Recent Progress and Future Directions in Protein-Protein Docking

David W. Ritchie*

Department of Computing Science, University of Aberdeen, Aberdeen, AB24 3UE, Scotland, U.K.

Abstract: This article gives an overview of recent progress in protein-protein docking and it identifies several directions for future research. Recent results from the CAPRI blind docking experiments show that docking algorithms are steadily improving in both reliability and accuracy. Current docking algorithms employ a range of efficient search and scoring strategies, including e.g. fast Fourier transform correlations, geometric hashing, and Monte Carlo techniques. These approaches can often produce a relatively small list of up to a few thousand orientations, amongst which a near-native binding mode is often observed. However, despite the use of improved scoring functions which typically include models of desolvation, hydrophobicity, and electrostatics, current algorithms still have difficulty in identifying the correct solution from the list of false positives, or decoys. Nonetheless, significant progress is being made through better use of bioinformatics, biochemical, and biophysical information such as e.g. sequence conservation analysis, protein interaction databases, alanine scanning, and NMR residual dipolar coupling restraints to help identify key binding residues. Promising new approaches to incorporate models of protein flexibility during docking are being developed, including the use of molecular dynamics snapshots, rotameric and off-rotamer searches, internal coordinate mechanics, and principal component analysis based techniques. Some investigators now use explicit solvent models in their docking protocols. Many of these approaches can be computationally intensive, although new silicon chip technologies such as programmable graphics processor units are beginning to offer competitive alternatives to conventional high performance computer systems. As cryo-EM techniques improve apace, docking NMR and X-ray protein structures into low resolution EM density maps is helping to bridge the resolution gap between these complementary techniques. The use of symmetry and fragment assembly constraints are also helping to make possible docking-based predictions of large multimeric protein complexes. In the near future, the closer integration of docking algorithms with protein interface prediction software, structural databases, and sequence analysis techniques should help produce better predictions of protein interaction networks and more accurate structural models of the fundamental molecular interactions within the cell.

Keywords: Protein-protein docking, protein-protein interactions, docking algorithms, data-driven docking, molecular dynamics, protein structure databases, protein interface prediction, CAPRI.

INTRODUCTION

Proteins play a central role in many cellular processes, ranging from enzyme catalysis and inhibition to signal transduction and gene expression. Proteins often perform their functions by interacting with other proteins to form protein-protein complexes. These complexes may exist as short-lived transitory associations, as in e.g. enzyme catalysis, or as long-lived multimeric systems such as the ribosome, transcription factors, cell surface and ion channel proteins. Using yeast two-hybrid (Y2H) and tandem-affinity-purification mass spectrometry (TAP-MS) techniques, large-scale functional genomic studies are producing interaction maps which describe complex networks of protein-protein interactions (PPIs) within a cell [1, 2]. High throughput Y2H and TAP-MS experiments have been applied on a genomic scale to yeast [3-7]. Bioinformatics approaches such as threading, phylogenetic profiling, gene neighbourhood and gene fusion analysis, and *in silico* two-hybrid methods are being used with increasing success to predict PPIs directly from gene sequences of yeast and other organisms (for recent reviews see e.g., references [8-13]).

*Address correspondence to this author at the Department of Computing Science, University of Aberdeen, Aberdeen, AB24 3UE, Scotland, UK; Tel: (+44)(–0)1224 272282; Fax: (+44)(–0)1224 273422;
E-mail: dritchie@csd.abdn.ac.uk

Protein docking is the task of calculating the three-dimensional (3D) structure of a protein complex starting from the individual structures of the constituent proteins. In other words, in contrast to the above approaches which determine or predict *which* proteins interact, protein docking aims to predict *how* proteins interact. Based on analyses of known protein structures, it has been estimated that the natural repertoir of protein folds may be of the order of 1,000 [14]. By applying similar reasoning to known yeast interactions, Aloy and Russell [15] estimate that each protein will have around 9 interaction partners and that most protein interactions will belong to one of around 10,000 basic types, of which we currently know only around 2,000. Therefore, there are potentially many thousands of as yet completely unknown PPIs. Crystallographic (X-ray) and nuclear magnetic resonance (NMR) structure determination techniques have improved dramatically in recent years, with around 12,000 protein structures having been deposited in the Protein Data Bank (PDB [16]). However, only a very small proportion of these structures correspond to protein-protein complexes. Due to a number of practical difficulties, it seems unlikely that it will become possible to solve the structures of protein complexes using high-throughput structural genomics techniques in the forseeable future [17]. Hence, computational techniques such as protein docking

1389-2037/08 $55.00+.00

will become an increasingly important way to help understand the molecular mechanisms of biological systems [18, 19]. Therapeutic drugs often modulate or block PPIs, and therefore PPIs represent an important class of drug target [20, 21].

Like all good scientific problems, the protein docking problem is easy to state but hard to solve. Almost 30 years ago, Wodak and Janin [22] described the first automated docking algorithm to predict the 3D interaction between bovine pancreatic trypsin and its natural inhibitor. Since then, protein docking has matured into a distinct computational discipline which brings together knowledge and techniques from a broad spectrum of sciences including physics, chemistry, biology, mathematics, and computing with the aim of modeling *in silico* how macromolecules such as proteins behave. Current docking algorithms employ a range of efficient search and energy-based scoring strategies, including e.g. fast Fourier transform (FFT) correlations, geometric hashing, and Monte Carlo (MC) techniques. These approaches generally produce a relatively small list of up to a few thousand putative docking orientations, amongst which a near-native binding mode is often observed. However, despite the use of improved scoring functions which typically include models of desolvation, hydrophobicity, and electrostatics, current algorithms still have difficulty in identifying the best solution from the list of false positives, or decoys. Hence many docking algorithms now use a two-step search and scoring procedure, in which *ab initio* techniques are used to generate an initial list of decoys which are then re-scored using available biophysical information (data-driven docking) and knowledge-based potentials derived from analyses of existing protein-protein interfaces [23-25].

There are several reviews of protein-protein docking techniques [26-35], and the performance of many current docking algorithms has been tested in the CAPRI (Critical Assessment of PRedicted Interactions) blind docking experiment [36-42]. The CAPRI experiment and its partner conference, Modeling of Protein Interactions in Genomes [43], have been instrumental in spurring new developments and providing a level playing field against which different docking algorithms may be tested and compared. This article gives an overview of recent progress in protein-protein docking and identifies several directions for future research. Incremental developments of the more established docking algorithms are generally not described here. Instead, the focus is on the salient or promising features of new approaches, several of which are data-based or data-driven, and many of which have not yet been tested in CAPRI.

AB INITIO RIGID BODY DOCKING

Many docking algorithms begin with a simplified rigid body representation of protein shape obtained by projecting each protein onto a regular 3D Cartesian grid, and by distinguishing grid cells according to whether they are near or intersect the protein surface, or are deeply buried within the core of the protein. Conceptually, a docking search is then performed by scoring the degree of overlap between pairs of grids in different relative orientations. However, performing a blind six-dimensional (6D) translational and rotational docking search typically entails evaluating in the order of billions ($O(10^9)$) of distinct grid overlaps. Hence, in practice, a variety of techniques are used to accelerate the calculation. For example, in the Fourier-based approaches the grid representations are first transformed into a set of orthogonal basis functions in order to perform the overlap calculations very efficiently using FFT techniques [44]. 3D FFT approaches have since been incorporated in several correlation-based docking algorithms [26, 45-51]. Eisenstein *et al.* [52] give a recent overview of the principles of grid-based FFT docking approaches. Grid overlaps may also be calculated rapidly using fast bit-wise arithmetic operations [53]. Unlike the 3D grid-based FFT correlation algorithms, the grid-free spherical polar Fourier (SPF) approach allows rotational rather than translational correlations to be calculated rapidly using one-dimensional (1D) FFTs [54]. In the geometric hashing approach, each protein surface is first pre-processed to give a list of a few hundred critical points ("pits", "caps", and "belts") which are then compared in a clique-detection algorithm to generate a relatively small number ($O(10^4)$) of trial docking orientations for grid scoring [55].

Solvation and desolvation effects are often considered as a surface phenomenon. All of the above *ab initio* docking algorithms incorporate an excluded volume model of shape complementarity, either explicitly using surface skins in the spherical polar Fourier (SPF) approach [54], or implicitly by assigning different values to surface and interior cells in the FFT grid representations [44]. The scoring functions in these algorithms favour orientations which occlude large surface cell volumes or bury large surface areas. Such approaches are largely consistent with the shell model of hydration [56]. However, as Elcock *et al.* [27] point out, most shape-based scoring functions generally do not discriminate between burial of different atom or side chain types because a single water probe radius or grid cell size is used to define the protein surface. Bhat *et al.* [57] demonstrated that using a variable radius probe sphere provides a straight-forward but superior way to represent the hydrophobicity of protein surface atoms. However, this has not yet been tested in existing docking algorithms.

The above approaches generally produce a list a few thousand candidate docking orientations which usually contains some near-native docking poses, provided the starting conformations are sufficiently similar to those of the complex. On a modern personal computer (PC), the calculation typically takes from a few minutes for the geometric hashing and polar Fourier approaches to a few hours for the FFT-based approaches. Hence, searching the 6D rigid-body space for putative docking orientations is not rate-limiting. However, existing scoring functions still have difficulty in distinguishing the near-native solutions from the list of decoys. Additionally, if the conformational changes on binding are large, then rigid body approaches can completely fail to produce any near-natives in the decoy list. Analysis of results in the CAPRI experiment shows that the best measure of target difficulty is the degree of conformational change between the bound and unbound protein structures [38, 42].

Fig. **1** shows the structures of two recent CAPRI targets, T21 and T26, which exemplify protein-protein complexes that are relatively hard and fairly easy to dock, respectively. Target T26 consists of a complex between a peptidoglycan-

associated lipoprotein (Pal), and the colicin tolerance-like protein (TolB) in which Pal binds across the bowl of the TolB C-terminal β-propeller domain, burying a large total solvent-accessible surface area of around 2,600Å^2 [58]. Although both Tolb, and Pal change conformation on binding, the backbone motions are relatively small (0.97Å and 0.41Å RMS for the TolB β-propeller and Pal domains, respectively), with much of the overall TolB conformational change appearing allosterically in the N-terminal domain [58]. There is some motion of the Pal residues E293, Y294, and E338, although only E338 changes rotameric conformation on binding. Hence, compared to many other CAPRI targets, the overall conformational changes in T26 are small. Additionally, there is considerable prior knowledge in the literature about the general mode of interaction between these proteins (e.g., [59]), which several predictor groups appear to have used. Overall, some 13 groups (including two solutions using Hex) obtained acceptable predictions, and 8 groups achieved medium accuracy predictions for this target, where the definitions of "acceptable," "medium," and "high accuracy" follow the assessment criteria of Méndez *et al.* [40]. These results indicate that, perhaps with the help of some prior knowledge, it is straight-forward for many current algorithms to make good docking predictions when the interface area of the complex is large and when the conformational differences between the unbound and bound structures are small.

In target T21, comprising a complex between the yeast origin recognition complex protein Orc1 and the silent information regulator protein Sir1 [60], the buried surface area is a relatively moderate ~1,300Å^2, but three Sir1 interfacial side chains (Y489, K522, and H524) change conformation on binding, and there are extensive conformational differences in the Orc1 small helical H domain (residues P97-A127) between the unbound and bound crystal structures (Cα deviation: 1.63Å RMS). Consequently, this complex proved to be rather difficult to predict well in CAPRI. For example, the Hex shape-based soft docking correlation produced many false-positive orientations with "side-to-side" domain contacts exhibiting much larger buried surface areas than that of the correct "head-to-head" orientation of the crystallographic complex (Fig. **1**), and it would appear that several other predictor groups encountered similar difficulties with this target. Nonetheless, 5 groups (Hirokawa, Weng, Vajda, Bonvin, and Gray) produced acceptable predictions, and 3 groups (Ten Eyck, Bonvin, and Gray) achieved medium predictions (M. F. Lensink and S. J. Wodak, personal communication). However, no high accuracy solutions were obtained despite the availability of considerable mutagenic evidence for likely interface residues on both protein partners [61]. Hence, the main challenges in protein docking today are to be able to generate reliably trial conformations which closely resemble those of the native complex, and to devise improved scoring functions which can correctly distinguish near-native docking orientations from a list of highly complementary decoys.

