Structural Bioinformatics

Atomic-level evolutionary information improves protein-protein interface scoring

Chloé Quignot[1], Pierre Granger[1], Pablo Chacón[2], Raphael Guerois[1,*] and Jessica Andreani[1,*]

[1]Université Paris-Saclay, CEA, CNRS, Institute for Integrative Biology of the Cell (I2BC), 91198, Gif-sur-Yvette, France., [2]Department of Biological Chemical Physics, Rocasolano Institute of Physical Chemistry C.S.I.C, Madrid, Spain.

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: The crucial role of protein interactions and the difficulty in characterising them experimentally strongly motivates the development of computational approaches for structural prediction. Even when protein-protein docking samples correct models, current scoring functions struggle to discriminate them from incorrect decoys. The previous incorporation of conservation and coevolution information has shown promise for improving protein-protein scoring. Here, we present a novel strategy to integrate atomic-level evolutionary information into different types of scoring functions to improve their docking discrimination.

Results: We applied this general strategy to our residue-level statistical potential from InterEvScore and to two atomic-level scores, SOAP-PP and Rosetta interface score (ISC). Including evolutionary information from as few as ten homologous sequences improves the top 10 success rates of individual atomic-level scores SOAP-PP and Rosetta ISC by respectively 6 and 13.5 percentage points, on a large benchmark of 752 docking cases. The best individual homology-enriched score reaches a top 10 success rate of 34.4%. A consensus approach based on the complementarity between different homology-enriched scores further increases the top 10 success rate to 40%.

Availability: All data used for benchmarking and scoring results, as well as a Singularity container of the pipeline, are available at http://biodev.cea.fr/interevol/interevdata/

Contact: jessica.andreani@cea.fr or guerois@cea.fr

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

1 Introduction

Proteins are key actors in a great number of cellular functions and often work in collaboration with others, thereby forming interaction networks. Knowledge of the detailed 3D structure of protein-protein interfaces can help to better understand the mechanisms they are involved in. Difficulties in the experimental determination of protein assembly structures have prompted the development of *in silico* prediction strategies such as molecular docking. When no homologous interface structure can be identified and used as a template, free docking is used instead, involving a systematic search where many interface conformations are sampled (Huang, 2014; Porter, et al., 2019). These interface models are then scored according to properties such as interface physics, chemistry, and statistics (Huang, 2015; Moal, et al., 2013). Guided docking approaches integrating complementary sources of information are also becoming increasingly popular (Koukos and Bonvin, 2019).

Over time, protein interfaces are submitted to evolutionary pressure to maintain functional interactions. Thus, protein interfaces tend to be more conserved than other regions on their surface (Mintseris and Weng, 2005; Teichmann, 2002) and signs of coevolution can be detected at protein interfaces, where potentially disrupting mutations are compensated for with mutations in contacting positions on the protein partner. These phenomena of conservation and coevolution can provide useful information in the analysis and prediction of their 3D interface structures (Andreani, et al., 2020). For example, evolutionary information is at the

heart of increasingly popular covariation-based approaches, such as statistical coupling analysis (SCA) (Socolich, et al., 2005) or direct coupling analysis (DCA) (Morcos, et al., 2011), for structural proximity prediction of residues based on multiple sequence alignments (MSAs). These approaches can be used to guide protein folding or to supplement predictions of macromolecular interactions (Cocco, et al., 2018; Simkovic, et al., 2017). The vast majority of protein interaction site predictors successfully use evolutionary information, be it by sequence conservation, sequence co-evolution, or through homologous structures (Andreani, et al., 2020).

Evolutionary information can also be especially useful to guide molecular docking (Geng, et al., 2019). The InterEvDock2 server implements a docking pipeline that uses evolutionary information (Quignot, et al., 2018; Yu, et al., 2016). It takes advantage of the spherical Fourier-based rigid-body docking programme FRODOCK2.1 (Ramírez-Aportela, et al., 2016) for the sampling step and hands out a set of ten most probable interfaces based on a consensus between three different scores, FRODOCK2.1's mostly physics-based score, SOAP-PP's atomic statistical potential (Dong, et al., 2013) and InterEvScore (Andreani, et al., 2013). InterEvScore extracts co-evolutionary information from joint multiple sequence alignments of the binding partners (called coMSAs), but unlike covariation-based approaches such as DCA cited above, InterEvScore needs only a small number of homologous sequences to improve discrimination between correct and incorrect models, by combining coMSAs with a multi-body residue-level statistical potential. As seen in the benchmarking of InterEvDock2, InterEvScore presents a high complementarity with SOAP-PP (Quignot, et al., 2018). As both scores are based on statistical potentials but SOAP-PP has an atomic level of detail, we hypothesised that a score integrating evolutionary information at an atomic scale might pick up on finer properties to better distinguish near-natives from the rest of the decoys.

In InterEvScore, evolutionary information is given implicitly at residue-level through coMSAs and is combined with a coarse-grained statistical potential. A major challenge in deriving evolutionary information to an atomic level of detail is finding a suitable way of representing residue-scale information from coMSAs at an atomic level. Here, we present a novel strategy to couple evolutionary information with atomic scores to improve model discrimination. We reconstruct an equivalent and hypothetical interfacial atomic contact network for each interface model and each pair of homologs present in the coMSAs, by using a threading-like strategy to generate explicit backbone and side-chain coordinates. These models can, in turn, be scored with non-evolutionary atomic-resolution scoring functions such as SOAP-PP (Dong, et al., 2013) or Rosetta interface score (ISC) (Chaudhury, et al., 2011; Gray, et al., 2003).

