 templates. Docking the ensemble substantially improves the results, and yields samples with less than 4.0-Å backbone RMSD.

samples are retained for each translation step. After the entire translation range is processed, the low-energy samples from individual translation steps are merged and resorted by energy value to select the K lowest energy samples that are presented as the final results.

It is important to note here that the sampling of the $S^2 \times SO(3)$ manifold [in practice probed as $(SO(3)\backslash S^1) \times SO(3)$], provided by the equispaced sampling of Euler angles, is inherently nonuniform. This becomes a significant problem if one seeks to obtain statistical information about the energy landscape of protein interaction, for example, to construct the partition function of the system. To battle this nonuniformity, a special procedure is used for the selection of low-energy scores. Specifically, once the 5D array of energy scores for a single translation step is acquired, the program starts selecting lowest-scoring conformations and excluding the samples corresponding to the surrounding region from further consideration. Here the "surrounding region" is defined as the subset of elements $\{(x, y)|x(\beta, \gamma) \subset S^2, y(\beta, \gamma, \alpha', \beta', \gamma') \subset SO(3)\}$ of the $S^2 \times SO(3)$ manifold, for which $(dist_{S^2}(x, x_{min}) < \Delta) \wedge (dist_{SO(3)}(y, y_{min}) < \Delta)$, where Δ is a cutoff parameter chosen to be 6.0°, which is slightly less than the grid step of $360°/59 = 6.1°$. This procedure ensures that the sampling explores a substantial fraction of the conformational space rather than producing structures very close to each other.

ACKNOWLEDGMENTS. This work was supported by National Science Foundation (NSF) Grants AF 1527292 and DBI 1458509, NIH, National Institute of General Medical Sciences, Grants R35 GM118078 and R01 GM093147, Russian Scientific Foundation Grant 14-11-00877, and US Israel Binational Science Foundation Grant 2009418.