SOFT DOCKING TECHNIQUES

While the ability to include protein flexibility in docking is obviously desirable, most docking algorithms have until recently been obliged to assume that the proteins are rigid, at least at the backbone level, as a matter of computational expediency. However, most rigid body algorithms are nonetheless able to accommodate a degree of conformational flexibility through the use of soft scoring functions. For example, the binary core/interior scoring function embodied in the Cartesian FFT algorithms acts as a simple step-like van der Waals potential [62]. Using a coarse FFT grid implies using low order correlations and also serves to soften the potential [46]. The grid-free SPF approach [54] generally uses relatively low order polynomial powers in the range N=25–32, whereas the grid-based FFT approaches typically use trigonometric powers of N=64 or N=128. Using a low-pass filter in high resolution FFT docking also softens the scoring function, and has been shown to improve the results for grid-based FFT docking [63]. One advantage of the SPF approach

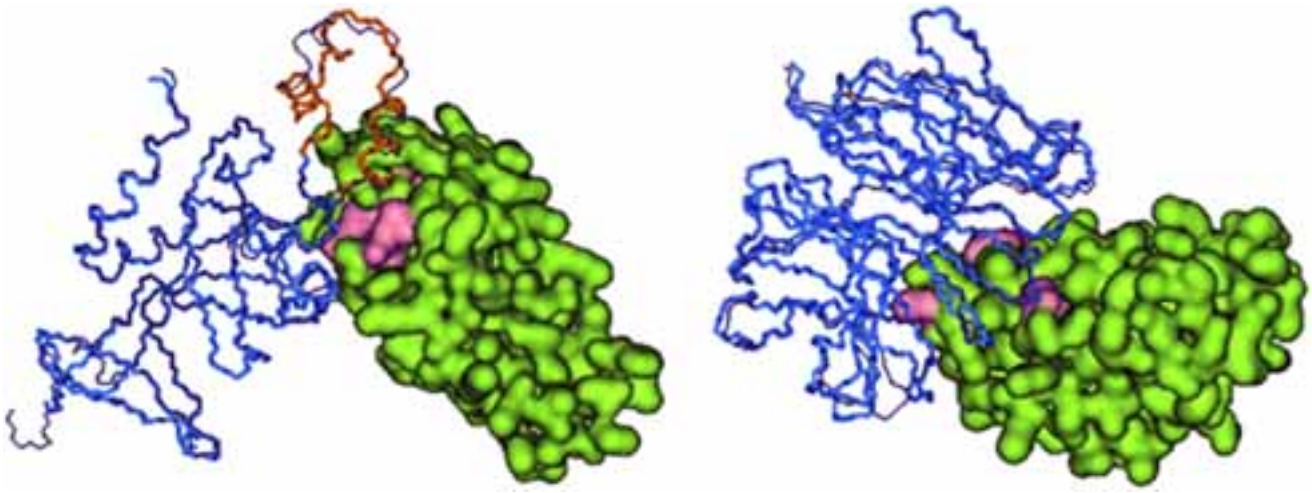

Fig. (1). CAPRI targets T21 (Orc1/Sir1) and T26 (TolB/Pal), as examples of relatively difficult and straight-forward complexes to dock, respectively. T21 (left) is coloured as grey: bound Orc1 backbone; blue: unbound Orc1 backbone (H domain in orange); yellow: unbound Sir1 van der Waals surface; pink: Sir1 surface patches corresponding to known interface residues V490, R493, D503, and L504 [61]. T26 (right) is coloured as grey: bound TolB backbone; blue: unbound TolB backbone; yellow: unbound Pal van der Waals surface; pink: Pal surface patches corresponding to known interface residues T93, G101, E102, and E130 [59].

compared to the grid-based FFT is that the polynomial order is independent of the search step size. Hence it is straight forward to calculate low order correlations with fine search increments. In a recent adaptation of the SPF approach, Sumikoshi *et al.* [64] perform very fast low order N=10–12 soft docking correlations. The rigid body search part of this approach is similar to the Hex algorithm [54]. However, Sumikoshi *et al.* use Legendre polynomials for the radial functions and calculate translations by numerical integration, whereas Hex uses Laguerre-Gaussian polynomials in order to calculate translations analytically [65]. With the help of some biochemical knowledge from the literature, the *ab initio* Hex correlation approach achieved 1 high accuracy, 1 medium accuracy, and 2 acceptable predictions in rounds 3–5 of CAPRI [66].

In order to incorporate a simple model of hydrophobicity into their FFT approach, Berchanski *et al.* [67] adapted the MolFit FFT algorithm to use the complex part of each grid cell value to give additional weight to complementary arrangements of hydrophobic residues. This was reported to improve significantly the rank of near native docking orientations for both tetrameric oligomers and hetero-dimers. Similarly, Heuser and Schomburg [68] modified their Ckordo FFT correlation algorithm to assign different shape complementarity weights to different amino acid types. The weights were determined by non-linear minimisation of docking scores from the Docking Benchmark complexes [69]. Different weights are used for different classes of complex but the general effect is to downgrade the contribution of flexible side chains such as Arg, Lys, Leu, Ser, and Thr, and to upgrade the docking score for hydrophobic side chains with high interface propensities such as e.g. Tyr and Trp. This approach is reported to give significant enrichment of near native orientations for all classes of complex [68]. Fig. **2** illustrates weighted soft docking Fourier correlations using colour-coded SPF shape density functions calculated for the T21 Orc1/Sir1 complex.

PREDICTING PROTEIN INTERACTION SURFACES

Rather than attempting to calculate protein docking interactions directly using *ab initio* approaches, it might be supposed that substantially fewer false-positive orientations would be obtained if one could first identify or predict the interaction surfaces on each protein partner. However, although the properties of protein-protein interfaces have been analysed in considerable detail [70-76], it remains a significant challenge to predict reliably the locations of protein-protein interaction surfaces using computational techniques alone [71, 75]. Nonetheless, some progress is being made. For example, Bogan and Thorn [72] compared differences in binding free energies following alanine scanning mutagenesis to show that often a small set of core interface residues contributes the majority of the binding free energy of a complex. They proposed a "hot-spot/O-ring" model of protein binding sites, in which the core hot spot residues are surrounded by a ring of energetically unimportant residues whose main role is to occlude bulk solvent from the hot spot. This study found that occlusion of solvent from the protein-protein interface is a necessary condition for binding, but there is no direct correlation between the experimentally determined binding free energy and buried surface area. However, the bulky side chains of Trp, Arg, and Tyr appear in hot spots with high frequencies (21%, 13%, and 12%, respectively), whereas Leu, Met, Ser, Thr, and Val residues are rarely observed in hot spots (3% frequency or less) [72].

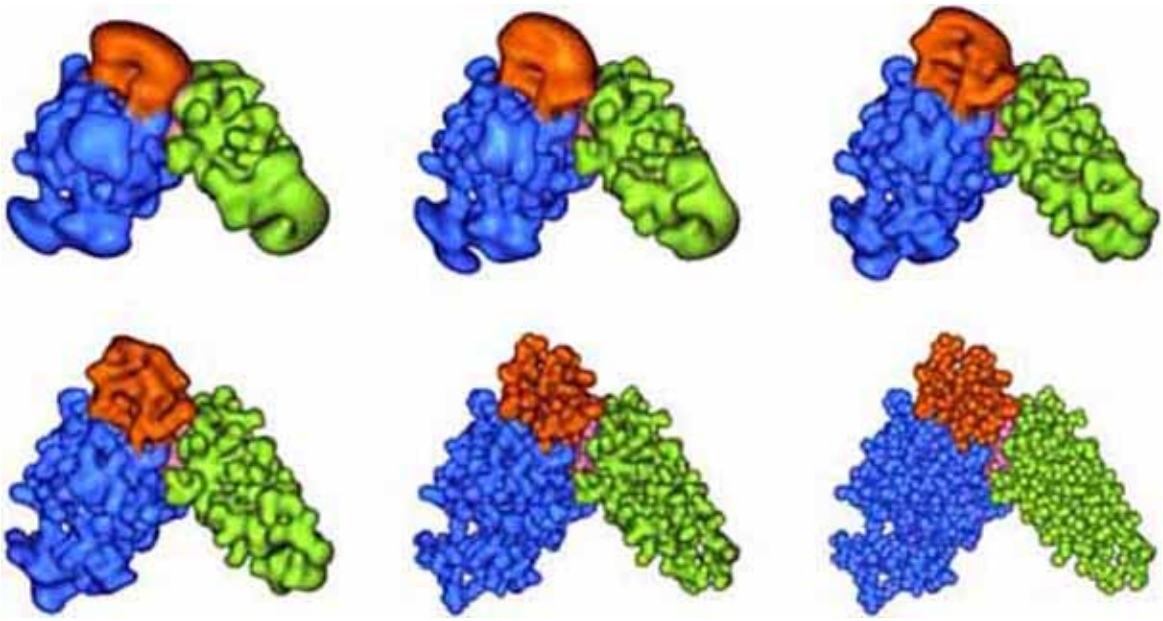

Fig. (2). Real 3D Fourier expansions of the bound conformations of Orc1 and Sir1 (CAPRI target T21), coloured as in Figure 1. Horizontally from top left to bottom right: the SPF steric density functions of Orc1 and Sir1 shown at polar expansion orders N=16, 20, 25, and 30, followed by a Gaussian van der Waals surface (bottom middle), and the van der Waals fused sphere representation (bottom right). It should be noted that SPF densities are 3D functions which are here contoured to give smooth 2D van der Waals surfaces for visualisation purposes. The Hex docking correlation algorithm typically uses N=16 and N=25 expansions [54]. The real part of a conventional complex Fourier shape representation, calculated by many FFT-based docking algorithms, would most closely resemble the N=30 image shown here (bottom left).

Studies of existing protein complexes by Chakrabarti and Janin [73] and Bahadur *et al.* [74] support a similar "core-rim" model of protein binding sites, in which the core residues resemble the composition of protein interiors and the rim residues, which remain to some extent solvent-accessible in the complex, resemble more the general composition of surface residues. Several studies have shown that alanine scanning hot spot residues correlate well with conserved residue locations in multiple sequence alignments (MSA) [77-79]. However, using such observations directly to predict putative docking epitopes requires a number of orthologous sequences to be available [80]. Interestingly, Halperin *et al.* [77] found that the local packing density around hot spot and conserved residues *across* binding interfaces is higher than expected, being reminiscent to that of protein cores, which suggests that good packing plays an important role in stabilising protein-protein interfaces. However, it should be noted that 60% of the alanine scan interfaces in this study belonged to homodimers. On the other hand, based on a study of 64 protein-protein interfaces (42 homodimers, 12 heterodimers and 10 transient complexes), Caffrey *et al.* [81] argue that interface surfaces are rarely significantly more conserved than other surface patches, and that using residue conservation alone is generally not sufficient for complete and accurate prediction of protein-protein interfaces.

Although the evidence to support the core-rim model is compelling, it is difficult to articulate simple rules with which unknown binding sites may be identified. Hence, machine learning techniques are being used to develop automated protein-protein interface prediction software [75,82-90]. These systems are typically trained using various combinations of e.g., buried surface areas, desolvation and electrostatic interaction energies, hydrophobicity scores, and residue conservation scores. Because different investigators used different learning datasets and because results are presented in different ways, it is difficult to make direct comparisons between individual approaches. However, as a very broad summary of recent results, current algorithms can generally predict the locations of protein interfaces with around 50% overlap between predicted and native interface residues in up to around 70% of complexes. Hence it would appear that such approaches are becoming useful and practical predictive tools. For example, Bradford and Westhead [85] used their SVM-based approach to predict, retrospectively, significant portions of the interfaces for 11 out of 15 selected CAPRI targets.

In many interface prediction algorithms, the manner in which surface patches are defined is critical. The optimal docking area (ODA) approach of Fernández-Recio *et al.* [88], which is based entirely on a desolvation model, is able to identify over 80% of protein interfaces in a test set of 66 non-obligate hetero-complexes. In a study of 97 such complexes, Burgoyne and Jackson [75] found that electrostatic complementarity showed little if any predictive capability, and that residue conservation had lower predictive power than expected, which tends to support the results of Caffrey *et al.* [81]. In agreement with the results of Fernández-Recio *et al.*, they found that cleft desolvation is the most strongly predictive characteristic of protein-protein interfaces [75]. Interestingly, Chung *et al.* [89] found that using crystallographic B factors to weight residue conservation scores was advantageous. It is also worth noting that de Vries *et al.* [91] showed that *intramolecular* surface contact propensities may also be used to infer protein interface regions.

