Focused Grid-Based Resampling for Protein Docking and Mapping

Artem B. Mamonov,[†[a]] Mohammad Moghadasi,[†[b]] Hanieh Mirzaei,[†[b]] Shahrooz Zarbafian,[[c]] Laurie E. Grove,[[d]] Tanggis Bohnuud,[[a]] Pirooz Vakili,[[b,c]] Ioannis Ch. Paschalidis,*[[b,e]] Sandor Vajda,*[[a,b,f]] and Dima Kozakov*[[a,g]]

The fast Fourier transform (FFT) sampling algorithm has been used with success in application to protein-protein docking and for protein mapping, the latter docking a variety of small organic molecules for the identification of binding hot spots on the target protein. Here we explore the local rather than global usage of the FFT sampling approach in docking applications. If the global FFT based search yields a near-native cluster of docked structures for a protein complex, then focused resampling of the cluster generally leads to a substantial increase in the number of conformations close to the native structure. In protein mapping, focused resampling of the selected hot spot regions generally reveals further hot spots that, while not as strong as the primary hot spots, also contribute to ligand binding. The detection of additional ligand binding regions is shown by the improved overlap between hot spots and bound ligands. © 2016 Wiley Periodicals, Inc.

DOI: 10.1002/jcc.24273

Introduction

The goal of protein–protein docking is to determine the structure of a complex in atomic detail, starting from the coordinates of the unbound component molecules.[1–4] In contrast, protein mapping is aimed at identifying the hot spots of proteins, i.e., regions that are major contributors to the binding free energy and hence are crucial to the binding of any ligand at that particular site. It was shown that such hot spots bind a large variety of fragment sized organic compounds, and hence determining where such probe molecules cluster on the protein surface can be used for finding the hot spots.[5–8] Although docking and mapping seem to substantially differ, what is common between them is that both require exhaustive sampling of the energy landscape over the conformational space of a receptor-ligand complex.[9,10] Assuming the validity of rigid body approximation, the fast Fourier transform (FFT) correlation algorithm provides a computationally very efficient approach to performing these calculations. Indeed, due to the pairwise character of interactions in most molecular mechanics force fields, the energy of the complex at the fixed position of the receptor and at a given rotation of the ligand can be calculated as the convolution integral of two functions, one defined on the receptor and the other on the ligand.[11–13] Considering the functions on grids, the integral is converted into a sum, which can be calculated as a product of the discrete Fourier transforms of the two functions, resulting in the so-called fast Fourier transform (FFT) correlation method. Due to its extreme computational efficiency, this approach can systematically evaluate billions of docked conformations within a few hours on a single processor, and the algorithm is easily parallelizable. FFT has been first used for protein-protein docking with a scoring function representing only shape complementarity,[11,14,15] but later it has been expanded to include electrostatic and solvation terms,[16–18] and more recently structure-based pairwise interaction potentials,[19,20] substantially improving the accuracy of the method. The power of using FFT is well demonstrated by the fact that a large number of best performing methods in the blinded protein-protein docking experiments CAPRI (Critical Assessment of PRotein Interactions)[21] are based on this approach.[22–26]

Protein mapping is useful for a number of applications including the identification of binding sites[27] and determining

[a] *Artem B. Mamonov, T. Bohnuud, S. Vajda, D. Kozakov*
Department of Biomedical Engineering, Boston University, Boston, Massachusetts 02215
[b] *M. Moghadasi, H. Mirzaei, P. Vakili, I. C. Paschalidis*
Center for Information and Systems Engineering, Boston University, Boston, Massachusetts 02215
E-mail: midas@laufercenter.org or vajda@bu.edu or yannisp@bu.edu
[c] *S. Zarbafian, P. Vakili*
Department of Mechanical Engineering, Boston University, Boston, Massachusetts 02215
[d] *L. E. Grove*
Department of Sciences, Wentworth Institute of Technology, Boston, Massachusetts 02115
[e] *I. C. Paschalidis*
Department of Electrical and Computer Engineering, Boston University, Boston, Massachusetts 02215
[f] *S. Vajda*
Department of Chemistry, Boston University, Boston, Massachusetts 02215
[g] *D. Kozakov*
Department of Applied Mathematics and Statistics, Stony Brook University, Stony Brook, New York 11790

[†]These authors contributed equally to this work.
Contract grant sponsor: National Institute of General Medical Sciences; Contract grant numbers: GM061867, GM093147, and GM064700; Contract grant sponsor: National Science Foundation; Contract grant numbers: AF 1527292 and DBI 1458509; Contract grant sponsor: Russian Scientific Fund; Contract grant number: 14-34-00017

the druggability of a site,[28,29] as well as for fragment-based drug discovery.[30–32] While the FFT algorithm is less widely used for mapping than for protein docking, it provides similar advantages.[10] Although in mapping each ligand used as a probe usually consists of less than 15 heavy atoms, identification of hot spots requires determining interactions with a number of different probes, and hence the efficiency of the search remains a concern. Increasing the number and variety of probes generally provides more specific information on the binding site in form of structure-based pharmacophores.[31,32] In addition to the improved numerical efficiency, a major advantage of using FFT is its ability to perform systematic sampling, which allows for the calculation of partition functions and thus determining important thermodynamic quantities such as free energy within the framework of rigid body grid approximation.[33,34]

In FFT based docking and mapping methods the global grid sampling is usually followed by some type of off-grid refinement, which can be simple energy minimization, Monte Carlo search, or Monte Carlo minimization.[9] Since the refinement is computationally much less efficient than the grid based sampling by FFT, it is necessarily restricted to certain regions of the conformational space, selected on the basis of the outcome of global search and frequently some *a priori* information. In view of the extreme efficiency of the grid sampling, the question we consider here is the potential utility of performing the refinement also by FFT, restricting the search to the region of interest but using enhanced sampling, either by employing a finer grid or by retaining more sample points in the region. Although such refinement still assumes rigid body association, it is important to note that both docking and mapping by FFT use "smooth" scoring functions that do not penalize some degree of steric overlaps, and hence neglecting some level of conformational change does not necessarily make the approach inferior to a refinement method that allows for some flexibility, but can perform only more limited sampling. In addition, the structures generated by rigid body resampling can be used as starting points for methods that account for flexibility. Accordingly, to demonstrate the relationship between focused resampling and Monte Carlo based refinement, we also study here the impact of short Monte Carlo minimization (MCM) runs.

To make the analysis of focused sampling meaningful we focus on correct regions of interest as usual at the refinement stage,[9] and hence we assume that such region can be selected based on the results of initial docking or mapping, or additional information is available to assist the selection. The information needed for such decisions may be very limited. Global protein-protein docking usually yields a few large clusters of low energy docked structures, and the knowledge of a single residue located in the interface, determined by site directed mutagenesis or crosslinking, is frequently enough to choose the right one among such clusters.[9] Moreover, in many cases one cluster is much larger than the others, facilitating finding the likely near-native cluster without additional data. It is also possible that structures in several clusters should be refined before the final decision can be made.[35] In protein mapping it is generally simpler to find the region of interest, as for most proteins some information is available on the approximate location of the functional site.[28,36] In this article we restrict consideration to the resampling of the most near-native large cluster in docking and to the region containing the known functional or drug target site in mapping.