Here, we show that including explicit evolutionary information improves the top 10 success rate of SOAP-PP and ISC by 6 and 13.5 percentage points respectively, on a large benchmark of 752 docking cases for which evolutionary information can be used (Yu and Guerois, 2016). We then use a consensus approach to take advantage of the complementarity between different scores. The top 10 success rate of a consensus integrating FRODOCK2.1 with InterEvScore and SOAP-PP increases from 32% to 36% when including the homology-enriched score variants. A more time-consuming consensus combining all scores with an explicit homolog representation reaches 40% top 10 success rate.

2 Methods

2.1 Docking benchmark

We evaluated docking methods using the large docking benchmark PPI4DOCK (Yu and Guerois, 2016), where unbound structures unavailable from experiments were modelled by homology from unbound homologous templates. This is especially important since bound docking is strongly biased due to shape complementarity (see supplementary Table S19). We excluded antigen-antibody interactions due to their specific interaction mode and evolutionary properties, leaving 1279 docking cases. Each case in PPI4DOCK is associated to a coMSA, i.e. a pair of joint MSAs for the two docking partners. Sampling was performed with FRODOCK2.1 using models from unbound structures as starting points (see supplementary methods) and keeping only the top 10,000 generated models. In the supplementary information, we show benchmarking performance using ZDOCK 3.0.2 (Pierce, et al., 2011) as an alternative sampling program. Near-native models were defined as being of Acceptable or better quality following the criteria from CAPRI (Critical Assessment of PRediction of Interactions) (Mendez, et al., 2003).

To focus the study on scoring performance and the usefulness of co-evolutionary information for this purpose, benchmarking results in the main figures and tables are shown on 752 cases that have more than 10 sequences in their coMSAs and at least one near-native within the top 10,000 FRODOCK2.1 models (supplementary Tables S1 and S2). In the supplementary information, we show benchmarking performance on the 1279 non-antigen-antibody cases from PPI4DOCK, as well as 230 cases from the protein docking benchmark version 5 (Vreven, et al., 2015).

The 1279 PPI4DOCK cases are split into five difficulty categories (supplementary Table S3). 74% of cases are amenable to rigid-body sampling but represent a challenge for scoring, including a majority of cases (55% of the total) in the 'easy' category, with moderate conformational changes between unbound and bound structure, and the other 19% in the 'very_easy' category, with small conformational changes. The remaining 26% ('hard', 'very_hard' and 'super_hard' categories) correspond to larger conformational changes, and sampling can generate an Acceptable model only in very few of those cases.

Performance was measured by top N success rate, i.e. the percentage of cases with at least one near-native in the top N ranked models. We especially focus on the top 10 success rate traditionally used as a docking metric, and the top 50 success rate since consensus computation typically involves the top 50 models of each score (see section 2.2.1).

2.2 Scoring functions

In addition to the sampling programme's integrated score, we rescored models and their threaded homologs with InterEvScore, SOAP-PP, and Rosetta ISC. In the supplementary information, we also show performance when rescoring ZDOCK models with ZRANK (Pierce and Weng, 2007).

InterEvScore combines co-evolutionary information taken from coMSAs with a residue-level statistical potential (Andreani, et al., 2013). It was re-implemented to accelerate scoring (see supplementary methods).

SOAP-PP is an atomic statistical-based score integrating distance-dependent potentials (Dong, et al., 2013). Here, we use a faster in-house implementation of this score (see supplementary methods).

Rosetta ISC includes a linear combination of non-bonded atom-pair interaction energies and empirical and statistical potentials (Chaudhury, et al., 2011; Gray, et al., 2003). ISC is calculated by subtracting the energy of both monomeric structures from the energy of the complex structure. Since Rosetta ISC is sensitive to small variations and clashes at the interface, we included high-resolution interface side-chain optimisation during ISC scoring (see supplementary methods). Models for which Rosetta scoring did not converge after 10 iterations were assigned the

worst score for that case. As Rosetta ISC scoring can take up to a couple of minutes per structure, we scored only the top 1,000 FRODOCK2.1 models (noted later 1k) per case rather than 10,000 (noted 10k).

2.2.1 Consensus scores

The aim of the consensus is to preferentially select models supported by several scores. Consensus calculations were performed similarly to InterEvDock2 (Quignot, et al., 2018) to obtain a set of 10 most likely models depending on the agreement between several scoring functions. Here, we apply consensus scoring to combinations of 3 to 5 different scoring functions. For a given set of scoring functions, ordered by their individual performances from best to worst performing, the top 10 models of each scoring function receive a convergence count based on the number of similar models (defined as L-RMSD ≤ 10 Å) that are found in the top 50 models of each other scoring function. The final 10 consensus models are selected iteratively by decreasing convergence count (if > 1). In the case of a tie, models are selected according to the ranking order of their respective scoring functions. Models are added to the top 10 consensus only if they are not structurally redundant with the already selected ones (L-RMSD > 10 Å). If necessary, the consensus list is completed up to 10 models by selecting the top 4, 3, 3 models for a consensus between three scoring functions (or the top 3, 3, 2, 2 or top 2, 2, 2, 2, 2 models for a consensus between four or five scoring functions, respectively).