STRUCTURAL PROTEIN-PROTEIN INTERACTION DATABASES

Databases of protein interactions are becoming important assets with which to predict the structures of protein complexes. Although not formally a database, the Protein-Protein Docking Benchmark [69], a collection of 84 non-redundant protein complexes for which both the bound and unbound structures are available, provides a valuable resource for testing new docking algorithms. Several further databases of structural protein-protein interaction data have been compiled, e.g., PQS [92, 93], DIP [94], BIND [95], DIMER [96], BID [97], 3DID [98], PIBASE [99], iPfam [100], Interdom [101], Interpare [102], 3D Complex [103], Dockground [104], I2I [105], SCOPPI [106], and PROTCOM [107]. Most of these databases provide at least the identities of domain interface residues, along with interface statistics such as residue propensities and buried surface areas. Some accept geometric queries to search for similar molecular surface patches [94, 105], or physico-chemically labeled patches [105]. The PQS system distinguishes interfaces between biologically active subunits from crystal packing contacts and hence provides a significantly richer source of protein-protein interface data than the original experimentally determined structures [93]. The comprehensive PIBASE [99], which is freely downloadable, draws data from the PDB, PQS, BIND, and DIP databases to provide data on around 160,000 domain pairs between 105,000 domains from 2,100 SCOP families [99]. The MULTIPROSPECTOR system uses a modified version of the PROSPECTOR threading algorithm to query the DIMER structural database in order to predict the interaction partners of a given query sequence [96]. The InterPreTS system [108] uses a similar approach with the 3DID database. The original version of this database (DBID) was relatively small, comprising 1,131 complexes, and in one of the examples described by the authors, only 35 out of a total of 2,590 putative yeast interaction pairs mapped to actual 3D structures [109]. The more extensive 3DID database (34,944 intermolecular domain interactions) has since been made available by the same group [98]. The above databases store information primarily on pair-wise protein interfaces. However, many proteins carry out their function as multimeric systems. Hence the recent 3D Complex database of Levy *et al.* [103] will be particularly useful for studying the structure, function, and evolutionary relationships of multimeric protein complexes.

Obviously, current databases contain only a very small fraction of all possible protein-protein complexes. However, as structural genomics initiatives continue to populate the space of protein 3D structures, it seems clear that using structural database systems to perform docking by homology will become an increasingly powerful approach. For example, Heuser *et al.* [101] used the Interdom database to rescore and improve FFT-based docking predictions for 16 out of 17 enzyme-inhibitor complexes and 2 out of 3 antibody-antigen complexes for which the structures of known homo-

logues existed. Korkin *et al.* [110] used comparative patch analysis queries on the PIBASE database [99] to predict correctly 70% of a test set of 20 complexes in the database compared to a 30% prediction rate using PatchDock [111]. Similarly, Kundrotas *et al.* [112] used structural queries to search the ProtCom database [107] in order to retrieve the correct domain partner of the query protein for 86% of the database of 418 complexes.

KNOWLEDGE-BASED DOCKING POTENTIALS

One natural way to exploit existing structural protein-protein interaction data is in the development of knowledge-based protein docking potentials. For example, Jiang *et al.* [113] developed a potential of mean force (PMF) approach based on hydrophobicity and hydrogen-bonding propensities, and parametrised using just 4 atoms types. This PMF model reproduces experimental binding energies for a test set of 28 complexes with a correlation coefficient of 0.75. Zhang *et al.* [114] developed the DFIRE (distance-scale finite ideal gas reference state) potential for scoring protein-protein, protein-DNA, and protein-ligand interactions. Using 19 atom types for protein-protein interactions, the DFIRE potential gives a good correlation (r=0.73) between the calculated and experimentally observed binding energies for a test set of 82 protein-protein complexes. Using a linear programming technique, Tobi and Bahar [115] developed a protein docking potential (PDP) based on 3 interaction centres (the side chain centroid, and backbone amide N and carbonyl O atoms) for each residue type. This PDP was able to identify a near-native conformation within the top 100 solutions in 10 out of 17 unbound-unbound test cases.

Although there is little doubt as to the importance of electrostatics in macromolecular interactions [116], and progress continues to be made in developing improved solvent models and fast Poisson-Boltzmann solvers [117], it appears that the electrostatic models used in current docking algorithms do not yet reliably help to identify near-native orientations. This limitation seems to be due, at least in part, to the need to use relatively simplistic physical models that afford rapid calculation over millions of trial orientations. For example, electrostatic correlations have been incorporated in several *ab initio* correlation search algorithms [118, 54, 47, 119, 120]. Gabb *et al.* [118] found that using a simple charge model for polar atoms improved the rank of near-native complexes in all cases tested. Similarly, Mandell *et al.* [47] found that their Poisson-Boltzmann electrostatic model was consistently beneficial when docking a set of complexes which are known to be electrostatically rate-accelerated. On the other hand, studies by Ritchie and Kemp [54] and Heifetz *et al.* [120], which used an accurate *in vacuo* SPF Coulomb representation, and the DELPHI linearised Poisson-Boltzmann model [121], respectively, both found that including electrostatics in the scoring function was beneficial in the majority of cases, but worsened the rank of near-native orientations in a significant number of others. Hence it would appear that the results of electrostatic calculations are rather sensitive to the type of model used and the specific complexes on which the model is tested. Sheinerman *et al.* [116] argue that due to desolvation of polar groups, protein-protein electrostatic interactions are generally net destabilizing. Hence we should perhaps not expect to see much benefit from using electrostatics in docking until we treat desolvation adequately. As an important step towards addressing this difficulty, Cerutti *et al.* [122] developed the ELSCA (energy by linear superposition of corrections approximation) knowledge-based potential method of including solvation effects into the linearised Poisson-Boltzmann/Surface-Area (PBSA) electrostatic model. This approach uses 16 basic atom types, each of which is endowed with Gaussian-type potential functions, parametrised using 45 protein-protein complexes from the Docking Benchmark [69]. Although the ELSCA potential has not yet been used in predictive docking, it reproduces calculated PBSA energies of the 45 bound and unbound test complexes with correlation coefficients of 0.96 and 0.79, respectively.

PCA OF KNOWLEDGE-BASED POTENTIALS

Two groups have described useful enhancements to Fourier-based rigid body search algorithms. Sumikoshi *et al.* [64] developed a fast low resolution SPF method of calculating docking energies using the ACE statistical potential of Zhang *et al.* [123]. By applying a principal component analysis (PCA) to the many cross terms in the ACE potential and by selecting only the 2 most significant eigenvector components, Sumikoshi *et al.* are able to calculate the most significant contributions to the ACE energy very efficiently [64]. On a test set of 6 unbound enzyme-inhibitor complexes using low order N=10–12 correlations, this approach is reported to give at least one near-native solution within the top 1,000 orientations in around 40 seconds on an ordinary PC. The PIPER program of Kozakov *et al.* [51] implements a similar PCA dimensionality reduction approach in the context of FFT-based docking. By counting the frequency of pair-wise atom occurrences in actual complexes compared to the corresponding frequencies found in a large number of *ab initio* decoys, Kozakov *et al.* [51] developed their "decoys as reference state" (DARS) knowledge-based potential. Applying a PCA to the cross terms in the DARS potential allows the leading contributions to be evaluated very efficiently *via* a small number of FFTs. In tests on the Docking Benchmark complexes, PIPER is reported to give up to 50% more near-native conformations than ZDOCK using the earlier atomic contact potential (ACP) scoring function [124].

DATA-DRIVEN DOCKING

If 3D structural information for a complex is not available, which is currently often the case, it is still extremely useful to be able to predict the location of a protein's functional site(s), or even just a single functional residue. Here, again, a variety of data-driven techniques are actively being developed. For example, the evolutionary trace (ET) approach of Lichtarge *et al.* [125] exploits the fact that functionally important residues are often conserved across species. ET techniques have been used successfully to identify protein functional sites [126] and to train support vector machines (SVMs) [127] or linear discriminant function (LDF) classifiers [128] to predict protein-protein interfaces. Sequence-based approaches are also able to identify protein-protein interface residues by locating correlated mutations in multiple sequence alignments for pairs of interacting proteins across different organisms [129, 8]. However, Halperin

et al. [130] suggest that such approaches may be limited to a relatively small number of protein families.

Nuclear Overhauser Effect (NOE) NMR has long been a powerful experimental tool with which to analyse the structures and dynamics of proteins in solution. Recent advances in the use of additional NMR data such as chemical shift perturbations (CSPs) and residual dipolar couplings (RDCs) allow new data-driven docking techniques to determine the solution structures of even relatively large and sometimes transient protein-protein and protein-DNA complexes [131]. In the HADDOCK approach [132], this experimental information is expressed in terms of ambiguous interaction restraints (AIRs) [133]. The generic $<r^{-6}>^{-1/6}$ form of an AIR acts like a potential energy which boosts the docking score whenever one or more pairs of restraint atoms occur close together across the protein-protein interface. Combining CSP and RDC AIRs with additional NMR information such as diffusion anisotropy relaxation data can also substantially improve data-driven docking [134, 135]. Other types of biochemical or biophysical data such as mutagenesis, H/D exchange [136], and ^{13}C-labeling data may also be usefully transformed into AIRs [32]. In favourable cases, as few as 3 restraints are sufficient to resolve the docked structure of a protein-protein complex [137, 138]. In recent rounds of CAPRI, the results obtained from HADDOCK have often been impressive, particularly for those targets for which experimental information was available [139]. Data-driven docking has also been applied successfully in protein-DNA docking [140]. Mass spectrometry radical probe shielding data has been used as a novel biophysical filter in the FFT-based PROXIMO docking algorithm [141]. Small-angle X-ray scattering data has been used to rank rigid body docking models of protein complexes in solution [142].

In addition to the general AIR formulation, several groups have developed a variety of strategies to incorporate biological information into their docking algorithms. For example, ZDOCK allows "blocking" residues to be defined, which are then given zero desolvation energy to bias those residues against appearing in the interface [143]. PatchDock allows known binding site residues to be specified in order to promote the scores for interfaces that contain a given percentage of those residues [111]. Smith *et al.* [144] use 3D conservation analysis [145] to identify putative interface residues for manual assessment of 3D-Dock predictions. Hex allows up to 2 search angle constraints to be specified to constrain its rotational correlation to remain near the starting orientation [54]. A similar cone angle constraint may be specified in 3D-Dock [144].

FFT correlation techniques are increasingly being used to fit high resolution X-ray protein structures into low resolution cryo-EM density maps [146-151]. Although the resolution of cryo-EM techniques is beginning to approach that of X-ray crystallography [152], such fitting or "interior docking" techniques are likely to remain very powerful approaches for determining atomic resolution structures of very large complexes which are unlikely to be solved using standard crystallographic techniques [148, 153]. It is interesting to note that interior docking algorithms are also beginning to incorporate biochemical knowledge from multiple biophysical sources in order to locate or anchor multimeric subunits in noisy low resolution EM density maps [154].

RE-SCORING DOCKING DECOYS

Several docking studies have indicated that low resolution scoring functions can often indicate the general location of a binding site, and that energetically favourable orientations tend to cluster around the native complex orientation [155, 156, 47, 157]. For example, Fernández-Recio *et al.* [157] mapped the distribution of predicted protein-protein docking orientations onto the receptor surface to show that highly populated regions often correspond to the actual binding site. Using an energy-based weight function, the contributions of surface residues to the interface was calculated to give a normalised interface propensity (NIP) for each residue. In a test set of 21 complexes, 80% of the predicted NIP residues were correctly located in native interfaces [157]. This approach has contributed to the very high success rate of the ICM software for many of the CAPRI targets [37, 40, 158]. Bernauer *et al.* [159] used a Voronoi tesselation representation of protein shape and a SVM-based machine learning approach to re-score successfully HADDOCK predictions for 4 out of 5 CAPRI targets. SVM techniques have also been used to re-score and discriminate RosettaDock energy funnels with encouraging results (O. Schueler-Furman, personal communication). Using the DFIRE knowledge-based potential, Zhang *et al.* [25] are able to place a near-native solution generally within the top 30 out of 2,000 ZDOCK decoys. When combined with clustering and manual selection based on biochemical knowledge, this approach gave reasonable predictions for 4 of the 6 targets in round 4 of CAPRI [40]. Although clustering is not normally considered as a scoring function *per se*, it has been shown that clustering uniformly sampled low energy *ab initio* FFT docking orientations to detect attractive energy basins can provide a simple but effective way to identify near-native binding orientations [47, 160, 161, 162, 163]. Marcia *et al*. [164] describe an iterative quadratic approximation method for finding the global docking minimum of a funnel-shaped energy landscape containing multiple local minima.

