(This is a sample cover image for this issue. The actual cover is not yet available at this time.)

Computational and Structural Biotechnology Journal 18 (2020) 114–124

COMPUTATIONAL AND STRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

A protocol to automatically calculate homo-oligomeric protein structures through the integration of evolutionary constraints and NMR ambiguous contacts

Davide Sala [a], Linda Cerofolini [b], Marco Fragai [a,c], Andrea Giachetti [b], Claudio Luchinat [a,c], Antonio Rosato [a,c,*]

[a] Magnetic Resonance Center (CERM), University of Florence, Via Luigi Sacconi 6, 50019 Sesto Fiorentino, Italy
[b] Consorzio Interuniversitario di Risonanze Magnetiche di Metallo Proteine, Via Luigi Sacconi 6, 50019 Sesto Fiorentino, Italy
[c] Department of Chemistry, University of Florence, Via della Lastruccia 3, 50019 Sesto Fiorentino, Italy

ARTICLE INFO

Article history:
Received 22 August 2019
Received in revised form 20 November 2019
Accepted 6 December 2019
Available online 26 December 2019

Keywords:
Homo-oligomers
Coevolution
NMR
Evolutionary constraints
Protein-protein interactions
Solid-state NMR

ABSTRACT

Protein assemblies are involved in many important biological processes. Solid-state NMR (SSNMR) spectroscopy is a technique suitable for the structural characterization of samples with high molecular weight and thus can be applied to such assemblies. A significant bottleneck in terms of both effort and time required is the manual identification of unambiguous intermolecular contacts. This is particularly challenging for homo-oligomeric complexes, where simple uniform labeling may not be effective. We tackled this challenge by exploiting coevolution analysis to extract information on homo-oligomeric interfaces from NMR-derived ambiguous contacts. After removing the evolutionary couplings (ECs) that are already satisfied by the 3D structure of the monomer, the predicted ECs are matched with the automatically generated list of experimental contacts. This approach provides a selection of potential interface residues that is used directly in monomer–monomer docking calculations. We validated the protocol on tetrameric L-asparaginase II and dimeric Sod1.

1. Introduction

Many proteins carry out their functional role acting as part of protein assemblies, i.e. a combination of different proteins (hetero-complexes) or of multiple copies of the same monomeric unit (homo-complexes). The assembly of the correct biological complex strongly depends upon specific protein–protein interactions (PPIs) that often are conserved among species [42,53]. Frequently, an initial step in the study of an assembly is to characterize the three-dimensional structure of its individual subunit components either by X-ray crystallography or NMR spectroscopy. Among NMR techniques, solid-state NMR (SSNMR) has been receiving increasing attention because it is not limited by protein size, solubility, crystallization problems, presence of inorganic/organic matrices or lack of long-range order that often make the application of other structural biology methods unsuitable. In particular, it is straightforward to extend SSNMR experiments designed for individual proteins to the investigation of protein assemblies [1,13,20,29,33,37], as the quality of SSNMR spectra does not decrease with increasing molecular weight, at variance with solution NMR.

A crucial step in the application of SSNMR to structure determination is the identification and assignment of through-space nucleus-nucleus interactions. DARR (Dipolar Assisted Rotational Resonance) is a commonly used pulse sequence for this purpose, which is based on ^{13}C–^{13}C magnetization transfer through proton-driven spin diffusion [56]. Tuning of experimental DARR parameters allows users to select the range of distances at which inter-nuclear interactions are sampled. Although solid-state resonance lines of protein complexes are narrow, spectral congestion makes the assignment of DARR peaks a challenging task. In practice, DARR experiments yield a list of ambiguous contacts in which the quaternary contacts must be separated from intra-monomeric contacts to determine the 3D structure of the complex. In hetero-complexes this problem can be alleviated by using different schemes for enrichment in stable NMR-active isotopes (^{13}C, ^{15}N) in the various subunits of the complex [23]; for instance, one subunit can be uniformly enriched while all other subunits are not. This approach may not be very effective for homo-complexes, and more complex and labor intensive strategies for the asymmetric enrichment of all subunits have been proposed [59]. Thus, the

* Corresponding author at: Via Luigi Sacconi 6, 50019 Sesto Fiorentino, Italy.
E-mail address: rosato@cerm.unifi.it (A. Rosato).

https://doi.org/10.1016/j.csbj.2019.12.002
2001-0370/

investigation of homo-complexes by SSNMR often remains a manual task, especially with respect to the identification of inter-subunit contacts.

Coevolution analysis assumes that evolutive pressure favors the preservation of protein function through the conservation of fundamental residue interactions [49]. This concept has been implemented, among others, in global coevolutionary or direct coupling analysis (DCA) methods [38,64]. These methods differ for the types of approximation used, from dimensional reduction [11] to pseudo-likelihood maximization [16] and others [8,30,52]. The information derived allows the identification of multiple protein conformational states [39,54] and the prediction of tertiary protein structures, either alone or in combination with experimental data [2,12,35,36,57]. Coevolution analysis can detect also ECs corresponding to inter-subunit contacts [26,27,40,46,51,55]. The identification of ECs consistent with PPIs for hetero-complexes requires the creation of a *joint* multiple sequence alignment (MSA) in which each line corresponds to an interacting protein pair [6,7,10,41]. This is a relatively complex task, especially due to the analysis required for the separation of orthologs and paralogs, prior to the construction of the MSA. Instead, the coevolution analysis of homo-complexes is based on a single protein sequence and thus on a single MSA. While this simplifies the construction of the alignment, it makes the identification of ECs belonging to intermolecular contacts much more complicated because such information is hidden among hundreds or thousands of ECs of which the majority are tertiary contacts [50,58,60]. The removal of tertiary contacts requires knowledge of the 3D structure of the monomeric protein. Notably, there is a relevant number (about 2000) of protein families annotated as forming homo-oligomeric assemblies *in vivo* with a deposited monomeric structure in the Protein Data Bank (PDB) [17,48]. These families potentially constitute an interesting target for homo-oligomeric structural predictions, also in the frame of drug discovery [3].

In the present work we developed a protocol to extract information on the protein–protein interface of homo-complexes by integrating SSNMR-derived ambiguous contact lists, which can be automatically generated, with coevolution analysis. All the ECs with a relevant probability to be true residue interactions in either the monomer (intra-monomeric contacts) or in the homo-oligomerization interface (inter-monomeric contacts) are considered. The removal of intra-monomeric ECs requires the availability of the structure of the monomer. The predicted ECs with possible matches to experimental peaks are used to identify candidate interface residues. The final list of such residues is used directly in protein–protein docking calculations. The same protocol can be also applied using only solution-state NMR data.

2. Results

Our protocol aims to predict the structure of homo-oligomeric complexes by using ambiguous NMR contacts to identify reliable inter-monomeric contacts within the list of ECs. NMR restraints can in principle be assigned to a specific pair of atoms. However, this part of the procedure for NMR-based structure determination of proteins is quite labor-intensive, so that current approaches mainly focus on the use of automatically generated lists of ambiguous contacts [47], as done here. The whole procedure, which is described in detail in the next section, can be divided in two main parts. First, intra-monomeric evolutionary couplings (ECs) are removed from the list of ECs based on the 3D structure of the monomer. Second, the list of ECs predicted to potentially be at the complex interface is compared with the list of ambiguous NMR contacts to extract all residue pairs matching both the predicted and the experimental dataset. The protocol was validated by predicting the tetrameric structure of *Escherichia coli* L-asparaginase II [9] (PDB ID: 6EOK), in which two distinct dimeric conformations must be recognized to reconstruct the functional complex (Fig. 1). Furthermore, the robustness of the procedure in the identification of complexes with small interface regions was tested by predicting the structure of dimeric human apo Sod1 [5] (PDB ID: 3ECU) (Fig. 1). For L-asparaginase II we used solid-state NMR data [9], whereas for Sod1 we used solution NMR data [5].

2.1. Description and application of the protocol

Our protocol calculates a list of putative interface residues to be used as input to HADDOCK for docking calculations. It needs four inputs (Fig. 2): one or more files with the list of ECs, the structure of the monomer, the experimental NMR-derived list of ambiguous contacts and the Naccess file (rsa format) with the per-residue relative solvent accessible area. The ECs of the target protein are obtained from so-called coevolution analysis. A number of servers performing coevolution analysis are available online. In this work, we employed three widely-used webservers to calculate ECs: Gremlin by the Baker group that uses an unsupervised approach [4], RaptorX [62] by the Xu group and ResTriplet by the Zhang group [32], both supervised. In general, they need the protein sequence as input to predict a contact map from multiple sequence alignments (MSAs). The output is a list of residue pairs scored for the probability that they are actually in contact in the monomeric or oligomeric structure. We apply a probability cutoff P to remove ECs with low probability of being true interactions. Coevolution analysis usually outputs from hundreds to thousands of ECs that cannot be assigned as intra-monomeric or inter-monomeric contacts without any structural information. As a consequence, our protocol calculates for each EC the corresponding Cα-Cα distance in the 3D structure of the monomer and all the ECs below the distance cutoff of 12 Å are classified as potentially intra-monomeric and removed.