2.3 Docking strategy to integrate evolutionary information

The proposed homology-enriched docking pipeline consists of four steps outlined in Figure 1. First, we dock query proteins A and B for which we are trying to predict the 3D structure of the complex, using FRODOCK2.1 (Ramírez-Aportela, et al., 2016). This results in a set of rotational and translational transforms that define a maximum of 10,000 complex models (Figure 1A). In parallel, we subsample coMSAs to a subset of M pairs of homologs (proteins A_1 and B_1, A_2 and B_2, ..., A_M and B_M, homologs of query proteins A and B respectively) (see section 2.3.1). We model the unbound structures of these M pairs of homologs, using the threading function from RosettaCM's pipeline (Song, et al., 2013) and the unbound query protein structures as templates (see Figure 1B and section 2.3.2). We then generate complex equivalents to each query model by applying the translational and rotational transforms obtained in the docking step to each pair of homologs. Figure 1C illustrates this reconstruction for the first pair of homologs (proteins A_1 and B_1). Finally, we average scores over the query model and its homolog models to obtain a final per-model score that integrates all the information (Figure 1D). Note that for one case, we have to compute (M+1) x N scores to obtain the final ranking of N models. The scoring functions we used are described in section 2.2. All steps of the pipeline are easily parallelisable to reduce end-user runtime, whether through MPI (sampling step) or by splitting over models (scoring steps).

2.3.1 Subsampling homologs in the coMSAs

Homologous sequences used in scoring were taken from the coMSAs provided with the PPI4DOCK benchmark, which contain homolog pairs with minimum 30% sequence identity and 75% coverage to the query complex. This aims to ensure that the interaction mode is conserved between homologous sequences (Andreani, et al., 2013). The coMSAs were reduced to maximum M=40, and then to M=10 sequences (plus the query sequence) to limit computational time. Indeed, it was already seen with InterEvScore that co-evolutionary information can be extracted from alignments with as few as 10 sequences (Andreani, et al., 2013). The sequences in the coMSAs are ordered by decreasing average sequence identity with the query sequences. This is important when sub-selecting sequences to keep a representative subset. Sequence selection was performed in three steps. First, the number of sequences was cut to keep at most 100 sequences with highest sequence identity to the query, as in the InterEvDock2 pipeline. Then the alignment was filtered with hhfilter 3.0.3 (Remmert, et al., 2011) from the hh-suite package. hhfilter was applied with the "-diff X" option on the concatenated coMSAs, adjusting the value of X for each case to return a reduced alignment with no more than 41 sequences (i.e. the query + 40 homologs). At this stage, we obtain coMSA[40], the first set of reduced coMSAs with maximum 40 sequences, representative of the diversity of the initial coMSAs. Finally, 11 equally distributed sequences (i.e. the query + 10 homologs) were uniformly selected within coMSA[40] to preserve sequence diversity compared to the initial coMSAs (see supplementary methods). The final set of reduced coMSAs is called coMSA[10].

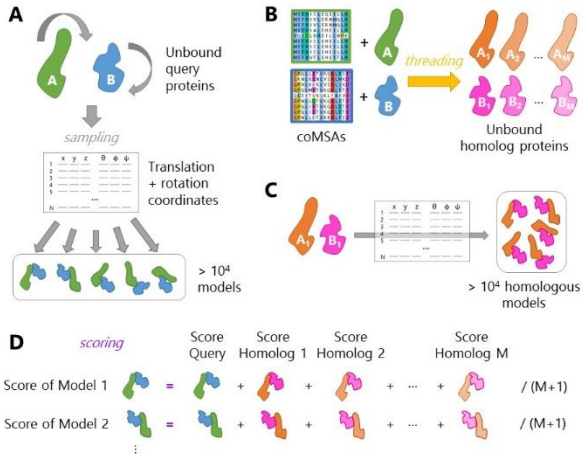


Fig. 1. Docking pipeline with explicit modelling of model homologs. (A) Upon docking of query unbound structures (proteins A and B in green and blue), FRODOCK2.1 outputs a rotation and translation matrix to reconstruct the corresponding models. (B) To generate their homologous counterparts, the unbound structures of each homolog (proteins A_1 and B_1, A_2 and B_2, ..., A_M and B_M, in various shades of orange and magenta) are threaded based on the query unbound structures and the homologous sequence alignments in the coMSAs of the query proteins. (C) For each homolog pair (such as homolog 1 illustrated here), models can be reconstructed using the same rotation and translation matrix as for the query. (D) The final score of each query model (left column) corresponds to the average score over itself and its M homologs for a given scoring function.

2.3.2 Threading models

Pairwise alignments between the template structure and the homolog sequence to be modelled were directly extracted from the reduced coMSAs. The templates used for threading were the unbound template structures provided in the PPI4DOCK benchmark (Yu and Guerois, 2016) so that no information about the bound structure is introduced when rescoring homologs (see supplementary methods).

Rosetta's threading protocol partial_thread, the first step in the RosettaCM pipeline (Song, et al., 2013), was used to thread the homologous sequences onto the template structure (Figure 1B). We used Rosetta 3.8 (version 2017.08.59291). Insertions (gaps) and termini that were missing from the template structure were not modelled. No refinement or side-chain optimisation was applied at that stage, since InterEvScore and SOAP-PP are not sensitive to small interface clashes, as both were developed to score rigid-body docking models. Reconstructed interface models (Figure 1C) were not relaxed, as this would be

computationally prohibitive and does not bring an obvious performance advantage (see supplementary Figure S9).