Two recent studies have investigated the use of protein-protein interface prediction algorithms as a way to re-score and filter conventional *ab inito* docking results. Gottschalk *et al.* [165] used the ProMate interface prediction algorithm [83] to re-rank the top 10,000 structures from FTDOCK runs on 21 unbound-unbound enzyme-inhibitor complexes. They compared the utility of their scoring function over random picking using a hypergeometric distribution. This combined docking and filtering approach produced at least 1 low RMS structure within the top 10 solutions in 15 of the 21 complexes, and the filter was found to enhance with statistical significance the FTDOCK scores in 77% of cases. In a similar study, Duan *et al.* [166] applied a residue conservation and physico-chemical scoring function to re-rank 10,000 FTDOCK structures for 59 unbound-unbound Docking Benchmark complexes. For the 48 complexes for which structural homologues exist, the filter was able to eliminate up to 86% of the FTDOCK structures while retaining the best near-native structure within the remaining list.

Using combinations of scoring functions can also improve docking results. For example, Murphy *et al.* [167] showed that using RPscore [168] and ACP [123] together gave better discrimination of near-native orientations. Liu *et al.* [169] developed the CFPScore function which was trained using 4 terms (PFM [113], packing density, contact size, and geometric complementarity) and which can distinguish true biological interfaces from crystal contact artifacts with an error rate of around 5%. In rounds 3–5 of CAPRI, Wiehe *et al.* [143] used biochemical information from the literature to define blocking residues to pre-filter FFT ZDOCK scans, and the top 2,000 decoys were then re-scored using the RDOCK desolvation and electrostatic model [24]. M-ZDOCK was used instead of ZDOCK for symmetrical targets. Overall, this approach produced 3 high accuracy, 3 medium accuracy, and 1 acceptable predictions [40]. The automated ClusPro server of Comeau *et al.* [170] uses ZDOCK or DOT for the FFT scan phase and then re-ranks the top 2,000 solutions using a greedy clustering algorithm. This approach achieved 1 high accuracy, 1 medium accuracy, and 2 acceptable CAPRI predictions. Using the unsupervised FFT-based GRAMM-X approach with conjugate gradient minimisation of a soft Lennard-Jones potential followed by re-scoring using evolutionary conservation analysis and phylogenetic residue contact preferences, Tovchigrechko and Vakser [163] achieved 2 medium accuracy CAPRI predictions. Tress *et al.* [171] used MSA and ET information to re-score GRAMM and Hex *ab initio* predictions for 7 out of 12 CAPRI targets. This gave 3 acceptable predictions [40], which is a rather impressive result for a non-structural sequence analysis based approach. After the CAPRI results were published, Camacho *et al.* [172] used CHARMm minimisation followed by re-scoring with the ACP potential [123] as implemented in the FastContact program [173] to re-rank the predictions of 6 targets from each participating group that achieved a near-native solution. Strikingly, the best FastContact score corresponded to the lowest ligand RMSD orientation in 16 out of 17 prediction sets.

MODELING SIDE-CHAIN FLEXIBILITY

When a pair of proteins form a complex, there is often a degree of structural rearrangement on going from the unbound to the bound conformations. Such induced fit effects can sometimes involve substantial changes of side chain torsion angles, particularly for flexible residues such as Lys and Arg. However, it is difficult to predict which side chains, if any, might change conformation on binding. Kimura *et al.* [174] argue that from a dynamical point of view there is insufficient time available during a collision encounter for extensive conformational rearrangements to take place. Using short time-scale simulations with explicit solvent, Kimura *et al.* showed that, when properly solvated, certain key interface residues generally adopt the same conformation in the unbound and bound structures whereas peripheral interface residues adopt a range of rotameric states. In other words, specific residues act as ready-made recognition motifs for docking [174]. Based on a subsequent molecular dynamics (MD) analysis of 11 complexes, Rajamani *et al.* [175] proposed a two-step binding model in which the first stage of complex formation often involves burial of one or more key "anchor residues" in a precursor encounter complex. These residues, typically located on the smaller partner, often correspond to alanine-scan hot spot residues, burial of which provides a significant proportion of the binding free energy of the complex. Once the anchors become docked in an encounter complex, the second stage of binding then involves peripheral interface or "latch" residues adjusting their conformations to provide the remainder of the binding free energy [175]. Camacho [176] used the notion of anchor residues to good effect for several CAPRI targets. Analysing side chain conformations after short MD simulations of the starting conformations allowed several hot spot interface residues to be identified as key anchor positions for the ClusPro/SmoothDock docking protocol [177]. This approach achieved 2 high accuracy, 2 medium accuracy, and 1 acceptable predictions in CAPRI rounds 3–5 [40].

Rather than attempting to predict interface residues *a priori*, the RosettaDock algorithm [178, 179] uses a multistage docking protocol which begins with a fast low resolution rigid body Monte Carlo (MC) rotation/translation search using simple residue-based potentials. This simulates the initial diffusional encounter between the proteins. Putative complexes are then refined using a further rigid body MC optimisation of fixed backbones with simultaneous sampling and minimisation of side chain conformations using a backbone-dependent rotamer library. Approximately $O(10)^5$ MC simulations are carried out per complex on a supercomputing cluster. The resulting solutions are scored using a detailed molecular mechanics (MM) energy function which includes solvation and hydrogen bond terms, and are then clustered for final ranking based on calculated energies and cluster size. In CAPRI rounds 3–5, this approach produced 2 high accuracy and 2 medium accuracy predictions for the targets attempted [40], with total computing times of around 50 CPU-days per target [178]. Wang *et al.* [180] use a modified version of RosettaDock which samples and minimises off-rotamer side chain conformations in order to achieve a better model of side chain flexibility than the former rigid-body plus rotamer-based approximation. Using this approach with the RosettaDock MC minimisation algorithm gave very good predictions for 6 out of 8 targets in CAPRI rounds 4 and 5 [181], which were amongst the best predictions over all CAPRI participants.

Carter *et al.* [50] used FTDOCK to perform an FFT scan of C[ε]-trimmed structures to provided 10,000 putative orientations which were scored using residue pair potentials [168], and the best 10 complexes were rebuilt using the Multidock rotamer refinement procedure [182]. Evolutionary Trace (ET) [125] analyses and the biochemical literature were used to help select predictions for some targets. This approach produced 2 medium accuracy and 3 acceptable predictions in CAPRI rounds 3–5 [40]. The ICM-DISCO algorithm [183, 184] also uses a two-stage search and refinement protocol. Initial encounter complexes are simulated using a rigid body pseudo Brownian motion algorithm with potentials pre-calculated on a 3D grid. The best 400 orientations are re-scored using a solvent-accessible area desolvation model, and side chain orientations are then optimised using biased probability minimisation. This approach, which takes from around 1 to up to 50 CPU-days of computation per complex [184], produced 2 high accuracy, 4 medium

accuracy, and 2 acceptable predictions in CAPRI rounds 3–5 [40].

The ATTRACT algorithm of Zacharias [185] incorporates a reduced protein model using just 2 or 3 pseudo atoms per residue, where each pseudo atom represents the locus of a 12-6 Lennard-Jones type potential. This simplified potential allows fast optimisation of rotational and translational space to be applied to multiple starting orientations. Flexibility is modeled using a side chain multi-copy approach. For unbound docking, ATTRACT is reported to produce a near-native orientation generally within the top 40 solutions. In rounds 3–5 of CAPRI, this approach gave 1 high accuracy and 2 medium accuracy predictions out of 5 targets attempted [186].

MODELING BACKBONE FLEXIBILITY

In a docking and MD study of the barnase/barstar complex, Ehrlich *et al.* [187] showed that even small backbone deformations can have as much impact on docking predictions as changes in side chain rotameric states, and that simultaneous treatment of backbone and side chain conformations is required for a complete picture of protein-protein binding. However, incorporating full backbone flexibility into protein docking simulations largely amounts to combining protein folding with protein docking, which is essentially an intractable task on current computer hardware. Nonetheless, several investigators are developing ways of introducing limited or simplified models of backbone flexibility into their docking algorithms. For example, flexible loops are commonly found at protein surfaces and often form a significant part of a protein-protein interface. Such loops are sometimes highly disordered in monomeric crystal structures and only become resolved in a fixed conformation in the complex. Bastard *et al.* [188] incorporated their multi-copy MC (MC2) approach for flexible protein-ligand docking [189] into the ATTRACT program. This modified approach was applied to 8 protein complexes, of which 4 corresponded to "difficult" targets in the Docking Benchmark [69]. In all but one case the approach was reported to improve docking results compared to docking only the unbound conformations. Schneidman *et al.* [111] extended the PatchDock geometric hashing approach to permit a simple model of backbone flexibility through the incorporation of hinge-bending regions (FlexDock) and to model cyclic symmetry (SymmDock). This suite of programs has consistently performed well in CAPRI, achieving acceptable or better predictions for 8 of the 9 rounds 3–5 targets, including remarkably good predictions for T8 (nidogen/flexible 3-domain laminin), and the challenging targets T9 (LicT homodimer) and T10 (TBEV trimer), both of which involved flexible docking of symmetrical subunits [190, 191]. In the HADDOCK protocol, several of the CAPRI target structures were subjected to MD simulations and typically 10–11 MD snapshots were selected for multi-copy docking [139]. Interfaces were initially predicted using PPISP algorithm of Chen and Zhou [86], which is available as a web service (pipe.scs.fsu.edu/ppisp.html). The HADDOCK approach produced 2 high accuracy and 2 medium accuracy predictions in CAPRI rounds 3–5 [40].

Several groups have used principal component analysis (PCA) of MD trajectories to generate protein conformations for docking [144,186,192-195]. The result of a PCA is a matrix of eigenvectors and a list of associated eigenvalues which together describe the principal components and amplitudes, respectively, of the internal motions within a protein. Often, these motions are concerted or collective. For example, one of the eigenvectors might correspond to the flexing of an entire α-helix about a hinge region. Typically, most of the internal motions within a protein can be adequately described by the first few eigenvectors [196, 197]. Hence a PCA analysis may be considered as a form of dimensionality reduction. Because the eigenvectors are orthogonal, they may be used to sample conformational space in a regular manner [66, 194]. This approach has been used successfully to model protein flexibility in protein-ligand docking [193, 194]. Although the computational cost of MD simulations will likely remain a bottleneck for docking purposes, MD-PCA would seem to provide a promising way to identify deformable residues or hinge-bending regions [198].

A detailed study by Smith *et al.* [195] showed that protein conformations from MD trajectories of unbound subunits generally sample part, but not all, of the conformational space corresponding to the bound proteins. MD conformations were docked in a multi-copy approach in which the starting conformations were clustered into 2 clusters for each protein. The central member of each cluster was taken as the representative structure to be docked along with the original unbound conformation. Hence, $((2+1)\times(2+1))=9$ cross dockings were performed for each complex. This approach was applied to 20 complexes. In some cases, docking MD structures gave better results than docking only the unbound structures, but in other cases the overall docking results were worse. However, this somewhat inconclusive result may be due to the small number of cross dockings performed per complex (FTDOCK computation times are around 1 day per docking), and because each cross docking run adds a lot of noise in the form of further false-positive orientations. Nonetheless, one very significant finding of Smith *et al.*'s study was that side-chain conformations in the core region of protein-protein interfaces were consistently less likely to change rotamer conformation than the peripheral interface residues. This is consistent with the explicit solvent MD results of Camacho *et al.* [174, 175]. Smith *et al.* [195] suggest a possible future strategy would be to perform fast rigid body core-core docking followed by MD on both proteins together, provided of course that the core regions can first be predicted with confidence. In CAPRI round 3–5, Smith *et al.* [144] used the 3D-Dock suite to dock representative MD structures for several targets. By using knowledge from the literature and conservation analysis to help identify good starting orientations, this combined approach generated 3 medium accuracy and 4 acceptable predictions out of 9 targets.