In order to determine whether the refinement is effective, we need test problems and baselines for comparison. In protein-protein docking we will consider a subset of the established docking benchmark set containing structures of protein pairs for which the X-ray structures of the complex is also available.[37] For establishing the baseline for docking success we first perform global FFT based docking using the PIPER program,[19] cluster the low energy structures as described for the ClusPro server,[18,34] select the cluster closest to the native state, and determine the fraction of structures that are within 5 Å interface root mean square deviation (IRMSD) from the native complex. The IRMSD value is calculated for the backbone atoms of the ligand that are within 10 Å of any receptor atom after superimposing the receptors in the X-ray and docked complex structures. The problem we consider here is determining whether this fraction can be increased if the region defined by the cluster is resampled on a finer grid and using the FFT algorithm that initially retains more low energy translation for each rotation of the ligand. For the analysis of the impact of resampling in protein mapping we consider a small benchmark set consisting of proteins with X-ray structures available for both apo and holo forms, the latter co-crystallized with a ligand. For each protein, the ligand-free structure is mapped, and the quality of mapping results is assessed on the basis of how well the hot spots overlap with the bound ligand. The results from this initial mapping, as well as *a priori* information available on the ligand binding site, will be used to determine the region of interest for focused mapping, and the enhanced hot spots will be analyzed as in the global mapping calculation. We have specifically selected proteins that have several hot spot regions, and hence the initial mapping does not provide a sufficiently accurate picture of the hot spots in the region of interest.

Methods

For a more detailed description of the methods, please refer to the Supporting Information.

Protein-protein docking

Global Docking Using the ClusPro Server. The ClusPro server[34] docks the two proteins (considered the receptor and the ligand, respectively) using the PIPER program based on the FFT correlation approach. As reported earlier,[19,34] the scoring function in PIPER is a weighted combination of three energy terms: a van der Waals term describing attractive and repulsive contributions, a truncated coulombic term for the electrostatic energy, and a structure-based pairwise potential primarily representing desolvation. Selecting the weights of the energy terms depends on the characteristics of the

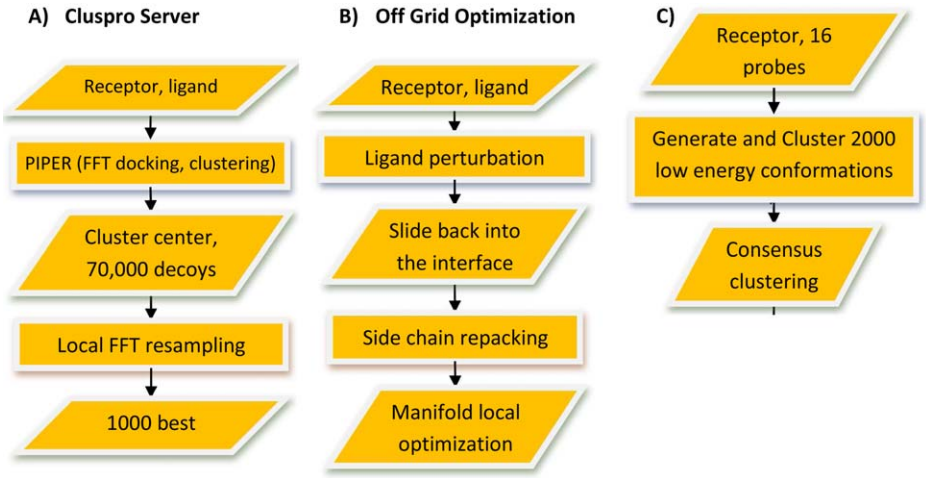

Figure 1. Flow charts for Cluspro server, off grid optimization and protein mapping protocols. a) Cluspro Server, b) off grid optimization, c) protein mapping.

protein-protein association (hydrophobicity driven, balanced and electrostatics driven).[34] Since the type of interaction primarily driving the association is generally not known before docking, a standard ClusPro run simultaneously performs docking with all three sets of coefficients (Fig. 1).[34] As will be described, we test resampling for type of complexes classified as "others" in the protein docking benchmark set.[37] This class includes complexes with highly variable properties that generally result in fairly challenging docking problem. To reduce the dependence of docking results on the selection of the energy weighting scheme, the 500 best scoring structures are retained for each of the three sets of weighting coefficients, and merged into the final set of 1500 structures.

The second step of ClusPro is clustering the 1500 structures using pairwise interface root mean square deviation (IRMSD) as the distance measure.[34,38] Clusters are ranked based on the number of structures that a ligand has within a (default) cluster radius of 9 Å IRMSD, and the ranked complexes are subjected to a straightforward (300 steps and fixed backbone) van der Waals minimization using the CHARMM potential to remove potential side chain clashes.[39]

As mentioned, in this work we explore refinement, and hence restrict consideration to the most near-native cluster. The region of interest, defined by this near-native cluster, is further restricted by removing the structures that are more than 10 Å IRMSD from the structure representing the center of the cluster. For comparison with our previous docking results,[19] we reduce the number of starting conformations to 1000 by selecting the 333 lowest energy structures from each of the three sets obtained by the different weighting coefficients, and add one more low energy structure. Measures are taken so that the final set has 1000 structures if any of the three sets has fewer than 333 structures.

Focused Resampling of the Near-Native Region. From a dense rotation set, rotations are chosen which are similar to the rotations of the 1000 structures retained from the initial docking. These rotations are then used for local resampling of the region by the PIPER program.[19] Translations were constrained to 10 Å distance from the geometric center of the structure representing the cluster center. Similarly to the initial docking, the conformations are docked using three sets of coefficients, and 1000 post-refinement conformations are retained.

Off-Grid Refinement with Rigid Body Energy Minimization and Side Chain Repacking. We have implemented an off-grid refinement based on a standard Monte Carlo Minimization (MCM) protocol.[40] Each MCM iteration consists of three steps: (1) the ligand conformation is (randomly) perturbed while keeping the receptor fixed; (2) we slide the proteins back into contact while repacking of interface side chains as described previously;[40] and (3) refine the resulting complex using a manifold-based local minimization algorithm,[41] allowing the side-chains to slightly move to off-rotamer positions. In contrast to the rigid body docking, steps 2 and 3 of MCM remove all steric clashes, and hence we can use a higher accuracy scoring function, which combines force-field and knowledge-based energy terms.[40] This MCM protocol will be used to further refine the structures generated by the focused sampling. As will be discussed, short MCM runs are sufficient to demonstrate the basic properties of the method, and hence we perform only five MCM iterations from each starting structure.

Benchmark Set of Complexes for Protein-Protein Docking. The docking, resampling, and off-grid refinement algorithms were applied to 49 protein pairs selected from the "other" type of complexes of version 4.0 of the docking benchmark set.[42] The "others" subset was chosen because it includes many complexes with limited shape complementarity and hence presents challenging problems for the refinement. Selected pairs of unbound proteins from this subset were docked. Although benchmark 4.0 includes 99 "other" type complexes, we considered only the 49 targets listed in Table 1 for which at least one of the clusters generated by ClusPro includes ten or more docked structures

Table 1. The number of docking decoys within the 5 Å IRMSD neighborhood of native complex at different stages of the refinement.