3 Results

3.1 Consensus approach with implicit homology scoring

In previous work, we integrated evolutionary information implicitly at the coarse-grained level by scoring models with residue-based InterEvScore (noted IES) (Andreani, et al., 2013). In IES, for each model, we enumerate all residue-level interface contacts. We then use a residue-level statistical potential to score models by considering all sequences in a coMSA and assuming the same contacts were conserved in all homologous interfaces.

We also combined InterEvScore with complementary scores FRODOCK2.1 and SOAP-PP (supplementary Figure S1A) in a three-way consensus score, denoted Cons^3, which preferentially selects models supported by several scores (section 2.2.1) (Quignot, et al., 2018; Yu, et al., 2016). Compared to individual scores, we observed a notable boost of about 8 points in the top 10 success rate using Cons^3, which captures a near-native in the top 10 models in 32% of the cases (Table 1 and Figure 2A).

This complementarity between scores, in particular SOAP-PP and InterEvScore, (supplementary Figure S1A), prompted us to attempt atomic-level integration of evolutionary information into the scores. Following the pipeline described in methods section 2.3 (Figure 1), in the next sections, we include evolutionary information into InterEvScore and SOAP-PP through explicit atomic-level homologous models.

Table 1: Performance of consensus scores including InterEvScore implicit homology scoring. Scores used in three-way consensus score Cons^3 were SOAP-PP on the top 10,000 FRODOCK2.1 models (SPP/10k), InterEvScore on full coMSAs and on the top 10,000 FRODOCK2.1 models (IES/10k) and FRODOCK2.1 (FD2.1). Performances of individual scores used in the consensus are reported in terms of top 10 and top 50 success rates since consensus calculation relies on the top 50 models ranked by each component score.

Score	Top 10 success rate	Top 50 success rate
FD2.1	164 (21.8%)	292 (38.8%)
IES/10k	182 (24.2%)	287 (38.2%)
SPP/10k	183 (24.3%)	328 (43.6%)
Cons^3	**241 (32.0%)**	/

3.2 InterEvScore with explicitly modelled homologs

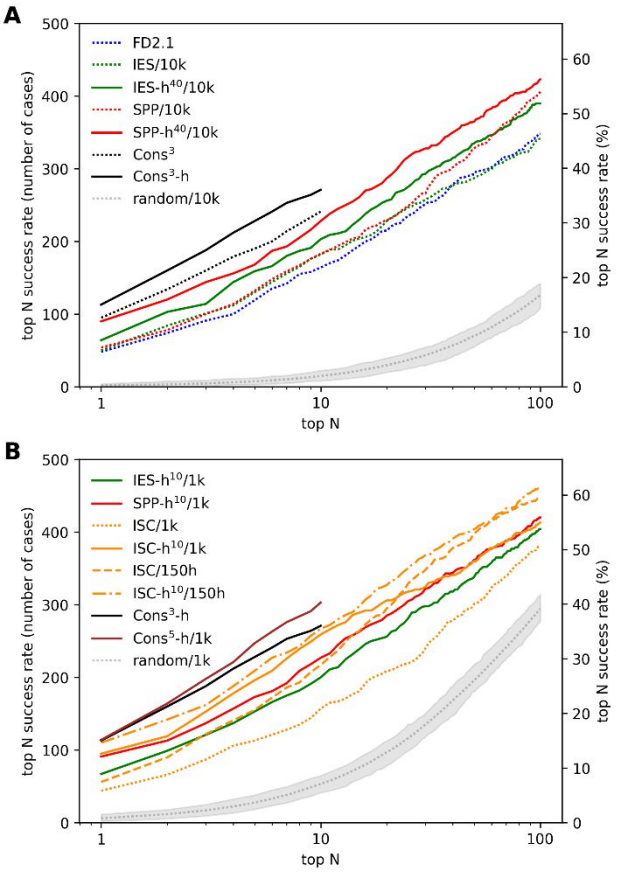
For efficiency, we represent homologs at atomic resolution by threading their sequences onto the query structure (section 2.3.2). As a first step to validate this new representation of evolutionary information, we test the performance of InterEvScore on these threaded models and compare it with the original InterEvScore. With the threaded models, atomic contacts are re-defined in each homolog at an explicit level, rather than implicitly deduced from the coMSAs as in the original InterEvScore. In practice, we calculate the threaded homolog version of InterEvScore (denoted IES-h) by scoring query interface models and their threaded homolog equivalents with the InterEvScore statistical potentials (section 2.3). The final score of each query model is the average over the query model and its homologs.

Table 2 and Figure 2A show the performance of IES-h^{40}, i.e. IES-h computed using threaded homologs from the set of reduced coMSAs with a maximum of 40 sequences (coMSA^{40}, see section 2.3.1). Results for the original InterEvScore with complete coMSAs (IES) and coMSAs^{40} (IES^{40}) are also shown for comparison. Reducing the number of sequences to 40 does not strongly affect the top 10 and top 50 success rates. However, the top 10 success rate increases from 23.8% to 27.0% when using explicit threaded models (IES-h^{40}) instead of implicit coMSA information (IES^{40}).

Table 2: Performance of InterEvScore using coMSAs without or with threaded models. Top 10 and top 50 success rates of InterEvScore on complete coMSAs (IES, reported in Table 1) and coMSA^{40} (IES^{40}) compared to InterEvScore using explicit threaded models of homologs in coMSA^{40} (IES-h^{40}) on 10,000 models (/10k). Performances were measured on 752 benchmark cases.