A similar MD and ensemble docking study by Grünberg *et al.* [199] used the fast shape-only Hex correlation function [54] to cross dock $((10+1)\times(10+1))=121$ principal component restrained MD (PCR-MD) and unbound structures for each of 17 protein-protein complexes. This study showed that even this relatively sparse coverage of conformations was able to given more and better near-native complexes

than docking the unbound structures alone. Remarkably, this enhancement appeared to be largely uncorrelated with the degree of similarity of the individual conformations to the bound form. In other words, the ensembles of structures appeared to contain multiple complementary conformations. Taking into consideration the physical rates of protein-protein collisions in solution, the rate at which proteins may dynamically exchange conformations, and the limited time available for proteins to reach their bound conformations during a collision event, Grünberg *et al.* argue that these results support the notion of a three-step docking mechanism of diffusion/collision, conformation selection, and induced fit [199].

As an alternative to performing expensive MD simulations, essential dynamics (ED) eigenvectors may be calculated using fast distance constraint ED (DCED) techniques [200]. We used Hex to dock multiple DCED-generated conformations for several of the CAPRI rounds 3–5 targets [66]. Our results for shape-only DCED multiconformer docking showed that the DCED approach gave a moderate but consistent improvement over docking unbound or model-built starting structures. This approach subsequently produced two acceptable predictions for CAPRI target T26 (TolB/Pal), the better of which had ligand and interface RMSDs of 3.35Å and 2.11Å, respectively. However, because this solution was not energy-minimised, it had a relatively high number of steric clashes. In our experience, DCED can be used to generate conformations which more closely resemble the complex than the starting unbound structure [66]. However, as the above example highlights, one drawback of the PCA-based structure generation approaches is that traversals along eigenvectors do not necessarily correspond to low energy conformational transitions. In other words, structures generated from PCA eigenvectors can violate standard bond length and torsion angle ranges, and hence need to be energy-minimised [201]. More recently, May and Zacharias [202] generated PCA eigenvectors from a Gaussian network model (GNM) of protein flexibility [203]. Computationally, this has the advantage that the eigenvectors may be derived directly from the GNM Hessian matrix [204], although it appears that the GNM eigenvectors do not always span the conformational space between the unbound and bound forms [202]. In any case, for practical docking purposes, it is not clear whether one should energy-minimise PCA-generated conformations and then dock them, or *vice versa*. The three-step docking model of Grünberg *et al.* [199] would support energy-minimising only in the final induced fit stage.

MODELING INTERFACIAL WATER

Although solvation and desolvation effects are crucially important in the thermodynamics of complex formation, most docking algorithms neglect to take into account the presence of water molecules at or around the protein-protein interface. In a study of 46 high resolution X-ray hetero-complexes, Rodier *et al.* [205] found that the majority of protein-protein interfaces are generally free of water, but that many interfaces have a peripheral hydration ring around the dry core. This is consistent with the O-ring or core-rim model [72, 73] of single patch interfaces. However, there are exceptions to this rule. Some interfaces can be significantly hydrated, especially in the case of protein-DNA complexes. On average, hetero-complexes have around 10–11 waters per buried 1,000Å^2 of interface, which corresponds to around 20 waters per complex, and the number of water-mediated polar interactions is similar to the number of interfacial protein-protein hydrogen bonds. Bound water molecules are therefore a general feature of protein-protein interactions [205].

Jiang *et al.* [206] describe a solvated rotamer library method of modeling interfacial waters, which was shown to help predict water locations in known complex structures. However, this approach has not been incorporated in a docking algorithm. As summarised above, Camacho [176] used short timescale explicit solvent MD simulations to identify key anchor side chains for subsequent rigid body docking. The only docking protocol to date that explicitly includes water molecules was described by van Dijk *et al.* [207]. In this approach, each starting structure is first solvated with a 5.5Å solvent shell in a short MD run. The solvated proteins are then rigidly docking using HADDOCK, and waters are iteratively removed from each encounter complex using a biased MC procedure until only 25% of the original interfacial waters remain, leaving from 6 to 12 waters per complex. This protocol was reported to give considerably better scores and RMS deviations than unsolvated docking for the majority of the 10 complexes studied, which included examples of both wet and dry interfaces. The predicted interfaces contained correctly buried waters in 17% of the acceptable solutions, and near-native water-mediated contacts were observed in from 30% to 66% of the near-native solutions [207]. Hence this approach to solvated docking would appear to be both feasible and indeed rather promising.

MULTIMERIC DOCKING AND DOCKING SERVERS

Because many proteins exist and function as multimers, there is a growing need to be able to model-build such complex macromolecular structures even if biophysical data is not available. Several groups have developed multimeric docking algorithms which typically apply symmetry operations to candidate dimers and reject those that produce intolerable steric clashes [208-211]. Inbar *et al.* [211, 212] showed that non-symmetrical multimeric complexes may be assembled using pair-wise docking techniques. Despite considerable uncertainties in individual dockings, the requirement that there must exist a mutually compatible set of pair-wise interactions serves as a very strong discriminator of the correct solution. For example, when tested on 5 complexes consisting of from 3 to 10 protein subunits, the CombDock combinatorial assembly algorithm was able to produce at least 1 near-native solution within the top 10 for each complex, starting from both bound and unbound conformations [211].

Several docking algorithms have been made available as internet servers (e.g., ClusPro [213]; PatchDock and SymmDock [111]; GRAMM-X [214]; M-ZDOCK [215]; HexServer [www.csd.abdn.ac.uk/hex_server]) or by electronic mail (SKE-Dock [216]). These services make docking calculations increasingly accessible to non-experts. A new Server section of CAPRI has been introduced to evaluate the performance of these completely automated docking services. Currently, all of the above servers employ *ab initio* prediction techniques rather than MD-based approaches. None

yet have links to a database, although some groups are working in this direction [104, 105]. As an alternative to server-based approaches, the Biskit platform provides a modular way to construct arbitrary workflows for sophisticated structural bioinformatics modelling and docking tasks [217].

CONCLUSIONS AND FUTURE DIRECTIONS

Recent docking and MD simulation studies support a picture of protein complexes being formed in at least a two-step process. In the initial collision encounter complex, recognition takes place through desolvation and burial of key hot spot anchor residues at the centre of the nascent interface, the conformations of which do not significantly change on binding. This is followed by a latching phase in which peripheral interface residues may adjust their rotameric conformations into complementary arrangements. Arguably, there is a final induced fit step in which interface side chains adjust their torsion angles to adopt off-rotamer conformations and interfacial waters become frozen into their crystallographically observable positions. Although this is clearly an idealised picture of a complex dynamical process, it is broadly compatible with the experimental and statistical studies of known protein interfaces reviewed here.

Existing rigid body search algorithms are now sufficiently fast that covering the 6D translation-rotation collision encounter space is no longer a rate-limiting step in protein-protein docking protocols. However, using explicit models of side-chain and backbone flexibility can involve computational costs of up to 50 CPU-days per complex, and using solvated MD simulations to locate hot spot anchor residues or to generate conformations for multi-copy docking is also very computationally expensive. PCA-based dimensionality reduction approaches seem to provide a promising way to generate candidate conformations for docking. However, PCA conformations can have poor internal geometries, which should be energy minimised, and cross docking multiple PCA conformations adds significantly to the computational load. It seems inevitable, therefore, that the use of more sophisticated *ab initio* flexible docking techniques will make increasingly heavy demands on computing resources. Although the cost of high performance computing facilities continues to fall, it is worth noting that modern graphics processing units (GPUs) offer potentially far greater arithmetic processing power than conventional CPUs, and a number of scientific calculations have been adapted to run on programmable GPUs [218]. For example, Buck *et al.* [219] have achieved an order of magnitude speed-up for Gromacs MD simulations in this way. Furthermore, 3D grid-based protein-ligand docking correlations have recently been implemented in low cost reconfigurable field programmable gate array (FPGA) devices, which are reported to give speed-ups of up to 3 orders of magnitude over the same calculations on ordinary PCs [220]. Similar speed-ups have been reported for FPGA-based MD simulations [221]. Thus it may soon be feasible to perform flexible protein-protein docking simulations using such hardware.

Structural PPI databases will become increasingly important resources for the development of docking-specific knowledge-based potentials and as training sets for machine learning based interface prediction software. Many of the CAPRI participants now use knowledge-based potentials to re-score *ab initio* solutions, and exploit biological and biophysical information to promote solutions that involve known interface residues. Using AIRs to express this information in a generic way seems particularly successful, although specifying lists of blocking residues and defining simple spatial search range parameters are also effective. The recent CAPRI results show that using experimental information to focus the docking search or to re-score *ab initio* decoys has become an integral component of many docking procedures, and can significantly improve the quality of docking predictions. In round 9 of CAPRI, a new Scorers section was introduced specifically to evaluate re-scoring techniques, although results from this section have not yet been published.

High throughput Y2H and TAP-MS experiments and bioinformatics techniques are beginning to generate entire networks of PPIs. However, such experimental and *in silico* results are difficult to validate and can contain many false-positives [11, 12, 222]. Hence, docking procedures could provide a potential way to filter physically implausible interactions. But is high throughput docking of predicted PPIs feasible? The answer will depend on the level of accuracy required. Assuming suitable template structures are available, Sánchez *et al.* [223] estimate that the structures of all of the ~6,400 yeast proteins can be comparatively modeled in a matter of days on a large PC cluster. Although such modeled structures would inevitably contain errors, the recent CAPRI experiments have shown that docking model-built structures is feasible. For example, targets T11, T14, and T19 each required a model building step, yet several groups produced medium accuracy or better predictions for each of these 3 targets [40]. However, in order to dock thousands of pairs of proteins, each pair-wise docking must be very fast, and current flexible docking protocols are therefore clearly impractical for high throughput purposes. Nonetheless, Tovchigrechko *et al.* [160] found that using low resolution structural models and FFT correlations was sufficient to recognise the gross structural features of PPIs as statistically significant clusters of orientations about the true binding site. Hence, using soft docking to detect low resolution energy funnels [224] could provide a useful way to enhance the reliability of experimental and *in silico* PPI predictions.

In summary, MD and flexible protein docking simulations are beginning to provide a convincing physical picture of how protein complexes are formed. Insights gained from these simulations are helping to inspire more reliable and practical docking algorithms. The use of symmetry and fragment assembly constraints are helping to make possible docking-based predictions of large multimeric complexes. Making better use of the increasing availability of structural, biological, and physico-chemical information about protein interactions is helping to improve significantly the quality of docking predictions. In the near future, the closer integration of docking algorithms with protein interface prediction software, structural databases, and sequence analysis techniques should help produce better predictions of PPI networks and more accurate structural models of the fundamental molecular interactions within the cell.

ABBREVIATIONS

1D	=	One Dimensional
2D	=	Two Dimensional
3D	=	Three Dimensional
6D	=	Six Dimensional
ACP	=	Atomic Contact Potential
AIR	=	Ambiguous Interaction Restraint
ANN	=	Artificial Neural Network
CAPRI	=	Critical Assessment of PRedicted Interactions
CPU	=	Central Processor Unit
DARS	=	Decoys As Reference State
DCED	=	Distance Constraint Essential Dynamics
ELSCA	=	Energy Linearised Superposition of Corrections Approximation
ED	=	Essential Dynamics
ET	=	Evolutionary Trace
FFT	=	Fast Fourier Transform
FPGA	=	Field Programmable Gate Array
FRM	=	Fast Rotational Matching
GNM	=	Gaussian Network Model
GPU	=	Graphics Processor Unit
PBSA	=	Poisson-Boltzmann Surface Area
PCA	=	Principal Component Analysis
PDP	=	Protein Docking Potential
PMF	=	Potential of Mean Force
PPI	=	Protein-Protein Interaction
LDF	=	Linear Discriminant Function
MC	=	Monte Carlo
MD	=	Molecular Dynamics
MS	=	Mass Spectrometry
MSA	=	Multiple Sequence Alignment
NMA	=	Normal Mode Analysis
NMR	=	Nuclear Magnetic Resonance
NOE	=	Nuclear Overhauser Effect
ODA	=	Optimal Docking Area
PC	=	Personal Computer
PDB	=	Protein Data Bank
RMS	=	Root Mean Squared
SPF	=	Spherical Polar Fourier
SVM	=	Support Vector Machine
TAP	=	Tandem Affinity Purification
Y2H	=	Yeast Two-Hybrid.

