PCCP

Accepted Manuscript

ROYAL SOCIETY OF CHEMISTRY

www.rsc.org/pccp

A New Scoring Function for Protein-Protein Docking that Identifies Native Structures with Unprecedented Accuracy

*Irina S. Moreira,[‡] João M. Martins,[‡,†] João T. S. Coimbra,[‡] Maria J. Ramos, Pedro A. Fernandes**

REQUIMTE/Departamento de Química e Bioquímica, Faculdade de Ciências da Universidade do Porto, Rua do Campo Alegre s/n, 4169-007 Porto- Portugal

ABSTRACT

Protein-protein (P-P) 3D structures are fundamental for structural biology and drug discovery. However, most of them have never been determined. Many docking algorithms were developed for that purpose, but they have a very limited accuracy in generating native-like structures and identifying the most correct one, in particular when a single answer is asked for. With such a low success rate it is difficult to point out one docked structure as being native-like. Here we present a new, high accuracy, scoring method to identify the 3D structure of P-P complexes among a set of trial poses. It incorporates Alanine scanning mutagenesis experimental data that needs to be obtained *a priori*. The scoring scheme works by matching the computational and the

experimental Alanine scanning mutagenesis results. The size of the trial P-P interface area is also taken into account. We show that the method ranks the trial structures and identifies the native-like structures with unprecedented accuracy (~94%), providing the correct P-P 3D structures that biochemists and molecular biologists need to pursue their studies. With such success rate, the bottleneck of Protein-Protein Docking moves from the scoring to the searching algorithms.

INTRODUCTION

The 3D structure of protein-protein (P-P) complexes is of the utmost importance for molecular biology and drug discovery, as these interactions form the basis of many phenomena. Most of the structures of the P-P complexes have never been solved, due to inherent difficulties of P-P structural determination by X-ray crystallography or NMR. There is an urgent request of computational procedures capable of reliably generating and identifying 3D P-P structures.[1] Protein docking is a method to assemble two separate proteins into their biologically relevant complex. It is the method of choice for this task nowadays.[2] It consists of two essential steps: search and scoring. The search step generates trial P-P geometries (poses). Usually, it begins by treating the proteins as rigid bodies and searches the translational and rotational 6D space to identify a set of promising poses using simple scoring functions, with shape complementarity playing a major role.[3] Full atomic detail is added subsequently, flexibilizing side-chains and, eventually, the backbone, with the help of energy-based scoring functions. If biological information on the location of the P-P interface is available, it should also be used as early as possible to direct the search.[4] The final purpose of the search stage is to find a set of poses as close as possible to the native pose and, ideally, with the true native pose included among them.

The next stage is scoring, which aims at ranking the poses according to their thermodynamic stability and identifying the most stable one, the true native structure (note that there are cases where the native structure is not the most stable, but instead a kinetically-trapped metastable structure, but these are not common). Scoring functions generally involve solvation potentials, empirical atom–atom or residue–residue contact energies or frequencies, and continuum electrostatics, among other techniques.[5] The problem is that the existent scoring functions neither provide an accurate identification of a native-like solution, nor do they rank correctly the trial poses. The community-wide experiment on the Critical Assessment of Predicted Interactions (CAPRI, http://www.ebi.ac.uk/msd-srv/capri) has been benchmarking the performance of protein-protein docking algorithms since 2001. The reports made over the years[6-8] show that the field is still in an early stage of development. If a molecular biologist solves the 3D structures of two proteins and wants to predict the correct 3D structure of the complex between them, using current docking algorithms, the probability of getting the right answer is exceedingly low. This means that current docking algorithms are still incapable of providing the answers that they were developed for, even though progresses are being made by very able people in the field.[6-8]

The scoring step is the subject of this work. We will show a way to identify the native pose among a group of trial poses with high accuracy and reliability. For that purpose we have incorporated our computational Alanine scanning mutagenesis method (cASM)[9, 10] in two scoring functions, which we called SF_{ASM} (scoring function of ASM) and $SF_{ASM:iA}$ (scoring function of ASM and interface Area, iA). Alanine scanning mutagenesis (ASM) is the trendiest method for mapping P-P interfaces as it measures the free energy contributions of individual side-chains to protein binding.[11] It identifies hot-spots (HS), which are the residues responsible for complex formation. We will show that a scoring function that compares the cASM results

with experimental ASM results (expASM), together with the value for the buried surface area upon complex formation (iA), can identify the native pose among a large set of decoys, with reliability and an accuracy much superior to any existing scoring function. This does not mean that our experiment-driven scoring functions are "better" than the many excellent ones available. Such comparison does not make sense, because the available scoring functions do not use or need expASM data. Our scoring functions use additional biochemical data that exists or alternatively it must be obtained experimentally *a priori*. Despite the extra effort, in the latter case, the impact of the result (*i.e.*, a reliable native P-P structure) completely pays off for the effort.

METHODS

Data-set. Our dataset (Table 1) was composed of 18 complex structures. Both X-ray structures and experimental values for the binding free energy changes upon Alanine mutation exist for all of them. The set includes four unbound-unbound cases (*i.e.*, the X-ray structure of the complex and of both isolated monomers is available), seven unbound-bound (*i.e.*, the X-ray structure of the complex and of just one of the isolated monomers is available) and seven bound-bound structures (*i.e.*, only X-ray structure of the complex is available or is used). The set was chosen to be diverse in nature and in size, in hydrophobicity, and in the number of hot-spots present. Benchmarks of docking algorithms usually use a larger number of protein complexes. Here, the size of the set was limited by the availability of accurate experimental results for Alanine scanning mutagenesis. Despite being not very large, its size is more than enough to illustrate the performance of the method.

Table 1. Data-set of the 18 protein-protein complexes under study. U:U stands for unbound-unbound, U:B stands for unbound-bound, B:B stands for bound-bound. The number of hot spots (HS) is given according to three typical definitions: residues that, upon being mutated by Alanine, increase the binding free energy (i) by > 4 kcal/mol, (ii) by > than 2 kcal/mol and (iii) by > 1.5kcal/mol. The total number of mutated residues (no. mut.), the X-ray resolution and the protein families following the SCOP[12] classification, is also given.

