BIOINFORMATICS ORIGINAL PAPER

Vol. 29 no. 14 2013, pages 1742–1749
doi:10.1093/bioinformatics/btt260

Structural bioinformatics

Advance Access publication May 6, 2013

InterEvScore: a novel coarse-grained interface scoring function using a multi-body statistical potential coupled to evolution

Jessica Andreani[1,2,3], Guilhem Faure[1,2,3] and Raphael Guerois[1,2,3,*]

[1]CEA, iBiTecS, Service de Bioenergetique Biologie Structurale et Mecanismes (SB2SM), Laboratoire de Biologie Structurale et Radiobiologie (LBSR), F-91191 Gif sur Yvette, France, [2]CNRS, UMR 8221, F-91191 Gif sur Yvette, France and [3]Université Paris Sud, UMR 8221, F-91401 Orsay, France

Associate Editor: Anna Tramontano

ABSTRACT

Motivation: Structural prediction of protein interactions currently remains a challenging but fundamental goal. In particular, progress in scoring functions is critical for the efficient discrimination of near-native interfaces among large sets of decoys. Many functions have been developed using knowledge-based potentials, but few make use of multi-body interactions or evolutionary information, although multi-residue interactions are crucial for protein–protein binding and protein interfaces undergo significant selection pressure to maintain their interactions.

Results: This article presents InterEvScore, a novel scoring function using a coarse-grained statistical potential including two- and three-body interactions, which provides each residue with the opportunity to contribute in its most favorable local structural environment. Combination of this potential with evolutionary information considerably improves scoring results on the 54 test cases from the widely used protein docking benchmark for which evolutionary information can be collected. We analyze how our way to include evolutionary information gradually increases the discriminative power of InterEvScore. Comparison with several previously published scoring functions (ZDOCK, ZRANK and SPIDER) shows the significant progress brought by InterEvScore.

Availability: http://biodev.cea.fr/interevol/interevscore

Contact: guerois@cea.fr

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

Received on February 14, 2013; revised on April 25, 2013; accepted on May 2, 2013

1 INTRODUCTION

Protein–protein interactions play a pivotal role in virtually all cellular processes (Goodsell and Olson, 2000). Despite the growing number of available 3D structures for protein complexes in the Protein Data Bank (Berman *et al.*, 2000), computational prediction of protein interfaces turns out as an important tool for the exploration of interactomes (Mosca *et al.*, 2013; Vakser, 2013; Zhang *et al.*, 2012). A large variety of methods have been developed in the past years to improve the success rate of computational protein–protein docking, and the combination of various methods can improve the accuracy of docking results (Janin, 2010; Stein *et al.*, 2011; Vajda and Kozakov, 2009). Docking methods can be classified in two broad and complementary categories: template-based (homology) docking (Kundrotas *et al.*, 2012; Sinha *et al.*, 2010) based on the availability of 3D complex structures similar to the complex of interest and template-free docking. In the present work, we focus on the second type of approach.

Classically, the template-free docking procedure can be divided in two steps (Vajda and Kozakov, 2009). The first step consists in sampling a large number of candidate interfaces (also called decoys). This sampling is often performed using a rigid-body method, which allows for fast search in the large decoy conformational space; in particular, Fast Fourier transform-based algorithms are highly efficient (Ritchie *et al.*, 2008). The most recent strategies often succeed in generating several near-native conformations; however, the second step of the docking process, which involves scoring and ranking the many decoys that have been generated, often fails to distinguish these near-native conformations from incorrect models, and there is still room for improvement of the scoring functions used in protein–protein docking (Lensink *et al.*, 2007; Lensink and Wodak, 2010).

Scoring functions can be divided in three categories: physics-based functions, which provide an estimate of the free energy of binding using physico-chemical principles, knowledge-based (empirical) potentials relying on statistics and hybrid potentials combining the two. Those methods are embodied in the programs of the most successful participants in the Critical Assessment of PRediction of Interactions (CAPRI) docking and scoring experiments (Lensink and Wodak, 2010). Among the successful programs, some rely mainly on physico-chemical scoring functions such as HADDOCK (Dominguez *et al.*, 2003), ATTRACT (Fiorucci and Zacharias, 2010), ZDOCK (Chen *et al.*, 2003; Mintseris *et al.*, 2007), pyDock (Cheng *et al.*, 2007) or FireDock (Andrusier *et al.*, 2007), whereas others such as ITScorePP rely on an iterative knowledge-based scoring function (Huang and Zou, 2010). Hybrid approaches were also implemented in scoring functions such as ZRANK (Pierce and Weng, 2007) or PIPER (Kozakov *et al.*, 2006) and in the switch from low to high resolution scoring functions in the Rosetta program (Chaudhury *et al.*, 2011).

An ideal scoring function would describe the energetics of the protein interface in atomic detail. However, in a context where low-resolution approaches are needed, given the highly combinatorial nature of protein docking, rigid-body sampling is largely favored as a first step. Coarse-grained knowledge-based potentials, less sensitive to atomic details such as side-chain

*To whom correspondence should be addressed.

positioning, seem appropriate for discriminating false interfaces from rigid-body near-native decoys and have been shown in several occasions to retain most of the useful information contained in all-atom potentials (Fitzgerald *et al.*, 2007; Zhang *et al.*, 2004).