	PIPER				Local PIPER resampling			Off-grid optimization		
			<5 Å			<5 Å			<5 Å	
PDB	Center IRMSD	Total number	Number	Fraction	Total number	Number	Fraction	Total number	Number	Fraction
1a2k	4.411	1000	211	0.211	1000	299	0.299	1000	87	0.087
1akj	6.229	605	165	0.27	1000	183	0.183	1000	152	0.152
1atn	11.527	1000	20	0.02	1000	0	0	1000	4	0.004
1azs	2.968	1000	441	0.441	1000	718	0.718	1000	564	0.564
1b6c	4.004	1000	230	0.23	1000	435	0.435	1000	512	0.512
1buh	3.901	1000	102	0.1	1000	144	0.144	1000	182	0.182
1E96	4.732	1000	118	0.12	1000	240	0.24	1000	320	0.32
1eer	6.362	698	6	0.01	1000	4	0.004	1000	30	0.03
1f51	10.671	1000	67	0.067	1000	87	0.087	1000	127	0.127
1ffw	8.275	1000	47	0.047	1000	52	0.052	1000	59	0.059
1gla	4.77	1000	144	0.144	1000	108	0.108	1000	104	0.104
1gpw	3.308	1000	272	0.272	1000	491	0.491	1000	511	0.511
1grn	5.489	653	86	0.132	1000	234	0.234	1000	170	0.17
1h9d	10.325	581	0	0	1000	0	0	1000	0	0
1he1	6.458	961	144	0.15	1000	285	0.285	1000	201	0.201
1i2m	10.796	1000	45	0.045	1000	42	0.042	1000	32	0.032
1j2j	8.239	1000	5	0.005	1000	3	0.003	1000	34	0.034
1jk9	3.566	1000	216	0.216	1000	355	0.355	1000	377	0.377
1jwh	4.717	1000	367	0.367	1000	414	0.414	1000	457	0.457
1jzd	11.359	728	15	0.021	1000	66	0.066	1000	40	0.04
1k5d	14.162	1000	0	0	1000	0	0	1000	0	0
1k74	3.313	1000	244	0.244	1000	478	0.478	1000	559	0.559
1kxp	3.86	1000	170	0.17	1000	396	0.396	1000	535	0.535
1lfd	13.246	1000	0	0	1000	0	0	1000	0	0
1ml0	4.89	1000	382	0.382	1000	661	0.661	1000	705	0.705
1ofu	3.712	711	224	0.315	1000	573	0.573	1000	532	0.532
1r6q	11.425	1000	19	0.019	1000	8	0.008	1000	14	0.014
1rlb	6.353	1000	105	0.105	1000	234	0.234	1000	528	0.528
1rv6	13.402	1000	0	0	1000	0	0	1000	0	0
1syx	5.261	1000	154	0.154	1000	214	0.214	1000	192	0.192
1wq1	7.908	840	59	0.07	1000	94	0.094	1000	103	0.103
1xd3	3.064	1000	296	0.296	1000	334	0.334	1000	309	0.309
1xqs	7.809	1000	126	0.126	1000	149	0.149	1000	101	0.101
1z0k	3.448	1000	92	0.092	1000	80	0.08	1000	93	0.093
1z5y	4.012	1000	143	0.143	1000	203	0.203	1000	163	0.163
1zhi	7.511	1000	175	0.175	1000	113	0.113	1000	104	0.104
2a5t	13.89	982	0	0	1000	0	0	1000	0	0
2ayo	5.547	1000	307	0.307	1000	565	0.565	1000	473	0.473
2btf	7.172	1000	171	0.171	1000	267	0.267	1000	200	0.2
2cfh	4.755	947	205	0.216	1000	515	0.515	1000	597	0.597
2g77	7.172	1000	70	0.07	1000	75	0.075	1000	60	0.06
2hle	8.572	1000	47	0.047	1000	106	0.106	1000	216	0.216
2hrk	6.388	1000	122	0.122	1000	154	0.154	1000	254	0.254
2i9b	5.637	1000	0	0	1000	0	0	1000	0	0
2nz8	10.336	902	0	0	1000	0	0	1000	0	0
2ot3	9.889	806	0	0	1000	0	0	1000	0	0
3bp8	8.798	1000	0	0	1000	0	0	1000	0	0
3cph	7.891	1000	5	0.005	1000	1	0.001	1000	1	0.001
3d5s	3.991	1000	353	0.353	1000	666	0.666	1000	737	0.737

within 10 Å IRMSD from the corresponding bound structures. We note that the fraction of such structures, obtained without any additional information on the bound complex, agrees well with the about 50% overall success rates seen in the rounds of the CAPRI protein docking experiment.[22–26]

Protein mapping

Global Mapping Using the FTMap Server. FTMap places small organic molecules as probes on a grid around the target protein. For each probe, an FFT correlation based algorithm generates bound positions using a detailed energy expression as the scoring function.[10] For each probe we select the 2000 lowest energy poses which are further minimized off the grid using the CHARMM potential[39] with a continuum electrostatic model.[43] The minimized probe conformations are clustered and ranked on the basis of their Boltzmann averaged energies, and the six clusters with the lowest average energies are retained for each probe. The clusters of different probes are then clustered into consensus clusters (CCs). The hot spots

Table 2. Consensus clusters in global and focused mapping.

	PDB ID		Consensus clusters from mapping	
Name	Unbound	Bound	Global	Focused
VHL	3zrf	3zrc	1(16)	2(17), 5(13), 8(2)
Survivin	1e31	3uih	7(4)	2(22), 5(8)
PCNA	1rwz	1rxz	3(17)	2(17), 8(4)
KEAP1	1x2j	4iqk	2(21)	1(25), 4(17), 5(6), 6(3)

are identified by these consensus clusters, and the consensus cluster with the highest number of cluster centers within 4 Å is selected as the main hot spot. Consensus clusters with fewer probe clusters identify secondary hot spots. We emphasize that global mapping is performed on ligand-free protein structures whenever such are available, and no information related to ligand binding is required. It was shown for many proteins that the main hot spot generally overlaps with the most important moiety of the bound ligand, and the ligand extends into secondary hot spots.[10,29,31,44–48]

Focused Mapping. Based on the results of the global mapping, we select the consensus clusters of interest. The resampling involves the same FTMap algorithm as in the global mapping, but this time we retain the best 2000 structures we find within the region just defined rather than anywhere around the protein as in global mapping. The clustering and finding of consensus clusters proceed as in the global case. However, due to retaining higher numbers of probes in the region of interest, the mapping generally provides much more detailed information on the location and shape of the hot spots.

Benchmark Set for Protein Mapping. As stated, ligands of proteins generally bind at the main hot spot, and the collection of hot spots overlap with the ligand.[49] For this study, however, we focus on four systems where this is not the case (Table 2). In all these cases we consider sites that bind small ligands, developed for inhibiting protein-protein interactions, but are not located at the main binding hot spot of the target protein. Since most probes cluster at the main binding hot spots, the global mapping provides limited information on the secondary hot spots that are located at such weaker binding sites, although the latter are known to interact with specific ligands. It is emphasized that we always map the structure of ligand-free proteins, and the global mapping does not involve any assumption on the binding. All water molecules, ions, or any ligands are removed before mapping.

Results and Discussion

Protein-protein docking

Results for the 49 complexes are shown in Table 1. First we show the global PIPER results. These include the IRMSD of the structure at the center at the selected low energy cluster from the native structure, the number of initial docked structures that were used as the starting points for the resampling, and the number and fraction of structures below 5 Å IRMSD. As shown, in a few cases we have less than 1000 starting structures that satisfy our selection condition of being within 10 Å IRMSD from the center of the near-native cluster. The next columns in Table 1 list the number of structures selected after resampling, which is always 1000, and the number and fraction of structures below 5 Å IRMSD.