After the removal of potentially intra-monomeric ECs, the resulting list is enriched in contacts across the interaction interface (inter-monomeric ECs). Nevertheless, it still contains a relevant number of false-positives. False-positives can be either ECs that do not correspond to a true residue-residue interaction or ECs that correspond to intra-monomeric interactions occurring in conformations sampled during the physiological conformational dynamics of the protein. The EC list thus cannot be used directly in docking calculations. We thought that the rate of false positives could be reduced by leveraging the information present in the list(s) of ambiguous contacts provided by NMR experiments. Indeed, NMR-derived contacts lists of protein complexes are affected by a high level of ambiguity caused by the accidental overlap of NMR resonances, making the extraction of reliable inter-monomeric contacts an arduous task. Our protocol addresses this bottleneck by matching the predicted inter-monomeric ECs with the experimental list to extract information present in both the datasets. In practice, residue pairs in the predicted inter-monomeric EC list are matched to ambiguous assignments in the experimental list [57], providing a list of interface residue pairs.

The number of residual false-positives in the matched list is further decreased by removing all the residues with a relative solvent accessibility lower than 40% in both main-chain and side-chain (i.e. buried residues). The remaining residues in the output list from our protocol can be used directly as ambiguous interaction restraints (AIRs) in monomer–monomer docking calculations with HADDOCK. The protocol can be run using the python script provided as supplementary material (SI Appendix).

We assessed the accuracy of the protocol in predicting residues at the homo-oligomeric interface for different probability cutoffs (Tables 1 and 2). Furthermore, we evaluated the NMR data

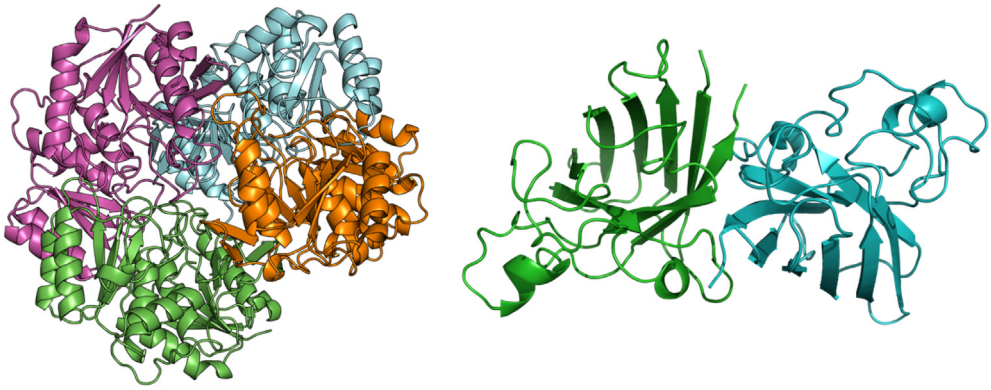

Fig. 1. Crystal structures of the tetrameric L-asparaginase II and the dimeric apo Sod1.

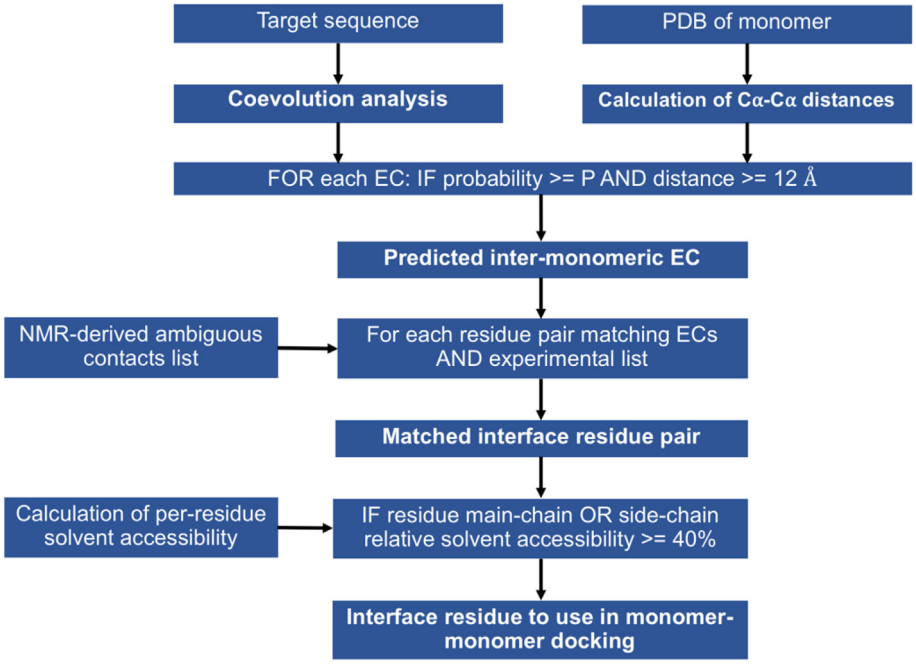

Fig. 2. Scheme of the protocol adopted to predict the structure of homo-oligomeric complexes using coevolution analysis and ambiguous NMR contacts.

Table 1
Number of residues predicted to make contacts across the L-asparaginase II homomeric interface. The protocol was applied as depicted in Fig. 2 with the ECs matched with the NMR data "ECs + NMR" and without the matching step with NMR data "ECs only". The number of TP contacts belonging to the AC or AD interface is in brackets (AC-AD). P indicates the probability threshold used to accept ECs. PPV = TP/(TP + FP).

P	L-asparaginase II					
	ECs only			ECs + NMR		
	TP + FP	TP (AC-AD)	PPV	TP + FP	TP (AC-AD)	PPV
0.90	13	10 (9-1)	0.8	3	3 (3-0)	1.0
0.85	23	20 (19-1)	0.9	3	3 (3-0)	1.0
0.80	30	21 (20-1)	0.8	3	3 (3-0)	1.0
0.75	34	24 (22-2)	0.8	3	3 (3-0)	1.0
0.70	38	27 (24-3)	0.8	4	4 (4-0)	1.0
0.65	41	30 (24-6)	0.8	4	4 (4-0)	1.0
0.60	47	31 (25-6)	0.7	4	4 (4-0)	1.0
0.55	51	33 (26-7)	0.7	4	4 (4-0)	1.0
0.50	60	36 (26-10)	0.7	4	4 (4-0)	1.0
0.45	73	42 (29-13)	0.7	4	4 (4-0)	1.0
0.40	84	47 (32-15)	0.6	5	5 (4-0)	1.0
0.35	97	52 (34-18)	0.6	7	7 (6-1)	1.0
0.30	105	54 (34-20)	0.6	9	8 (6-2)	0.9
0.25	121	60 (37-23)	0.6	19	16 (10-6)	0.8
0.20	128	60 (37-23)	0.6	34	28 (18-10)	0.8

Table 2
Number of residues predicted to make contacts across the Sod1 homomeric interface.

P	Sod1					
	ECs only			ECs + NMR		
	TP + FP	TP	PPV	TP + FP	TP	PPV
0.90	0	0	NA	0	0	NA
0.85	0	0	NA	0	0	NA
0.80	0	0	NA	0	0	NA
0.75	4	3	0.7	0	0	NA
0.70	4	3	0.7	0	0	NA
0.65	4	3	0.7	0	0	NA
0.60	8	4	0.6	2	0	NA
0.55	10	4	0.4	2	0	0.0
0.50	17	5	0.2	4	0	0.0
0.45	23	7	0.3	5	1	0.2
0.40	29	9	0.3	5	1	0.2
0.35	38	12	0.3	9	3	0.3
0.30	50	14	0.3	18	7	0.4
0.25	68	17	0.2	27	7	0.3
0.20	74	17	0.2	48	9	0.2

contribution to the prediction accuracy by comparing the results obtained with or without (“ECs + NMR” and “ECs only”, respectively) matching with the NMR data. A residue correctly predicted at the complex interface is defined as a true-positive (TP) residue. More in detail, we defined a true-positive (TP) residue as having at least one atom with a distance < 7 Å from any atom located on a different chain in the crystal structure of the complex.

