Modeling Protein Complexes and Molecular Assemblies Using Computational Methods

Romain Launay, Elin Teppa, Jérémy Esque, Isabelle André

Romain Launay, Elin Teppa, Jérémy Esque, Isabelle André. Modeling Protein Complexes and Molecular Assemblies Using Computational Methods. Kumar Selvarajoo. Computational Biology and Machine Learning for Metabolic Engineering and Synthetic Biology, 2553, Springer, pp.57-77, 2022

Modeling Protein Complexes and Molecular Assemblies Using Computational Methods

Romain Launay, Elin Teppa, Jérémy Esque, and Isabelle André

Abstract

Many biological molecules are assembled into supramolecular complexes that are necessary to perform functions in the cell. Better understanding and characterization of these molecular assemblies are thus essential to further elucidate molecular mechanisms and key protein-protein interactions that could be targeted to modulate the protein binding affinity or develop new binders. Experimental access to structural information on these supramolecular assemblies is often hampered by the size of these systems that make their recombinant production and characterization rather difficult. Computational methods combining both structural data, molecular modeling techniques, and sequence coevolution information can thus offer a good alternative to gain access to the structural organization of protein complexes and assemblies. Herein, we present some computational methods to predict structural models of the protein partners, to search for interacting regions using coevolution information, and to build molecular assemblies. The approach is exemplified using a case study to model the succinate-quinone oxidoreductase heterocomplex.

Key words Protein-protein interaction, PPI, Molecular assembly, Protein structure prediction, Protein-protein docking, Sequence coevolution

1 Introduction

Protein-Protein Interactions (PPIs) play an important role in the functioning of living cells, including cell-to-cell interactions and metabolic and developmental control [1, 2]. Most cellular functions are mediated by the assembly of proteins as more than 80% of the proteins operate in vivo in the form of homo- or hetero-oligomers [3] whose constituents assemble/disassemble dynamically [4]. Interaction between the proteins can be permanent or transient. While permanent interactions will form a stable protein

Romain Launay and Elin Teppa contributed equally with all other contributors.

Supplementary Information The online version contains supplementary material available at [https://doi.org/10.1007/978-1-0716-2617-7_4].

complex, the transient interactions are rather involved in signaling and regulation pathways or substrate/metabolite channeling [2, 5, 6]. Better understanding these molecular assemblies and PPIs is thus of major importance to further elucidate molecular mechanisms of cellular processes, engineer synthetic metabolic pathways for synthetic biology, or identify drug targets for biomedical applications [5].

PPIs can be investigated at different levels. In vivo, yeast two-hybrid (Y2H, Y3H) techniques enable to detect protein interactions, while in vitro, a variety of methods can be used such as tandem affinity purification, affinity chromatography, coimmunoprecipitation, protein arrays, protein fragment complementation, phage display, and mass spectrometry [6–8] among others. At the structural level, investigation of PPIs has largely benefited from the growing number of protein-protein complexes solved in recent years using different biophysical techniques, such as X-ray crystallography, nuclear magnetic resonance spectroscopy, and cryo-electron microscopy [7]. To complete this arsenal of approaches, in silico molecular modeling based on a combination of template-based methods and docking approaches that can integrate experimental restraints (i.e., coevolution information) has also emerged as a powerful technique to investigate protein assemblies, in particular when experimental data are lacking [3, 9].

In this chapter, we provide a brief introduction to computational methods that allow to predict structural models of proteins, to search for interacting regions using inter-protein coevolution information, and to model and analyze molecular assemblies. The use of some of these methods and tools is illustrated for the modeling of the succinate-quinone oxidoreductase heterocomplex as a case study.

2 Methods for Building a 3D Model of a Protein

Predicting the three-dimensional structure of a protein based on its sequence is still an open problem in research. Protein structure prediction methods on the basis of protein sequences are based on two principles: (i) protein structure is more conserved across evolution than protein sequence, and (ii) there is a finite and relatively small (less than 10,000) number of unique protein folds in Nature [10].

Structure prediction methods are broadly classified into two categories: (a) template-based modeling (which uses one or several known structure(s) as template(s)) and (b) template-free modeling (which predicts a protein structure without using a significant template). There are also hybrid approaches that combine the two kinds of methods.