A limitation of most statistical potentials developed so far is the decomposition of interface contacts into pairs, leaving out multi-body interactions, which are known to be important for binding. A few statistical potentials developed for protein structure prediction take into account three-body and four-body interactions (Feng *et al.*, 2007; Krishnamoorthy and Tropsha, 2003; Li and Liang, 2005; Ngan *et al.*, 2006). Most recently, the SPIDER scoring function has been developed for protein interface scoring based on multi-residue patterns with no theoretical size limitation (Khashan *et al.*, 2012). One of the major difficulties toward the development of multi-body docking potentials is the scarcity of interface contact statistics. We recently designed the InterEvol database to explore the structure and evolution of protein complexes (Faure *et al.*, 2012). This database contains >18 000 non-redundant interfaces, including >1500 non-obligate heteromeric interfaces, which form a large and robust basis for the extraction of reliable two- and three-body interaction statistics.

A complementary source of information about protein interfaces comes from the conservation of quaternary structures and binding modes, which is observed above 30% sequence identity (Aloy *et al.*, 2003; Faure *et al.*, 2012; Levy *et al.*, 2008). To predict the location of binding sites and the structure of complexes, different types of evolutionary information were used. They consist either in the analysis of correlated mutations, in the mapping of evolutionary rates or in the detection of residue contact complementarities. The use of correlated mutations, successful in some cases (Tress *et al.*, 2005), was found to have strong limitations (Halperin *et al.*, 2006), and sophisticated statistical treatments together with a large number of sequences were found beneficial to reach high precision (Weigt *et al.*, 2009). Scores integrating conservation profiles (Ofran and Rost, 2007; Res *et al.*, 2005) or residue contacts likelihood extracted from multiple sequence alignments (MSAs) (de Vries *et al.*, 2006; Engelen *et al.*, 2009; Zellner *et al.*, 2012) were found to improve the prediction of binding sites at protein surfaces. As for docking, scoring functions have also been developed that bias interface detection toward evolutionarily conserved residues (Akbal-Delibas *et al.*, 2012; Kanamori *et al.*, 2007). The development of the SCOTCH method (Madaoui and Guerois, 2008) highlighted a remarkable plasticity in the way interface physicochemical complementarities are maintained through evolution. Moreover, in a recent study of >1000 pairs of homologous heteromeric interfaces, we showed that although interface contacts are highly versatile, some contact conservation signal can be extracted, in particular when considering specific interface descriptors (Andreani *et al.*, 2012).

In the present article, we introduce InterEvScore, a novel scoring function for protein docking, which combines a multi-body statistical potential with evolutionary information derived from coupled MSAs for each partner in the complex and from our previous observations on the evolution of homologous interfaces (Andreani *et al.*, 2012).

To our knowledge, InterEvScore is the first scoring function for docking applications relying on both two- and three-body statistical potentials combined with the scoring of interface contacts inferred from MSAs. This way of integrating evolutionary information was found significantly superior to solely accounting for conserved positions. We also show that InterEvScore achieves significant improvement compared with several scoring functions, namely, ZDOCK, ZRANK and SPIDER, in the detection of near-native conformations among large sets of decoys from the widely used protein–protein docking benchmark (Hwang *et al.*, 2010). InterEvScore is freely available as a standalone Python program and can be run using different modes to take advantage of available structural and evolutionary information.

2 METHODS

The global InterEvScore workflow is represented in Figure 1.

2.1 Datasets

The training set for the statistical potential part of InterEvScore was derived from the InterEvol database (Faure *et al.*, 2012). First, we collected all the non-redundant dimeric interfaces that were predicted as both biologically relevant and non-obligate by the NOXclass algorithm, a machine-learning based classifier, which distinguishes between obligate, non-obligate and crystal packing interactions on the basis of structural interface properties (Zhu *et al.*, 2006). To avoid biases in the training and evaluation, this set of 1554 dimeric interfaces was then filtered to exclude all interfaces, which had >30% sequence identity (on both chains) with the interfaces in the test set (176 complexes from the protein–protein docking benchmark version 4.0, see later in the text); the remaining 1398 interfaces were clustered to reduce internal redundancy of the training set below 40% sequence identity using Uclust (Edgar, 2010). The final training dataset contained 1289 interfaces.

The test set was initially composed of all 176 complexes from the protein–protein docking benchmark version 4.0 (Hwang *et al.*, 2010). For each of these complexes, the benchmark includes 54 000 decoys generated by ZDOCK 3.0 (Mintseris *et al.*, 2007) using the structures of the unbound partners. At least one near-native decoy (cf. definition in Section 2.5) was generated for 131 of these 176 complexes. However, evolutionary information could be extracted only for 85 complexes (cf. Section 2.3 for details), among which 31 with no near-native decoy generated by ZDOCK; therefore, the main test dataset used in the article contains 54 complexes with both near-native decoys and evolutionary

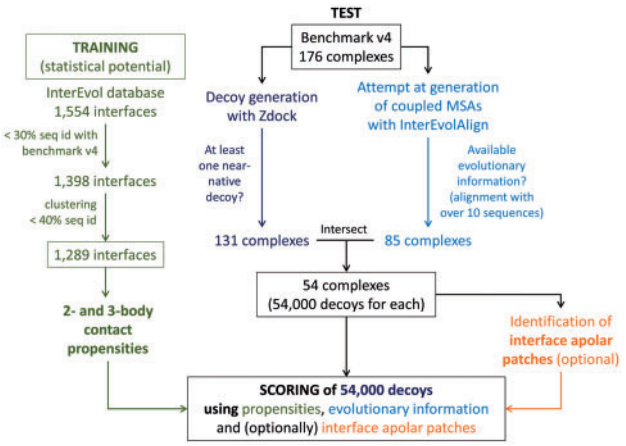

Fig. 1. Workflow for the training and testing of InterEvScore

information available. A subset of 43 complexes (excluding 11 complexes from the benchmark, which are redundant with the SPIDER training set) is used for comparison with SPIDER.