In the case of the L-asparaginase II protein, the crystallographic complex is formed by four subunits with a D_2 symmetry. Thus, the ensemble of all TP residues contains the amino acids at both dimeric interfaces, where AC is the largest interface formed by chain A and chain C while AD is the smaller interface formed by chain A and chain D. For this system, the inclusion of NMR data enhances the positive predictive value (PPV, also known as Recall), defined as true-positive (TP) residue predictions over all predictions [TP/(TP + FP)], at all the probability cutoffs assessed (Table 1). In fact, on the basis of the “ECs only” analysis the absolute number of TP residues present in the prediction is significantly higher than the number of TP obtained after the match with NMR data. However, the same “ECs only” analysis also outputs a much greater number of FPs. Moreover, the residues belonging to the smaller interface A-D consistently appear only at P < 0.70, precluding the possibility to solve the whole tetrameric complex at higher P. Consequently, the “ECs + NMR” analysis features a PPV of 100% for $P \geq 0.35$; the PPV remains very high (≥80%) even at low probabilities (P < 0.35) and the number of predicted interface residues belonging to both the interfaces is sufficient to successfully drive docking calculations (see next section).

A complementary case was provided by the Sod1 complex, which contains two subunits with a C_2 symmetry and a small protein–protein interface. As a consequence, in the central part of the interface the inter-monomeric contacts involve also residue pairs whose intra-monomer distance is lower than the 12 Å threshold that we used to remove intra-monomeric ECs. In practice, this structural organization significantly reduces the number of detectable TPs because the aforementioned inter-monomeric contacts are discarded. Furthermore, small interfaces are harder to predict computationally and also provide a lower number of NMR-detectable contacts. All these features make the Sod1 system challenging but useful to test the limits of the protocol. When considering the Sod1 protein, the “ECs only” protocol yielded a reasonable PPV for $P \geq 0.55$, but with only a handful of TPs in the prediction (Table 2). Instead, the match with NMR data removed the signal for $P \geq 0.45$ while retaining information at lower P values, especially for P = 0.30.

These results suggest that the quality of the initial EC prediction is quite important for the performance of our protocol, leading to a larger enhancement of the PPV when the prediction includes a larger number of TPs. When the EC data yielded is weaker and mixed with noise, our protocol retains a good part of the available information but the PPV is mostly unchanged.

2.2. HADDOCK calculations for L-asparaginase II starting from the crystal structure

The ECs at the P cutoff of 0.25 were matched with a solid-state 2D ^{13}C-^{13}C DARR dataset (mixing time 200 ms) holding 4937 ambiguous assignments, resulting in 19 surface residues predicted to be at the protein–protein interface (corresponding to 14% of the whole protein surface). The final 200 water-refined models generated by HADDOCK were analyzed by measuring the RMSD from the structure with the lowest HADDOCK score. The clustering algorithm grouped the models in 7 clusters (Fig. 3A). The first cluster was the most populated and included the models with the lowest score. Indeed, the lowest HADDOCK score model of the first cluster was a dimer with an RMSD of 0.7 Å from the crystallographic dimer formed by chain A and chain C of the tetrameric protein (Fig. 3B). In addition to the HADDOCK score, the desolvation energy calculated using empirical atomic solvation parameters proved to be an useful scoring function [19], allowing the identification of the correct A-C dimer (Fig. S1).

Both the predicted inter-monomeric ECs and the experimental NMR inter-monomeric contacts include residue pairs belonging to all the pairs of chains effectively in contact in the functional complex. In the case of the tetrameric L-asparaginase II, besides the largest A-C interface also chains A and D share a relevant number of contacts. Thus, in a single docking run one might expect to sample both relevant dimeric configurations (A-C and A-D) in two different clusters. Indeed, by checking the position of the 19 predicted interface residues within the crystal structure, it appears that the A-C and A-D interfaces were both mapped (Fig. 4). In fact, the largest portion of residues effectively in contact belonged to dimer A-C and the smallest portion to dimer A-D.

However, the structural configuration present in the other clusters did not correspond to the A-D dimer. This could be easily verified by observing that the superimposition of the two dimers on the common chain A resulted in evident steric clashes between the subunits, as shown for the cluster 3 (Fig. 5). If the two dimers actually corresponded to the A-C and A-D dimers of the tetrameric

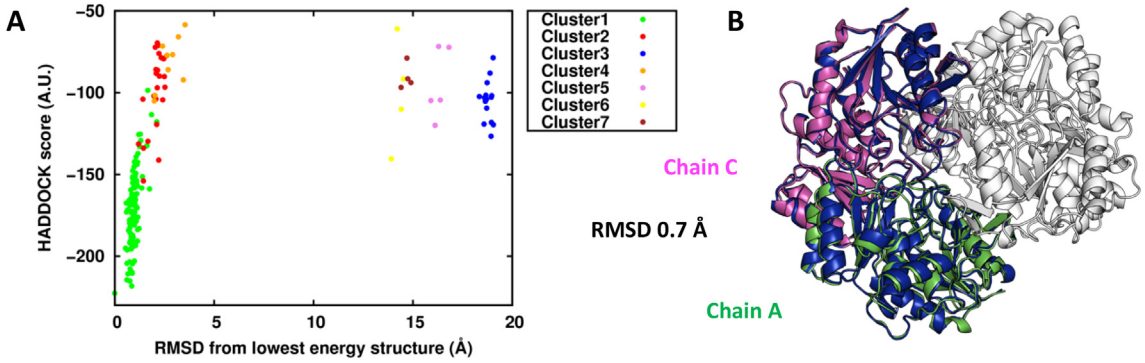

Fig. 3. L-asparaginase II monomer–monomer docking. A) Plot of the HADDOCK score vs RMSD clusters distribution with respect to the lowest HADDOCK score model. B) Structural alignment between the lowest HADDOCK score model (in blue) of the first cluster and the crystal structure. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

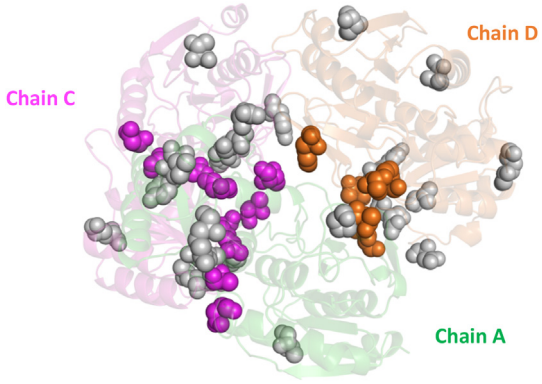

Fig. 4. Projection on the crystal structure of the L-asparaginase II residues used to generate AIRs in the docking calculation. The residues making inter-monomeric contacts are shown as colored spheres (A–C interface in purple; A–D interface in orange). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

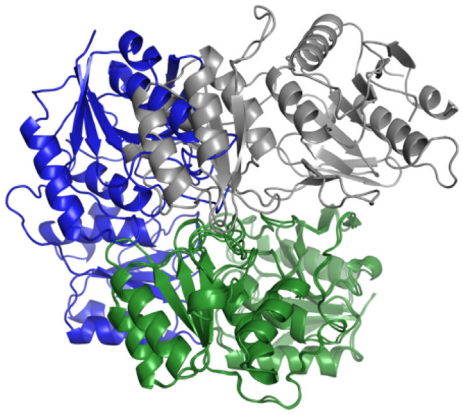

Fig. 5. Superimposition on chain A (in green) of the third (in gray) and the best (in blue) dimer configurations in the first run. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

structure, the superimposition on the A chain would have caused no significant clashes.

In principle, the absence of the second compatible dimer in calculations can be due to two reasons. First, the interface residues belonging to the second configuration were not present in the AIRs dataset. Second, the residues belonging to the second interface region were present, but the correct structural configuration had a HADDOCK score worse than the wrong sampled configurations. In the present case, the latter reason was the relevant one. In fact, the wrong dimer models in general contained some contacts from both interface regions, thus satisfying a higher number of AIRs than the correct dimer A-D.

To obtain a model of the A-D dimer, we performed a second docking run in which the restraints already satisfied in the best cluster (containing the most favored configuration) of the first run were removed from the input dataset. To this end, we looked at the violation analysis of HADDOCK, and retained all contacts that were not satisfied by the majority of the members of the first cluster by at least 3 Å. This resulted in 9 residues being used as input to a second monomer–monomer docking run. As in the previous calculation, the first cluster was the largest and contained the models with the best HADDOCK score and desolvation energy (Fig. 6A and S2). Superimposing the lowest HADDOCK score water-refined model with the crystal structure resulted in an RMSD of 0.9 Å from the dimer A-D (Fig. 6B). This whole procedure did not assume or require any previous knowledge of the A-D structure.