Figure 2 shows the fraction of structures below 5 Å IRMSD among the conformations generated by the global docking as provided by the ClusPro server, and the fraction of such structures after the refinement by focused resampling followed by off grid optimization. Supporting Information Figure S2 also shows the fraction when the off grid optimization is performed without the focused resampling. For nine targets the

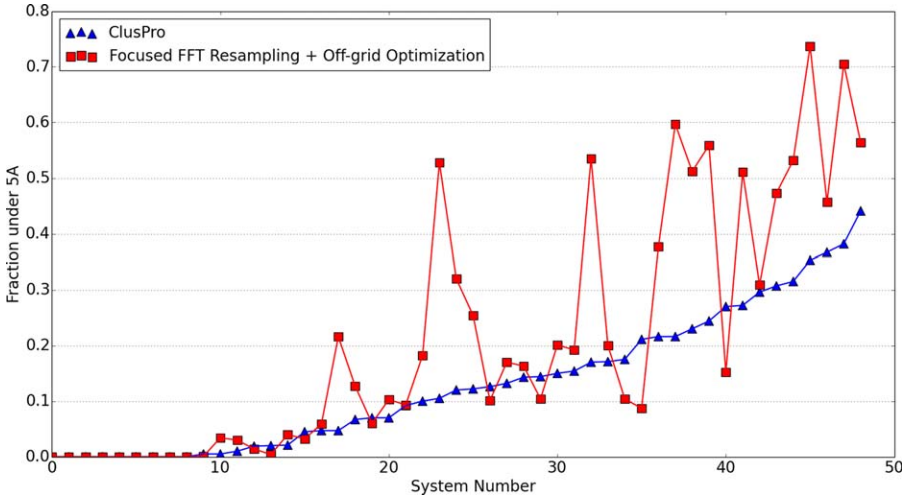

Figure 2. Fraction of docked structures with less than 5 Å IRMSD among the structures generated by ClusPro (blue triangles) and focused FFT resampling followed by off grid optimization (red squares) of the near-native cluster obtained by the global docking. Results are shown for the 49 complexes listed in Table 1. Additionally, results of off grid optimization without focused FFT resampling are shown in Supporting Information Figures S2 to S4 and Table S1. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

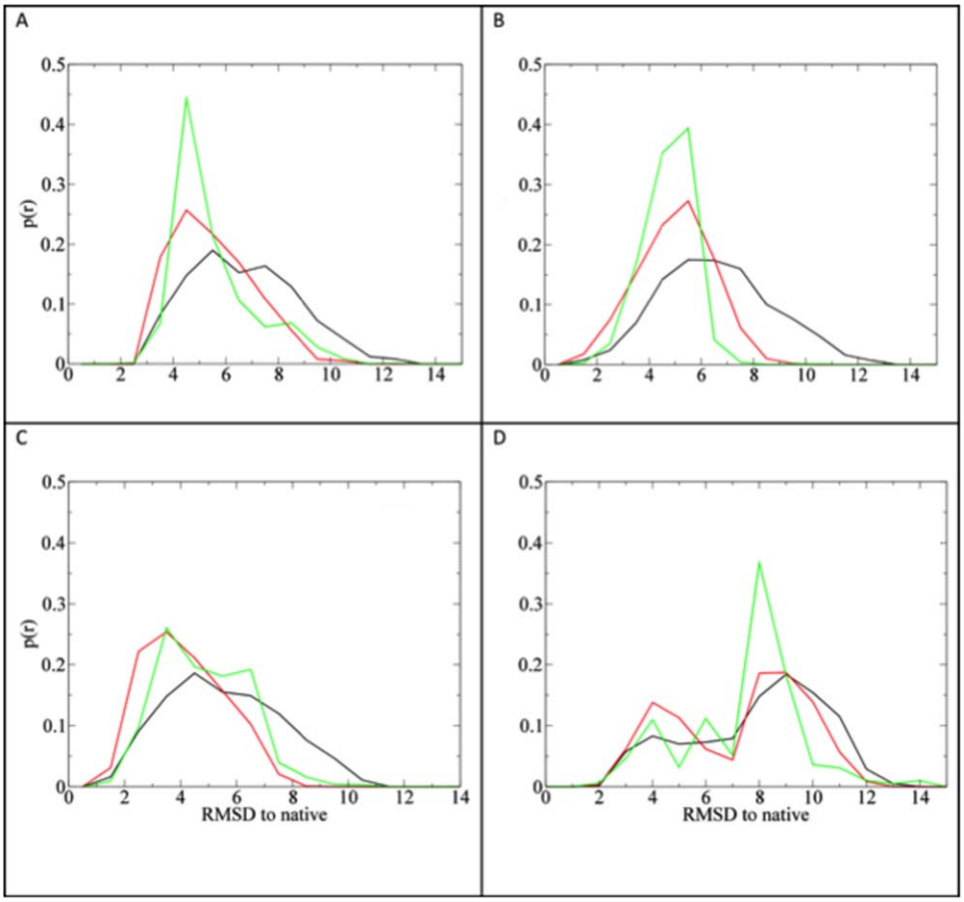

Figure 3. Distribution of IRMSD values from the native complex for the docked structures of four different complexes. For each complex, IRMSD distributions are shown for the results from the global docking using PIPER (black curve); after resampling the most near-native cluster of PIPER generated structures (red curve); and after offgrid refinement by five steps of the Monte Carlo minimization (MCM) MCM algorithm of the structures from the focused resampling (green curve). a) Complex of the cytoplasmic domain of the type I TGFα receptor and FKBP12, PDB ID: 1B6C; b) human PPARγ and RXRα ligand binding domains, PDB ID: 1K74; c) adenylyl cyclase in a complex with its stimulatory heterotrimeric G protein alpha subunit, PDB ID: 1AZS; and d) disulfide-linked complex between the N-terminal domain of the electron transfer catalyst DsbD and the cytochrome c biogenesis protein CcmG, PDB ID: 1Z5Y.

selected low energy clusters do not include any structure with less than 5 Å IRMSD. It is well known that docking methods may have difficulties for a number of different reasons.[50] In particular, rigid body methods such as PIPER do not work well if the association is accompanied by substantial backbone conformational change, which is the case for targets 2I9B, 2NZ8, and 2OT3.[42,51–53] In other cases the interface is too small and results in weak interaction, the target is part of a larger complex, or the complex is stabilized by strong electrostatic interactions that heavily depend on the conformation of some charged side chains.[50] The lack of low energy near-native structures implies that, according to the PIPER scoring function, the lowest energy regions are not close to the native state. As expected, resampling does not help in such cases, and no higher accuracy structures emerge (Fig. 2). However, resampling increases the fraction of structures with less than 5 Å IRMSD for 30 complexes that already have such structures from the global docking. The improvement is substantial for 19 complexes, where "substantial" is defined as a change of at least 0.05 in the fraction of structures below 5 Å IRMSD, i.e., adding at least 50 higher accuracy structures. Using the same definition, the resampling substantially decreases the fraction of near-native structures only for 3 of the 49 complexes.