REFERENCES

[1] Marcotte, E.M., Pellegrini, P., Thompson, M.J., Yeates, T.O. and Eisenberg, D. (**1999**) *Nature*, *402*, 83-86.

[2] Eisenberg, D., Marcotte, E.M., Xenarios, I. and Yeates, T.O. (**2000**) *Nature*, *405*, 823-826.

[3] Uetz, P., Giot, L., Cagney, G., Mansfield, T.A., Judson, R.S., Knight, J.R., Lockshon, D., Narayan, V., Srinivasan, M., Pochart, P., Qureshi-Emili, A., Li, Y., Godwin, B., Conover, D., Kalbfleisch, T., G. Vijayadamodar, Yang, M.J., Johnston, M., Fields, S. and Rothberg, J.M. (**2000**) *Nature*, *403*, 623-671.

[4] Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M. and Sakaki, Y. (**2001**) *Proc. Natl. Acad. Sci.*, *98*, 4569-4574.

[5] Ho, Y., Gruhler, A., Heilbut, A., Bader, G.D., Moore, L., Adams, S.L., Millar, A., Taylor, P., Bennett, B., Boutilier, K., Yang, L., Wolting, C., Donaldson, I., Schandorff, S., Shewnarane, J., Vo, M., Taggart, T., Goudreault, M., Muskat, B., Alfarano, C., Dewar, D., Lin, Z., Michalickova, K., Willems, A.R., Sassi, H., Nielsen, P.A., Rasmussen, K.J., Andersen, J.R., Johansen, L.E., Hansen, L.H., Jespersen, H., Podtelejnikov, A., Nielsen, E., Crawford, J., Poulsen, V., Sørensen, B.D., Matthiesen, J., Hendrickson, R.C., Gleeson, F., Pawson, T., Moran, M.F., Durocher, D., Mann, M., Hogue, C.W.V., Figeys, D. and Tyers, M. (**2002**) *Nature*, *415*, 180-183.

[6] Gavin, A.C., Aloy, P., Grandi, P., Krause, R., Boesche, M., Marzioch, M., Rau, C., Jensen, L.J., Bastuck, S., Dumpelfeld, B., Edelmann, A., Heurtier, M.A., Hoffman, V., Hoefert, C., Klein, K., Hudak, M., Michon, A.M., Schelder, M., Schirle, M., Remor, M., Rudi, T., Hooper, S., Bauer, A., Bouwmeester, T., Casari, G., Drewes, G., Neubauer, G., Rick, J.M., Kuster, B., Bork, P., Russell, R.B. and Superti-Furga, G. (**2006**) *Nature*, *440(7084)* 631-636.

[7] Krogan, N.J., Cagney, G., Yu, H., Zhong, G., Guo, X., Ignatchenko, A., Li, J., Pu, S., Datta, N., Tikuisis, A.P., Punna, T., Peregrín-Alvarez, J.M., Shales, M., Zhang, X., Davey, M., Robinson, M.D., Paccanaro, A., Bray, J.E., Sheung, A., Beattie, B., Richards, D.P., Canadien, V., Lalev, A., Mena, F., Wong, P., Starostine, A., Canete, M.M., Vlasblom, J., Wu, S., Orsi, C., Collins, S.R., Chandran, S., Haw, R., Rilstone, J.J., Gandi, K., Thompson, N.J., Musso, G., St Onge, P., Ghanny, S., Lam, M.H.Y., Butland, G., Altaf-Ul, A.M., Kanaya, S., Shilatifard, A., Shea, E.O., Weissman, J.S., James Ingles, C., Hughes, T.R., Parkinson, J., Gerstein, M., Wodak, S.J., Emili, A. and Greenblatt, J.F. (**2006**) *Nature*, *440*, 637-643.

[8] Valencia, A. and Pazos, F. (**2002**) *Curr. Op. Struct. Biol.*, *12*, 368-373.

[9] Franzot, G. and Carugo, O. (**2003**) *Journal of Structural and Functional Genomics*, *4*, 245-255.

[10] Janin, J. and Séraphin, B. (**2003**) *Curr. Op. Struct. Biol.*, *13*, 383-388.

[11] Sawinski, L. and Eisenberg, D. (**2003**) *Curr. Op. Struct. Biol.*, *13*, 377-382.

[12] Bork, P., Jensen, L.J., von Mering, C., Ramani, A.K., Lee, I. and Marcotte, E.M. (**2004**) *Curr. Op. Struct. Biol.*, *14*, 292-299.

[13] Reš, I. and Lichtarge, O. (**2005**) *Phys. Biol.*, *2*, S36-S43.

[14] Vitkup, D., Melamud, E., Moult, J. and Sander, C. (**2001**) *Nature Struct. Biol.*, *8*, 559-566.

[15] Aloy, P. and Russell, R.B. (**2004**) *Nature Biotechnology*, *22*, 1317-1321.

[16] Sussman, J.L., Lin, D., Jiang, J., Manning, N.O, Prilusky, J., Ritter, O. and Abola, E.E. (**1998**) *Acta Cryst.*, *D54*, 1078-1084.

[17] Russell, R.B., Alber, F., Aloy, P., Davis, F.P., Korkin, D., Pichaud, M., Topf, M. and Sali, A. (**2004**) *Curr. Op. Struct. Biol.*, *14*, 313-324.

[18] Aloy, P., Pichaud, M. and Russell, R.B. (**2005**) *Curr. Op. Struct. Biol.*, *15*, 15-22.

[19] Aloy, P. and Russell, R.B. (**2006**) *Nat. Rev. Mol. Cell Biol.*, *7*, 188-197.

[20] Arkin, M.R. and Wells, J.A. (**2004**) *Nat. Rev. Drug Discov.*, *3*, 301-317.

[21] González-Ruiz, D. and Gohlke, H. (**2006**) *Curr. Med. Chem.*, *13*, 2607-2625.

[22] Wodak, S.J. and Janin, J. (**1978**) *J. Mol. Biol.*, *124*, 323-342.

[23] Camacho, C.J., Gatchell, D.W., Kimura, S.R. and Vajda, S. (**2000**) *Proteins: Struct. Func. Genet.*, *40*, 525-537.

[24] Li, L., Chen, R. and Weng, Z. (**2003**) *Proteins: Struct. Func. Genet.*, *53*, 693-707.

[25] Zhang, C., Liu, S. and Zhou, Y. (**2005**) *Proteins: Struct. Func. Bioinf., 60*, 314-318.
[26] Sternberg, M.J.E., Gabb, H.A. and Jackson, R.M. (**1998**) *Curr. Op. Struct. Biol., 8*, 250-256.
[27] Elcock, A.H., Sept, D. and McCammon, J.A. (**2001**) *J. Phys. Chem., 105*, 1504-1518.
[28] Camacho, C. J. and Vajda, S. (**2002**) *Curr. Op. Struct. Biol., 12*, 36-40.
[29] Halperin, I., Ma, B., Wolfson, H. and Nussinov, R. (**2002**) *Proteins: Struct. Func. Genet., 47*, 409-443.
[30] Smith, G.R.. and Sternberg, M.J.E. (**2002**) *Curr. Op. Struct. Biol., 12*, 28-35.
[31] Deremble, C. and Lavery, R. (**2005**) *Curr. Op. Struct. Biol., 15*, 171-175.
[32] van Dijk, A.D.J., Boelens, R. and Bonvin, A.M.J.J. (**2005**) *FEBS J., 272*, 293-312.
[33] Bonvin, A.M.J.J. (**2006**) *Curr. Op. Struct. Biol., 16*, 194-200.
[34] Gray, J.J. (**2006**) *Curr. Op. Struct. Biol., 16*, 183-193.
[35] Pierce, B. and Weng, Z., Xu, Y., Xu, D. and Liang, J., Eds. *Computational Methods for Protein Structure Prediction and Modeling, 2*, pp. 109-134, New York, Springer.
[36] Janin, J., Henrick, K., Moult, J., Ten Eyck, L., Sternberg, M.J.E., Vajda, S., Vakser, I. and Wodak, S.J. (**2003**) *Proteins: Struct. Func. Genet., 52*, 2-9.
[37] Méndez, R., Leplae, R., De Maria, L. and Wodak, S.J. (**2003**) *Proteins: Struct. Func. Genet., 52*, 51-67.
[38] Vajda, S. and Camacho, C.J. (**2004**) *Trends in Biotechnology, 22*, 110-116.
[39] Wodak, S.J. and Méndez, R. (**2004**) *Curr. Op. Struct. Biol., 14*, 242-249.
[40] Méndez, R., Leplae, R., Lensink, M.F. and Wodak, S.J. (**2005**) *Proteins: Struct. Func. Bioinf., 60*, 150-169.
[41] Janin, J. (**2005**) *Protein Sci., 14*, 278-283.
[42] Vajda, S. (**2005**) *Proteins: Struct. Func. Bioinf., 60*, 176-180.
[43] Vajda, S., Vakser, I.A., Sternberg, M.J.E. and Janin, J. (**2002**) *Proteins: Struct. Func. Genet., 47(4)* 444-446.
[44] Katchalski-Katzir, E., Shariv, I., Eisenstein, M., Friesem, A.A. and Aflalo, C. (**1992**) *Proc. Natl. Acad. Sci., 89*, 2195-2199.
[45] Vakser, I.A. and Aflalo, C. (**1994**) *Proteins: Struct. Func. Genet., 20*, 320-329.
[46] Vakser, I.A. (**1995**) *Protein Eng., 8(4)* 371-377.
[47] Mandell, J.G., Roberts, V.A., Pique, M.E., Kotlovyi, V., Mitchell, J.C., Nelson, E., Tsigelny, I. and Ten Eyck, L.F. (**2001**) *Protein Eng., 14(2)* 105-113.
[48] Del Carpio-Muñoz, C.A., Ichiishi, E., Yoshimori, A. and Yoshikawa, T. (**2002**) *Proteins: Struct. Func. Genet., 48*, 696-732.
[49] Chen, R., Li, L. and Weng, Z. (**2003**) *Proteins: Struct. Func. Genet., 52*, 80-87.
[50] Carter, P., Lesk, V.A., Islam, S.A. and Sternberg, M.J.E. (**2005**) *Proteins: Struct. Func. Bioinf., 60*, 281-288.
[51] Kozakov, D., Brenke, R., Comeau, S.R. and Vajda, S. (**2006**) *Proteins: Struct. Func. Bioinf., 65*, 392-406.
[52] Eisenstein, M. and Katchalski-Katzir, E. (**2004**) *Comptes Rendus Biologies, 327*, 409-420.
[53] Palma, P.N., Krippahl, L., Wampler, J.E. and Moura, J.J.G. (**2000**) *Proteins: Struct. Func. Genet., 39*, 372-384.
[54] Ritchie, D.W. and Kemp, G.J.L. (**2000**) *Proteins: Struct. Func. Genet., 39(2)* 178-194.
[55] Fischer, D., Lin, S., Wolfson, H.L. and Nussinov, R. (**1995**) *J. Mol. Biol., 248*, 459-477.
[56] Kang, Y.K., Némethy, G. and Scheraga, H.A. (**1987**) *J. Phys. Chem., 91*, 4105-4109.
[57] Bhat, S. and Purisima, E.O. (**2006**) *Proteins: Struct. Func. Bioinf., 62*, 244-261.
[58] Bonsor, D.A., Grishkovskaya, I,, Dodson, E.J. and Kleanthous, C. (**2007**) *J. Am. Chem. Soc., 129*, 4800-4807.
[59] Clavel, T., Germon, P., Vianney, A., Portalier, R. and Lazzaroni, J.C. (**1998**) *Mol. Microbiol., 29*, 359-367.
[60] Hou, Z., Bernstein, D.A., Fox, C.A. and Keck, J.L. (**2005**) *Proc. Natl. Acad.. Sci., 102*, 8489-9494.
[61] Bose, M.E., McConnell, K.H., Gardner-Aukema, K.A., Müller, U., Weinreich, U., Keck, J.L. and Fox, C.A., (**2004**) *Mol. Cell. Biol., 24*, 774-786.
[62] Vakser, I.A. (**1996**) *Protein Eng., 9(9)* 741-744.
[63] Segal, D. and Eisenstein, M. (**2005**) *Proteins: Struct. Func. Bioinf., 59*, 580-591.
[64] Sumikoshi, K., Tereda, T., Nakamura, S. and Shimizu, K. (**2005**) *Genome Inform., 16*, 161-173.
[65] Ritchie, D.W. (**2005**) *J. Appl. Cryst., 38*, 808-818.
[66] Mustard, D. and Ritchie, D.W. (**2005**) *Proteins: Struct. Func. Bioinf., 60*, 269-274.
[67] Berchanski, A., Shapira, B. and Eisenstein, M. (**2004**) *Proteins: Struct. Func. Bioinf., 56*, 130-142.
[68] Heuser, P. and Schomburg, D. (**2006**) *BMC Bioinformatics*, 7, 344.
[69] Mintseris, J., Wiehe, K., Pierce, B. anderson, R., Chen, R., Janin, J. and Weng, Z. (**2005**) *Proteins: Struct. Func. Genet., 60*, 214-216.
[70] Jones, S. and Thornton, J.M. (**1997**) *J. Mol. Biol., 272*, 121-132.
[71] Jones, S. and Thornton, J.M. (**1997**) *J. Mol. Biol., 272*, 133-143.
[72] Bogan, A.A. and Thorn, K.S. (**1998**) *J. Mol. Biol., 280*, 1-9.
[73] Chakrabarti, P. and Janin, J. (**2002**) *Proteins: Struct. Func. Genet., 47*, 334-343.
[74] Bahadur, R.P., Chakrabarti, P., Rodier, F. and Janin, J. (**2003**) *Proteins: Struct. Func. Genet., 53*, 708-917.
[75] Burgoyne, N.J. and Jackson, R.M. (**2006**) *Bioinformatics, 22(11)* 1335-1342.
[76] Janin, J., Rodier, F., Chakrabarti, P. and Bahadur, R.P. (**2007**) *Acta Cryst., D63*, 1-8.
[77] Halperin, I., Wolfson, H. and Nussinov, R. (**2004**) *Structure, 12*, 1027-1038.
[78] Li, X., Keskin, O., Ma, B., Nussinov, R. and Liang, J. (**2004**) *Structure, 12*, 1027-1038.
[79] Keskin, O., Ma, O. and Nussinov, R. (**2005**) *J. Mol. Biol., 345*, 1281-1294.
[80] Chelliah, V., Blundell, T.L. and Fernández-Recio, J. (**2006**) *J. Mol. Biol., 357*, 1669-1682.
[81] Caffrey, D.R., Somaroo, S., Hughes, J.D., Mintseris, J. and Huang, E.S. (**1998**) *Prot. Sci., 13*, 190-202.
[82] Koike, A. and Takagi, T. (**2004**) *Protein Engineering, Design & Selection, 17*, 165-173.
[83] Neuvirth, H., Raz, R. and Schreiber, G. (**2004**) *J. Mol. Biol., 338*, 181-199.
[84] Bordner, A.J. and Abagyan, R. (**2005**) *Proteins: Struct. Func. Bioinf., 60*, 353-366.
[85] Bradford, J.R. and Westhead, D.R. (**2005**) *Bioinformatics, 21*, 1487-1494.
[86] Chen, H. and Zhou, H.X. (**2005**) *Proteins: Struct. Func. Bioinf., 61*, 21-35.
[87] Del Sol, A. and Meara, P.O. (**2005**) *Proteins: Struct. Func. Bioinf., 58*, 672-682.
[88] Fernández-Recio, J., Totrov, M., Skorodumov, C. and Abagyan, R. (**2005**) *Proteins: Struct. Func. Bioinf., 58*, 134-143.
[89] Chung, J.L., Wang, W. and Bourne, P.E. (**2006**) *Proteins: Struct. Func. Bioinf., 62*, 630-640.
[90] Porollo, P. and Meller, J. (**2007**) *Proteins: Struct. Func. Bioinf., 66*, 630-645.
[91] De Vries, .S.J. and Bonvin, A.M.J.J (**2006**) *Bioinformatics, 22*, 2094-2098.
[92] Henrick, K. and Thornton, J.M. (**1998**) *Trends in Biochemical Sciences, 23*, 358-361.
[93] Jefferson, E.R., Walsh, T.P. and Barton, G.J. (**2006**) *J. Mol. Biol., 23*, 1118-1129.
[94] Preißner, R., Goede, A. and Frömmel, C. (**1998**) *J. Mol. Biol., 280*, 535-550.
[95] Bader, G.D. and Hogue, C.W.V. (**2000**) *Bioinformatics, 15*, 465-477.
[96] Lu, L. and Skolnick, J. (**2002**) *Proteins: Struct. Func. Genet., 49*, 350-364.
[97] Fischer, T.B., Arunachalam, K.V., Bailey, D., Mangual, V., Bakhru, S., Russo, R., Huang, D., Paczkowski, M., Lalchandani, V., Ramachandra, C., Ellison, B., Galer, S., Shapley, J., Fuentes, E. and Tsai, J. (**2003**) *Bioinformatics, 19*, 1453-1454.
[98] Stein, A., Russell, R.B. and Aloy, P. (**2005**) *Nucleic Acids Res., 33*, D413-D417.
[99] Davis, F.P. and Sali, A. (**2005**) *Bioinformatics, 21*, 1901-1907.
[100] Finn, R.D., Marshall, M. and Bateman, A. (**2005**) *Bioinformatics, 21*, 410-412.
[101] Heuser, P., Baù, D. and Schomburg, D., (**2005**) *Proteins: Struct. Func. Bioinf., 61*, 1059-1067.
[102] Gong, S., Park, C., Choi, H., Ko, J., Jang, I., Lee, J., Molser, D.M., Oh, D., Kim, D.S. and Bhak, J. (**2005**) *BMC Bioinformatics, 6*, 207.
[103] Levy, E.D., Pereira-Leal, J.B., Chothia, C. and Teichmann, S.A. (**2006**) *PLOS Comp. Biol., 2(11)* e155.