complex (PDBID)	resolution / Å	monomer A (PDBID)	monomer B (PDBID)	monomer A / monomer B protein family	type	no. HS (i)	no. HS(ii)	no. HS (iii)	no. mut.	ref.
1BRS[13]	2.00	Barnase (1A2P)[14]	Barstar (1A19)[15]	bacterial ribonucleases / barstar-related	U:U B:B	6	9	10	14	16, 17
1VFB[18]	1.80	Igg1 Kappa D1.3 FV (1VFA)[18]	Hen Egg white lysozyme (2VB1)[19]	V set domains / C-type lysozyme	U:U B:B	1	4	5	27	20, 21
1DFJ[22]	2.50	RNaseA (3EUX)[23]	RNaseA Inhibitor (2BNH)[24]	ribonuclease A-like / 28-residue LRR	U:U B:B	1	4	5	14	25, 26
1FLT[27]	1.70	VEGF (2VPF)[28]	FLT-1 (1QSV)[29]	platelet-derived growth factor-like / I set domains	U:U	0	4	5	20	30
1DVF[31]	1.90	Igg1 Kappa D1.3 FV (1VFA)[18]	Igg1 Kappa E5.2 FV (1DVF)[31]	V set domains / V set domains	U:B	5	8	16	22	20, 32
2EKS[33]	2.00	HyHEL-10 (2EKS)[33]	HEL (2VB1)[19]	- / C-type lysozyme	U:B	3	5	6	16	34, 35
1FCC[36]	3.20	Igg1 FC (1H3T)[37]	Streptococcal protein G (1FCC)[36]	C1 set domains / immunoglobulin-binding domains	U:B B:B	1	4	4	8	38
1FQ9[39]	3.00	Fibroblast Growth Factor 2[40]	Fibroblast Growth Factor Receptor I[39]	FGF / I set domains	U:B B:B	1	4	4	18	41, 42
1A22[43]	2.60	hGH (2AEW)[44]	hGHbp (1A22)[43]	Fibronectin type III / long-chain cytokines	U:B	0	3	5	23	45-47
1F47[48]	1.95	Bacterial cell division ZipA (1F46)[48]	Ftsz (1F47)[48]	cell division protein ZipA, C-terminal domain / -	U:B B:B	0	3	4	9	48
1IAR[49]	2.30	IL-4 (1RCB)[50]	IL4bp (1IAR)[49]	short-chain cytokines / fibronectin type III	U:B	0	2	3	18	51
1EMV[52]	1.70	Im9 (1EMV)[52]	Colicin E9 (1EMV)[52]	Colicin E immunity proteins / HNH-motif	B:B	2	6	7	25	53

Generation of trial poses - the search step. We used HADDOCK (High Ambiguous Driven DOCKing) for the "search" as it one of the most widely used softwares, has a high number of citations per year and performs consistently well in reputable tests.[2] HADDOCK starts with a randomization of orientations and rigid body energy minimization (1000 solutions), followed by semi-rigid simulated annealing in torsion angle space (200 solutions), and final refinement in Cartesian space with explicit solvent (200 solutions). It uses biological information to drive the docking by introducing AIRs (Ambiguous Interaction Restrains).[54] We have used the known HS residues (2.0 kcal/mol cutoff) as the "active residues" in HADDOCK in order to drive the search stage to generate native-like structures as much as possible. The final 200 solutions were evaluated according to their i-RMSDs (Root-Mean-Square-Deviations of the Interface in comparison with the X-ray structure) and fnat (fraction of native contacts). We have used the following classification for the 200 structures, inspired in the CAPRI criteria:[6-8] High (i-RMSD lower than 1 Å and fnat higher than 0.50), Medium (i-RMSD between 1 and 2 Å and fnat between 0.30 and 0.50), Acceptable (i-RMSD between 2 and 4 Å and fnat between 0.1 and 0.3) and Incorrect (i-RMSD higher than 4 Å or fnat lower than 0.1). We have used an in house script to reduce the 200 solutions to a smaller subset of 20 structures, which is more manageable for computational Alanine scanning calculations. The script automatically selected 5 structures for each of the 4 categories mention before. In the cases that the criterion was not fulfilled, we automatically selected a higher number of solutions for the remaining categories.

Generating the two new scoring functions for Protein-Protein docking. The MM-PBSA (Molecular Mechanics Poisson Boltzmann Surface Area) script[55] integrated into the AMBER 10.0 package[56] was used to calculate the binding free energy difference upon Alanine mutation.

It combines a continuum approach to model solvent interactions with a MM-based approach to atomistically model protein-protein interactions. This provides speed and accuracy and has been quite used in the last years.[9, 10, 55, 57-66] The MM-PBSA approach first developed by Massova et al.,[55] was improved by Moreira et al.[10] through the introduction of residue-specific dielectric constants to be used in ASM, and can now be applied for this task with an accuracy of 1 kcal/mol. It was shown that the use of residue-specific dielectric constants in ASM calculations makes MM-PBSA to be as accurate as more time consuming methods such as thermodynamic integration.[67] The mutant complexes were generated by a single truncation of the mutated side chain, replacing C_γ with a hydrogen atom and setting the C_β-H direction to that of the former C_β-C_γ. The difference in binding free energy upon Alanine mutation ($\Delta\Delta G$) is defined as the difference between the mutant and wild type complexes defined as:

$$\Delta\Delta G = \Delta G_{cpx-mutant} - \Delta G_{cpx-wild\ type} \quad (1)$$

Typical contributions to the free energy include the internal energy (bond, dihedral, and angle), the electrostatic and the van der Waals interactions, the free energy of polar solvation, the free energy of nonpolar solvation, and the entropic contribution:

$$G_{molecule} = E_{internal} + E_{electrostatic} + E_{vdW} + G_{polar\ solvation} + G_{non-polar\ solvation} - TS \quad (2)$$

Although, molecular dynamics proved to give the best results, as it is computational expensive to be applied to every docking trial pose, it was not used. Instead, we used a 1,000 steps minimization in a water box of 10 Å, which has already demonstrated to give reasonable

results.[10] For the calculations of relative free energies between closely related complexes, it is assumed that the total entropic term in equation 2 is negligible as the partial contributions essentially cancel each other.[64] The first three terms of equation 2 were calculated with no cutoff. The $G_{polar\ solvation}$ was calculated by solving the Poisson-Boltzmann equation with the software DELPHI.[68, 69] In this continuum method, the protein is modelled as a dielectric continuum of low polarizability embedded in a dielectric medium of high polarizability. We used a set of values for the DELPHI parameters that proven in a previous study to constitute a good compromise between accuracy and computing speed.[70] We used a value of 2.5 grids/Å for scale (the reciprocal of the grid spacing); a value of 0.001 kT/c for the convergence criterion; a 90% for the fill of the grid box; and the coulombic method to set the potentials at the boundaries of the finite-difference grid. The dielectric boundary was taken as the molecular surface defined by a 1.4 Å probe sphere and by spheres centered on each atom with radii taken from the Parse[71] vdW radii parameter set. The key aspect of the new improved approach is the use of a three dielectric constant set of values (ε=2 for nonpolar residues, ε=3 for polar residues and ε=4 for charged residues plus histidine) to mimic the expected rearrangement upon alanine mutation (the method has been previously described).[9, 10] It is important to highlight that we used only one trajectory for the computational energy analysis as it has been proven to give the best results.[10] Side-chain reorientation was implicitly included in the formalism by raising the internal dielectric constant. The nonpolar contribution to the solvation free energy due to van der Waals interactions between the solute and the solvent was modelled as a term dependent of the solvent accessible surface area (SASA) of the molecule. It was estimated by 0.00542×SASA+0.92 using the molsurf program developed by Mike Connolly.[72] As a systematic mutation of residues of typical protein-protein interfaces is a fastidious and time consuming methodological approach we have recently

developed a VMD[73] plugin (http://compbiochem.org/Software/compasm/Home.html).[74] This plugin has an easy-to-use graphical interface to prepare the input files, run the calculations and analyze the final results.

The performance of the ASM method can be assessed by the use of the F1 score (equation 3), which is defined as a function of Precision (P, also called specificity, equation 4) and Recall (R, also called sensitivity, equation 5). F1 score, P and R can be defined as:

$$F1 = \frac{2PR}{P+R} \quad (3)$$

$$P = \frac{TP}{TP+FP} \quad (4)$$

$$R = \frac{TP}{TP+FN} \quad (5),$$

in which TP stands for true positive (predicted HS that are an actual HS), FP stands for false positive (predicted HS that are not an actual HS), and FN stands for false negative (non-predicted HS that are an actual HS). Accuracy is defined as the ratio of number of correctly predicted residues to number of all predicted residues as in equation 6:

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \quad (6),$$

in which TN are the true negatives (correctly predicted NS). Specificity is another measure of performance and is formulated as:

$$Specificity = \frac{TN}{TN + FP} \quad (7)$$

Average and maximum $\left|\Delta\Delta G_{MM-PBSA} - \Delta\Delta G_{experimental}\right|$ values were also calculated for the interfacial residues in all interfaces.