New modeling methods or corrections to existing methods continually emerge. There are several ways to keep up with the best existing methods, identify the progress over time, and recognize where future efforts may be most productively focused. One way is to be aware of CASP results (the Critical Assessment of Protein Structure Prediction, www.predictioncenter.org) conducted every 2 years since 1994. Another way is to check the Continuous Automated Model EvaluatiOn (CAMEO; www.cameo3d.org) project that provides weekly follow-ups for three different aspects of the prediction by web servers: (a) homology modeling, (b) model quality estimation, and (c) contact prediction.

In recent years, machine learning approaches have contributed tremendously to improve the accuracy of structural prediction, even when no similar structure is known [11]. Particularly in the recent CASP14, the AlphaFold2 method [11] outperformed most methods by predicting structures with high accuracy.

2.1 Template-Based Methods

The methods referred to as template-based modeling include threading techniques and comparative modeling. Template-based modeling predicts the 3D structure of a query protein through the sequence alignment between the query and one or several proteins with known structures. When query and template sequences have been derived from a common ancestor, the method is referred to as homology modeling. However, proteins from different evolutionary origins may still adopt a similar structure; in this case, threading methods are used to identify structural templates.

Generally, the process of comparative modeling involves four steps: (a) template identification, (b) sequence alignment, (c) model building, and (d) model refinement and validation. If the model is not satisfactory, some or all of the steps can be repeated. As such, the success of homology modeling depends on the ability to identify the closely homologous templates based on sequence identity and to generate an accurate query-template alignment. The goal of the alignment is to map the one-dimensional target sequence onto corresponding three-dimensional positions of the template structure correctly, ideally with only substitutions and small insertions/deletions. Broadly speaking, comparative modeling produces a good result if the query-template alignment has a global sequence identity $\geqq$30%. As the sequence identity decreases, a correct template identification is more difficult and prone to misaligned regions. When query-template sequence identity is between 20% and 30%, they fall in the twilight zone; the evolutionary relatedness of proteins becomes uncertain [12, 13]. In this case, the threading technique may help to identify remote homology, leaving the ab initio method as the last alternative for protein structure prediction.

2.2 Template-Free or Ab Initio Methods

For query proteins that have no structurally related protein in the PDB library, the structure must be built from scratch. This procedure is called ab initio modeling, de novo modeling, or template-free modeling. An ab initio method conducts an exhaustive search to identify the minimum energy conformation through optimization algorithms, such as Monte Carlo [14] or molecular dynamics [15], using knowledge-based scoring or physics-based energy functions. This procedure generates several putative conformations (also called decoys), and final models are selected from them. A successful ab initio modeling depends on three factors:

(a) An accurate energy function that scores the native structure of a protein as being the most thermodynamically stable state, compared to all possible decoy structures

(b) An efficient search method that can quickly identify the low-energy states through conformational search

(c) A strategy that can select near-native models from a pool of decoy structures

2.3 Servers for Protein Structure Prediction and Related Databases

Hereafter are presented some servers and databases used for protein structure prediction based on various strategies and using, in some cases, sequence coevolution information and artificial intelligence-derived methods.

2.3.1 MODELLER via ModWeb and ModBase

MODELLER is one of the most widespread comparative modeling methods for prediction of protein structures [16]. Models are obtained by satisfying spatial restraints derived from the query-template alignment.

These restraints include:

(a) Cα-Cα and backbone N-O distances and dihedral angles restraints

(b) Stereochemical restraints from the CHARMM-22 force field

(c) Statistical preferences for dihedral angles and non-bonded inter-atomic distances derived from representative sets of known protein structures

Optionally, it is possible to add manually additional restraints. MODELLER is available free of charge only to academic nonprofit institutions at https://salilab.org/modeller/.

Several servers based on MODELLER have been developed such as ModWeb or ModBase.

ModWeb server (https://modbase.compbio.ucsf.edu/modweb/) offers the possibility to use MODELLER online.

ModBase (http://salilab.org/modbase) is a database containing fold assignments, sequence-structure alignments, models, and model assessments for all sequences related to a known structure [17]. The models are derived by ModPipe, an automated modeling

pipeline relying on the programs PSI-BLAST [18] and MODELLER. ModBase also includes binding site prediction for small ligands and a set of predicted interactions between pairs of modeled sequences from the same genome that are predicted to interact with each other.