Score	Top 10 success rate	Top 50 success rate
IES/10k	182 (24.2%)	287 (38.2%)
IES^{40}/10k	179 (23.8%)	284 (37.8%)
IES-h^{40}/10k	**203 (27.0%)**	**335 (44.5%)**

Fig. 2. Success rate as a function of the number of ranked models for individual and consensus scores. We plot success rates on 752 PPI4DOCK cases, as a function of the number of top N models with N going from 1 to 100, for (A) FRODOCK2.1 (FD2.1), SOAP-PP (SPP) and InterEvScore (IES) individual and consensus scores (dotted lines) on 10,000 models (10k) and their homology-enriched variants on coMSA^{40} (-h^{40}, solid lines); (B) Rosetta ISC score (dotted line) together with homology-enriched variants of individual scores on coMSA^{10} (-h^{10}) and 1,000 models (1k) or the homology-enriched subset of 150 models (150h) and homology-enriched consensus scores. Consensus scores produce only a selection of 10 models, hence they stop at N=10. We obtained a reference performance curve with 95% confidence interval (in grey) by shuffling at random the top 10,000 (A) or top 1,000 models (B) and assessing the first N models (see supplementary methods).



The difference in performance between IES^{40}/10k and IES-h^{40}/10k can

be explained by the fact that, in IES-h^{40}, contacts are not extrapolated from the query interface network anymore but are redefined in each homolog based on their modelled interface structure.

3.3 Homology-enriched SOAP-PP

Having explicit structures at atomic resolution corresponding to each homolog enables us to score them directly using an atomic potential such as SOAP-PP (Dong, et al., 2013), which might be able to better exploit the atomic detail of homologs for the final ranking of query interface models. As for the threaded version of InterEvScore, homology-enriched SOAP-PP (SPP-h^{40}) consists in the average SOAP-PP score over all homologs.

SPP-h^{40} performs better than SOAP-PP on the query interface models alone (Table 3 and Figure 2A). Using threaded homology models in this way gives a large performance boost to SOAP-PP (+6 percentage points on the top 10 success rate). SPP-h^{40} also outperforms InterEvScore and IES-h^{40} (section 3.2) as well as the FRODOCK2.1 score (section 3.1).

Table 3: Performance of SOAP-PP against SPP-h^{40}. Top 10 and top 50 success rates of SOAP-PP (SPP) compared to its homology-enriched version SPP-h^{40} over sequences in coMSA^{40} on 10,000 models (/10k). Performances were measured on 752 benchmark cases.

Score	Top 10 success rate	Top 50 success rate
SPP/10k	183 (24.3%)	328 (43.6%)
SPP-h^{40}/10k	**228 (30.3%)**	**365 (48.5%)**

3.4 Homology-enriched Rosetta interface score (ISC)

Since we build atomic-level homologous interface models, we can score them explicitly using a physics-based score such as Rosetta ISC. As Rosetta scoring is much more computationally expensive (about 750 times slower) than SOAP-PP and InterEvScore, to compute homology-enriched ISC, the number of models was reduced to 1,000 (as ranked by FRODOCK2.1) and the number of homologs to 10 (coMSA^{10}, section 2.3.1).

As above, homology-enriched ISC consisted in the average score of the query and its homologous interface models (ISC-h^{10}). For easier comparison, homology-enriched InterEvScore and SOAP-PP were evaluated in the same conditions (*i.e.* 1,000 models and coMSAs^{10}) (Table 4 and Figure 2B). Their success rates are very similar to those with 10,000 models and coMSAs^{40} (supplementary Table S4). Even though ISC on query models performs worse than SPP-h and IES-h, ISC-h^{10} largely outperforms the best-performing individual score, SPP-h^{10}, with 34.4% top 10 success rate (259 cases) compared to 30.2% (227). With only 165 successful cases in common, SPP-h^{10} and ISC-h^{10} remain very complementary (supplementary Figure S1B).

Note that for scores calculated on the top 1,000 FRODOCK2.1 models, success rates are technically capped to 77.1%, as only 580 cases out of the 752 in our benchmark have a near-native within this subset of models. In light of this, the ISC-h^{10}/1k performance is all the more remarkable.

Table 4: Scoring performance of Rosetta homology-enriched ISC. Scoring performance of ISC on query interface models only and using the threaded homology models (ISC-h^{10}) on top 1,000 FRODOCK2.1 models (1k) and coMSA^{10} as well as the performance of SPP-h^{10} and IES-h^{10} on 1,000 FRODOCK2.1 models with coMSA^{10} for easier comparison. Performances were measured as the top 10 and top 50 success rates on 752 benchmark cases.

Score	Top 10 success rate	Top 50 success rate
IES-h^{10}/1k	200 (26.6%)	338 (44.9%)
SPP-h^{10}/1k	227 (30.2%)	**362 (48.1%)**
ISC/1k	157 (20.9%)	301 (40.0%)
ISC-h^{10}/1k	**259 (34.4%)**	360 (47.9%)

3.4.1 Using ISC to re-score homology-enriched interface models

ISC-h^{10} showed the highest top 10 success rate from all scores tested above, but scoring 1,000 x 11 models with Rosetta ISC is excessively time consuming in a generalised docking context as it takes approximatively 137 CPU hours per case (supplementary Table S5). One way to alleviate the total scoring time is to score only a pre-selected amount of interface models, using Rosetta ISC as a second step in the scoring pipeline.