We also fitted, using a linear regression, each group data against each corresponding $\Delta\Delta G_{binding}$ value in order to ascertain the ability of the various descriptors to score the different docking solutions.

RESULTS AND DISCUSSION

cASM was used to score docking poses provided by the HADDOCK[54, 75, 76] docking software for 18 P-P complexes (described in the Methods section), for which reliable X-ray structures and significant expASM data were available (216 distinct Ala mutations in total). This dataset is large enough to show, beyond doubt, the advantages of the method proposed here. Larger datasets could eventually be used, but they would not bring significant differences to the results and conclusions, and the testing of much larger sets would be very much CPU-demanding and time consuming. Moreover, the number of protein-protein complexes to be used is greatly limited by the availability of accurate experimental data for alanine scanning mutagenesis. The very popular HADDOCK docking software was used for the search because it consistently gives good results at wide comparative experiments.[77] Search poses with other software (pyDock,[78] PatchDock,[79, 80] FireDock,[81, 82] Cluspro 2.0,[83-85] Gramm-X,[86, 87] Hex,[88-95] ZDOCK,[96-98] and SwarmDock[99-101]) was preliminarily tried, giving basically similar results (data not shown). Our diverse dataset is composed of 11 unbound-unbound/unbound-bound and 7 bound-bound systems. The first corresponds to the cases where at least one (unbound-bound) or the two

(unbound-unbound) protein monomers were crystallized independently, in the free uncomplexed conformations. The second (bound-bound) includes the monomers crystallized within the complex, already in a bound conformation. The first set reflects the "real problem" that structural biology is facing. The second represents a good test set because it allows evaluating the scoring performance (which is the purpose of this work) with less complications arising from the limitations in the current searching algorithms (*i.e.* the limits in the capacity of generating good trial poses).

The strictest definition of HS is that of a residue that increases the binding free energy by > 4.0 kcal/mol upon Alanine mutation. Smaller free energy thresholds, such as 2.0 kcal/mol and 1.5 kcal/mol are also used commonly.[59, 61, 62, 66] The 18 diverse P-P complexes have a number of hot-spots (> 4 kcal/mol) per complex ranging from zero to six (see Table 1).

It is obvious that, among the trial P-P poses, the closest to the native one will have the better agreement between the expASM and the cASM results. The rationale for this is the following: i) interactions among residues across the P-P interface depend on the pose; ii) incorrect poses lead to mismatch in the P-P contacts, changing inter-residue interactions; iii) ASM measures these individual interactions; iv) consequently, the ASM values will strongly depend on the poses; v) the correct pose will have the correct contacts, thus the correct ASM results; vi) incorrect trial poses will have incorrect contacts, thus cASM values computed for those incorrect poses will be different from the true expASM results measured in the true pose, with the true contacts.

In summary, the closest the pose from the native one, the closest the cASM and expASM results will be.

Another factor that we expect to be important to identify the native pose is the interface area (iA). Most of the protein-protein binding energy comes from dispersive interactions, and these

are proportional to the contact surface. Polar residues interacting across the interface contribute less, due to desolvation penalties.[13] Even though the value of the interface area does not distinguish electrostatic mismatches, it may help in the fine distinction between similar poses, for which electrostatic interactions have been taken into account previously.

A set of nine structural quality indicators commonly used to compare expASM with cASM methods, have been tested to build a simple scoring function; all trials are summarized in supplementary data SI2-SI3. Bound-bound docking is particularly useful for these tests because it allows the testing of the scoring without introducing complications arising from imperfect searching of the conformational space. The correlation between the indicators shown in supplementary data SI2 and pose accuracy is low, even though there is an increase of the correlation for those complexes with a higher number of hot-spots, such as 1BRS.[102] A careful analysis of supplementary data SI2 shows that most indicators do not have enough discriminative power. Those that are more predictive are the number of hot spots correctly detected (TP), which makes sense, as they are the ones that confer the directionality for the complex formation event. We have defined HS in three typical ways, *i.e.* residues that increase the binding free energy upon Alanine mutation by (i) more than 4.0 kcal/mol, (ii) more than 2.0 kcal/mol and (iii) more than 1.5 kcal/mol. From (i) to (iii) we move from a more stringent to a broader HS definition. Analysis of supplementary data SI2 suggested us that 1) ranking the poses by the number of true positives detected by cASM according to definition (i); 2) applying definition (ii) subsequently to rank poses where condition (i) gave equivalent results; 3) applying definition (iii) for further differentiation of equivalent results, when the application of definitions (i) and (ii) still gave equivalent results, leads to the most correct pose ranking that can be generated with the used indicators. In principle, multiple regression analysis could retrieve a mathematical equation,

involving further indicators, or a different weighting scheme, with superior predictive power for these 18 complexes. However, we wanted a scoring function/method as simple as possible, rooted in logical physical ground, to be both physically informative and with broad applicability beyond the test set. Consistently, we did not try to "train a function" to get the "best fit" in a subset of complexes, but instead derive the "best scoring rationale" based on the physics of the P-P systems. By using this hierarchic energetic scheme, we are weighting more the contacts most important for binding (with larger free energy contributions) from a more restrictive to a more broad definition, with a subsequent increase in the number of residues that fulfill the specific conditions. The resulting scoring function is the same whenever one uses the whole 18 complexes to derive it, or use a subset of 12 complexes (a training set) to derive it and 6 complexes (a test set) to test it, because it is not fitted, but instead it is build up using physical grounds.

As this scoring method is based on a binary logical classification scheme, it can lead to draws between ranked structures. To further distinguish between similar top-ranked poses we have used the area of the interface as discriminating criterion. The rationale is that, among poses with good electrostatic matches, and overall correct poses, the one with the largest buried area will be the one that binds better, as hydrophobic forces are the driving force for the binding, and they are roughly proportional to the buried area. The importance of the diminution of solvent accessibility of HS was already demonstrated for a variety of complexes.[66, 102] When we applied the interface area (iA) to discriminate between equivalent poses we realized that it was so successful in identifying native conformations that we introduced another variant of the scoring function, using a consensus between SF_{ASM} and iA (*i.e.*, averaging the ranking positions given by both, after applying each one independently), which we have called $SF_{ASM:iA}$.

Despite being known that scoring functions can not be used as binding affinity predictors, it is informative to determine the correlation between a given scoring function and the binding free energy. However, here is not possible to do so. SF_{ASM} and $SF_{ASM:iA}$ provide a ranking position, and not a continuous variable as value. This means that the best-scored complex will always have a score of "first", whichever is the binding free energy of the protein-protein complex.

cASM was applied to the 20 trial poses (plus the experimental X-ray structures of the complexes) generated by HADDOCK for each of the 18 P-P complexes. The results for each complex are given in Tables 2 and 3. The number of structures (among the 20 trial poses of each complex) that are incorrect, and that have high accuracy (i-RMSD < 1 Å), as well as the respective averages, are given for the reader to understand the merits and limitations of the searching algorithms and the quality of the set of structures that is ranked by our scoring functions.