In summary, the two correct dimeric conformations A-C and A-D were obtained performing two distinct docking runs, the first one with the whole AIRs dataset and the second one with the subset resulting from the removal of the AIRs satisfied in the best cluster of the first run. Crucially, this procedure provided us with two compatible non-overlapping dimeric models that, for symmetry, can be used to reconstruct the tetrameric model (Fig. 7). This step strictly depended by the correct identification of the structural model on which the distance violation analysis was carried out. In fact, selecting the third cluster of Fig. 3 to perform the violation analysis instead of the best one resulted in a second docking run that sampled again the dimer A-C in the two best clusters and not-compatible structural configurations in the others (Fig. S3).

2.3. HADDOCK calculations for L-asparaginase II starting from homology models

Extracting the monomer from the PDB of the complex results in a protein model with the side chains oriented in a contact-ready state that favors the correct assembly, in terms of both docking score and RMSD from the experimental structure, with respect to incorrect docking poses. Thus, to test our protocol in a more realistic condition we generated 15 homology models of L-asparaginase II using the structure of the homolog from *Wolinella succinogenes* [34] as the structural template (PDB ID 1WSA, chain A). The homology models had a backbone RMSD lower than 1 Å from the crystal structure of the *E. coli* protein, but widely

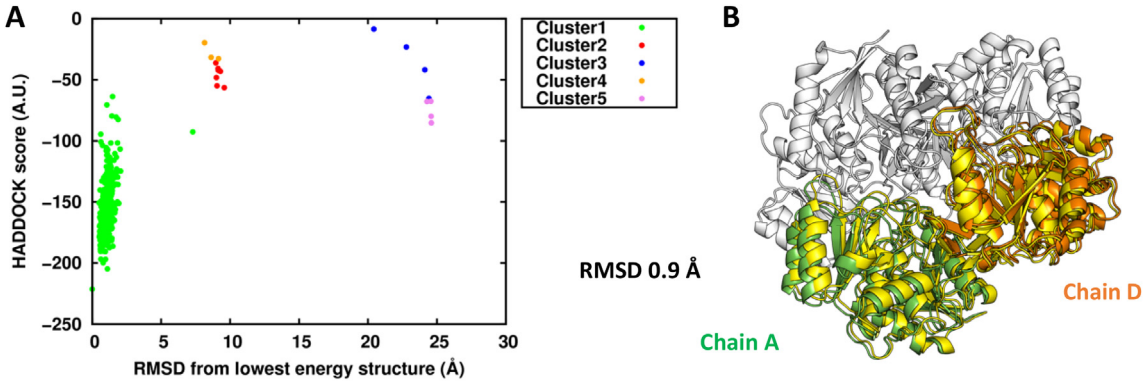

Fig. 6. L-asparaginase II monomer–monomer docking using AIRs violated in the A-C dimeric model. A) Plot of the HADDOCK score vs RMSD clusters distribution with respect to the lowest HADDOCK score model B) Structural alignment between the lowest HADDOCK score model (in yellow) of the first cluster and the crystal structure. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

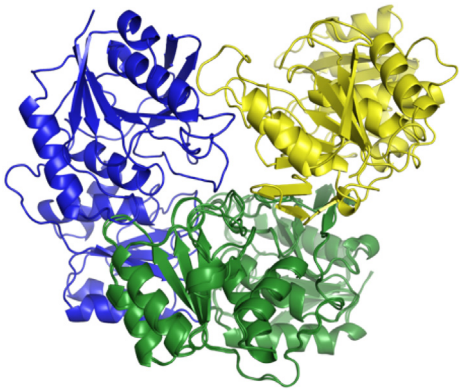

Fig. 7. Superimposition on the chain A (in green) of the best structural configurations in the second run (in yellow) and in the first run (in blue). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

differing in the orientation of the surface side chains. Each model was used in protein–protein docking with the same input AIRs of the "crystal P 0.25" runs, for both the A-C and A-D dimers. The results of Table 3 show the significant influence of the orientation of side chains on the ability of the docking calculations to sample the correct dimer in the best cluster. Based on the HADDOCK score of the best cluster for each model, the AC runs pointed out that the five runs with the best score also had the lowest RMSD from the crystal A-C dimer, (green gradient in the table). However, for these five models the second calculation with the AIRs providing the A-D dimer resulted in wrong dimeric conformations. Nevertheless, by inspecting the results for all models (Table 3), it turned out that the runs with the best HADDOCK scores (for their first clusters) indeed provided conformations close to the crystallographic A-D dimer (in particular models 6 and 15). For further comparison, we performed a docking run of the crystallographic monomer with the 34 residues (25% of the whole protein surface) output by the protocol run at a P cutoff of 0.20. Changing the AIRs dataset with a larger one having the same PPV did not significantly affect the results.

Overall, the results described above pointed out the importance of generating a sufficiently large number of homology models to sample many different side chain orientations, thus increasing the probability to capture the orientation permitting residue-residue contacts across the monomeric interface. The best clusters of the two crystal runs showed that ideal side chain orientations provided the top HADDOCK score values. In line with this, the models that had the best HADDOCK scores resulted in the configurations closest to the crystal structure, with a backbone RMSD between 1 and 3 Å from it. For these models, the HADDOCK scores themselves were similar to the values observed for the runs starting from the crystal monomer. Indeed, superimposing on the chain A the AC dimer of model 13 and the AD dimer of model 15 or model 6 showed two compatible dimeric models that, taken together, can be used to reconstruct the tetrameric structure (Fig. S4)

2.4. HADDOCK calculations for Sod1

The predicted inter-monomeric ECs at P = 0.30 were matched with 7611 ambiguous assignments from solution-state 3D ^{1}H ^{15}N NOESY-HSQC spectrum. The protocol yielded 18 putative interface residues, corresponding to 23% of the whole monomer surface. By comparing the prediction to the crystal structure, it appeared that 7 out of 18 residues effectively formed inter-monomeric contacts (Fig. S5).

From the docking calculation starting with the crystal monomer we obtained 7 clusters with comparable HADDOCK score values (Fig. 8A). However, the distribution of the desolvation energies discriminated the second cluster as the most favored (Fig. 8B). Indeed, the structural alignment of the best model of this cluster with the experimental dimer revealed an impressive RMSD of 0.6 Å (Fig. S6A). Instead, the same superimposition on the crystal structure of the first cluster resulted in a dimer in which one of the two monomeric units was rotated by 180° with respect to the corresponding experimental monomer, while preserving the same interface region (Fig. S6B).

2.5. HADDOCK calculations without NMR data

The protocol was applied also without the matching step with NMR data, i.e. only taking advantage of the 3D structure to remove intra-monomeric contacts and residues not at the protein surface from the lists of ECs. In this regard, we performed three docking calculations starting from the monomeric crystal structure of L-asparaginase II using the residues extracted at P values of 0.85, 0.65 and 0.50, respectively (Table 4). At P = 0.85 only 1 out of 20 TP residues belonged to the smaller A-D interface. Consequently, the dimer AC was successfully sampled in the best HADDOCK score cluster (RMSD 0.8 Å), whereas the dimer AD was not detected. Instead, the dimer AD was successfully sampled using the datasets at P 0.65 and P 0.50, where a significant number of residues belonging to the smaller interface is present. These results suggest that the two interfaces have a different degree of conservation within the family. In general, such a factor, in the absence of prior

Table 3
Docking results for homology models of L-asparaginase II. The two "Crystal" runs were performed using the chain A of the crystal structure. Each model mainly differs in the orientation of side chains. For each run the HADDOCK score of the best cluster (calculated as the average value of the 4 best structures of the cluster) and the RMSD of its best structure from the experimental dimer are reported.