In Cons^3, we pre-selected the top 50 models of FRODOCK2.1, InterEvScore, and SOAP-PP. Similarly, here we use the top 50 models of the top-performing homology-enriched score variants tested above, namely SPP-h^{40}/10k and IES-h^{40}/10k, as well as FRODOCK2.1. These scores have a high complementarity in terms of top 10 success rate with only 67 cases found in common between all three (supplementary Figure S1C). Using this subset of 150 pre-selected models for ISC scoring (referred to with /150h) reduced scoring times approximately by a factor 7. We enrich near-natives in this set of 150 models since they were pre-selected by three already well-performing scores, but only 476 out of 752 cases in our benchmark possess a near-native in this subset.

In terms of the top 10 success rate, both ISC-h^{10} and ISC perform better on 150 than 1,000 models with 35.5% and 28.9% top 10 success rate instead of 34.4% and 20.9%, respectively (Tables 4 and 5 and Figure 2B). Here again, the addition of evolutionary information to ISC through the threaded homolog models remarkably increases its performance. ISC-h^{10}/150h has the best performance of all tested scores so far, for a much lower computational cost than ISC-h^{10}/1k.

Table 5: Performance of ISC and ISC-h^{10} on 150 pre-selected models. Below are summarised the top 10 success rates of ISC and ISC-h^{10}. Top 10 success rates of ISC/150h and ISC-h^{10}/150h were calculated after a pre-selection of a maximum of 150 models taken from the 3 x top 50 models of IES-h^{40}/10k, SPP-h^{40}/10k, and FRODOCK2.1. Scoring was performed on all 752 benchmark cases.

Score	Top 10 success rate	Top 50 success rate
ISC/150h	217 (28.9%)	398 (52.9%)
ISC-h^{10}/150h	**267 (35.5%)**	**404 (53.7%)**

3.5 Homology-enriched consensus scoring

As a first step, we calculate Cons^3-h, the homology-enriched variant of the Cons^3 consensus. Calculating a three-way consensus using higher-performing homology-enriched variants (Cons^3-h) instead of their original counterparts (Cons^3) increases the top 10 success rate from 32% to 36% (Table 6 and Figure 2A). Consensus Cons^3-h performs as well as ISC-

h[10]/150h, while calculated on the same top 150 models, and computation is about 20 times faster for Cons[3]-h than for ISC-h[10]/150h.

Out of the 271 and 267 successful cases for Cons[3]-h and ISC-h[10]/150h, only 198 cases are in common. Moreover, ISC and ISC-h[10] remain complementary to SPP-h[40]/10k, IES-h[40]/10k, and FRODOCK2.1 (supplementary Figure S1D and S1E). This led us to test four- and five-way consensus approaches to combine ISC optimally with other homology-enriched scores. We tested two four-way consensuses that integrate ISC without homology on 1,000 or 150 models (Cons[4]-h/1k and Cons[4]-h/150h respectively) and two five-way consensuses that integrate ISC both with and without homology on 1,000 or 150 models (Cons[5]-h/1k and Cons[5]-h/150h respectively). Performances are reported in Figure 2B and Table 6, together with time estimates when parallelising the whole pipeline on 4 CPUs.

With five-way consensus Cons[5]-h/1k, the top 10 success rate rises to 303 cases (40.3%). Unfortunately, computation time strongly increases, since we have to compute ISC-h[10] on 1,000 models. The most time-effective consensus, Cons[3]-h, has 36.0% top 10 success rate and the same top 1 success rate as Cons[5]-h/1k (Figure 2B and supplementary Figure S2).

Table 6: Performance of homology-enriched consensus scores. Performance of three-, four- and five-way consensus scores in terms of top 10 success rates on 752 benchmark cases and approximate timescales for the whole pipeline (including sampling with FRODOCK2.1, homology model generation, scoring steps, and consensus calculation). Scores used in Cons[3] were SOAP-PP/10k, InterEvScore/10k, and FRODOCK2.1. Scores used in all homology-based consensuses (Cons[X]-h) were FRODOCK2.1, SPP-h[40]/10k, IES-h[40]/10k, ISC and ISC-h[10]. The three-way consensus included the first three scores, four-way consensuses included all scores up to ISC and five-way consensuses included all of them. Cons[X]-h/150h included ISC scores over 150 models only and Cons[X]-h/1k included ISC scores over 1k models.

Score	Top 10 success rate	Whole pipeline time estimates on 4 Intel® Xeon® E5 CPU*
Cons[3]	241 (32.0%)	**15 min**
Cons[3]-h	271 (36.0%)	**15 min**
Cons[4]-h/150h	273 (36.3%)	45 min
Cons[4]-h/1k	282 (37.5%)	3 h 15
Cons[5]-h/150h	289 (38.4%)	5 h 30
Cons[5]-h/1k	**303 (40.3%)**	34 h 30

* all steps are parallelisable using MPI (sampling) or over the models (scoring)

4 Discussion

In InterEvScore (Andreani, et al., 2013), evolutionary information improved protein-protein scoring performance when given implicitly through coMSAs and coupled with a residue-level statistical potential. Combining InterEvScore with complementary scoring functions FRODOCK2.1 and SOAP-PP by computing a consensus (Quignot, et al., 2018; Yu, et al., 2016) improved over the individual scores, reaching 32% top 10 success rate (Table 1). However, this strategy did not take full advantage of the three scores' complementarity. We thus decided to combine directly evolutionary information from coMSAs with atomic scores such as SOAP-PP. To this aim, we scored threaded homologous interface models together with each query interface model.