Table 2. cASM results for each complex and considering the U:U/U:B situations. The root mean square deviation for the interfacial residues (i-RMSD) was measured relative to the X-ray solution. Incorrect poses were defined as poses with i-RMSD > 4 Å.

complex	i-RMSD of the 1st ranked structure / Å	i-RMSD of 1st ranked structure (without X-Ray) / Å	Position of the 1st incorrect pose	no. structures with i-RMSD > 4	no. structures with i-RMSD <1	average i-RMSD
U:U/U:B structures, SF_{ASM} scoring function						
1BRS	0.0	1.8	12	5	1	4.3
1DVF	1.1	1.1	13	4	5	4.1
2EKS	0.0	7.4	2	19	1	8.4
1VFB	0.0	3.6	7	8	1	4.2
1DFJ	0.0	4.1	2	11	1	4.7
1FCC	0.0	0.9	5	5	6	3.5
1FQ9	0.9	0.9	14	4	5	3.2
1FLT	1.9	1.9	2	5	1	4.7
1A22	3.9	3.9	3	16	1	10.9
1F47	0.0	2.4	3	17	1	9.8
1IAR	0.0	4.0	4	9	1	4.3
average	*0.7*	*2.9*	*6*	*9.4*	*2.2*	*5.6*
U:U/U:B structures, $SF_{ASM:iA}$ scoring function						
1BRS	0.0	1.8	12	5	1	4.3
1DVF	0.0	1.1	13	4	5	4.1
2EKS	0.0	7.6	2	19	1	8.4
1VFB	0.0	3.4	7	8	1	4.2
1DFJ	0.0	3.9	2	11	1	4.7
1FCC	0.0	0.9	5	5	6	3.5
1FQ9	0.9	0.9	14	4	5	3.2
1FLT	0.0	2.8	2	5	1	4.7
1A22	0.0	16.8	3	16	1	10.9
1F47	0.0	2.4	3	17	1	9.8
1IAR	0.0	4.0	4	9	1	4.3
average	*0.1*	*4.2*	*6*	*9.4*	*2.2*	*5.6*

Table 3. cASM results for each complex and considering the B:B situations. The root mean square deviation for the interfacial residues (i-RMSD) was measured relative to the X-ray solution. Incorrect poses were defined as poses with i-RMSD > 4 Å.

complex	i-RMSD of the 1st ranked structure / Å	i-RMSD of 1st ranked structure (without X-Ray) / Å	Position of the 1st incorrect pose	no. structures with i-RMSD > 4	no. structures with i-RMSD <1	average i-RMSD
B:B structures, SF_{ASM} scoring function						
1BRS	1.0	1.0	10	4	5	3.2
1EMV	0.7	0.7	19	1	10	1.6
1DFJ	0.9	0.9	-	0	2	2.7
1FCC	1.4	1.4	10	4	6	2.6
1FQ9	0.0	3.5	3	9	1	6.0
1VFB	0.9	0.9	7	6	8	4.0
1F47	2.1	2.1	10	5	1	3.8
average	*1.0*	*1.5*	*10*	*4.1*	*4.7*	*3.4*
B:B structures, $SF_{ASM:iA}$ scoring function						
1BRS	0.7	0.7	10	4	5	3.2
1EMV	0.7	0.7	20	1	10	1.6
1DFJ	0.9	0.9	-	0	2	2.7
1FCC	0.9	0.9	6	4	6	2.6
1FQ9	0.0	3.5	3	9	1	6.0
1VFB	0.9	0.9	7	6	8	4.0
1F47	2.1	2.1	9	5	1	3.8
average	*0.9*	*1.4*	*9*	*4.1*	*4.7*	*3.4*

Table 4 summarizes the performance of the method, showing the root mean square deviations between the interfacial residues (i-RMSD) for the top ranked poses (among the 21 possible) against the X-ray structures, averaged over the 18 complexes. Detailed results for every complex separately are given in supplementary data SI3.

Table 4. Interfacial RMSD (i-RMSD) between the X-ray structures and the docking results, considering i) the best ranked pose, ii) the best pose but excluding the X-ray solution and iii) the position of the first incorrect solution (with i-RMSD > 4 Å) in the ranking results for the two SF developed here. The results correspond to the average for all the unbound-unbound (U:U) and unbound-bound (U:B) complexes (top) and bound-bound (B:B) complexes (bottom). Note that the values of i-RMSD for the 18 complexes do not follow a parametric distribution, which means that it is not statistically adequate to express the spread of the i-RMSD values through a single parameter (*i.e.* a standard error). Instead, the reader shall look into the list of values in table SI3.

	SF_{ASM}	$SF_{ASM:iA}$
	U:U and U:B test set (11 complexes)	
i-RMSD of 1st ranked structure / Å	0.7	0.1
i-RMSD of 1st ranked structure (no X-ray) / Å	2.9	4.2
Position of the 1st structure with i-RMSD > 4 Å	6	6
	B:B test set (7 complexes)	
i-RMSD of 1st ranked structure / Å	1.0	0.9
i-RMSD of 1st ranked structure (no X-ray) / Å	1.5	1.4
Position of the 1st structure with i-RMSD > 4 Å	10	9

Other quality criteria, such as the fraction of native contacts (shown in supplementary data SI2) or the RMSD of the whole complex could be used. We used only the i-RMSD for simplicity, but given the large number of cases where the X-ray structures was ranked first, and in particular the very large number of high-resolution structures ranked first, we realized that using all these quality measurements would give equivalent results as using only the i-RMSD.

In a large number of cases (12 out of 18 with SF_{ASM} and 17 out of 18 with $SF_{ASM:iA}$) the top ranked pose was within 1 Å of the X-ray structure. $SF_{ASM:iA}$ was the most successful scoring function, with a success in identifying the 3D correct structure for the P-P complex of ~94%.

In the unbound-unbound/unbound-bound set, $SF_{ASM:iA}$ ranked the X-ray 1st, in 10 out of 11 cases; in the remaining case, it ranked first a structure of high accuracy (i-RMSD < 1 Å), with an i-RMSD of 0.9 Å. In the bound-bound set (7 complexes), it ranked the X-ray structure 1st in one case, ranked high-accuracy structures 1st in 5 cases, and in a single case it ranked first a structure with an i-RMSD of 2.1 Å. In summary, in 17 out of 18 cases the structure delivered by $SF_{ASM:iA}$ was highly accurate and the answer was systematically reliable.

The results show that, as far as the search algorithms retrieving good poses, our method is very robust. However, current search algorithms are not perfect enough, and in some cases many of the poses they generate are not accurate. It is obvious that we cannot identify a true pose with our scoring function if the true pose is not present in the pose set at all.

The quality of the search algorithm was very much case-dependent. An overview can be seen in Table 5, where the number of high-resolution structures and the number of incorrect structures (i-RMSD > 4 Å) among the 20 trial poses per complex, as well as the average i-RMSD of the trial poses, is depicted. The same numbers can be seen in Tables 2 and 3 but in a case-by-case basis. Table 5 shows that, on average, and excluding the X-ray structure, there was only another

high accuracy structure in the unbound-unbound/unbound-bound set, highlighting how sensitive a scoring function has to be, to identify it. Among the same set of 20 structures there were on average 9.4 incorrect structures, and the average i-RMSD corresponded to 5.4 Å.