2.2 Statistical two- and three-body potential

We developed several scoring schemes based on two- and three-body potentials, which are illustrated in Figure 2. They are all residue based, most importantly because evolutionary information is available only at the residue level. Moreover, they rely on an analytical reference state independent from any particular set of decoys. Derivation of the two- and three-body interface contact propensities was similar to the derivation described in (Li *et al.*, 2003; Li and Liang, 2005) for a folding statistical potential. Atomic contacts in all 1289 interfaces of the training dataset were calculated on the basis of an α-shape representation of the interface. We excluded contacts involving backbone atoms (C, N, O). Contacts from all 1289 interfaces were then pooled to derive the observed number counts of two- (pairwise contacts) and three-body interactions (Fig. 2A). For two-body interactions, only inter-molecular contacts (one residue in each partner protein) were retained; for three-body interactions, only contacts involving residues from both partner proteins were retained (contacts involving three residues from the same protein were excluded).

The propensity $p(i,j)$ of a given two-body interaction between residue types i and j was defined as the ratio of the observed frequency $o(i,j)$ and the expected frequency $e(i,j)$, derived as the random probability to pick an interaction between residue types i and j:

$$p(i,j) = \frac{o(i,j)}{e(i,j)}$$

with

$$o(i,j) = \frac{N_{ij}}{N}$$

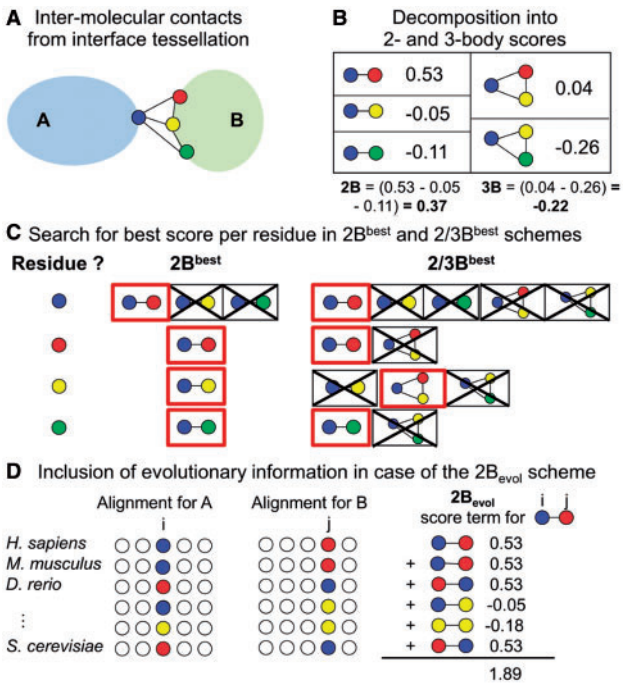

Fig. 2. Explanatory diagrams for the $2B^{best}$, $3B^{best}$, $2/3B^{best}$ and $2B_{evol}$ scoring schemes. Different scores are assigned depending residue types as illustrated from the different shades

where N_{ij} is the number count of atomic contacts between residue types i and j, N is the total number count of atomic contacts participating in all two-body interactions and,

$$\begin{cases} e(i,j) = S_i S_j \cdot \left(\frac{ninj}{n(n-ni)} + \frac{ninj}{n(n-nj)}\right) & \text{when } i \neq j \\ e(i,i) = S_i(S_i - 1) \cdot \left(\frac{nini}{n(n-ni)}\right) & \text{when } i = j \end{cases}$$

with $n = \sum_l Sl \cdot nl$

where S_i is the number of surface residues of amino acid type i at the surface of the protein (calculated over the 1289 pairs of proteins from the training dataset), and n_i is the average number of atomic contacts in which residue type i is involved. n_i is defined as the number of atomic contacts involving all residues of type i over the total number of atomic contacts. This term replaces the less accurate number of atoms per residue, which was used in (Li *et al.*, 2003), as only the average number of atoms engaged in inter-molecular interactions is counted for each residue.

For a given interface, the two-body interaction score (2B) was defined as the sum of $\ln(p(i,j))$ over all inter-molecular contacts between residues i and j (Fig. 2B):

$$2B = \sum_{(i,j)} [\ln p(i,j)]$$

For three-body interactions, amino acids were grouped into six residue types according to their physico-chemical properties as in (Li *et al.*, 2003; Li and Liang, 2005): basic (K, R), acidic (D, E), aromatic (F, W, Y), polar (H, N, Q, S, T), alkyl (C, I, L, M, V) and small (A, G, P). Three-body propensities $p(g_i,g_j,g_k)$ were defined for any three amino acid groups (g_i, g_j, g_k) in a manner similar to two-body propensities, and non-additivity (cooperativity) coefficients $c(g_i,g_j,g_k)$ were defined as:

$$c(g_i,g_j,g_k) = \frac{p(g_i,g_j,g_k)}{p(g_i,g_j)p(g_j,g_k)p(g_i,g_k)}$$

Bootstrap calculations performed with the same method as in (Li *et al.*, 2003) showed that the three-body propensities and non-additivity coefficients were more reliable when derived for the six amino acid groups than for the 20 amino acid types (see Supplementary Material). Consequently, for a given interface, the three-body interaction score (3B) was defined as the sum over all interactions between residues (i,j,k) belonging to amino acid groups (g_i,g_j,g_k) of three 2-body terms and the non-additivity term calculated with the reduced amino acid alphabet:

$$3B = \sum_{(i,j,k)\in(g_i,g_j,g_k)} \left[\ln p(i,j) + \ln p(j,k) + \ln p(i,k) + \ln c(g_i,g_j,g_k)\right]$$

A first set of scoring schemes was tested by summing the propensities over all the contacts of an interface as described earlier in the text (2B and 3B scores). Given the modular organization of interfaces (Reichmann *et al.*, 2007), we also explored a scoring scheme in which only one score per residue is counted, chosen as the highest one among all the contacts the residue is involved in (Fig. 2C). These scores were short-named $2B^{best}$ and $3B^{best}$ depending whether two- and three-body scores were used, respectively. Finally, a score (short-named $2/3B^{best}$) was derived by choosing for each residue the best two- or three-body contact. The two-body contact propensities are represented in Supplementary Figure S1.