Figure 2 emphasizes the substantial variations in the impact of focused mapping among the complexes. In fact, as already mentioned, focused mapping does not improve the results for the nine complexes for which the global PIPER search fails to find any structure below 5 Å IRMSD, whereas the improvements are substantial for a number of other complexes.

The results of off grid optimization without focused resampling are also presented in Supporting Information Figures S2–S4 and Table S1. According to these results, performing off grid optimization after focused rigid resampling always yields more near-native structures (with less than 5 Å IRMSD) than off grid optimization without the focused rigid resampling. This result demonstrates that resampling is useful.

Further insight on the differences between complexes can be obtained by examining how resampling affects the distributions of the docked structure's IRMSDs from the native bound structure. Results for four complexes are shown in Figure 3. Figure 3 also shows the impact of focused FFT resampling before off-grid optimization. The resampling significantly

increases the number of near-native structures and shifts the distribution toward the native structure for the complex between cytoplasmic domain of the type I TGFα receptor and FKBP12, PDB ID 1B6C (Fig. 3a), and also for the complex between human PPARγ and RXRα ligand binding domains, PDB ID 1K74 (Fig. 3b). Both these complexes have well-defined structures and relatively high affinity, and the scoring function used in PIPER and based on molecular mechanics has its global minimum close to the native structures. This is further emphasized by the result that the refinement using MCM further increases the fraction of structures within 5 Å IRMSD from the native state (Figs. 3a and 3b). The third system, adenylyl cyclase with its stimulatory heterotrimeric G protein alpha subunit, PDB ID 1AZS, forms a weaker complex as it is modulated by ATP *versus* ADP binding. Nevertheless, the focused resampling increases the number of structures below 5 Å IRMSD (Table 1) and shifts the distribution toward the native structure, substantially reducing the fraction of structures with IRMSD values in excess of 6 Å (Fig. 3c). Interestingly, the subsequent Monte Carlo minimization actually slightly reduces the number of structures below 5 Å, and according to Figure 3c most low energy structures are between 3.8 Å and 6.3 Å from the native, with a substantial fraction exceeding the selected threshold of 5 Å IRMSD. The fourth complex, formed by the N-terminal domain of the electron transfer catalyst DsbD and the cytochrome C biogenesis protein CcmG, is weak, and is stabilized by a disulfide bridge. For this fourth complex, most structures generated by the docking are between 8 Å and 11 Å IRMSD from the one in the x-ray structure (Fig. 3d). The resampling and particularly the Monte Carlo refinement shifts the distribution toward these large IRMSD values, indicating that the energy function, not accounting for the impact of the disulfide link, has its lowest minima far from the X-ray structure. Thus, depending whether or not the PIPER scoring function yields low values near the native state, different complexes exhibit very different distributions of IRMSD values. This statement generally also applies to the outcome of the refinement by the MCM algorithm, in spite of the local refinement that allows for the use of a more detailed energy function. We note that Table 1 and Figure 3 show the results of performing only five MCM steps from each of the structures generated by the focused resampling, and further analysis of refinement by MCM has been presented separately.[54]

According to Figure 2, resampling increases the number of near-native structures for the majority of complexes, and the important question is how these results can be used. As shown in Figures 3a–3c, in such cases the IRMSD distribution generally shifts toward the native state (Figures 3a–3c). Since the clustering method used in ClusPro identifies clusters with the highest density of structures, we expected that using a smaller clustering radius the cluster centers would be closer to the native structures than the centers of the original near-native clusters identified by the global docking. While this was true for a number of complexes, overall this strategy did not yield significant improvement.

In fact, as shown in Figure 3d, the cluster center may shift farther from the native state. However, there are several applications where the densely sampled energy landscape is useful. First, we have developed the medium-range optimization method SDU that can use these data for the construction of an initial underestimator function.[55] After finding the minimum of the underestimator, the method attempts to iteratively expand the region of interest toward lower energies by adding further sample points and updating the underestimator.[55] However, we admit that SDU is unlikely to improve the accuracy of docked structures if the energy minimum is far from the native complex. The second application is stability analysis, which is used if the near-native cluster cannot be identified among the ones produced by the docking. In this case focused sampling is applied to all potential clusters. Stability analysis is based on the hypothesis that clusters of near-native structures are located in broad energy funnels.

In our current methodology this hypothesis is tested by starting short Monte Carlo minimization (MCM) simulations from randomly selected structures of the cluster.[35] However, if the clusters are finely sampled, the MCM runs can be replaced by a more efficient approach based on stochastic roadmap simulations (SRSs) to determine the escape tendencies of the ligand from the low energy regions.[56] To perform SRS one needs first to build a network called stochastic roadmap. Once the roadmap is constructed, simple calculations can be used to determine escape times and thus the "stability" of each cluster. Although the method is extremely elegant, it requires very dense sampling, and hence it has rarely been used in applications. However, after resampling we will have all the energy values needed, and thus can set up and apply the method with little additional effort. The Markov chain model description, recently used for cluster discrimination in protein-protein docking, provides a similar approach and also can take advantage of the finely sampled energy landscape.

Protein mapping

Figure 4 shows both global and focused mapping results for the four proteins listed in Table 2. We discuss the results for each protein separately.

Von Hippel-Lindau (VHL) Protein. The von Hippel-Lindau (VHL) protein is the substrate recognition subunit of an E3 ubiquitin ligase.[57] The primary substrate of VHL is hypoxia-inducible factor 1α (HIF-1α), which binds in a shallow groove on the surface of the VHL protein. The VHL/HIF-1α is interaction is currently under investigation as a target for the treatment of chronic anemia associated with chronic kidney disease and chemotherapy.[58–61]

Global mapping identifies three CCs as main hot spots, all three with 16 probe clusters, in different regions of VHL as it interacts with a variety of proteins. All other CCs are weaker. One of the main hot spots overlaps with the central core of known inhibitors, the N-acetyl-L-Hydroxyproline N-methylamide (L-Hyp) moiety (Fig. 4a), but no hot spot overlaps with the rest of the inhibitor.[62] Having 16 probe clusters in a CC is the minimum number required for druggability,[29,49] in agreement with the observation that the binding sites in the VHL/HIF-1α interface are relatively flat and difficult to target with