Table 5. Average number of high accuracy poses and of incorrect poses among the 20 trial poses generated by the search algorithm of the docking software, for each complex. The average i-RMSD is shown as well. The results for each of the individual complexes can be consulted in Table SI4.

	U:U/U:B set	B:B set
no. of high accuracy trial poses	1.2	4.1
no. of incorrect trial poses	9.4	3.7
Average i-RMSD for the trial	5.6	3.4

The situation is much better in the bound-bound set, where an average of 3.7 high resolution structures are present (as well as 4.1 incorrect ones) and the average i-RMSD was 3.4 Å. This clearly highlights the difficulties that search algorithms have in dealing with the rearrangement that the monomers undergo upon binding.

The low quality of the trial poses set causes problems for the overall result. If we exclude the X-ray solution from the trial set and use $SF_{ASM:iA}$, we move from an average i-RMSD for the 1st ranking solution of 0.1 Å to 4.1 Å in the unbound-unbound/unbound-bound set (from 0.7 Å to 2.9 Å with SF_{ASM}), which is not as satisfactory as before (but still much better than the results of current scoring algorithms). The situation is much less problematic in the bound-bound set, where we move from an average i-RMSD of 0.9 Å to 1.4 Å (from 1.0 Å to 1.5 Å with SF_{ASM}). This confirms again that the scoring problem is mostly solved with $SF_{ASM:iA}$ or SF_{ASM}, and the remaining inaccuracies come from the existing search algorithms.

One of the most important aspects of prediction is to identify the failure. We registered the position of the first incorrect structure in Table 2-4 for that purpose. Obviously, an incorrect structure is only problematic if it is ranked 1st (and "uncomfortable" if it is ranked 2nd). On average, the first incorrect structure is ranked 5th-6th, so there is no problem with this… on average. But on a case-by-case basis (Tables 2 and 3), a few problems may eventually come up.

The most important thing is to realize that an incorrect structure was never ranked 1st, in any single case of both unbound-unbound/unbound-bound and bound-bound sets, with any of the two scoring functions we have developed.

Cases in which incorrect structures have been ranked 2nd in the unbound-unbound/unbound-bound set, were rare (only 2 cases among 18 complexes, structures 2EKS and 1A22 with $SF_{ASM:iA}$, and structures 2EKS and 1DFJ with SFASM). With 2EKS, the problem was unavoidable because there was only a non-incorrect structure among the trial poses, and it was ranked 1st. In 1DFJ there were 11 incorrect structures, and just one high accuracy structure, in the whole trial poses set. In the case of 1A22, there were 16 incorrect structures and just one high accuracy structure within the whole trial poses set. Again, limitations came obviously from search, not from scoring. A final emphasis on this comes from the results of the bound-bound set, where search is simpler. Within this set, an incorrect structure was never chosen for the 1st or 2nd ranking places, with any of the two scoring functions.

CONCLUSIONS

In this work we have used 18 protein:protein complexes with different sizes, hydrophobicity, and hot spots number, to derive and validate two high-reliability, physically based, scoring function for protein:protein docking (here named SF_{ASM} and $SF_{ASM:iA}$). They calculate the binding free energy differences upon alanine mutation of interfacial residues (using a calibrated MM-PBSA approach), and compare the computed values with the corresponding experimental ones. Therefore, they do need experimental binding free energies upon ASM as input. The best match between computed and experimental values identifies the true pose. The interface area serves as an additional criterion to distinguish between equivalent experimental/theoretical matches.

The trial structures were generated with widely used protein:protein docking software. Care was taken to include decoy poses among the set to be tested by SF_{ASM} and $SF_{ASM:iA}$.

SF_{ASM} and $SF_{ASM:iA}$ identified the native-like structures with high accuracy and reliability (~94% success). Experimental ASM results may eventually need to be measured *a priori* if they are not known from the beginning. This can be seen as a limitation of the method. However, despite time and labor that this task involves, it is obvious that the outcome of a reliable P-P structure pays off the effort.

Several machine learning and knowledge-based statistical method have been used to identify native structures[103-105] and hot-spots.[106-126] For example, Zhao *et al.* used a feature-based learning method based on tridimensional protein-protein interface features to distinguish between native and non-native structures with very remarkable success.

It is possible that the integration of energy-based and knowledge-based scoring functions could yield better results than any of the strategies alone (a definite answer for this need further

studies). We shall note that the outcome of this "consensus scoring scheme" had to be characterized as an empirical, knowledge/training based scoring scheme because the consensus scheme has to be based on training.

The limitation of machine-learning and statistical approaches is that they are not grounded on rigorous first-principles physics. As such, their predictability capacity will always depend on the examples used to train/fit/parameterize them, and they will easily fail when dealing with unusual structural features, which many times are the most interesting ones. The only way to have a truly predictable method is to root it in basic physics. Physics-based methods will naturally predict unusual (or even unique) structural features with the same accuracy as the most common approaches. We are still far from a purely physics-based efficient method for protein:protein docking, and this is the most important limitation of physically-based methods. Our scoring functions try to move in that direction. It is our opinion that "parameterized" methods are very useful for today, as they provide answers for many current problems. In contrast, the future probably belongs to the fully-predictive, first principles, physics methods, and that is why we are investing on them. However, how distant is this "future" is something that we can not really know by sure.

When using SF_{ASM} and $SF_{ASM:iA}$ the bottleneck that remains in the process is the search algorithm. The accuracy of the pose identified by SF_{ASM} and $SF_{ASM:iA}$ improves markedly with the accuracy of the trial structure set generated by current search algorithms. Ultimately, one can only find a native-like structure if it is present at all in the trial poses. Therefore, the results presented here move the bottleneck of P-P docking from the scoring of poses to the search for good trial poses. We may say that a new step was given in the difficult world of predictive docking.

ASSOCIATED CONTENT

Electronic Supplementary Information (ESI) available: Table S1 - List of protein-protein complexes and number of hot spots of each complex. Table S2 - Detailed data on the descriptors used to score the cASM methodology. Table S3 – Discriminative results for each complex in regards to 1) the i-RMSD values and 2) the docking poses scored by SF_{ASM} and $SF_{ASM:iA}$. See DOI: 10.1039/b000000x/

AUTHOR INFORMATION

‡ Authors contributed equally.

† Present address: Structural Biology and NMR Laboratory, Department of Biology, University of Copenhagen, Ole Maaløes Vej 5, 2200, Copenhagen, Denmark

Corresponding Author

* E-mail: pafernan@fc.up.pt (PAF)

ACKNOWLEDGMENT

This work received financial support from the European Union (FEDER funds through COMPETE) and National Funds (FCT, Fundação para a Ciência e Tecnologia) through project Pest-C/EQB/LA0006/2013. The work also received financial support from project EXCL/QEQ-COM/0394/2012. To all financing sources the authors are greatly indebted.