	A-C dimer		A-D dimer	
	HADDOCK score	RMSD	HADDOCK score	RMSD
Crystal P 0.25	-218	0.7	-206	0.9
Crystal P 0.20	-170	0.7	-187	1.3
model1	-204	1.4	-121	16.1
model2	-93	17.9	-116	9
model3	-101	21.9	-104	14.5
model4	-72	16.6	-109	3.2
model5	-141	16.1	-91	4.3
model6	-95	14.4	-159	1
model7	-166	1.3	-113	14.5
model8	-72	16.6	-109	3.2
model9	-106	18.9	-120	8.3
model10	-184	2.2	-123	8.8
model11	-187	1.5	-132	11.6
model12	-117	18.9	-124	7.4
model13	-204	1.4	-121	16.1
model14	-101	21.9	-104	11
model15	-134	18.5	-169	2.6

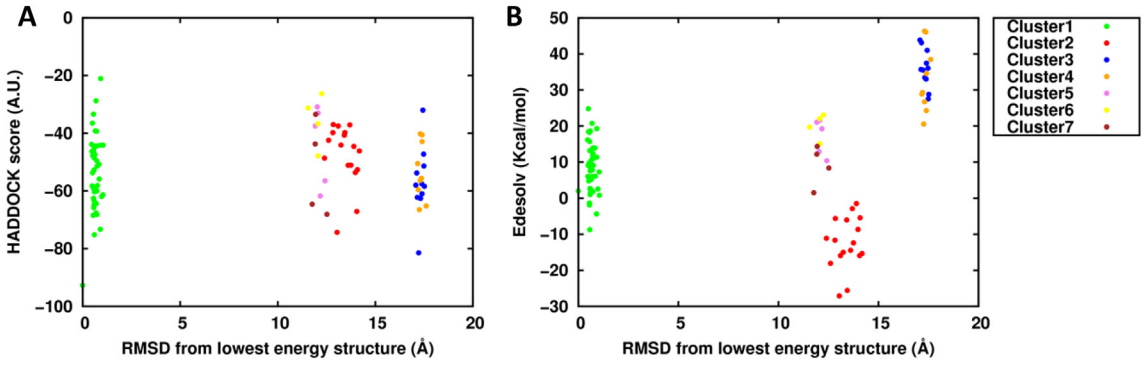

Fig. 8. Sod1 clusters distribution with respect to the lowest HADDOCK score model. A) HADDOCK score distribution. B) Desolvation energy distribution.

Table 4
Number of residues predicted without NMR data to make contacts across the L-asparaginase II interface. The number of TP contacts belonging to the AC or AD interface in the L-asparaginase II protein is in brackets (AC-AD). RMSD from AC and AD crystal dimers is reported in two separated columns.

P	L-asparaginase II – ECs only				
	TP + FP	TP	PPV	RMSD (AC)	RMSD (AD)
0.85	23	20 (19-1)	0.9	0.8 Å	>10 Å
0.65	41	30 (24-6)	0.8	0.7 Å	1.3 Å
0.50	60	36 (26-10)	0.7	0.6 Å	1.6 Å

Table 5
Number of residues predicted without NMR data to make contacts across the Sod1 dimeric interface.

P	Sod1 – ECs only			
	TP + FP	TP	PPV	RMSD
0.75	4	3	0.7	4.4 Å
0.60	8	4	0.6	1.3 Å
0.50	17	5	0.2	>10 Å

knowledge on the evolutionary history of the family, may complicate the choice of the cutoff value for P.

In the case of the Sod1 protein the P values chosen to collect the residues were 0.75, 0.60 and 0.50 (Table 5). At P = 0.75 only 4 residues were retrieved of which 3 were actually at the interface (TP). Despite the high PPV, the low number of input residues caused the output clusters to have very similar HADDOCK scores and desolvation energies. The best solution featured a 4.4 Å RMSD from the crystal (Fig. S8). At P = 0.60 we found a lower PPV but one additional TP residue was included in the input to HADDOCK, making it possible to identify a satisfactory solution based on desolvation energy (cluster 3, Fig. S9). Finally, at P = 0.50 our protocol identified a further additional TP residue, but the low PPV (0.2) prevented the sampling of the correct complex (Fig. S10).

2.6. Step-by-step procedure for a structurally uncharacterized oligomer

In this section we describe a step-by-step procedure for the application of our protocol with suggestions on how maximize the information that can be extracted from your data. Since this protocol was developed to use the NMR data in combination with ECs, we assume that the user already knows the stoichiometry of the complex (NMR experiments require purification of the protein and then the stoichiometry can be straightforwardly determined, e.g. by size exclusion chromatography), in addition to the structure of the monomer.

1. Perform coevolution analysis of your protein. To do this you can either download a suitable software or use online servers. In both cases, the protein sequence is the only required input to get ECs. The lists of ECs accepted by our python script (see next step) should be in RR format (the description of RR format is at http://predictioncenter.org/casprol/index.cgi?page=format#RR) or Gremlin format.
2. Compute the per-residue relative solvent accessible area with Naccess [28] (freely available for academic users at http://wolf.bms.umist.ac.uk/naccess/). Use the command "Naccess <pdbfile>", giving the PDB structure of the monomer as input; one of the outputs is the .rsa file required in the next step.
3. Run the ecnmr.py python script (downloadable at https://github.com/davidesala/ecnmr) using as input files the PDB structure of the monomer, the lists of ECs, the list of ambiguous NMR contacts in CYANA format [25] and the .rsa file from Naccess. The script has two cutoff values that can be adjusted: the probability P and the distance D (Fig. 2). Their default values are 0.2 and 12 Å, respectively. Besides the residues predicted to be at the interface, the script outputs the percentage of the total solvent accessible area represented by the list of residues. This information can be useful for selecting relevant P and D values. For a monomer with a rather spherical shape in a dimeric complex, a reasonable number of interface residues could be about 15–20% of all solvent-accessible residues. This fraction may be higher for higher-order multimers. Thus, a practical approach could be to keep D fixed at values such as 12 Å or 10 Å, which were already proven to be excellent general thresholds to reduce the number of false positive contacts [60] and explore your data as a function of P. The script is fast (computing time less than a minute even for big proteins), thus it is possible to scan P values very quickly. In general, a relevant strategy could be select two or three datasets to drive the docking calculations and check the convergence of your results toward the most sampled solution (see step 5), i.e. the same complex structure.
4. Since the accuracy of protein–protein docking is strictly dependent on the rotameric state of side chains at the interface (Table 3) we suggest building several models of your monomeric structure (10 or more) with the same backbone conformation but side chains randomly oriented. This step can be done using common software for protein modelling as Modeller [18] or using a webserver as Swiss-Model [63] (https://swissmodel.expasy.org) or even with a short MD simulation. Subsequently, docking calculations starting from each model can be performed on the HADDOCK webserver (http://haddock.science.uu.nl/services/HADDOCK/haddock.php), using the residues selected in step #3 as AIRs. The rationale for this approach is to extend the conformational sampling by HADDOCK by creating a larger variety of "encounter complexes" in the first part of the docking simulation. This, together with the conformational flexibility inherently sampled by the HADDOCK protocol [14], could, in principle, accommodate small-scale structural rearrangements in the monomer structure upon formation of the oligomer also when starting from a homology model [45,61].
5. The identification of the most probable complex is based on the comparative analysis of the best HADDOCK score cluster of each calculation as shown in Table 3. In practice, the user should check if the representative structures of similarly scored clusters (or the clusters featuring the best desolvation energy if the distribution of the HADDOCK scores over the various clusters is narrow) share the same configuration of the complex (as is the case for the green cells in Table 3). For multimers, the identification of the smaller/less conserved interface can be achieved by performing a second docking calculation for each starting model with the AIRs not satisfied by the interface identified in the first run (deviation > 3 Å in the ana_dist_viol_all.lis file located in the analysis folder of the HADDOCK output).

3. Discussion

Solid-state NMR is an attractive technique to study large protein assemblies as even systems with high molecular weight can provide very good spectra. However, the determination of their 3D structure involves two very time-consuming steps: the assignment of the side chains in contact at the interface between the subunits and, for homo-oligomeric complexes, the discrimination of intra- vs inter-monomer contacts. Both steps are labor-intensive and require extensive efforts by an experienced user. From the bioinformatics point of view, focusing on homo- rather than hetero-oligomers makes the interpretation of coevolution signals harder. In fact, the difficult step in the coevolution analysis of hetero-oligomers is the proper pairing of orthologs of interacting proteins and the corresponding removal of paralogs. Once this has been achieved, the creation of a *joint* MSA in which each line contains a pair of interacting proteins allows the straightforward use of predicted inter-protein contacts as restraints to drive the modelling of the quaternary structure [6,26,40]. Instead, the coevolution analysis of homo-oligomers is based on a single protein MSA, which is relatively effortless to build. Unfortunately, the availability of the three-dimensional structure of the monomeric unit is necessary to successfully separate intra-monomeric and inter-monomeric ECs [60]. In this work, we developed a protocol to integrate ECs with NMR–derived ambiguous contacts in order to identify interface residues in homo-oligomers. The input lists of ambiguous contacts can be automatically generated from solution or solid-state NMR spectra. Our protocol was validated by predicting two difficult cases: the tetrameric L-asparaginase II, in which two distinct dimeric conformations must be recognized to reconstruct the functional complex and the dimeric Sod1, in which the interface region is comparatively small.