2.3 Integration of evolutionary information

For each of the 176 complexes in the test set, we automatically derived couples of MSAs for both chains using the InterEvolAlign server (Faure *et al.*, 2012). These alignments contain couples of most likely orthologs for each partner in a number of species. For various reasons (notably the presence in benchmark v4 of 25 antibody/antigen complexes and 33 interfaces involving proteins from different superkingdoms, e.g. a

human protein and a viral protein), sufficient evolutionary information (with a minimum number of 10 sequences in the alignments) could be derived for a restricted set of 85 complexes. The detailed procedure of alignment derivation and filtering can be found in Supplementary Material.

In a first approach based on conservation analysis, we derived a scoring scheme termed 'cons' in which the conservation score of each residue was used to weight the statistical scores. For each score described earlier in the text (2B, 3B, $2B^{best}$, $3B^{best}$, $2/3B^{best}$), a corresponding score including conservation was derived ($2B_{cons}$, $3B_{cons}$, $2B^{best}_{cons}$, $3B^{best}_{cons}$ and $2/3B^{best}_{cons}$). Using the MSAs derived for each partner, conservation scores were calculated with the Rate4Site program (Mayrose *et al.*, 2004) running behind the ConSurf server (Landau *et al.*, 2005). The 'cons' scoring scheme is biased toward interfaces involving more conserved residues.

We also implemented evolutionary information in a novel way taking into account co-evolution between contacting residues rather than simple residue conservation. In this scoring scheme termed 'evol', for each (two- or three-body) contact detected in the interface, we extracted the (two or three) corresponding residues in each couple of sequences from the MSAs, and we calculated the evolutionary-based score as the sum over all species of individual scores for each derived contact (Fig. 2D). We then summed obtained scores over all contacts. In this way, we derived the $2B_{evol}$, $3B_{evol}$, $2B^{best}_{evol}$, $3B^{best}_{evol}$ and $2/3B^{best}_{evol}$ scores. For 'best' scores, only contacts derived from the best contact for each residue were summed.

2.4 Focus on apolar patches

In our recent study of the evolution of homologous interfaces, we found that apolar patches constituted a major conserved interface descriptor featuring a higher conservation of inter-residues contacts (Andreani *et al.*, 2012). We included a mode in InterEvScore in which evolutionary information was used only for residues belonging to apolar patches. In this mode, the score was calculated by adding the sum of evolutionary-derived scores for interface contacts with at least one residue involved in an apolar patch and the sum of statistical scores for other contacts (with no inclusion of evolutionary information). Apolar patches were detected using the method described in (Andreani *et al.*, 2012). Details about the patches mode can be found in Supplementary Material.

2.5 Metrics for evaluation and comparison

The predictions were evaluated by calculating the interface root mean square deviation (iRMSD) on α-carbon atoms between the native (bound) complex and the decoys generated by ZDOCK (Chen *et al.*, 2003). Among the 54 000 decoys generated for each complex, near-native decoys (also named hits) were defined as predictions with Cα iRMSD below 2.5 Å. An alternative definition of hits based on CAPRI criteria (Mendez *et al.*, 2005) yields similar results (see Supplementary Material).

InterEvScore was compared with the widely used methods ZDOCK (Mintseris *et al.*, 2007) and ZRANK (Pierce and Weng, 2007) and to the recently published multi-body interaction scoring function SPIDER (Khashan *et al.*, 2012).

The capacity of each scoring method to correctly rank hits with respect to other decoys was assessed using several metrics. Success rate (the proportion of cases with at least one hit in the top N predictions) and hit rate (the overall proportion of hits among the top N predictions) were calculated for N between 1 and 1000. The integrated success rate (ISR) and integrated hit rate (IHR) were derived from the plots of these rates against log(N) for N between 1 and 1000 predictions: the ISR and IHR represent the area under the curve normalized between 0 and 1 (Supplementary Fig. S2). The IHR was corrected to account for the maximum number of hits that can be present in the top N predictions.

The top 1000 predictions were clustered on the basis of ligand RMSD (lRMSD) using a cutoff of 7.5 Å. The distribution of hits among the 10 largest clusters was assessed. Details about the clustering (choice of the cutoff and alternative ranking) can be found in Supplementary Material.

The ISR was recently developed for performance comparison of scoring functions (Vreven *et al.*, 2011). The IHR is the corresponding metric in terms of hit rate (instead of success rate). A high ISR accounts for a good capacity to spot at least one hit for each case, whereas a high IHR accounts for the ability to correctly rank as many hits as possible. To increase the robustness of a prediction and of the clustering steps, not only a high ISR is required, but also a high IHR that guarantees that a maximum of near-native structures can be recovered.

The statistical significance of the results presented in this article (Tables 1–3) is assessed in the Supplementary Material.

3 RESULTS

3.1 Performance of InterEvScore

3.1.1 Statistical potential First, we assessed the scoring efficiency of the statistical potentials contained in InterEvScore on the subset of 131 complexes from the docking benchmark with at least one hit. The results are summarized in Table 1. The metrics used to evaluate the predictive power of the various scoring schemes are the ISR and the IHR, based on a recent performance comparison study (Vreven *et al.*, 2011) and which provide a normalized and synthetic evaluation of scoring methods (see Section 2). Supplementary Table S1 provides complementary metrics that were also used in other studies such as (Khashan *et al.*, 2012; Pierce and Weng, 2007).