With this explicit implementation of evolutionary information, a variant of InterEvScore where we scored models and their homologs with a residue-level statistical potential (IES-h) had a slightly improved success rate compared to the implicit homology version (Table 2). The explicit representation of homologous models enabled us to build homology-enriched versions of atomic scores SOAP-PP (SPP-h) and Rosetta ISC (ISC-h). For both, adding homology drastically improved top 10 success rates (Table 3 and Table 4) even when coMSAs were down-sampled to at most 10 homologous sequences. The Rosetta homology-enriched version, ISC-h[10], had outstanding performances, but it also was the most time-consuming score, about 750 times slower than SOAP-PP or InterEvScore. The first compromise between computation time and performance was to run ISC-h[10] on a pre-selection of 150 models defined by the top 50 models of SPP-h[40]/10k, IES-h[40]/10k, and FRODOCK2.1 (Table 5). This score (ISC-h[10]/150h) had a similar top 10 success rate (36%) to a much faster consensus score involving the same top 150 models. Taking further advantage of their complementarity, different four- and five-way consensuses managed top 10 success rates from 36.3% to 40.3% at runtimes ranging from 45 minutes to 34.5 hours on four CPUs (Table 6).

We further tried to understand the origin of the large performance improvements obtained through homology enrichment. We found that the performance improvement of the homology-enriched scores is driven positively by better recognition of correct models (up-weighted by conserved homologous interfaces), rather than negatively by the down-ranking of incorrect models due to clashing or incomplete homologous interfaces (since insertions in reference to the query structures were not modelled). Indeed, the number of gaps or the number of clashes (heteroatom contacts under 1.5 Å) at the interface of homologous interface models do not strongly correlate with the ISC-h[10] score. Additionally, ranking using only the repulsive van der Waals component of the Rosetta score (fa_rep) performs extremely poorly in comparison to other scoring schemes (supplementary Table S6). Finally, IES-h, SPP-h, or ISC-h score variants where only the worst-scored homologous interface models are used showed systematically worse performance than using the full range of homologous models (supplementary Table S6).

Improvement of SOAP-PP and Rosetta ISC by homology enrichment is significant (supplementary Figure S3), robust to a change in evaluation metrics (supplementary Table S7), and consistent over difficulty categories (supplementary Table S8). The strongest relative gain for homology-enriched scores occurs on "very_easy" and "easy" cases, which correspond to small to moderate conformational changes between unbound and bound structure (supplementary Tables S8 and S12). Consensus scoring also consistently improves results over the "very_easy", "easy" and "hard" categories, in order of decreasing improvement. We hypothesise that correct ranking of very_easy and easy models mainly depends on the ability to score positively native-like models. More difficult cases typically are out of the scope of rigid-body sampling and would require integration of flexibility, an ongoing challenge of protein docking (Desta, et al., 2020; Torchala, et al., 2013). Even in those difficult cases, when a near-native model can be generated in the sampling step, we see a positive effect of our improved methodology using atomic-level homology information and consensus approaches.

We demonstrate the robustness of our approach when sampling with ZDOCK 3.0.2 (Pierce, et al., 2011) instead of FRODOCK: when evolutionary information is included explicitly, we also observed a clear improvement in the success rates (supplementary Figure S6, Tables S13-S14). Additionally, homology enrichment also improves performance of ZRANK (Pierce and Weng, 2007) for rescoring ZDOCK models; the increase in top 10 success rate is comparable for ZRANK and SOAP-PP (supplementary Figure S6). Interestingly, combining top 5 Cons[3]-h models from the pipeline applied to ZDOCK and FRODOCK models is more successful than using the top 10 Cons[3]-h models from any of the two sampling programmes (supplementary Table S15).

Finally, we demonstrate the robustness of our strategy by applying it to the widely-used protein docking benchmark (supplementary Figure S8, Tables S16-S17). This also enables comparison with iScore (Geng, et al., 2019): the top 10 Cons^3-h models combined from ZDOCK and FRODOCK compare favourably to iScore over the subset of 21 cases from benchmark version 5 that was presented in the iScore publication (supplementary Table S18).

In this work, we developed a strategy to enrich scoring functions with evolutionary information by including atomic-level models for as few as ten homologs. This strategy improves the performance of several scores with different properties: InterEvScore, SOAP-PP, Rosetta ISC, and ZRANK. We provide a Singularity container (Kurtzer, et al., 2017) as a powerful means to re-run our docking pipeline. The container packages our docking tools and internally supports parallelisation. Thanks to the container, our homology enrichment strategy could be extended to other scores, as the container also allows users to generate all models (including homolog models) for rescoring with a different scoring function.

The homology enrichment strategy that we propose can in principle be applied to any scoring function with at most a ten-fold increase in runtime. This enrichment works with a very small number of sequences compared e.g. to the large MSAs needed by covariation methods to pick up coevolutionary signal, highlighting complementarity between the two approaches, which may be exploited by using additional DCA-derived constraints, e.g. in intermediate cases with a few hundred homologous sequences (Cong, et al., 2019; Simkovic, et al., 2017).

The docking success boost also opens interesting perspectives regarding the large-scale application of structural prediction to interaction networks, although sampling remains difficult for cases with large conformational changes upon binding. Evolutionary information can also be used to predict structures of interfaces between a globular protein and a peptide or a disordered region (Andreani, et al., 2020). Extension of the atomic-level homology enrichment strategy to these interaction types would require careful analysis, as evolutionary signals are more difficult to extract for low complexity regions. Finally, with the rise of machine learning techniques in computational biology, one can expect interesting future developments using these approaches to further enhance the extraction of (co)evolutionary signal from coMSAs.