REFERENCES

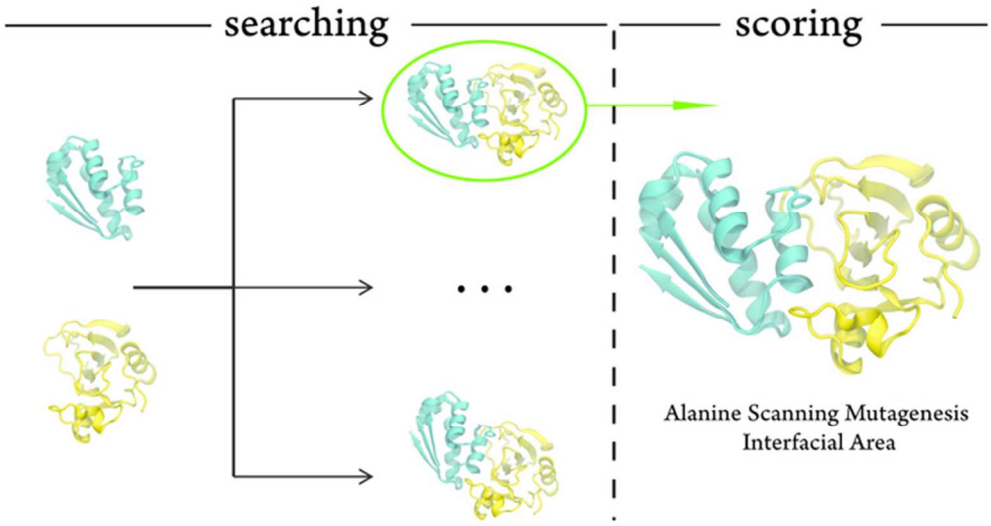
1. S. J. Wodak and J. Janin, in *Protein Modules and Protein-Protein Interactions* 2003, vol. 61, pp. 9-73.
2. I. S. Moreira, P. A. Fernandes and M. J. Ramos, *J. Comput. Chem.*, 2010, 31, 317-342.
3. D. W. Ritchie, *Curr. Protein Peptide Sci.*, 2008, 9, 1-15.
4. A. D. J. van Dijk, R. Boelens and A. M. J. J. Bonvin, *FEBS J.*, 2005, 272, 293-312.
5. P. L. Kastritis and A. M. J. J. Bonvin, *J. Proteome Res.*, 2010, 9, 2216-2225.
6. M. F. Lensink, R. Méndez and S. J. Wodak, *Proteins: Struct. Funct. Bioinform.*, 2007, 69, 704-718.
7. M. F. Lensink and S. J. Wodak, *Proteins: Struct. Funct. Bioinform.*, 2010, 78, 3073-3084.
8. M. F. Lensink and S. J. Wodak, *Proteins: Struct. Funct. Bioinform.*, 2013, 81, 2082-2095.
9. I. S. Moreira, P. A. Fernandes and M. J. Ramos, *Theor. Chem. Acc.*, 2007, 117, 99-113.
10. I. S. Moreira, P. A. Fernandes and M. J. Ramos, *J. Comput. Chem.*, 2007, 28, 644-654.
11. I. S. Moreira, P. A. Fernandes and M. J. Ramos, *Proteins: Struct. Funct. Bioinform.*, 2007, 68, 803-812.
12. A. G. Murzin, S. E. Brenner, T. Hubbard and C. Chothia, *J. Mol. Biol.*, 1995, 247, 536-540.
13. A. M. Buckle, G. Schreiber and A. R. Fersht, *Biochemistry (Mosc).* 1994, 33, 8878-8889.
14. C. Martin, V. Richard, M. Salem, R. Hartley and Y. Mauguen, *Acta Crystallogr. Sect. D*, 1999, 55, 386-398.
15. G. S. Ratnaparkhi, S. Ramachandran, J. B. Udgaonkar and R. Varadarajan, *Biochemistry (Mosc).* 1998, 37, 6958-6966.
16. G. Schreiber and A. R. Fersht, *J. Mol. Biol.*, 1995, 248, 478-486.
17. G. Schreiber and A. R. Fersht, *Biochemistry (Mosc).* 1993, 32, 5145-5150.
18. T. N. Bhat, G. A. Bentley, G. Boulot, M. I. Greene, D. Tello, W. Dallacqua, H. Souchon, F. P. Schwarz, R. A. Mariuzza and R. J. Poljak, *Proc. Natl. Acad. Sci. U. S. A.*, 1994, 91, 1089-1093.
19. J. Wang, M. Dauter, R. Alkire, A. Joachimiak and Z. Dauter, *Acta Crystallogr. Sect. D*, 2007, 63, 1254-1268.
20. W. DallAcqua, E. R. Goldman, E. Eisenstein and R. A. Mariuzza, *Biochemistry (Mosc).* 1996, 35, 9667-9676.
21. R. E. Hawkins, S. J. Russell, M. Baier and G. Winter, *J. Mol. Biol.*, 1993, 234, 958-964.
22. B. Kobe and J. Deisenhofer, *Nature*, 1995, 374, 183-186.
23. M. Dechene, G. Wink, M. Smith, P. Swartz and C. Mattos, *Proteins: Struct. Funct. Bioinform.*, 2009, 76, 861-881.
24. B. Kobe and J. Deisenhofer, *J. Mol. Biol.*, 1996, 264, 1028-1043.
25. R. Shapiro, M. Ruiz-Gutierrez and C.-Z. Chen, *J. Mol. Biol.*, 2000, 302, 497-519.
26. C.-Z. Chen and R. Shapiro, *Proc. Natl. Acad. Sci.*, 1997, 94, 1761-1766.
27. C. Wiesmann, G. Fuh, H. W. Christinger, C. Eigenbrot, J. A. Wells and A. M. de Vos, *Cell*, 91, 695-704.
28. Y. A. Muller, H. W. Christinger, B. A. Keyt and A. M. de Vos, *Structure*, 5, 1325-1338.
29. M. A. Starovasnik, H. W. Christinger, C. Wiesmann, M. A. Champe, A. M. de Vos and N. J. Skelton, *J. Mol. Biol.*, 1999, 293, 531-544.