The scoring schemes derived by including only the best contact for each residue perform significantly better than the schemes summing over all contacts. This might be due to the low resolution of the rigid-body docking approach where some unfavorable interface contacts in generated decoys may appear by chance, whereas local flexibility would remove them in the real interface. By choosing only the best contact for each residue, we retain only the most favorable sub-network of interface contacts for each decoy. It is noteworthy that these scoring schemes do not particularly bias InterEvScore toward larger interfaces, as illustrated in Supplementary Figure S3 and Supplementary Results.

We see in Table 1 that the three-body interaction scores appear less efficient than the two-body interaction scores. A possible explanation might be that by forcing the potential to include only three-body terms, we disfavored local situations only favorable in a pairwise context. Performance similar to that of the $2B^{best}$ score is recovered for the $2/3B^{best}$ score when each residue is given the opportunity to count only in its most favorable local structural environment, whether this environment involves a pairwise or a three-body contact.

Table 1. Benchmark v4 results of InterEvScore (statistical potential only) on 131 complexes with at least one hit

Metric	2B	$2B^{best}$	3B	$3B^{best}$	$2/3B^{best}$
ISR	0.172	**0.243**	0.174	0.214	**0.242**
IHR	0.070	**0.088**	0.044	0.077	**0.093**

Note: Highest ISR/IHR values in bold

Table 2. Benchmark v4 results of InterEvScore ($2B^{best}$ and $2/3B^{best}$ without evolutionary information, $2B^{best}_{cons}$ and $2/3B^{best}_{cons}$ with residue-based conservation, $2B^{best}_{evol}$ and $2/3B^{best}_{evol}$ with interface co-evolution) compared with ZDOCK, ZRANK and SPIDER on 54 complexes with at least one hit and available coupled MSAs

Metric	$2B^{best}$	$2/3B^{best}$	$2B^{best}_{cons}$	$2/3B^{best}_{cons}$	$2B^{best}_{evol}$	$2/3B^{best}_{evol}$	$2/3B^{best}_{evol}$ *(patches)*	ZDOCK	ZRANK	SPIDER
ISR	0.195	0.195	0.239	0.254	0.288	**0.321** (0.391)	**0.316**	0.259	0.241	(0.265)
IHR	0.103	0.111	0.119	0.138	0.147	**0.180** (0.199)	**0.206**	0.127	0.079	(0.071)

Note: Results between brackets correspond to the subset of 43 complexes not redundant with the SPIDER training set. Highest ISR/IHR values in bold.

Table 3. Benchmark results of the scoring functions: number of successful test cases among 54 test cases, defined by at least one hit in the top 10 predictions or at least one hit in the 5 or 10 largest clusters for each scoring function

Predictions	$2/3B^{best}_{evol}$	ZDOCK	ZRANK	SPIDER
Top 10 predictions	**14 (14)**	9	7	(7)
5 largest clusters	**21 (20)**	18	14	(10)
10 largest clusters	**25 (23)**	20	19	(12)

Note: Results between brackets correspond to the subset of 43 complexes not redundant with the SPIDER training set. Best results are in bold.

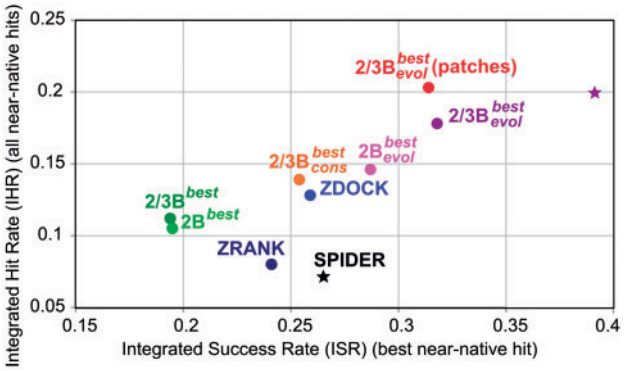

Fig. 3. Benchmark results in terms of ISR and IHR. Circles represent results on 54 test complexes. Stars represent results on the 43 complexes not redundant with the SPIDER training set (corresponding to values between brackets in Table 2); in particular, the star on the far right of the panel corresponds to results for $2/3B^{best}_{evol}$ on 43 test complexes

3.1.2 Inclusion of evolutionary information In the restricted set of 54 complexes for which evolutionary information can be collected, the inclusion of this information results in a large improvement in InterEvScore performance (see Table 2 and Fig. 3). The leftmost columns in Table 2 and Supplementary Table S2 show that the 'cons' scoring schemes (involving simple weighting of interface scores depending on residue conservation) perform better than the scoring schemes with no evolution. However, the 'evol' scoring scheme, which accounts for contact (not only residue) conservation, performs even better. For instance, the $2/3B^{best}_{evol}$ score ranks at least one hit in the top 10 predictions for 14 of 54 cases, versus five cases for the $2/3B^{best}$ score and seven for the $2/3B^{best}_{cons}$ score.

As shown in Table 2, Supplementary Table S3 and Figure 3, a major contribution is brought by the $2/3B^{best}_{evol}$ scheme over the $2B^{best}_{evol}$ and $3B^{best}_{evol}$ schemes. This probably reflects the advantage of providing local plasticity opportunities to each residue when including evolution.