Acknowledgements

Benchmarking was done partly through granted access to the HPC resources of CCRT under the allocations 2018-7078 and 2019-7078 by GENCI (Grand Equipement National de Calcul Intensif). We thank Arnaud Martel for his help with setting up the data web page and the Singularity container.

Funding

This work was supported by Agence Nationale de la Recherche [ANR-15-CE11-0008 to R.G., ANR-18-CE45-0005 to J.A]; IDEX Paris-Saclay [IDI 2017 to C.Q.]; MINECO [BFU2016-76220-P to P.C.]; and AEI/FEDER, UE [PID2019-109041GB-C21 to P.C.].

Conflict of Interest: none declared.

References

Andreani, J., Faure, G. and Guerois, R. InterEvScore: a novel coarse-grained interface scoring function using a multi-body statistical potential coupled to evolution. *Bioinformatics* 2013;29(14):1742-1749.

Andreani, J., Quignot, C. and Guerois, R. Structural prediction of protein interactions and docking using conservation and coevolution. *Wires Comput Mol Sci* 2020.

Chaudhury, S., *et al.* Benchmarking and analysis of protein docking performance in Rosetta v3.2. *PLoS One* 2011;6(8):e22477.

Cocco, S., *et al.* Inverse statistical physics of protein sequences: a key issues review. *Rep Prog Phys* 2018;81(3):032601.

Cong, Q., *et al.* Protein interaction networks revealed by proteome coevolution. *Science* 2019;365(6449):185-189.

Desta, I.T., *et al.* Performance and Its Limits in Rigid Body Protein-Protein Docking. *Structure* 2020;28(9):1071-1081 e1073.

Dong, G.Q., *et al.* Optimized atomic statistical potentials: assessment of protein interfaces and loops. *Bioinformatics* 2013;29(24):3158-3166.

Geng, C., *et al.* iScore: A novel graph kernel-based function for scoring protein-protein docking models. *Bioinformatics* 2019.

Gray, J.J., *et al.* Protein–Protein Docking with Simultaneous Optimization of Rigid-body Displacement and Side-chain Conformations. *Journal of Molecular Biology* 2003;331(1):281-299.

Huang, S.Y. Search strategies and evaluation in protein-protein docking: principles, advances and challenges. *Drug Discov Today* 2014;19(8):1081-1096.

Huang, S.Y. Exploring the potential of global protein-protein docking: an overview and critical assessment of current programs for automatic ab initio docking. *Drug Discov Today* 2015;20(8):969-977.

Koukos, P.I. and Bonvin, A. Integrative modelling of biomolecular complexes. *J Mol Biol* 2019.

Kurtzer, G.M., Sochat, V. and Bauer, M.W. Singularity: Scientific containers for mobility of compute. *PLoS One* 2017;12(5):e0177459.

Mendez, R., *et al.* Assessment of blind predictions of protein-protein interactions: current status of docking methods. *Proteins* 2003;52(1):51-67.

Mintseris, J. and Weng, Z. Structure, function, and evolution of transient and obligate protein-protein interactions. *Proc Natl Acad Sci U S A* 2005;102(31):10930-10935.

Moal, I.H., *et al.* The scoring of poses in protein-protein docking: current capabilities and future directions. *BMC Bioinformatics* 2013;14:286.

Morcos, F., *et al.* Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proc Natl Acad Sci U S A* 2011;108(49):E1293-1301.

Pierce, B. and Weng, Z. ZRANK: reranking protein docking predictions with an optimized energy function. *Proteins* 2007;67(4):1078-1086.

Pierce, B.G., Hourai, Y. and Weng, Z. Accelerating protein docking in ZDOCK using an advanced 3D convolution library. *PLoS One* 2011;6(9):e24657.

Porter, K.A., *et al.* What method to use for protein-protein docking? *Curr Opin Struct Biol* 2019;55:1-7.

Quignot, C., *et al.* InterEvDock2: an expanded server for protein docking using evolutionary and biological information from homology models and multimeric inputs. *Nucleic Acids Res* 2018;46(W1):W408-W416.

Ramírez-Aportela, E., López-Blanco, J.R. and Chacón, P. FRODOCK 2.0: Fast Protein-Protein docking server. *Bioinformatics* 2016:btw141.

Remmert, M., *et al.* HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nat Methods* 2011;9(2):173-175.

Simkovic, F., *et al.* Applications of contact predictions to structural biology. *IUCrJ* 2017;4(Pt 3):291-300.

Socolich, M., *et al.* Evolutionary information for specifying a protein fold. *Nature* 2005;437(7058):512-518.

Song, Y., *et al.* High-resolution comparative modeling with RosettaCM. *Structure* 2013;21(10):1735-1742.

Teichmann, S.A. The constraints protein-protein interactions place on sequence divergence. *J Mol Biol* 2002;324(3):399-407.

Torchala, M., *et al.* SwarmDock: a server for flexible protein-protein docking. *Bioinformatics* 2013;29(6):807-809.

Vreven, T., *et al.* Updates to the Integrated Protein-Protein Interaction Benchmarks: Docking Benchmark Version 5 and Affinity Benchmark Version 2. *J Mol Biol* 2015;427(19):3031-3041.

Yu, J. and Guerois, R. PPI4DOCK: large scale assessment of the use of homology models in free docking over more than 1000 realistic targets. *Bioinformatics* 2016;32(24):3760-3767.

Yu, J., *et al.* InterEvDock: a docking server to predict the structure of protein-protein interactions using evolutionary information. *Nucleic Acids Res* 2016;44(W1):W542-549.