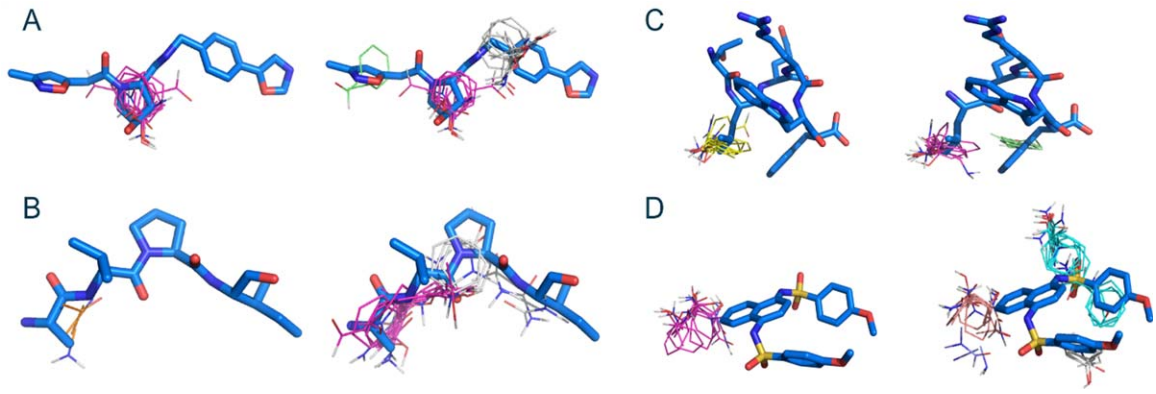

Figure 4. Results of global and focused mapping for the proteins listed in Table 2. For each target protein, a bound ligand is superimposed on the mapping results for reference (blue sticks). Each probe cluster in the consensus clusters is represented by the single lowest energy bound conformation (thin lines). All such representative probe conformations within each consensus cluster have the same color. In each figure, global mapping is shown on the left, focused mapping on the right. The PDB IDs of the unbound structures mapped and the bound structures containing the ligands, and the numbers of probe clusters are listed in Table 2. a) Mapping results for the von Hippel-Lindau (VHL) protein, superimposed on an inhibitor of the interaction between VHL and the hypoxia-inducible factor 1α (HIF-1α). b) Results for survivin, superimposed on a Smac/DIABLO N-terminal peptide. c) Results for the proliferating cell nuclear antigen (PCNA) superimposed on the PCNA-binding motif of FEN-1. d) Results for KEAP1, superimposed on a small molecular inhibitor of the interaction the KEAP1 Kelch domain and Nrf2.

small molecule inhibitors. Nevertheless, by using focused FFT resampling we are able to detect additional hot spots in the vicinity of this CC, both of which are used by the inhibitor. These CCs are CC5 with 13 probe clusters and CC8 with 2 probe clusters (Table 2 and Fig. 4a). In agreement with the existence of such additional hot spots, recent optimization of the inhibitor on the two sides of the central L-Hyp moiety resulted in several inhibitors with high nanomolar affinities.[59] Nevertheless, the importance of the main hot spot is shown by the fact that all these inhibitors retain the L-Hyp central core.[62]

Survivin. Survivin is an anti-apoptotic protein that is overexpressed in most human tumors.[63,64] Downregulation of survivin expression or its function can sensitize tumor cells to chemotherapeutics.[65] Survivin belongs to the inhibitor of apoptosis (IAP) protein family. Proteins within the IAP family share a common baculovirus IAP repeat (BIR) domain that contains a peptide-binding groove on the surface, in many cases interacting with the Smac/DIABLO N-terminal peptide.[57] However, survivin has much lower affinity for the peptide than other members of the IAP family, and the protein has a much stronger binding site for small molecules at the dimerization domain, where the main hot spot is found.[63,66] Within the peptide-binding groove, global mapping finds only CC7 with four probe clusters, indicating a very weak binding site.[49] CC7 overlaps with the Ala1 residue of the N-terminus of the Smac/DIABLO peptide (Fig. 4b). The majority of probe clusters are located in the dimerization domain, which is known to include a binding site for high affinity ligands.[63,66] Nevertheless, survivin very weakly binds the Smac/DIABLO peptide, and focused resampling of the peptide binding region yields improvement in terms of better identifying the ligand coverage area. The two new CCs, CC2 and CC5 with 22 and 8 probe clusters, respectively, overlap with the peptide (Table 2 and Fig. 4b).

Proliferating Cell Nuclear Antigen (PCNA). PCNA is a highly conserved eukaryotic protein that is essential for DNA replication and for both DNA excision and mismatch repair.[67] PCNA has recently been identified as a target for the binding of several proteins. All these interactions occur through the conserved motif QXX(L/I/M)XX(F/Y/W)(F/Y), contacting the same site on PCNA.[67] In particular, the replication endonuclease FEN-1 is recognized via its LXXFF motif.[68] Global mapping of PCNA places CC3 with 17 probe clusters in the PCNA pocket that binds Leu7 in the PCNA-binding motif of FEN-1, but the Phe binding pocket is not detected. Focused FFT resampling finds another consensus cluster, CC8 with 4 probe clusters, in the pocket that binds Phe11 of FEN-1 (Fig. 4c).

KEAP1. Kelch-like ECH-associated protein 1 (KEAP1) is a facilitator of ubiquitination that negatively regulates the transcription factor, Nrf2 (Nuclear factor-erythroud 2-related factor 2).[69] Nrf2 directs the transcription of protective genes related to oxidative stresses and thus a small molecule inhibitor that blocks the KEAP1/Nrf2 interaction may be relevant to a number of diseases, including cancer, Alzheimer's disease, and arthritis.[70] KEAP1 possesses a large positively charged central cavity and it is within this cavity that most of the mapping probes are found upon initial mapping with FTMap. However, a fairly high affinity ligand (IC50 = 2.7 μM) is found to bind only at the bottom of this central cavity and has minimal overlap with the probe clusters; CC2 with 21 probe clusters is the closest CC to this ligand (Fig. 4d). Focused FFT resampling of this region of interest yields improved results, identifying 4 CCs that overlap with the ligand: CC1 (25 probe clusters), CC4 (17 probe clusters), CC5 (6 probe clusters), and CC6 (3 probe clusters).

One goal of mapping is to obtain information on the regions of the binding site that are most likely to accommodate ligands.[71] Such analysis is important, because we map ligand-free proteins, and thus the results provide information

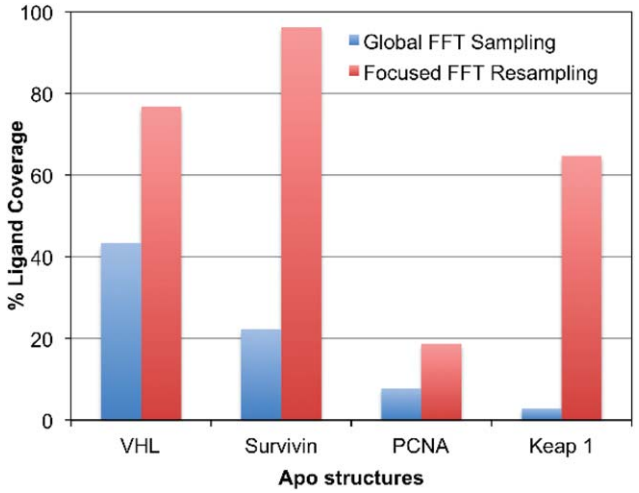

Figure 5. Percent ligand coverage by hot spots from global and focused mapping. As also shown in Figures 4a, 4b, and 4d, the probe positions obtained by focused mapping delineate the ligand binding sites much better in VHL, surviving, and KEAP1. The coverage is not as good for PCNA, but the peptide bound to this protein substantially extends beyond the binding site that accommodates only two hydrophobic side chains. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

on the size and shape of potential ligands. The success of the mapping can be assessed by ligand coverage, which measures the overlap between the region occupied by the probes and the region occupied by a ligand of interest. Therefore, to determine the impact of focused FFT resampling method, we calculated the percent ligand coverage for each target, defined as the number of ligand heavy atoms covered by probes, divided by the number of total ligand heavy atoms. A ligand atom is considered covered by probes if it is within a radius of 2 Å of any heavy atom of any probe in a consensus cluster. The results for the four protein targets listed in Table 2 are shown in Figure 5. For all four cases there is substantial improvement in the percent ligand coverage using the focused FFT resampling technique. Indeed, as shown in Figure 4, the hot spots obtained by focused mapping delineate fairly well the ligands bound to three targets, i.e., the von Hippel-Lindau (VHL) protein, survivin, and KEAP1. For PCNA (PDB ID 1rwz) the percent ligand coverage is only 18.75% even when using the results of the focused FFT resampling, which, while double the percent ligand coverage obtained using the global mapping (7.8%), is still quite low. However, in this case, a very large fraction of the pentapeptide ligand extends beyond the binding site, with only Leu and Phe occupying binding pockets. Therefore, we expect that only a relatively small fraction of the ligand overlaps with hot spots, resulting in moderate ligand coverage. Nevertheless, the focused resampling still improves the accuracy of the information provided by the mapping.