Individual IHRs for each of the 54 test complexes are represented in Supplementary Figure S4. Strikingly, in the vast majority of these cases, the inclusion of evolutionary information leads to a visible improvement of the predictive capacity. In the rare cases where the individual IHR for the $2/3B^{best}_{evol}$ score is lower than the IHR for $2/3B^{best}$, either both IHRs are low or the difference between the two IHRs is small. In other words, the inclusion of evolutionary information is often helpful and never detrimental for the 54 cases tested here.

3.1.3 Focus on apolar patches InterEvScore can be used either in standard 'evol' mode (with all interface residues contributing similarly to the total score) or in an alternative mode (termed 'patches mode') in which evolutionary information is used only for contacts including at least one residue belonging to an apolar patch. In patches mode, there is an overall improvement of the scoring efficiency in terms of hit rate (see IHR values in Table 2). The fact that the IHR is improved implies that the profiles of the scores versus iRMSDs are more likely to display a funnel-like shape (Gray *et al.*, 2003). As illustrated by the comparison between individual IHR values in Supplementary Figure S4 and profiles of $2/3B^{best}_{evol}$ scores versus iRMSD in Supplementary Figure S5, we can define a cutoff of 0.1 for the individual IHR, which corresponds to a funnel-like shape. In 8 of 17 cases with individual IHR >0.1, a clear improvement by >20% of the IHR is observed, whereas in the remaining nine cases, inclusion of patches has a neutral effect. Overall, InterEvScore in patches mode has similar or better performance compared with standard mode. Results are summarized in Table 2 and Supplementary Table S4, and individual IHR values are plotted in Supplementary Figure S4.

3.2 Comparison with other scoring functions

3.2.1 Overall comparison In Table 2, Supplementary Tables S4 and S5 and Figure 3, InterEvScore is shown to perform significantly better than the previously published scoring functions ZDOCK, ZRANK and SPIDER. Results between brackets correspond to scores restricted to the subset of 43 (of 54) complexes in the test dataset, which are not redundant with the SPIDER training dataset; indeed, SPIDER was tested on benchmark v3

and thus trained on a dataset filtered for redundancy with benchmark v3, but not with v4 so that 11 cases had to be filtered out for proper comparison.

3.2.2 Clustering results In protein–protein docking, clustering is generally applied to reduce the complexity of the conformation space. For each scoring function, the top 1000 predictions were clustered on the basis of their mutual lRMSD values. The 10 largest clusters were retrieved and scanned for the presence of hits. The clustering results were compared with the hit and success rates.

For the $2/3B^{best}_{evol}$ scoring scheme, there are 14 cases with at least one hit in the top 10 predictions before the clustering step is applied (Table 3). Among these 14 cases, 11 have some hits among the five largest clusters, one has hits in the seventh largest cluster and only two have no hit in the 10 largest clusters. Moreover, owing to the clustering, we recover an additional 10 cases with hits in the five largest clusters and three cases with hits in the 6th to 10th largest clusters. Analyzing the five (respectively 10) largest clusters thus leads to identification of 21 (respectively 25) cases with hits. This illustrates the ability of InterEvScore to score consistently well close groups of decoys containing near-natives; this ability is precisely what the IHR metrics evaluates. Indeed, the low-resolution character of InterEvScore enables near-native interfaces that do not necessarily share common atomic details to obtain similarly favorable scores.

In Table 3, we compare InterEvScore with other scoring functions using the presence of at least one hit in either the top 10 predictions or the 5 or 10 largest clusters. The relative performance of the different scoring functions using these criteria is completely in keeping with the view provided by the ISR and IHR metrics.

Complete clustering results are provided in Supplementary Table S6. Interestingly, if we combine the three largest clusters obtained using InterEvScore, ZDOCK and ZRANK (nine clusters in total), these clusters contain a hit in 27 of the 54 test cases (50%). This illustrates the complementarity of our method with existing scoring functions.

4 DISCUSSION

To take advantage of both structural and evolutionary information toward the prediction of protein interfaces, we have developed InterEvScore, a novel scoring function using a multi-body potential coupled to contact evolution, which achieves significant improvement over traditional scoring functions on the 54 test cases from the docking benchmark with available coupled MSAs and near-native decoys.

The statistical potential in InterEvScore was trained on a large non-redundant dataset of non-obligate protein interfaces and derived using an analytical reference state, without fitting any parameters and with no optimization on a particular set of decoys. This derivation method strongly reduces the risk of overfitting.

InterEvScore is most efficient when it makes use of both two- and three-body contact information derived through coupled sequence alignments, retaining only the best contact for each interface residue in a fashion that gives each position the opportunity to count only in its most favorable environment. In our study of the evolution of homologous heteromeric interfaces (Andreani *et al*., 2012), we had identified a large amount of plasticity in the way contacts were conserved in interface evolution. This depends on sequence identity between homologous interfaces; hence, here we limit the divergence of our alignments by using the InterEvolAlign threshold of 35%. We see that despite the plasticity, evolutionary information derived about interface contacts provides significant improvement of our scoring function. Moreover, when the use of evolutionary contact information is targeted toward residues involved in apolar patches, we keep or reinforce the evolutionary signal; this might seem counter-intuitive because we restrict the set of considered interface residues, but it probably corresponds to a trade-off between depletion of the signal and exclusion of noisy information (because contacts between apolar patches are more conserved than others).

However, the inclusion of other interface descriptors that we had previously identified as important in the conservation of interface contacts proved difficult because of the low resolution character of our approach. Indeed, the identification of anchor residues or the definition of core, rim and support regions in the interface depend strongly on the interface itself, and as near-native decoys do not necessarily share these properties with the native interface, there is no discriminative power of such features in a rigid-body coarse-grained context. The steps of higher resolution docking such as those used in perturbation methods (Chaudhury *et al*., 2011) might benefit from the use of other conserved interface features as well. In contrast to other features, apolar patches can be defined on the surface of the protein independently of interface regions and including apolar patches proved useful in the discrimination of near-native decoys from others even at the low resolution docking step.