1. Ritchie DW (2008) Recent progress and future directions in protein-protein docking. *Curr Protein Pept Sci* 9(1):1–15.
2. Andrusier N, Mashiach E, Nussinov R, Wolfson HJ (2008) Principles of flexible protein-protein docking. *Proteins* 73(2):271–289.
3. Vajda S, Kozakov D (2009) Convergence and combination of methods in protein–protein docking. *Curr Opin Struct Biol* 19(2):164–170.
4. Smith GR, Sternberg MJE (2002) Prediction of protein–protein interactions by docking methods. *Curr Opin Struct Biol* 12(1):38–35.
5. 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(6):2195–2199.
6. Gabb HA, Jackson RM, Sternberg MJ (1997) Modelling protein docking using shape complementarity, electrostatics and biochemical information. *J Mol Biol* 272(1): 106–120.
7. Chen R, Li L, Weng Z (2003) ZDOCK: An initial-stage protein-docking algorithm. *Proteins* 52(1):80–87.
8. Kozakov D, Brenke R, Comeau SR, Vajda S (2006) PIPER: An FFT-based protein docking program with pairwise potentials. *Proteins* 65(2):392–406.
9. Mintseris J, et al. (2007) Integrating statistical pair potentials into protein complex prediction. *Proteins* 69(3):511–520.
10. Mandell JG, et al. (2001) Protein docking using continuum electrostatics and geometric fit. *Protein Eng* 14(2):105–113.
11. Vakser IA (1996) Low-resolution docking: prediction of complexes for underdetermined structures. *Biopolymers* 39(3):455–464.
12. Ravikant DVS, Elber R (2010) PIE—Efficient filters and coarse-grained potentials for unbound protein–protein docking. *Proteins* 78(2):400–419.
13. Crowther R (1972) *The Molecular Replacement Method*, ed Rossmann MG (Gordon and Breach, New York), pp 173–178.
14. Ritchie DW, Kemp GJ (2000) Protein docking using spherical polar Fourier correlations. *Proteins* 39(2):178–194.
15. Ritchie DW, Kozakov D, Vajda S (2008) Accelerating and focusing protein–protein docking correlations using multi-dimensional rotational FFT generating functions. *Bioinformatics* 24(17):1865–1873.
16. Kovacs JA, Chacón P, Cong Y, Metwally E, Wriggers W (2003) Fast rotational matching of rigid bodies by fast Fourier transform acceleration of five degrees of freedom. *Acta Crystallogr D Biol Crystallogr* 59(Pt 8):1371–1376.
17. Garzon JI, et al. (2009) FRODOCK: A new approach for fast rotational protein-protein docking. *Bioinformatics* 25(19):2544–2551.
18. Ritchie DW, Venkatraman V (2010) Ultra-fast FFT protein docking on graphics processors. *Bioinformatics* 26(19):2398–2405.
19. Kostelec PJ, Rockmore DN (2008) FFTs on the rotation group. *J Fourier Anal Appl* 14(2):145–179.
20. Kozakov D, et al. (2013) How good is automated protein docking? *Proteins* 81(12): 2159–2166.
21. Gray JJ, et al. (2003) Protein–protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations. *J Mol Biol* 331(1):281–299.
22. Zare RN (2013) *Angular Momentum: Understanding Spatial Aspects in Chemistry and Physics* (Wiley-Interscience, New York).
23. Driscoll J, Healy D (1994) Computing Fourier transforms and convolutions on the 2-sphere. *Adv Appl Math* 15(2):202–250.
24. Ritchie DW (2005) High-order analytic translation matrix elements for real-space six-dimensional polar Fourier correlations. *J Appl Cryst* 38(5):808–818.
25. Hwang H, Vreven T, Janin J, Weng Z (2010) Protein–protein docking benchmark version 4.0. *Proteins* 78(15):3111–3114.
26. Frigo M, Johnson SG (2005) The design and implementation of FFTW3. *Proc IEEE* 93(2): 216–231.
27. Chuang GY, Kozakov D, Brenke R, Comeau SR, Vajda S (2008) DARS (Decoys As the Reference State) potentials for protein-protein docking. *Biophys J* 95(9):4217–4227.
28. Murzin AG, Brenner SE, Hubbard T, Chothia C (1995) SCOP: A structural classification of proteins database for the investigation of sequences and structures. *J Molec Biol* 247(4):536–540.
29. Dominguez C, Boelens R, Bonvin AMJJ (2003) HADDOCK: A protein–protein docking approach based on biochemical or biophysical information. *J Am Chem Soc* 125(7): 1731–1737.
30. Garrett DS, Seok YJ, Peterkofsky A, Clore GM, Gronenborn AM (1997) Identification by NMR of the binding surface for the histidine-containing phosphocarrier protein HPr on the N-terminal domain of enzyme I of the *Escherichia coli* phosphotransferase system. *Biochemistry* 36(15):4393–4398.
31. Hwang H, Vreven T, Weng Z (2014) Binding interface prediction by combining protein–protein docking results. *Proteins* 82(1):57–66.
32. Fernández-Recio J, Totrov M, Abagyan R (2004) Identification of protein–protein interaction sites from docking energy landscapes. *J Mol Biol* 335(3):843–865.
33. Eswar N, et al. (2006) Comparative protein structure modeling using MODELLER. *Current Protocols in Bioinformatics* (Wiley, New York), pp 5.6.1–5.6.32.
34. Biedenharn L, Louck J (1981) *Angular Momentum in Quantum Physics* (Addison-Wesley, Reading, MA).
35. Potts D, Prestin J, Vollrath A (2009) A fast algorithm for nonequispaced Fourier transforms on the rotation group. *Numer Algorithms* 52(3):355–384.
36. Rabiner L (1979) On the use of symmetry in FFT computation. *IEEE Trans Acoust Speech Sig Proc* 27(3):233–239.