30. B. A. Keyt, H. V. Nguyen, L. T. Berleau, C. M. Duarte, J. Park, H. Chen and N. Ferrara, *J. Biol. Chem.*, 1996, 271, 5638-5646.
31. B. C. Braden, B. A. Fields, X. Ysern, W. Dall'Acqua, F. A. Goldbaum, R. J. Poljak and R. A. Mariuzza, *J. Mol. Biol.*, 1996, 264, 137-151.
32. E. R. Goldman, W. DallAcqua, B. C. Braden and R. A. Mariuzza, *Biochemistry (Mosc).* 1997, 36, 49-56.
33. T. Nakanishi, K. Tsumoto, A. Yokota, H. Kondo and I. Kumagai, *Protein Sci.*, 2008, 17, 261-270.
34. A. Rajpal, M. G. Taylor and J. F. Kirsch, *Protein Sci.*, 1998, 7, 1868-1874.
35. K. Tsumoto, K. Ogasahara, Y. Ueda, K. Watanabe, K. Yutani and I. Kumagai, *J. Biol. Chem.*, 1995, 270, 18551-18557.
36. A. E. Sauer-Eriksson, G. J. Kleywegt, M. Uhlén and T. A. Jones, *Structure*, 1995, 3, 265-278.
37. S. Krapp, Y. Mimura, R. Jefferis, R. Huber and P. Sondermann, *J. Mol. Biol.*, 2003, 325, 979-989.
38. D. J. Sloan and H. W. Hellinga, *Protein Sci.*, 1999, 8, 1643-1648.
39. J. Schlessinger, A. N. Plotnikov, O. A. Ibrahimi, A. V. Eliseenkova, B. K. Yeh, A. Yayon, R. J. Linhardt and M. Mohammadi, *Mol. Cell*, 2000, 6, 743-750.
40. A. E. Eriksson, B. W. Matthews and L. S. Cousens, *Protein Sci.*, 1993, 2, 1274-1284.
41. H. Zhu, K. Ramnarayan, J. Anchin, W. Y. Miao, A. Sereno, L. Millman, J. Zheng, V. N. Balaji and M. E. Wolff, *J. Biol. Chem.*, 1995, 270, 21869-21874.
42. B. A. Springer, M. W. Pantoliano, F. A. Barbera, P. L. Gunyuzlu, L. D. Thompson, W. F. Herblin, S. A. Rosenfeld and G. W. Book, *J. Biol. Chem.*, 1994, 269, 26879-26884.
43. T. Clackson, M. H. Ultsch, J. A. Wells and A. M. de Vos, *J. Mol. Biol.*, 1998, 277, 1111-1128.
44. R. J. Brown, J. J. Adams, R. A. Pelekanos, Y. Wan, W. J. McKinstry, K. Palethorpe, R. M. Seeber, T. A. Monks, K. A. Eidne, M. W. Parker and M. J. Waters, *Nat. Struct. Mol. Biol.*, 2005, 12, 814-821.
45. S. H. Bass, M. G. Mulkerrin and J. A. Wells, *Proc. Natl. Acad. Sci. U. S. A.*, 1991, 88, 4498-4502.
46. T. Clackson and J. A. Wells, *Science*, 1995, 267, 383-386.
47. B. C. Cunningham and J. A. Wells, *J. Mol. Biol.*, 1993, 234, 554-563.
48. L. Mosyak, Y. Zhang, E. Glasfeld, S. Haney, M. Stahl, J. Seehra and W. S. Somers, *EMBO J.*, 2000, 19, 3179-3191.
49. T. Hage, W. Sebald and P. Reinemer, *Cell*, 1999, 97, 271-281.
50. A. Wlodaver, A. Pavlovsky and A. Gustchina, *FEBS Lett.*, 1992, 309, 59-64.
51. Y. H. Wang, B. J. Shen and W. Sebald, *Proc. Natl. Acad. Sci. U. S. A.*, 1997, 94, 1657-1662.
52. U. C. Kühlmann, A. J. Pommer, G. R. Moore, R. James and C. Kleanthous, *J. Mol. Biol.*, 2000, 301, 1163-1178.
53. R. Wallis, K.-Y. Leung, M. J. Osborne, R. James, G. R. Moore and C. Kleanthous, *Biochemistry (Mosc).* 1998, 37, 476-485.
54. C. Dominguez, R. Boelens and A. M. J. J. Bonvin, *J. Am. Chem. Soc.*, 2003, 125, 1731-1737.
55. S. Huo, I. Massova and P. A. Kollman, *J. Comput. Chem.*, 2002, 23, 15-27.

56. D. A. Case, T. A. Darden, I. Cheatham, T.E., C. L. Simmerling, J. Wang, R. E. Duke, R. Luo, M. Crowley, R. C. Walker, W. Zhang, K. M. Merz, B. Wang, S. Hayik, A. Roitberg, G. Seabra, I. Kolossváry, K. F. Wong, F. Paesani, J. Vanicek, X. Wu, S. R. Brozell, T. Steinbrecher, H. Gohlke, L. Yang, C. Tan, J. Mongan, V. Hornak, G. Cui, D. H. Mathews, M. G. Seetin, C. Sagui, V. Babin and P. A. Kollman, 2008.
57. I. S. Moreira, P. A. Fernandes and M. J. Ramos, *Proteins: Struct. Funct. Bioinform.*, 2006, 63, 811-821.
58. I. S. Moreira, P. A. Fernandes and M. J. Ramos, *J. Phys. Chem. B*, 2006, 110, 10962-10969.
59. I. S. Moreira, P. A. Fernandes and M. J. Ramos, *Int. J. Quantum Chem*, 2007, 107, 299-310.
60. I. S. Moreira, P. A. Fernandes and M. J. Ramos, *J. Chem. Theory Comput.*, 2007, 3, 885-893.
61. I. S. Moreira, P. A. Fernandes and M. J. Ramos, *J. Phys. Chem. B*, 2007, 111, 2697-2706.
62. I. S. Moreira, P. A. Fernandes and M. J. Ramos, *Theor. Chem. Acc.*, 2008, 120, 533-542.
63. L. T. Chong, Y. Duan, L. Wang, I. Massova and P. A. Kollman, *Proc. Natl. Acad. Sci. U. S. A.*, 1999, 96, 14330-14335.
64. P. A. Kollman, I. Massova, C. Reyes, B. Kuhn, S. H. Huo, L. Chong, M. Lee, T. Lee, Y. Duan, W. Wang, O. Donini, P. Cieplak, J. Srinivasan, D. A. Case and T. E. Cheatham, *Acc. Chem. Res.*, 2000, 33, 889-897.
65. R. T. Bradshaw, B. H. Patel, E. W. Tate, R. J. Leatherbarrow and I. R. Gould, *Protein Eng. Des. Sel.*, 2011, 24, 197-207.
66. I. S. Moreira, J. M. Martins, R. M. Ramos, P. A. Fernandes and M. J. Ramos, *Bba-Proteins Proteom*, 2013, 1834, 404-414.
67. S. A. Martins, M. A. S. Perez, I. S. Moreira, S. F. Sousa, M. J. Ramos and P. A. Fernandes, *J. Chem. Theory Comput.*, 2013, 9, 1311-1319.
68. W. Rocchia, S. Sridharan, A. Nicholls, E. Alexov, A. Chiabrera and B. Honig, *J. Comput. Chem.*, 2002, 23, 128-137.
69. W. Rocchia, E. Alexov and B. Honig, *J. Phys. Chem. B*, 2001, 105, 6507-6514.
70. I. S. Moreira, P. A. Fernandes and M. J. Ramos, *J. Mol. Struc. - Theochem*, 2005, 729, 11-18.
71. D. Sitkoff, K. A. Sharp and B. Honig, *J. Phys. Chem.*, 1994, 98, 1978-1988.
72. M. L. Connolly, *J. Appl. Crystallogr.*, 1983, 16, 548-558.
73. W. Humphrey, A. Dalke and K. Schulten, *J. Mol. Graphics*, 1996, 14, 33-38.
74. J. V. Ribeiro, N. M. F. S. A. Cerqueira, I. S. Moreira, P. A. Fernandes and M. J. Ramos, *Theor. Chem. Acc.*, 2012, 131.
75. S. J. De Vries, A. D. J. van Dijk, M. Krzeminski, M. van Dijk, A. Thureau, V. Hsu, T. Wassenaar and A. M. J. J. Bonvin, *Proteins: Struct. Funct. Bioinform.*, 2007, 69, 726-733.
76. S. J. De Vries, M. van Dijk and A. M. J. J. Bonvin, *Nat Protoc*, 2010, 5, 883-897.
77. J. Janin, *Protein Sci.*, 2005, 14, 278-283.
78. T. M.-K. Cheng, T. L. Blundell and J. Fernandez-Recio, *Proteins: Struct. Funct. Bioinform.*, 2007, 68, 503-515.
79. D. Duhovny, R. Nussinov and H. Wolfson, in *Algorithms in Bioinformatics*, eds. R. Guigó and D. Gusfield, Springer Berlin Heidelberg, 2002, vol. 2452, ch. 14, pp. 185-200.