[104] Douguet, D., Chen, H.C., Tovchigrechko, A. and Vakser, I.A. (**2006**) *Bioinformatics, 22(21)* 2612-2618.
[105] Mintz, S., Shulman-Peleg, A., Wolfson, H.J. and Nussinov, R. (**2006**) *Proteins: Struct. Func. Bioinf., 61*, 6-20.
[106] Winter, C., Henschel, A., Kim, W.K. and Schroeder, M. (**2006**) *Nucleic Acids Research, 34*, D310-D314.
[107] Kundrotas, P.J. and Alexov, E. (**2007**) *Nucleic Acids Research, 35*, D575-D579.
[108] Aloy, P. and Russell, R.B. (**2003**) *Bioinformatics, 19*, 161-162.
[109] Aloy, P. and Russell, R.B. (**2003**) *Proc. Natl. Acad. Sci., 99*, 5896-5901.
[110] Korkin, D., Davis, F.P., Alber, F., Luong, T., Shen, M.Y., Lucic, V., Kennedy, M.B. and Sali, A. (**2006**) *PLOS Comp. Biol., 2(11)*, e153.
[111] Schneidman-Duhovny, D., Inbar, Y., Nussinov, R. and Wolfson, H.J. (**2005**) *Nucleic Acids Res., 33*, W363-W367.
[112] Kundrotas, P.J. and Alexov, E. (**2006**) *Biochim. Biophys. Acta, 1764*, 1498-1511.
[113] Jiang, L., Gao, Y., Mao, F., Li, Z. and Lai, L. (**2002**) *Proteins: Struct. Func. Genet., 46*, 190-196.
[114] Zhang, C., Liu, S., Zhu, Q. and Zhou, Y. (**2005**) *J. Med. Chem., 48*, 2325-2335.
[115] Tobi, D. and Bahar, I. (**2006**)*Proteins: Struct. Func. Bioinf., 62*, 970-810.
[116] Sheinerman, F.B., Norel, R. and Honig, B. (**2000**) *Curr. Op. Struct. Biol., 10*, 153-159.
[117] Koehl, P. (**2006**) *Curr. Op. Struct. Biol., 16*, 142-151.
[118] Gabb, H.A., Jackson, R.M. and Sternberg, M.J.E. (**1997**) *J. Mol. Biol., 272(1)* 106-120.
[119] Chen, R. and Weng, Z. (**2002**) *Proteins: Struct. Func. Genet., 47*, 281-294.
[120] Heifetz, A., Katchalski-Katzir, E. and Eisenstein, M. (**2002**) *Protein Sci., 11*, 571-587.
[121] Nicholls, A. and Honig, B. (**1991**) *J. Comp. Chem., 12*, 435-445.
[122] Cerutti, D.S., L. Ten Eyck, F. and McCammon, J.A. (**2005**) *J. Chem. Theory Comp., 1*, 143-152.
[123] Zhang, C., Vasmatzis, G.A., Cornette, J.L. and DeLisi, C., (**1997**) *J. Mol. Biol., 267(3)*, 707-726.
[124] Keskin, O., Bahar, I., Badretdinovabd, A.Y., Ptitsyn, O.B. and Jernigan, R.L. (**1998**) *Prot. Sci., 7*, 2578-2586.
[125] Lichtarge, O., Bourne, H.R. and Cohen, F.E., (**1996**) *J. Mol. Biol., 257*, 342-358.
[126] Yao, H., Kristensen, D.M., Mihalek, I., Sowa, M.E., Shaw, C., Kimmel, M., Kavraki, L. and Lichtarge, O. (**2003**) *J. Mol. Biol., 326*, 255-261.
[127] Reš, I., Mihalek, I. and Lichtarge, O. (**2005**) *Bioinformatics, 21*, 2496-2501.
[128] Li, J.J., Huang, D.S., Wang, B. and Chen, P. (**2006**) *Int. J. Biol. Macromol., 38*, 241-247.
[129] Pazos, F., Helmer-Citterich, M., Ausiello, G. and Valencia, A. (**1997**) *J. Mol. Biol., 271*, 511-523.
[130] Halperin, I., Wolfson, H. and Nussinov, R. (**2006**) *Proteins: Struct. Func. Bioinf., 63*, 832-845.
[131] Bonvin, A.M.J.J., Boelens, R. and Kaptein, R., (**2005**) *Curr. Op. Struct. Biol., 9*, 501-508.
[132] Dominguez, C., Boelens, R. and Bonvin, A.M.J.J. (**2003**) *J. Am. Chem. Soc., 125*, 1731-1737.
[133] Nilges, M. (**1995**) *J. Mol. Biol., 245*, 645-660.
[134] Van Dijk, A.D.J., Fushman, D. and Bonvin, A.M.J.J. (**2005**) *Proteins: Struct. Func. Bioinf., 60*, 367-381.
[135] Van Dijk, A.D.J., Kaptein, R., Boelens, R. and Bonvin, AM.J.J (**2006**) *J. Biomol. NMR, 34*, 237-244.
[136] Anand, G.S., Law, D., Mandell, J.G., Snead, A.N., Tsigelny, I., Taylor, S.S., Ten Eyck, L.F. and Komives, E.A. (**2003**) *Proc. Natl. Acad. Sci., 100*, 13264-13269.
[137] Clore, M. (**2000**) *Proc. Natl. Acad. Sci., 97*, 9021-9025.
[138] Tang, C. and Clore, M. (**2006**) *J. Biomol. NMR, 36*, 37-44.
[139] Van Dijk, A.D.J., De Vries, S.J., Dominguez, C., Chen, H., Zhou, H.X . and Bombin, A.M.J.J. (**2005**) *Proteins: Struct. Func. Bioinf., 60*, 232-238.
[140] Van Dijk, M., Van Dijk, A.D.J., Hsu, V., Boelens, R. and Bombin, A.M.J.J. (**2006**) *Nucleic Acids Res., 34*, 3317-3325.
[141] Gerega, S.K. and Downard, K.M. (**2006**) *Bioinformatics, 22*, 1702-1709.
[142] Petoukhov, M.V. and Svergun, D.I. (**2005**) *Biophys. J., 89*, 1237-1250.
[143] Wiehe, K., Pierce, B., Mintseris, J., Tong, M.W.W. anderson, R.., Chen, R. and Weng, Z. (**2005**) *Proteins: Struct. Func. Bioinf., 60*, 207-213.
[144] Smith, G.R., Fitzjohn, P.W., Page, C.S. and Bates, P.A. (**2005**) *Proteins: Struct. Func. Bioinf., 60*, 263-268.
[145] Landgraf, R., Xenarios, I. and Eisenberg, D. (**2001**) *J. Mol. Biol., 307*, 1487-1502.
[146] Wriggers, W., Milligan, R.A. and McCammon, J.A. (**1999**) *J. Struct. Biol., 125*, 185-195.
[147] Roseman, A.M. (**2000**) *Acta Cryst., D56*, 1332-1340.
[148] Rossmann, M.G. (**2000**) *Acta Cryst., D56*, 1341-1349.
[149] Wriggers, W. and Chacón, P. (**2001**) *J. Appl. Cryst., 34*, 773-776.
[150] Kovacs, J.A., Chacon, P., Cong, Y., Metwally, E. and Wriggers, W. (**2003**) *Acta Cryst., D59*, 1371-1376.
[151] Ceulemans, H. and Russell, R.B. (**2004**) *J. Mol. Biol., 338*, 783-793.
[152] Stowell, M.H.B., Miyazawa, A. and Unwin, N. (**1998**) *Curr. Op. Struct. Biol., 8*, 606-611.
[153] Rossmann, M.G., Arisaka, F., Battisti, A.J., Bowman, V.D., Chipman, P.R., Fokine, Halfstein, S., Kanamura, S., Kostyuchemko, V.A., Mesyanzhinov, V.V., Schneider, M.M., Morais, M.C., Leiman, P.G., Palermo, L.M., Parrish, C.R. and Xiao.C. (**2007**) *Acta Cryst., D63*, 9-16.
[154] Birmanns, S. and Wriggers, W. (**2007**) *J. Sruct. Biol., 157*, 271-280.
[155] Zhang, Z., Chen, J. and DeLisi, C. (**1999**) *Proteins: Struct. Func. Genet., 34(2)* 255-267.
[156] Camacho, C.J. and Vajda, S. (**2001**) *Proc. Natl. Acad. Sci., 98*, 10636-10641.
[157] Fernández-Recio, J., Totrov, M. and Abagyan R. (**2004**) *J. Mol. Biol., 335*, 843-865.
[158] Fernández-Recio, J., Abagyan and Totrov, M. (**2005**) *Proteins: Struct. Func. Bioinf., 60*, 308-313.
[159] Bernauer, J., Azé, J., Janin, J. and Poupon, A. (**2007**) *Bioinformatics, 23*, 555-562.
[160] Tovchigrechko, A., Wells, C.A. and Vakser, I.A. (**2002**) *Protein Sci., 11*, 1888-1896.
[161] Comeau, S.R., Gatchell, D.W., Vajda, S. and Camacho, C.J. (**2004**) *Bioinformatics, 20*, 45-50.
[162] Kozakov, D., Brenke, R., Comeau, S.R. and Vajda, S. (**2005**) *Biophys. J., 89*, 867-875.
[163] Tovchigrechko, A. and Vakser, I.A. (**2005**) *Proteins: Struct. Func. Bioinf., 60*, 296-301.
[164] Marcia, R.F., Mitchell, J.C. and Rosen, J.B. (**2005**) *Computational Optimization and Applications, 32*, 285-297.
[165] Gottschalk, K.E., Neuvirth, H. and Schreiber, G. (**2004**) *Prot. Eng. Des. Sel., 17*, 183-189.
[166] Duan, Y., Reddy, B.V.B. and Kaznessis, Y.N. (**2005**) *Prot. Sci., 14*, 316-328.
[167] Murphy, J., Gatchell, D.W., Prasad, J.C. and Vajda, S. (**2003**) *Proteins: Struct. Func. Genet., 53*, 840-854.
[168] Moont, G., Gabb, H.A. and Sternberg, M.J.E. (**1999**) *Proteins: Struct. Func. Genet., 35*, 364-373.
[169] Liu, S., Li, Q. and Lai, L. (**2006**) *Proteins: Struct. Func. Bioinf., 64*, 68-78.
[170] Comeau, S.R., Vajda, S. and Camacho, C.J. (**2005**) *Proteins: Struct. Func. Bioinf., 60*, 239-244.
[171] Tress, S., De Juan, D., Graña, P., Gómez, M.J., Gómez-Puertas, P., González, J.M., López, G. and Valencia, A. (**2005**) *Proteins: Struct. Func. Bioinf., 60*, 275-280.
[172] Camacho, C.J., Ma, H. and Champ, C. (**2006**) *Proteins: Struct. Func. Bioinf., 63*, 868-877.
[173] Camacho, C.J. and Zhang, C. (**2005**) *Bioinformatics, 21*, 2534-2536.
[174] Kimura, S., Brower, R.C., Vajda, S. and Camacho, C.J. (**2001**) *Biophys. J., 80*, 635-642.
[175] Rajamani, D., Thiel, S., Vajda, S. and Camacho, C.J. (**2004**) *Proc. Natl. Acad. Sci., 101*, 11287-11292.
[176] Camacho, C.J. (**2005**) *Proteins: Struct. Func. Bioinf., 60*, 245-251.
[177] Camacho, C.J. and Gatchell, D.W. (**2003**) *Proteins: Struct. Func. Bioinf., 52*, 92-97.
[178] Gray, J.J., Moughan, S., Wang, C., Schueler-Furman, O., Kuhlman, B., Rohl, C.A. and Baker, D. (**2003**) *J. Mol. Biol., 331*, 281-299.
[179] Daily, M.D., Masica, D., Sivasubramanian, A., Somarouthu, S. and Gray, J.J. (**2005**) *Proteins: Struct. Func. Bioinf., 60*, 181-186.