The correct identification of interface residues was readily verified by comparing the output of the protocol with the known

interfaces in the crystal structures of the two systems (Tables 1 and 2). This analysis showed that NMR data can be beneficial by enriching the predictions in true contacts (i.e. higher PPV). This improvement comes at the cost of reducing the absolute number of predicted residues, which however did not limit the subsequent docking calculations. The requisite for the integration of ECs and NMR data to be successful is that the initial list of potential inter-monomeric ECs contains enough information. This is clearly exemplified by the case of Sod1, for which the absolute number of predictions, after removing all contacts that could be satisfied within the monomer, was quite low. Consequently, many NMR signals could not be matched and the benefit in PPV was modest. Nevertheless, when the total number of predicted interface residues is in a reasonable range (15%–20% of all surface residues, i.e. 12–16 residues for Sod1) the prediction resulting from the integration of ECs and NMR data is more reliable than that based only on ECs.

To generate a 3D structural model of the oligomer, the output of our protocol can be exploited in docking calculations. As a proof-of-principle, we run these calculations starting from the monomer conformation observed in the crystal structure. This is an ideal case, where all the side chains at the protein–protein interface are already in the correct rotameric state to engage in the formation of the complex. Nevertheless, it was important to perform this step to ensure that the output of our protocol contained enough information to successfully drive the docking. This was indeed the case for the main dimer of L-asparaginase II (A-C) as well as for Sod1. The calculation with the complete AIR dataset could not identify the A-D dimer even though the dataset contained contacts belonging to both interfaces. The A-D interface is somewhat smaller than the A-C interface; as HADDOCK aims to satisfy the highest number of AIRs, the situation where the second chain of the dimer is positioned in between the two interfaces, thus partly satisfying both subsets of AIRs, was favored over the situation in which all of the A-D and none of the A-C AIRs are satisfied. To circumvent this bottleneck, it is necessary to separate the residues belonging to each interface. This was done by removing the contacts already satisfied in the first docking calculation to run a second calculation only with the unsatisfied restraints. The best cluster of the second run indeed matched closely the A-D dimer of the tetramer (Fig. 6). Intriguingly, the AIRs derived only from ECs at a P cutoff of 0.8 (Table 1), whose number was similar to the number of AIRs used in the "ECs + NMR" calculations, did not contain information on the A-D dimer interface. Thus, the information provided by ECs at high levels of confidence was not balanced over the two interfaces, possibly due to the evolutionary history of the system. This makes it necessary to use data at lower P cutoffs, which is efficiently filtered by the ambiguous contacts provided by solid-state NMR. The experimental data in fact contain information on both interfaces and thus permit the identification of true contacts of both interfaces within the list of ECs. In the case of Sod1, only EC predictions at intermediate P cutoffs allowed successful docking calculations, due to the low number of ECs predicted at restrictive P values and the low PPV at low (i.e. permissive) P cutoffs. It is thanks to the filter of NMR data, and the corresponding increase in PPV, that it became possible to reliably use the latter set of predictions. It is important to point out that the difficulties highlighted in our analysis of calculations based on ECs only would have been difficult to anticipate solely from the output of the EC predictors. Thus the integration of NMR data can be regarded as a generally viable approach to improve the reliability of the whole procedure, avoiding unforeseen pitfalls.

In a realistic scenario one would use a homology model of the monomer as the input structure to docking calculations. We tested this scenario by generating 15 different models of L–asparaginase II (Table 3) and using the same input AIRs used in the docking of the crystal monomer for all calculations, so that the structure was the only source of variability. For the A-C dimer, we observed that in four cases the best model of the adduct was within 2 Å from the crystal structure, while an additional calculation provided a model with a RMSD of 2.2 Å. The A-D dimer resulted in a similar situation, with two structures within 3 Å and another two at 3.2 Å. Remarkably, there was a very good correlation between the HADDOCK score and the RMSD, allowing the more accurate models to be identified quite straightforwardly. It is also noteworthy that the best results obtained with the homology models had scores close to those obtained with the crystal monomer, which can be reasonably assumed to represent the best possible score. It thus appears that sampling a relatively extensive ensemble of different conformations is an important factor to obtain accurate models of the oligomer in a real-life setting. Previous studies demonstrated that good homology models, as indicated by their predicted RMSD to the true structure, result in accurate complexes, provided that the interface information used as input is of high quality [45].

In summary, our protocol allowed us to predict homo-oligomeric structure in multimers and in presence of a small homodimerization interface. Notably, this goal was achieved with a minimal user effort, making the determination of the 3D structure of the complex faster than using experimental data alone. The only parameter that must be decided by the user is the probability cutoff P below which the ECs are removed. In our hands selecting a P cutoff such that the number of predicted interface residues was 15%–20% of the number of surface residues in the monomer worked well. The results of our protocol clearly depend upon the quality of the ECs obtained from the online servers. Their integration with NMR data serves two different purposes, namely enriching the input AIRs in true contacts when working at low P cutoffs and removing biases among different regions of the protein. From the point of view of NMR spectroscopists, the present work provides a methodology to analyze homo-oligomers with reduced manual effort.

4. Methods

4.1. Computational aspects

The protocol described in the "results" section can be carried out running the python script provided (SI Appendix). The script needs four inputs: the ECs files, the PDB structure of the monomeric protein, the experimental ambiguous NMR contacts list and the Naccess file (rsa format) with the relative solvent accessibility of the residues. Details about inputs preparation, script steps, and docking protocol adopted for the L-asparaginase II and Sod 1 are described below.

The ECs for both proteins were collected using 3 servers available online: Gremlin [40] (http://gremlin.bakerlab.org), RaptorX [62,65] (http://raptorx.uchicago.edu/) and ResTriplet [66] (https://zhanglab.ccmb.med.umich.edu/ResTriplet/). The last two methods are supervised but the PDBs used in this work were not present in the training sets. The MSA in the Gremlin server was generated with the Jackhmmer method and default options [15]. Using different servers adopting different methods in the ECs generation can result in multiple copies of the same EC with different computed likelihood probability. If this is the case, the EC with the highest probability is kept.

The reference protein structures were retrieved from the Protein Data Bank: *E. coli* L-asparaginase II corresponds to PDB ID 6EOK, whereas human apo-Sod1 has the PDB ID 3ECU. Inter-monomeric ECs were identified by removing from the full EC lists all residue pairs with a corresponding Cα-Cα distance <12 Å in chain A of the structures. This distance was already proved as an

appropriate threshold for the selection of true contacts across the interface [60]. In addition, we verified that choosing D = 10 Å or 12 Å produces only a shift of the profile of residue predictions versus P (Table S1). Using the smaller cutoff results in roughly the same number of predicted contacts and a slightly worse PPV but at a greater P value than we used for calculations (0.3 vs 0.25).

The per-residue relative solvent accessible area for both main chain and side chain was calculated with Naccess [28]. Our python script requires the Naccess file in the rsa format to automatically remove all the residues with a relative solvent accessible area below 40% for both the side chain and the main chain. We used the same criteria used in the HADDOCK software to define "active" residues, i.e. residues directly involved in the formation of contacts across interface in protein assemblies.

The monomer–monomer docking calculations were carried out with the HADDOCK software [14]. The residues chosen to drive the docking run were given as active residues (directly involved in the interaction) to generate ambiguous interaction restraints (AIRs) with the default upper distance limit of 2 Å. The water-refined models were clustered based on the fraction of common contacts [44], FCC = 0.75, and the minimum number of elements in a cluster of 4. For the docking run starting from crystal structures, chain A was used as the input monomer. The number of models generated for each step of the HADDOCK docking procedure were set as follow: 10,000 for rigid-body energy minimization, 400 for semi-flexible simulated annealing and 400 for refinement in explicit solvent. The distance violation analysis was performed on the best cluster and the corresponding output written in the ana_dist_viol_all.lis file. In this file we selected all the residues with a violation larger than 3 Å to generate a subset of AIRs to drive a second docking run. Thus, the second docking run was performed using exactly the same conditions as the first one.