The docking benchmark exclusively contains non-obligate complexes (Hwang *et al*., 2010). In contrast with previous studies showing that the evolutionary signal was generally too weak to be used in predictive approaches (Halperin *et al*., 2006; Mintseris and Weng, 2005) or that many sequences were needed to detect an evolutionary signal (Weigt *et al*., 2009), here we find that the use of evolution is never detrimental and often useful in the discrimination of near-native interfaces, even though the number of sequences in the coupled alignments is limited (between 10 and 100 species with an average of 35). In addition, unlike a previous docking study using evolutionary trace analysis (Kanamori *et al*., 2007), we do not find the scoring improvement on inclusion of evolutionary data to be limited to certain categories of complexes. Apart from antigen-antibody complexes to which our approach is not applicable, InterEvScore is successful on a variety of cases.

The strength of our approach thus relies on the use of available evolutionary information coupled to a knowledge-based potential. Several fields of application can be considered for this scoring function, especially in a context where low-resolution strategies remain important (Vakser, 2013), and structural information is starting to be used on a large-scale with a view toward the prediction of interactomes (Zhang *et al*., 2012).

5 AVAILABILITY AND IMPLEMENTATION

InterEvScore is freely available as a Python script, whose output contains relevant scores (with or without the inclusion of

evolutionary information). The script can be run in standard or patches mode. The scoring of large decoy sets can be easily parallelized. Scoring relies on the calculation of α-shape contacts but InterEvScore can also be run in degraded mode using distance-based contacts (see Supplementary Data). A script to cluster results based on lRMSD values is also provided with InterEvScore.

ACKNOWLEDGEMENTS

The authors thank A. Martel for his help in setting up the InterEvScore web page and F. Ochsenbein for critical reading of the manuscript.

Funding: CEA, the ANR-SVSE3-2010-DYNAMOPHAGE, the ANR-IAB-2011-BIP:BIP and the French Infrastructure for Integrated Structural Biology (FRISBI) ANR-10-INSB-05-01 grants.

Conflict of Interest: none declared.

REFERENCES

Akbal-Delibas,B. *et al.* (2012) An evolutionary conservation-based method for refining and reranking protein complex structures. *J. Bioinform. Comput. Biol.*, **10**, 1242002.

Aloy,P. *et al.* (2003) The relationship between sequence and interaction divergence in proteins. *J. Mol. Biol.*, **332**, 989–998.

Andreani,J. *et al.* (2012) Versatility and invariance in the evolution of homologous heteromeric interfaces. *PLoS Comput. Biol.*, **8**, e1002677.

Andrusier,N. *et al.* (2007) FireDock: fast interaction refinement in molecular docking. *Proteins*, **69**, 139–159.

Berman,H.M. *et al.* (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.

Chaudhury,S. *et al.* (2011) Benchmarking and analysis of protein docking performance in Rosetta v3.2. *PLoS One*, **6**, e22477.

Chen,R. *et al.* (2003) ZDOCK: an initial-stage protein-docking algorithm. *Proteins*, **52**, 80–87.

Cheng,T.M. *et al.* (2007) pyDock: electrostatics and desolvation for effective scoring of rigid-body protein-protein docking. *Proteins*, **68**, 503–515.

de Vries,S.J. *et al.* (2006) WHISCY: what information does surface conservation yield? Application to data-driven docking. *Proteins*, **63**, 479–489.

Dominguez,C. *et al.* (2003) HADDOCK: a protein-protein docking approach based on biochemical or biophysical information. *J. Am. Chem. Soc.*, **125**, 1731–1737.

Edgar,R.C. (2010) Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, **26**, 2460–2461.

Engelen,S. *et al.* (2009) Joint evolutionary trees: a large-scale method to predict protein interfaces based on sequence sampling. *PLoS Comput. Biol.*, **5**, e1000267.

Faure,G. *et al.* (2012) InterEvol database: exploring the structure and evolution of protein complex interfaces. *Nucleic Acids Res.*, **40**, D847–D856.

Feng,Y. *et al.* (2007) Four-body contact potentials derived from two protein datasets to discriminate native structures from decoys. *Proteins*, **68**, 57–66.

Fiorucci,S. and Zacharias,M. (2010) Binding site prediction and improved scoring during flexible protein-protein docking with ATTRACT. *Proteins*, **78**, 3131–3139.

Fitzgerald,J.E. *et al.* (2007) Reduced C(beta) statistical potentials can outperform all-atom potentials in decoy identification. *Protein Sci.*, **16**, 2123–2139.

Goodsell,D.S. and Olson,A.J. (2000) Structural symmetry and protein function. *Annu. Rev. Biophys. Biomol. Struct.*, **29**, 105–153.

Gray,J.J. *et al.* (2003) Protein-protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations. *J. Mol. Biol.*, **331**, 281–299.

Halperin,I. *et al.* (2006) Correlated mutations: advances and limitations. A study on fusion proteins and on the Cohesin-Dockerin families. *Proteins*, **63**, 832–845.

Huang,S.Y. and Zou,X. (2010) MDockPP: a hierarchical approach for protein-protein docking and its application to CAPRI rounds 15-19. *Proteins*, **78**, 3096–3103.

Hwang,H. *et al.* (2010) Protein-protein docking benchmark version 4.0. *Proteins*, **78**, 3111–3114.

Janin,J. (2010) Protein-protein docking tested in blind predictions: the CAPRI experiment. *Mol. Biosyst.*, **6**, 2351–2362.