[180] Wang, C., Schueler-Furman, O. and Baker, D. (**2005**) *Prot. Sci., 14*, 1328-1339.
[181] Schueler-Furman, O., Wang, C. and Baker, D. (**2005**) *Proteins: Struct. Func. Bioinf., 60*, 187-194.
[182] Jackson, R.M., Gabb, H.A. and Sternberg, M.J.E. (**1998**) *J. Mol. Biol., 276*, 265-285.
[183] Fernández-Recio, J., Totrov, M. and Abagyan, R. (**2002**) *Protein Sci., 11*, 280-291.
[184] Fernández-Recio, J., Totrov, M. and Abagyan, R. (**2003**) *Proteins: Struct. Func. Bioinf., 52*, 113-117.
[185] Zacharias, M. (**2003**) *Prot. Sci., 12*, 1271-1282.
[186] Zacharias, M. (**2005**) *Proteins: Struct. Func. Bioinf., 60*, 252-256.
[187] Ehrlich, L.P., Nilges, M. and Wade, R.C. (**2005**) *Proteins: Struct. Func. Bioinf., 58*, 126-133.
[188] Bastard, K., Prévost, C. and Zacharias, M. (**2006**) *Proteins: Struct. Func. Bioinf., 62*, 956-969.
[189] Bastard, K., Thureau, A,. Lavery, R. and Prévost, C. (**2003**) *J. Comp. Chem., 24*, 1910-1920.
[190] Inbar, Y., Schneidman-Duhovny, D., Oron, A., Nussinov, R. and Wolfson, H.J. (**2005**) *Proteins: Struct. Func. Bioinf., 60*, 217-223.
[191] Schneidman-Duhovny, D., Inbar, Y., Nussinov, R. and Wolfson, H.J. (**2005**) *Proteins: Struct. Func. Bioinf., 60*, 224-231.
[192] Zacharias, M. and Sklenar, H. (**1999**) *J. Comp. Chem., 20(3)* 287-300.
[193] Zacharias, M. (**2004**) *Proteins: Struct. Func. Bioinf., 54*, 759-767.
[194] Cavasotto, C.N., Kovacs, J.V. and Abagyan, R.A. (**2005**) *J. Am. Chem. Soc., 127*, 9632-9640.
[195] Smith, G.R., Sternberg, M.J.E. and Bates, P.A. (**2005**) *J. Mol. Biol., 347*, 1077-1101.
[196] Amadei, A., Linssen, A.B.M. and Berendsen, H.J.C. (**1993**) *Proteins: Struct. Func. Genet., 17*, 412-425.
[197] May, A. and Zacharias, M. (**2005**) *Biochim. Biophys. Acta, 1754*, 225-231.
[198] Kovacs, J.A., Chacon, P. and Abagyan, R. (**2004**) *Proteins: Struct. Func. Bioinf., 56*, 661-668.
[199] Grünberg, R., Leckner, L. and Nilges, M. (**2004**) *Structure, 12*, 2125-2136.
[200] De Groot, B.L., Van Aalten, D.M.F., Scheek, R.M., Amadei, A., Vriend, G. and Berendsen, H.J.C. (**1997**) *Proteins: Struct. Func. Genet., 29*, 240-251.
[201] Lei, M., Zavodoszsky, M.I., Kuhn, L.A. and Thorpe, M.F. (**2004**) *J. Comp. Chem., 25*, 1133-1148.
[202] May, A. and Zacharias, M. (**2007**) *In press*.
[203] Tirion, M. (**1996**) *Phys. Rev. Lett., 77*, 1905-1908.
[204] Hinsen, K. (**1998**) *Proteins: Struct. Func. Genet., 33*, 417-427.
[205] Rodier, F., Bahadur, R.P., Chakrabarti, P. and Janin, J. (**2005**) *Proteins: Struct. Func. Bioinf., 60*, 36-45.
[206] Jiang, L., Kuhlman, B., Kortemme, R. and Baker, D. (**2005**) *Proteins: Struct. Func. Bioinf., 58*, 893-904.
[207] Van Dijk, A.D.J. and Bonvin, A.M.J.J. (**2006**) *Bioinformatics, 22*, 2340-2347.
[208] M. Eisenstein, I. Shirav, G. Koren, A. A. Friesem and E. Katchalski-Katzir. (**1997**) *J. Mol. Biol., 266*, 135-143.
[209] Berchanski, A. and Eisenstein, M. (**2003**) *Proteins: Struct. Func. Genet., 53*, 817-829.
[210] Comeau, S.R. and Camacho, C.J. (**2004**) *J. Struct. Biol., 150*, 233-244.
[211] Inbar, Y., Benyamini, H., Nussinov, R. and Wolfson, H.J. (**2005**) *J. Mol. Biol., 349*, 435-447.
[212] Inbar, Y., Benyamini, Nussinov, R. and Wolfson, H.J. (**2005**) *Phys. Biol., 2*, S156-S165.
[213] Comeau, S.R., Gatchell, D.W., Vajda, S. and Camacho, C.J. (**2004**) *Nucleic Acids Res., 32*, W96-W99.
[214] Tovchigrechko, A. and Vakser, I.A. (**2005**) *Nucleic Acids Research, 34*, W310-W314.
[215] Pierce, B., Tong, W. and Weng, Z. (**2005**) *Bioinformatics, 21*, 1472-1478.
[216] Terashi, G., Takeda-Shitaka, M., Takaya, D., Komatsu, K. and Umeyama, H. (**2005**) *Proteins: Struct. Func. Bioinf., 60*, 289-295.
[217] Grünberg, R., Nilges, M. and Leckner, J. (**2007**) *Bioinformatics, 23*, 769-770.
[218] Owens, J.D., Luebke, D., Govindaraju, N., Harris, M., Krüger, J., Lefohn, A.E. and Purcell, T.J. (**2007**) *Computer Graphics Forum, 26*, 80-113.
[219] Buck, I., Foley, T., Horn, D., Sugerman, J., Fatahalian, K., Houston, M. and Hanrahan, P. (**2004**) *ACM Trans. Graph., 23*, 777-786.
[220] VanCourt, T., Gu, Y., Mundada, V. and Herbordt, M. (**2006**) *EURASIP Journal on Applied Signal Processing, 2006*, 1-10.
[221] Gu, Y., VanCourt, T. and Herbordt, M.C. (**2006**) *IEE Proc. Comp. Dig. Tech., 153*, 189-195.
[222] Hart, G.T., Ramani, A.K. and Marcotte, E.M. (**2006**) *Genome Biology, 7*, 120.
[223] Sánchez, R., Pieper, U., Melo, F., Eswar, N., Martí-Renom, M.A., Madhusudhan, M.S., Mirkovic, N. and Šali, A. (**2000**) *Nat. Struct. Biol., 7*, 986-990.
[224] Tovchigrechko, A. and Vakser, I.A. (**2001**) *Protein Sci., 10*, 1572-1583.

Received: February 13, 2007 **Revised: July 11, 2007** **Accepted: July 11, 2007**