We generated 15 models of monomeric *E. coli* L-asparaginase II using the structure of *Wolinella succinogenes* L-asparaginase [34] as a template (PDB ID 1WSA, chain A) using Modeller [18]. The two proteins have 55% sequence identity. The resulting template-based models featured a very similar backbone conformation, lower than 1 Å from the *E. coli* crystal, but different side chain orientations. Each model was assessed in protein–protein docking using the same AIRs used in the "crystal P 0.25" runs, with all the AIRs (A-C dimer calculation) and after the removal of the ones already satisfied by the A-C dimer (A-D dimer calculation), respectively. The number of models generated for each step were reduced as follow: 1000 for rigid-body energy minimization, 200 for semi-flexible simulated annealing and 200 for refinement in explicit solvent.

All the RMSD values reported in this work were measured on the Cα atoms.

4.2. Solid- and solution-state NMR data

The L-asparaginase II protein [U- ^{13}C, ^{15}N] was expressed and purified as previously reported [9,21,22,43], freeze-dried and packed (ca. 20 mg) into a Bruker 3.2 mm zirconia rotor. The material was rehydrated with a solution of 9 mg/mL NaCl in MilliQ H_2O; the hydration process was monitored through 1D {^1H}-^{13}C cross-polarization SSNMR spectrum and stopped when the resolution of the spectrum did not change any further for successive additions of the solution [21,22,43]. Silicon plug, (courtesy of Bruker Biospin) placed below the turbine cap, was used to close the rotor and preserve hydration.

SSNMR experiments were recorded on a Bruker AvanceIII spectrometer operating at 800 MHz (19 T, 201.2 MHz ^{13}C Larmor frequency) equipped with Bruker 3.2 mm Efree NCH probe-head. All spectra were recorded at 14 kHz MAS frequency and the sample temperature was kept at ≈290 K.

Standard ^{13}C–^{13}C correlation spectra (Dipolar Assisted Rotational Resonance, DARR) with different mixing times (50, 200 and 400 ms) were acquired using the pulse sequences reported in the literature [56]. Pulses were 2.6 μs for ^1H, 4 μs for ^{13}C; the spectral width was set to 282 ppm; 2048 and 1024 points were acquired in the direct and indirect dimensions, respectively; 128 scans were acquired; the inter-scan delay was set to 1.5 s in all the experiments.

All the spectra were processed with the Bruker TopSpin 3.2 software package and analyzed with the program CARA [31].

The assignment of the carbon resonances of the 2D ^{13}C–^{13}C DARR spectra of rehydrated freeze-dried ANSII was easily obtained by comparison with the 2D ^{13}C–^{13}C DARR spectrum collected on the crystalline and PEGylated preparations of L-asparaginase II [9,43].

The experimental data used for the Sod1 protein were taken from deposited solution-state 3D ^1H-^{15}N NOESY-HSQC spectrum [5].

Ambiguous assignment lists of the 2D ^{13}C–^{13}C DARR and 3D ^1H-^{15}N NOESY-HSQC peaks were generated with the program ATNOS/CANDID [1,24] by setting the chemical-shift–based assignment tolerances to 0.25 ppm and 0.025 ppm, respectively.

The ambiguous contacts list of the L-asparaginase II protein contains 4937 resonance assignments grouped in 1634 peaks. Among the residue pairs in the ambiguous contacts list, 27 out of 325 true positive interface contacts, i.e. present in the crystal structure, were found in the list

The ambiguous contacts list of the Sod1 protein contains 7611 resonance assignments grouped in 2878 peaks. Among the residue pairs in the ambiguous contacts list, 32 out of 72 true positive interface contacts were found in the list.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

Financial support was provided by the European Commission (project no. 777536) and by the Interuniversity Consortium for Magnetic resonance of MetalloProteins (CIRMMP).

We thank Prof. Gaetano Montelione for many useful discussions.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.csbj.2019.12.002.

References

[1] Andreas LB, Jaudzems K, Stanek J, Lalli D, Bertarello A, Le Marchand T, et al. Structure of fully protonated proteins by proton-detected magic-angle spinning NMR. Proc Natl Acad Sci 2016;113:9187–92.
[2] Anishchenko I, Ovchinnikov S, Kamisetty H, Baker D. Origins of coevolution between residues distant in protein 3D structures. Proc Natl Acad Sci 2017;114:9122–7.
[3] Bai F, Morcos F, Cheng RR, Jiang H, Onuchic JN. Elucidating the druggable interface of protein–protein interactions using fragment docking and coevolutionary analysis. Proc Natl Acad Sci 2016;113:E8051–8.
[4] Balakrishnan S, Kamisetty H, Carbonell JG, Lee S-I, Langmead CJ. Learning generative models for protein fold families. Proteins Struct Funct Bioinforma 2011;79:1061–78.
[5] Bertini I, Cantini F, Vieru M, Banci L, Girotto S, Boca M, et al. Structural and dynamic aspects related to oligomerization of apo SOD1 and its mutants. Proc Natl Acad Sci 2009;106:6980–5.