Conclusions

Due to the pairwise character of the potential function terms representing nonbonded interactions in most molecular mechanics force fields, the interaction energy between two molecules can be written in the form of a correlation function. The fast Fourier transform (FFT) correlation algorithm provides an extremely efficient approach to evaluating such energy expressions. The algorithm has been used with great success both for protein-protein docking and protein mapping, two applications that require global sampling of the energy landscape. In view of the computational efficiency of the FFT algorithm, in this work we explored whether the same approach can provide additional information when applied to selected regions of interest in a focused manner. In protein-protein docking, resampling of near-native clusters, identified by the global docking, substantially increased the number of docked structures within 5 Å IRMSD from the native structure for the majority of complexes considered. However, clustering of these conformations, the usual method of selecting models in ClusPro (see Methods), did not lead to significant overall improvement. Nevertheless, we discuss that several refinement and cluster discrimination algorithm can take advantage of the finely sampled energy landscape. In protein mapping the focused resampling of the hot spot regions identified by the global mapping provide information on secondary hot spots that are frequently missed by the initial mapping. This additional information better shows the regions that participate in ligand binding. In addition, as shown previously with applications to fragment based drug design, the knowledge of secondary hot spots facilitates the extension of fragment hits into larger and higher affinity ligands.[31]

Keywords: protein-protein docking · protein mapping · fast Fourier transform · systematic sampling · binding hot spots

How to cite this article: A. B. Mamonov, M. Moghadasi, H. Mirzaei, S. Zarbafian, L. E. Grove, T. Bohnuud, P. Vakili, I. Ch. Paschalidis, S. Vajda, D. Kozakov. *J. Comput. Chem.* **2016**, *37*, 961–970. DOI: 10.1002/jcc.24273

Additional Supporting Information may be found in the online version of this article.

[1] I. Halperin, B. Ma, H. Wolfson, R. Nussinov, *Proteins* **2002**, *47*, 409.
[2] G. R. Smith, M. J. Sternberg, *Curr. Opin. Struct. Biol.* **2002**, *12*, 28.
[3] C. J. Camacho, S. Vajda, *Curr. Opin. Struct. Biol.* **2002**, *12*, 36.
[4] D. W. Ritchie, *Curr. Protein Peptide Sci.* **2008**, *9*, 1.
[5] C. Mattos, D. Ringe, *Nat. Biotechnol.* **1996**, *14*, 595.
[6] K. N. Allen, C. R. Bellamacina, X. C. Ding, C. J. Jeffery, C. Mattos, G. A. Petsko, D. Ringe, *J. Phys. Chem. Us.* **1996**, *100*, 2605.
[7] P. J. Hajduk, J. R. Huth, S. W. Fesik, *J. Med. Chem.* **2005**, *48*, 2518.
[8] W. L. DeLano, *Curr. Opin. Struct. Biol.* **2002**, *12*, 14.
[9] S. Vajda, D. Kozakov, *Curr. Opin. Struct. Biol.* **2009**, *19*, 164.
[10] R. Brenke, D. Kozakov, G. Y. Chuang, D. Beglov, D. Hall, M. R. Landon, C. Mattos, S. Vajda, *Bioinformatics* **2009**, *25*, 621.
[11] E. Katchalski-Katzir, I. Shariv, M. Eisenstein, A. A. Friesem, C. Aflalo, I. A. Vakser, *Proc. Natl. Acad. Sci. USA* **1992**, *89*, 2195.
[12] D. W. Ritchie, G. J. L. Kemp, *Proteins* **2000**, *39*, 178.
[13] M. J. Sternberg, H. A. Gabb, R. M. Jackson, G. Moont, *Methods Mol. Biol.* **2000**, *143*, 399.
[14] I. A. Vakser, *Biopolymers* **1996**, *39*, 455.
[15] I. A. Vakser, *Protein Eng.* **1996**, *9*, 37.