80. D. Schneidman-Duhovny, Y. Inbar, V. Polak, M. Shatsky, I. Halperin, H. Benyamini, A. Barzilai, O. Dror, N. Haspel, R. Nussinov and H. J. Wolfson, *Proteins: Struct. Funct. Bioinform.*, 2003, 52, 107-112.
81. N. Andrusier, R. Nussinov and H. J. Wolfson, *Proteins: Struct. Funct. Bioinform.*, 2007, 69, 139-159.
82. E. Mashiach, D. Schneidman-Duhovny, N. Andrusier, R. Nussinov and H. J. Wolfson, *Nucleic Acids Res.*, 2008, 36, W229-W232.
83. S. R. Comeau, D. W. Gatchell, S. Vajda and C. J. Camacho, *Bioinformatics*, 2004, 20, 45-50.
84. D. Kozakov, R. Brenke, S. R. Comeau and S. Vajda, *Proteins: Struct. Funct. Bioinform.*, 2006, 65, 392-406.
85. D. Kozakov, D. R. Hall, D. Beglov, R. Brenke, S. R. Comeau, Y. Shen, K. Li, J. Zheng, P. Vakili, I. C. Paschalidis and S. Vajda, *Proteins: Struct. Funct. Bioinform.*, 2010, 78, 3124-3130.
86. A. Tovchigrechko and I. A. Vakser, *Proteins: Struct. Funct. Bioinform.*, 2005, 60, 296-301.
87. A. Tovchigrechko and I. A. Vakser, *Nucleic Acids Res.*, 2006, 34, W310-W314.
88. D. Ritchie, *J. Appl. Crystallogr.*, 2005, 38, 808-818.
89. D. W. Ritchie, PhD PhD Thesis, University of Aberdeen, 1998.
90. D. W. Ritchie, *Proteins: Struct. Funct. Bioinform.*, 2003, 52, 98-106.
91. D. W. Ritchie and G. J. L. Kemp, *J. Comput. Chem.*, 1999, 20, 383-395.
92. D. W. Ritchie, D. Kozakov and S. Vajda, *Bioinformatics*, 2008, 24, 1865-1873.
93. D. W. Ritchie and V. Venkatraman, *Bioinformatics*, 2010, 26, 2398-2405.
94. G. Macindoe, L. Mavridis, V. Venkatraman, M.-D. Devignes and D. W. Ritchie, *Nucleic Acids Res.*, 2010, DOI: 10.1093/nar/gkq311, W445-W449.
95. D. Mustard and D. W. Ritchie, *Proteins: Struct. Funct. Bioinform.*, 2005, 60, 269-274.
96. B. G. Pierce, Y. Hourai and Z. P. Weng, *Plos One*, 2011, 6.
97. J. Mintseris, B. Pierce, K. Wiehe, R. Anderson, R. Chen and Z. Weng, *Proteins: Struct. Funct. Bioinform.*, 2007, 69, 511-520.
98. R. Chen, L. Li and Z. Weng, *Proteins: Struct. Funct. Bioinform.*, 2003, 52, 80-87.
99. X. Li, I. H. Moal and P. A. Bates, *Proteins: Struct. Funct. Bioinform.*, 2010, 78, 3189-3196.
100. I. H. Moal and P. A. Bates, *Int. J. Mol. Sci.*, 2010, 11, 3623-3648.
101. M. Torchala, I. H. Moal, R. A. G. Chaleil, J. Fernandez-Recio and P. A. Bates, *Bioinformatics*, 2013, 29, 807-809.
102. I. S. Moreira, R. M. Ramos, J. M. Martins, P. A. Fernandes and M. J. Ramos, *J. Biomol. Struct. Dyn.*, 2014, 32, 186-197.
103. R. Esmaielbeiki and J. C. Nebel, *BMC Bioinformatics*, 2014, 15.
104. O. G. Othersen, A. G. Stefani, J. B. Huber and H. Sticht, *J. Mol. Model.*, 2012, 18, 1285-1297.
105. N. Zhao, B. Pang, C. R. Shyu and D. Korkin, *Proteomics*, 2011, 11, 4321-4330.
106. L. Deng, J. H. Guan, X. M. Wei, Y. Yi, Q. C. Zhang and S. G. Zhou, *J. Comput. Biol.*, 2013, 20, 878-891.
107. K. I. Cho, D. Kim and D. Lee, *Nucleic Acids Res.*, 2009, 37, 2672-2687.
108. J. S. Mora, S. A. Assi and N. Fernandez-Fuentes, *Plos One*, 2010, 5.
109. J. F. Xia, X. M. Zhao, J. N. Song and D. S. Huang, *BMC Bioinformatics*, 2010, 11.

110. L. Wang, Z. P. Liu, X. S. Zhang and L. N. Chen, *Protein Eng. Des. Sel.*, 2012, 25, 119-126.
111. Q. Liu, S. C. H. Hoi, C. K. Kwoh, L. Wong and J. Y. Li, *BMC Bioinformatics*, 2014, 15.
112. B. Xu, X. M. Wei, L. Deng, J. H. Guan and S. G. Zhou, *Bmc Syst Biol*, 2012, 6.
113. D. R. Koes and C. J. Camacho, *Bioinformatics*, 2012, 28, 784-791.
114. P. Ozbek, S. Soner and T. Haliloglu, *Plos One*, 2013, 8.
115. L. Li, B. Zhao, Z. Cui, J. Gan, M. K. Sakharkar and P. Kangueane, *Bioinformation*, 2006, 1, 121-126.
116. T. Geppert, B. Hoy, S. Wessler and G. Schneider, *Chem. Biol.*, 2011, 18, 344-353.
117. N. Tuncbag, O. Keskin and A. Gursoy, *Nucleic Acids Res.*, 2010, 38, W402-W406.
118. E. Guney, N. Tuncbag, O. Keskin and A. Gursoy, *Nucleic Acids Res.*, 2008, 36, D662-D666.
119. Y. Ofran and B. Rost, *PLoS Comp. Biol.*, 2007, 3, 1169-1176.
120. S. Grosdidier and J. Fernandez-Recio, *BMC Bioinformatics*, 2008, 9.
121. M. Guharoy, A. Pal, M. Dasgupta and P. Chakrabarti, *Journal of structural and functional genomics*, 2011, 12, 33-41.
122. H. L. Chen and H. X. Zhou, *Proteins-Structure Function and Bioinformatics*, 2005, 61, 21-35.
123. P. Chen, J. Y. Li, L. Wong, H. Kuwahara, J. H. Z. Huang and X. Gao, *Proteins-Structure Function and Bioinformatics*, 2013, 81, 1351-1362.
124. J. M. Martins, R. M. Ramos, A. C. Pimenta and I. S. Moreira, *Proteins-Structure Function and Bioinformatics*, 2014, 82, 479-490.
125. X. L. Zhu and J. C. Mitchell, *Proteins-Structure Function and Bioinformatics*, 2011, 79, 2671-2683.
126. P. L. Kastritis, J. P. G. L. M. Rodrigues, G. E. Folkers, R. Boelens and A. M. J. J. Bonvin, *J. Mol. Biol.*, 2014, 426, 2632-2652.



TOC figure
60x32mm (300 x 300 DPI)

TABLE OF CONTENTS

Protein-protein 3D structures are fundamental for structural biology and drug discovery. Here we present a new, high accuracy, scoring method to discover the native 3D structure of protein-protein complexes. This methodology incorporates Alanine scanning experimental data previously known. The protein-protein interface area is also included in the scheme.