[6] Bitbol A-F, Dwyer RS, Colwell LJ, Wingreen NS. Inferring interaction partners from protein sequences. Proc Natl Acad Sci 2016;113:12180–5.
[7] Burger L, van Nimwegen E. Accurate prediction of protein–protein interactions from sequence alignments using a Bayesian method. Mol Syst Biol 2008;4:165.
[8] Burger L, van Nimwegen E. Disentangling direct from indirect co-evolution of residues in protein alignments. PLoS Comput Biol 2010;6:e1000633.
[9] Cerofolini L, Giuntini S, Carlon A, Ravera E, Calderone V, Fragai M, et al. Characterization of PEGylated Asparaginase: new opportunities from NMR analysis of large PEGylated therapeutics. Chem Eur J 2019;25:1984–91.
[10] Cheng RR, Morcos F, Levine H, Onuchic JN. Toward rationally redesigning bacterial two-component signaling systems using coevolutionary information. Proc Natl Acad Sci 2014;111:E563–71.
[11] Cocco S, Monasson R, Weigt M. From Principal component to direct coupling analysis of coevolution in proteins: low-eigenvalue modes are needed for structure prediction. PLoS Comput Biol 2013;9:e1003176.
[12] Dago AE, Schug A, Procaccini A, Hoch JA, Weigt M, Szurmant H. Structural basis of histidine kinase autophosphorylation deduced by integrating genomics, molecular dynamics, and mutagenesis. Proc Natl Acad Sci 2012;109:E1733–42.
[13] Demers J-P, Fricke P, Shi C, Chevelkov V, Lange A. Structure determination of supra-molecular assemblies by solid-state NMR: practical considerations. Prog Nucl Magn Reson Spectrosc 2018;109:51–78.
[14] Dominguez C, Boelens R, Bonvin AMJJ. HADDOCK: a protein-protein docking approach based on biochemical or biophysical information. J Am Chem Soc 2003;125:1731–7.
[15] Eddy SR. Profile hidden Markov models. Bioinformatics 1998;14:755–63.
[16] Ekeberg M, Lövkvist C, Lan Y, Weigt M, Aurell E. Improved contact prediction in proteins: using pseudolikelihoods to infer Potts models. Phys Rev E 2013;87:012707.
[17] El-Gebali S, Mistry J, Bateman A, Eddy SR, Luciani A, Potter SC, et al. The Pfam protein families database in 2019. Nucleic Acids Res 2019;47:D427–32.
[18] Eswar N, Webb B, Marti-Renom MA, Madhusudhan MS, Eramian D, Shen M, Pieper U, Sali A. In: Comparative protein structure modeling using MODELLER. In current protocols in protein science. Hoboken, NJ, USA: John Wiley & Sons, Inc.; 2007. p. 2.9.1–2.9.31.
[19] Fernández-Recio J, Totrov M, Abagyan R. Identification of protein-protein interaction sites from docking energy landscapes. J Mol Biol 2004;335:843–65.
[20] Gauto DF, Estrozi LF, Schwieters CD, Effantin G, Macek P, Sounier R, et al. Integrated NMR and cryo-EM atomic-resolution structure determination of a half-megadalton enzyme complex. Nat Commun 2019;10:2697.
[21] Giuntini S, Cerofolini L, Ravera E, Fragai M, Luchinat C. Atomic structural details of a protein grafted onto gold nanoparticles. Sci Rep 2017;7:17934.
[22] Giuntini S, Balducci E, Cerofolini L, Ravera E, Fragai M, Berti F, et al. Characterization of the conjugation pattern in large polysaccharide-protein conjugates by NMR spectroscopy. Angew Chemie Int Ed 2017;56:14997–5001.
[23] Göbl C, Madl T, Simon B, Sattler M. NMR approaches for structural analysis of multidomain proteins and complexes in solution. Prog Nucl Magn Reson Spectrosc 2014;80:26–63.
[24] Guerry P, Herrmann T. Comprehensive Automation for NMR Structure Determination of Proteins (Clifton, NJ). In: Methods in Molecular Biology. p. 429–51.
[25] Güntert P, Buchner L. Combined automated NOE assignment and structure calculation with CYANA. J Biomol NMR 2015;62:453–71.
[26] Hopf TA, Schärfe CPI, Rodrigues JPGLM, Green AG, Kohlbacher O, Sander C, et al. Sequence co-evolution gives 3D contacts and structures of protein complexes. Elife 2014;3:1–45.
[27] Hu J, Liu HF, Sun J, Wang J, Liu R. Integrating co-evolutionary signals and other properties of residue pairs to distinguish biological interfaces from crystal contacts. Protein Sci 2018;27:1723–35.
[28] Hubbard SJ, Thornton JM. NACCESS, 1993.
[29] Jehle S, van Rossum B, Stout JR, Noguchi SM, Falber K, Rehbein K, et al. alphaB-crystallin: a hybrid solid-state/solution-state NMR investigation reveals structural aspects of the heterogeneous oligomer. J Mol Biol 2009;385:1481–97.
[30] Jones DT, Buchan DWAA, Cozzetto D, Pontil M. PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. Bioinformatics 2012;28:184–90.
[31] Keller R. The computer aided resonance tutorial. 2007, p. 81.
[32] Li Y, Zhang C, Bell EW, Yu D, Zhang Y. Ensembling multiple raw coevolutionary features with deep residual neural networks for contact-map prediction in CASP13. Proteins Struct Funct Bioinforma 2019.
[33] Loquet A, Sgourakis NG, Gupta R, Giller K, Riedel D, Goosmann C, et al. Atomic model of the type III secretion system needle. Nature 2012;486:276–9.
[34] Lubkowski J, Palm GJ, Gilliland GL, Derst C, Röhm KH, Wlodawer A. Crystal structure and amino acid sequence of Wolinella succinogenes L-asparaginase. Eur J Biochem 1996;241:201–7.
[35] Marks DS, Colwell LJ, Sheridan R, Hopf TA, Pagnani A, Zecchina R, et al. Protein 3D structure computed from evolutionary sequence variation. PLoS One 2011;6:e28766.
[36] Marks DS, Hopf TA, Sander C. Protein structure prediction from sequence variation. Nat Biotechnol 2012;30:1072–80.
[37] Meier BH, Riek R, Böckmann A. Emerging structural understanding of amyloid fibrils by solid-state NMR. Trends Biochem Sci 2017;42:777–87.
[38] Morcos F, Pagnani A, Lunt B, Bertolino A, Marks DS, Sander C, et al. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. Proc Natl Acad Sci 2011;108:E1293–301.
[39] Morcos F, Jana B, Hwa T, Onuchic JN. Coevolutionary signals across protein lineages help capture multiple protein conformations. Proc Natl Acad Sci 2013;110:20533–8.
[40] Ovchinnikov S, Kamisetty H, Baker D. Robust and accurate prediction of residue-residue interactions across protein interfaces using evolutionary information. Elife 2014;2014:1–21.
[41] Procaccini A, Lunt B, Szurmant H, Hwa T, Weigt M. Dissecting the specificity of protein-protein interaction in bacterial two-component signaling: orphans and crosstalks. PLoS One 2011;6:e19729.
[42] Qian W, He X, Chan E, Xu H, Zhang J. Measuring the evolutionary rate of protein-protein interaction. Proc Natl Acad Sci USA 2011;108:8725–30.
[43] Ravera E, Ciambellotti S, Cerofolini L, Martelli T, Kozyreva T, Bernacchioni C, et al. Solid-state NMR of PEGylated proteins. Angew Chemie Int Ed 2016;55:2446–9.
[44] Rodrigues JPGLM, Trellet M, Schmitz C, Kastritis P, Karaca E, Melquiond ASJ, et al. Clustering biomolecular complexes by residue contacts similarity. Proteins Struct Funct Bioinforma 2012;80:1810–7.
[45] Rodrigues JPGLM, Melquiond ASJ, Karaca E, Trellet M, van Dijk M, van Zundert GCP, et al. Defining the limits of homology modeling in information-driven protein docking. Proteins Struct Funct Bioinforma 2013;81:2119–28.
[46] Rodriguez-Rivas J, Marsili S, Juan D, Valencia A. Conservation of coevolving protein interfaces bridges prokaryote–eukaryote homologies in the twilight zone. Proc Natl Acad Sci 2016;113:15018–23.
[47] Rosato A, Vranken W, Fogh RH, Ragan TJ, Tejero R, Pederson K, et al. The second round of critical assessment of automated structure determination of proteins by NMR: CASD-NMR-2013. J Biomol NMR 2015;62:413–24.
[48] Rose PW, Prlić A, Bi C, Bluhm WF, Christie CH, Dutta S, et al. The RCSB Protein Data Bank: views of structural biology for basic and applied research and education. Nucleic Acids Res 2015;43:D345–56.
[49] Salinas VH, Ranganathan R. Coevolution-based inference of amino acid interactions underlying protein function. Elife 2018;7.
[50] dos Santos RN, Morcos F, Jana B, Andricopulo AD, Onuchic JN. Dimeric interactions and complex formation using direct coevolutionary couplings. Sci Rep 2015;5:13652.
[51] Schug A, Weigt M, Onuchic JN, Hwa T, Szurmant H. High-resolution protein complexes from integrating genomic information with molecular simulation. Proc Natl Acad Sci USA 2009;106:22124–9.
[52] Skwark MJ, Elofsson A. PconsD: ultra rapid, accurate model quality assessment for protein structure prediction. Bioinformatics 2013;29:1817–8.
[53] Sun MGF, Kim PM. Evolution of biological interaction networks: from models to real data. Genome Biol 2011;12:235.
[54] Sutto L, Marsili S, Valencia A, Gervasio FL. From residue coevolution to protein conformational ensembles and functional dynamics. Proc Natl Acad Sci 2015;112:13567–72.
[55] Szurmant H, Weigt M. Inter-residue, inter-protein and inter-family coevolution: bridging the scales. Curr Opin Struct Biol 2018;50:26–32.
[56] Takegoshi K, Nakamura S, Terao T, Nakamura S. 13C–1H dipolar-assisted rotational resonance in magic-angle spinning NMR. Chem Phys Lett 2001;344:631–7.
[57] Tang Y, Huang YJ, Hopf TA, Sander C, Marks DS, Montelione GT. Protein structure determination by combining sparse NMR data with evolutionary couplings. Nat Methods 2015;12:751–4.
[58] Tian P, Boomsma W, Wang Y, Otzen DE, Jensen MH, Lindorff-Larsen K. Structure of a functional amyloid protein subunit computed using sequence variation. J Am Chem Soc 2015;137:22–5.
[59] Traaseth NJ, Verardi R, Veglia G. Asymmetric methyl group labeling as a probe of membrane protein homo-oligomers by NMR spectroscopy. J Am Chem Soc 2008;130:2400–1.
[60] Uguzzoni G, John Lovis S, Oteri F, Schug A, Szurmant H, Weigt M. Large-scale identification of coevolution signals across homo-oligomeric protein interfaces by direct coupling analysis. Proc Natl Acad Sci 2017;114:E2662–71.
[61] de Vries SJ, van Dijk ADJ, Krzeminski M, van Dijk M, Thureau A, Hsu V, et al. HADDOCK versus HADDOCK: new features and performance of HADDOCK2.0 on the CAPRI targets. Proteins Struct Funct Bioinforma 2007;69:726–33.
[62] Wang S, Sun S, Li Z, Zhang R, Xu J. Accurate de novo prediction of protein contact map by ultra-deep learning model. PLoS Comput Biol 2017;13:e1005324.
[63] Waterhouse A, Bertoni M, Bienert S, Studer G, Tauriello G, Gumienny R, et al. SWISS-MODEL: homology modelling of protein structures and complexes. Nucleic Acids Res 2018;46:W296–303.
[64] Weigt M, White RA, Szurmant H, Hoch JA, Hwa T, White RA, et al. Identification of direct residue contacts in protein-protein interaction by message passing. Proc Natl Acad Sci 2008;106:67–72.
[65] Xu J, Zhang R, Wang S, Li W, Liu S. CoinFold: a web server for protein contact prediction and contact-assisted protein folding. Nucleic Acids Res 2016;44:W361–6.
[66] Li Y, Zhang C, Dongjun Yu YZ. Contact prediction by stacked fully convolutional residual neural network using coevolution features from deep multiple sequence alignment. CASP13 Abstr B 2018;154.