Kanamori,E. *et al.* (2007) Docking of protein molecular surfaces with evolutionary trace analysis. *Proteins*, **69**, 832–838.

Khashan,R. *et al.* (2012) Scoring protein interaction decoys using exposed residues (SPIDER): a novel multibody interaction scoring function based on frequent geometric patterns of interfacial residues. *Proteins*, **80**, 2207–2217.

Kozakov,D. *et al.* (2006) PIPER: an FFT-based protein docking program with pairwise potentials. *Proteins*, **65**, 392–406.

Krishnamoorthy,B. and Tropsha,A. (2003) Development of a four-body statistical pseudo-potential to discriminate native from non-native protein conformations. *Bioinformatics*, **19**, 1540–1548.

Kundrotas,P.J. *et al.* (2012) Templates are available to model nearly all complexes of structurally characterized proteins. *Proc. Natl Acad. Sci. USA*, **109**, 9438–9441.

Landau,M. *et al.* (2005) ConSurf 2005: the projection of evolutionary conservation scores of residues on protein structures. *Nucleic Acids Res.*, **33**, W299–W302.

Lensink,M.F. *et al.* (2007) Docking and scoring protein complexes: CAPRI 3rd edition. *Proteins*, **69**, 704–718.

Lensink,M.F. and Wodak,S.J. (2010) Docking and scoring protein interactions: CAPRI 2009. *Proteins*, **78**, 3073–3084.

Levy,E.D. *et al.* (2008) Assembly reflects evolution of protein complexes. *Nature*, **453**, 1262–1265.

Li,X. *et al.* (2003) Simplicial edge representation of protein structures and alpha contact potential with confidence measure. *Proteins*, **53**, 792–805.

Li,X. and Liang,J. (2005) Geometric cooperativity and anticooperativity of three-body interactions in native proteins. *Proteins*, **60**, 46–65.

Madaoui,H. and Guerois,R. (2008) Coevolution at protein complex interfaces can be detected by the complementarity trace with important impact for predictive docking. *Proc. Natl Acad. Sci. USA*, **105**, 7708–7713.

Mayrose,I. *et al.* (2004) Comparison of site-specific rate-inference methods for protein sequences: empirical Bayesian methods are superior. *Mol. Biol. Evol.*, **21**, 1781–1791.

Mendez,R. *et al.* (2005) Assessment of CAPRI predictions in rounds 3-5 shows progress in docking procedures. *Proteins*, **60**, 150–169.

Mintseris,J. *et al.* (2007) Integrating statistical pair potentials into protein complex prediction. *Proteins*, **69**, 511–520.

Mintseris,J. and Weng,Z. (2005) Structure, function, and evolution of transient and obligate protein-protein interactions. *Proc. Natl Acad. Sci. USA*, **102**, 10930–10935.

Mosca,R. *et al.* (2013) Interactome3D: adding structural details to protein networks. *Nat. Methods*, **10**, 47–53.

Ngan,S.C. *et al.* (2006) A knowledge-based scoring function based on residue triplets for protein structure prediction. *Protein Eng. Des. Sel.*, **19**, 187–193.

Ofran,Y. and Rost,B. (2007) ISIS: interaction sites identified from sequence. *Bioinformatics*, **23**, e13–e16.

Pierce,B. and Weng,Z. (2007) ZRANK: reranking protein docking predictions with an optimized energy function. *Proteins*, **67**, 1078–1086.

Reichmann,D. *et al.* (2007) The molecular architecture of protein-protein binding sites. *Curr. Opin. Struct. Biol.*, **17**, 67–76.

Res,I. *et al.* (2005) An evolution based classifier for prediction of protein interfaces without using protein structures. *Bioinformatics*, **21**, 2496–2501.

Ritchie,D.W. *et al.* (2008) Accelerating and focusing protein-protein docking correlations using multi-dimensional rotational FFT generating functions. *Bioinformatics*, **24**, 1865–1873.

Sinha,R. *et al.* (2010) Docking by structural similarity at protein-protein interfaces. *Proteins*, **78**, 3235–3241.

Stein,A. *et al.* (2011) Three-dimensional modeling of protein interactions and complexes is going 'omics. *Curr. Opin. Struct. Biol.*, **21**, 200–208.

Tress,M. *et al.* (2005) Scoring docking models with evolutionary information. *Proteins*, **60**, 275–280.

Vajda,S. and Kozakov,D. (2009) Convergence and combination of methods in protein-protein docking. *Curr. Opin. Struct. Biol.*, **19**, 164–170.

Vakser,I.A. (2013) Low-resolution structural modeling of protein interactome. *Curr. Opin. Struct. Biol.*, pii, S0959-440X(12)00199-6.

Vreven,T. *et al.* (2011) Integrating atom-based and residue-based scoring functions for protein-protein docking. *Protein Sci.*, **20**, 1576–1586.

Weigt,M. *et al.* (2009) Identification of direct residue contacts in protein-protein interaction by message passing. *Proc. Natl Acad. Sci. USA*, **106**, 67–72.

Zellner,H. *et al.* (2012) PresCont: predicting protein-protein interfaces utilizing four residue properties. *Proteins*, **80**, 154–168.

Zhang,C. *et al.* (2004) An accurate, residue-level, pair potential of mean force for folding and binding based on the distance-scaled, ideal-gas reference state. *Protein Sci.*, **13**, 400–411.

Zhang,Q.C. *et al.* (2012) Structure-based prediction of protein-protein interactions on a genome-wide scale. *Nature*, **490**, 556–560.

Zhu,H. *et al.* (2006) NOXclass: prediction of protein-protein interaction types. *BMC Bioinformatics*, **7**, 27.