[16] H. A. Gabb, R. M. Jackson, M. J. E. Sternberg, *J. Mol. Biol.* **1997**, *272*, 106.
[17] R. Chen, L. Li, Z. Weng, *Proteins* **2003**, *52*, 80.
[18] S. R. Comeau, D. W. Gatchell, S. Vajda, C. J. Camacho, *Bioinformatics* **2004**, *20*, 45.
[19] D. Kozakov, R. Brenke, S. R. Comeau, S. Vajda, *Proteins* **2006**, *65*, 392.
[20] J. Mintseris, B. Pierce, K. Wiehe, R. Anderson, R. Chen, Z. Weng, *Proteins* **2007**, *69*, 511.
[21] J. Janin, K. Henrick, J. Moult, L. T. Eyck, M. J. Sternberg, S. Vajda, I. Vakser, S. J. Wodak, *Proteins* **2003**, *52*, 2.
[22] R. Mendez, R. Leplae, L. De Maria, S. J. Wodak, *Proteins* **2003**, *52*, 51.
[23] R. Mendez, R. Leplae, M. F. Lensink, S. J. Wodak, *Proteins* **2005**, *60*, 150.
[24] M. F. Lensink, S. J. Wodak, *Proteins* **2010**, *78*, 3073.
[25] M. F. Lensink, R. Mendez, S. J. Wodak, *Proteins* **2007**, *69*, 704.
[26] M. F. Lensink, S. J. Wodak, *Proteins* **2013**, *81*, 2082.
[27] C. H. Ngan, D. R. Hall, B. Zerbe, L. E. Grove, D. Kozakov, S. Vajda, *Bioinformatics* **2012**, *28*, 286.
[28] M. R. Landon, D. R. Lancia, Jr., J. Yu, S. C. Thiel, S. Vajda, *J. Med. Chem.* **2007**, *50*, 1231.
[29] D. Kozakov, D. R. Hall, G. Y. Chuang, R. Cencic, R. Brenke, L. E. Grove, D. Beglov, J. Pelletier, A. Whitty, S. Vajda, *Proc. Natl. Acad. Sci. USA* **2011**, *108*, 13528.
[30] D. A. Erlanson, *Top. Curr. Chem.* **2012**, *317*, 1.
[31] D. R. Hall, C. H. Ngan, B. S. Zerbe, D. Kozakov, S. Vajda, *J. Chem. Inf. Model* **2012**, *52*, 199.
[32] D. Joseph-McCarthy, A. J. Campbell, G. Kern, D. Moustakas, *J. Chem. Inf. Model* **2014**, *54*, 693.
[33] S. Vajda, D. R. Hall, D. Kozakov, *Proteins* **2013**, *81*, 1874.
[34] D. Kozakov, D. Beglov, T. Bohnuud, S. E. Mottarella, B. Xia, D. R. Hall, S. Vajda, *Proteins* **2013**, *81*, 2159.
[35] D. Kozakov, O. Schueler-Furman, S. Vajda, *Proteins* **2008**, *72*, 993.
[36] S. Vajda, F. Guarnieri, *Curr. Opin. Drug Discov. Dev.* **2006**, *9*, 354.
[37] T. Vreven, I. H. Moal, A. Vangone, B. G. Pierce, P. L. Kastritis, M. Torchala, R. Chaleil, B. Jiménez-García, P. A. Bates, J. Fernandez-Recio, A. M. Bonvin, Z. Weng, *J. Mol. Biol.* **2015**, *427*, 3031.
[38] D. Kozakov, K. H. Clodfelter, S. Vajda, C. J. Camacho, *Biophys. J.* **2005**, *89*, 867.
[39] B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan, M. Karplus, *J. Comput. Chem.* **1983**, *4*, 187.
[40] M. Moghadasi, H. Mirzaei, A. Mamonov, P. Vakili, S. Vajda, I. Paschalidis, D. Kozakov, *J. Chem. Inf. Model* **2015**, *55*, 872.
[41] H. Mirzaei, S. Zarbafian, E. Villar, S. Mottarella, D. Beglov, S. Vajda, I. C. Paschalidis, P. Vakili, D. Kozakov, *J. Chem. Theory Comput.* **2015**, *11*, 1063.
[42] H. Hwang, T. Vreven, J. Janin, Z. P. Weng, *Proteins* **2010**, *78*, 3111.
[43] M. Schaefer, M. Karplus, *J. Phys. Chem. Us.* **1996**, *100*, 1578.
[44] M. R. Landon, R. L. Lieberman, Q. Q. Hoang, S. Ju, J. M. Caaveiro, S. D. Orwig, D. Kozakov, R. Brenke, G. Y. Chuang, D. Beglov, S. Vajda, G. A. Petsko, D. Ringe, *J. Comput. Aid. Mol. Des.* **2009**, *23*, 491.
[45] G. Y. Chuang, D. Kozakov, R. Brenke, D. Beglov, F. Guarnieri, S. Vajda, *Biophys. J.* **2009**, *97*, 2846.
[46] D. Kozakov, G. Y. Chuang, D. Beglov, S. Vajda, *Trends Biochem. Sci.* **2010**, *35*, 471.
[47] D. H. Hall, L. E. Grove, C. Yueh, C. H. Ngan, D. Kozakov, S. Vajda, *J. Am. Chem. Soc.* **2011**, *133*, 20668.
[48] G. Buhrman, C. O'Connor, B. Zerbe, B. M. Kearney, R. Napoleon, E. A. Kovrigina, S. Vajda, D. Kozakov, E. L. Kovrigin, C. Mattos, *J. Mol. Biol.* **2011**, *413*, 773.
[49] D. Kozakov, D. R. Hall, R. L. Napoleon, C. Yueh, A. Whitty, S. Vajda, *J. Med. Chem.* **2015**, *58*, 9063.
[50] S. Vajda, *Proteins* **2005**, *60*, 176.
[51] R. Chen, J. Mintseris, J. Janin, Z. Weng, *Proteins* **2003**, *52*, 88.
[52] J. Mintseris, K. Wiehe, B. Pierce, R. Anderson, R. Chen, J. Janin, Z. Weng, *Proteins* **2005**, *60*, 214.
[53] H. Hwang, B. Pierce, J. Mintseris, J. Janin, Z. Weng, *Proteins* **2008**, *73*, 705.
[54] Y. Shen, R. Brenke, D. Kozakov, S. R. Comeau, D. Beglov, S. Vajda, *Proteins* **2007**, *69*, 734.
[55] Y. Shen, I. Paschalidis, P. Vakili, S. Vajda, *PLoS Comput. Biol.* **2008**, *4*, e1000191.
[56] M. S. Apaydin, C. E. Guestrin, C. Varma, D. L. Brutlag, J. C. Latombe, *Bioinformatics* **2002**, *18*, 18.
[57] J. Du, A. E. Kelly, H. Funabiki, D. J. Patel, *Structure* **2012**, *20*, 185.
[58] D. M. Dias, I. Van Molle, M. G. J. Baud, C. Galdeano, C. F. G. C. Geraldes, A. Ciulli, *ACS Med. Chem. Lett.* **2014**, *5*, 23.
[59] C. Galdeano, M. S. Gadd, P. Soares, S. Scaffidi, I. Van Molle, I. Birced, S. Hewitt, D. M. Dias, A. Ciulli, *J. Med. Chem.* **2014**, *57*, 8657.
[60] I. Van Molle, A. Thomann, D. L. Buckley, E. C. So, S. Lang, C. M. Crews, A. Ciulli, *Chem. Biol.* **2012**, *19*, 1300.
[61] D. L. Buckley, I. Van Molle, P. C. Gareiss, H. S. Tae, J. Michel, D. J. Noblin, W. L. Jorgensen, A. Ciulli, C. M. Crews, *J. Am. Chem. Soc.* **2012**, *134*, 4465.
[62] D. Kozakov, D. R. Hall, S. Jehle, L. Luo, S. O. Ochiana, E. V. Jones, M. Pollastri, K. N. Allen, A. Whitty, S. Vajda, *Proc. Natl. Acad. Sci. USA* **2015**, *112*, E2585.
[63] M. D. Wendt, C. H. Sun, A. Kunzer, D. Sauer, K. Sarris, E. Hoff, L. P. Yu, D. G. Nettesheim, J. Chen, S. Jin, K. M. Comess, Y. H. Fan, S. N. Anderson, B. Isaac, E. T. Olejniczak, P. J. Hajduk, S. H. Rosenberg, S. W. Elmore, *Bioorg. Med. Chem. Lett.* **2007**, *17*, 3122.
[64] N. K. Sah, Z. Khan, G. J. Khan, P. S. Bisen, *Cancer Lett.* **2006**, *244*, 164.
[65] R. K. Kanwar, C. H. Cheung, J. Y. Chang, J. R. Kanwar, *Curr. Med. Chem.* **2010**, *17*, 1509.
[66] L. Chantalat, D. A. Skoufias, J. P. Kleman, B. Jung, O. Dideberg, R. L. Margolis, *Mol. Cell* **2000**, *6*, 183.
[67] E. Warbrick, *Bioessays* **1998**, *20*, 195.
[68] B. R. Chapados, D. J. Hosfield, S. Han, J. Z. Qiu, B. Yelent, B. H. Shen, J. A. Tainer, *Cell* **2004**, *116*, 39.
[69] R. S. Balaban, S. Nemoto, T. Finkel, *Cell* **2005**, *120*, 483.
[70] D. Marcotte, W. K. Zeng, J. C. Hus, A. McKenzie, C. Hession, P. Jin, C. Bergeron, A. Lugovskoy, I. Enyedy, H. Cuervo, D. P. Wang, C. Atmanene, D. Roecklin, M. Vecchi, V. Vivat, J. Kraemer, D. Winkler, V. Hong, J. H. Chao, M. Lukashev, L. Silvian, *Bioorg. Med. Chem.* **2013**, *21*, 4011.
[71] D. Kozakov, L. E. Grove, D. R. Hall, T. Bohnuud, S. E. Mottarella, L. Luo, B. Xia, D. Beglov, S. Vajda, *Nat. Protoc.* **2015**, *10*, 733.

Received: 29 March 2015
Revised: 31 August 2015
Accepted: 26 September 2015
Published online on 2 February 2016