2.3.2 PHYRE2

PHYRE2 (http://www.sbg.bio.ic.ac.uk/phyre2) is designed to predict a protein three-dimensional structure from a protein sequence [19]. The server uses a powerful strategy to detect remote homology combining PSI-BLAST alignment with hidden Markov models (HMM) via HHsearch for template detection. The primary algorithmic strategy is composed of four steps. In the first step, homologous sequences of the query are searched using HHblits. The resulting alignment is used to predict secondary structure. In the second step, HHsearch is performed against a database of HMMs of protein of known structures. The top-scored alignments are used to construct the protein model backbone. In the third step, the loops are modeled, and in the last step, the side chains are added to generate the final model. When the intensive mode is used, a step is added to use an ab initio folding simulation called Poing[2] to model regions of the query protein with no detectable homology to known structures.

2.3.3 I-TASSER

I-TASSER (Iterative Threading ASSEmbly Refinement) is a hierarchical approach to protein structure and function predictions from their amino acid sequences [20]. I-TASSER is accessible via a web server (https://zhanglab.dcmb.med.umich.edu/I-TASSER) and a stand-alone package. Starting from an amino acid sequence, the algorithm tries to retrieve protein templates of similar fold from the Protein Data Bank (PDB: https://www.rcsb.org) using a meta-threading approach called LOMETS (https://zhanggroup.org/LOMETS/). In the next step, the continuous fragments taken from the PDB templates are reassembled into full-length models. For cases where no appropriate template is identified, I-TASSER builds the whole structure by ab initio modeling. SPICKER identifies the low free-energy states through clustering the simulation decoys (https://zhanggroup.org/SPICKER/). In the third step, a second iteration of the fragment assembly simulation is performed again to remove the steric clash and refine the global topology of the cluster centroids. The decoys generated are then clustered, and the lowest energy structures are selected followed by an optimization of the hydrogen-bonding network. The final model is used to predict the protein biological function by matching the model with other known proteins using the enzyme classification (EC number), gene ontology vocabulary, and ligand binding sites. More recently, an I-TASSER-derived method called D-I-TASSER has been developed for distance-guided protein structure prediction (https://zhanggroup.org//D-I-TASSER/). This

method integrates inter-residue contacts predicted by deep neural network and has been reported to significantly enhance accuracy of models compared to I-TASSER.

2.3.4 trRosetta

trRosetta (transform-restrained Rosetta) is an algorithm for protein structure prediction using a deep neural network to predict the inter-residue distances [9]. The algorithm is available in a standalone version and a web server (https://yanglab.nankai.edu.cn/trRosetta/). The input is the amino acid sequence or a multiple sequence alignment of the query protein. A deep neural network is applied to predict the inter-residue distances and orientation distributions between residues. Some of the features used in the convolutional layers of the networks include amino acid frequencies, entropies, and coevolutionary couplings.

Predicted inter-residue distances and orientations are used as restraints to guide the Rosetta method to build three-dimensional structure models based on direct energy minimization.

Recently, the algorithm was modified to include the option to use templates. It is recommended to run the algorithm including homologous templates, which are used to add restraints to Rosetta.

2.3.5 AlphaFold2 Method and Structural Database

Given a query sequence, AlphaFold2 [11] searches for related sequences in three databases: UbiRef90, BFD, and MGnify. Then, potential templates are searched using HHsearch against the PDB70 database [21]. The input sequence, multiple sequence alignment, and template hits are used as inputs for the deep learning-based method that produces a variety of predictions including distances, torsions, and atom coordinates. Then, the predicted 3D model is relaxed using restrained gradient descent with the Amber ff99SB force field [22] integrated in OpenMM [23].

AlphaFold2 produces a per-residue confidence metric called the predicted local distance difference test (pLDDT) on a scale from 0 to 100, to estimate how well the prediction agrees with an experimental structure considering the Cα. A pLDDT >90 is considered as a highly accurate prediction; in addition to a good backbone prediction, the side chains are often correctly oriented (χ1 rotamers are 80% correct). Regions with pLDDT between 70 and 90 indicate a generally good backbone prediction. Regions with pLDDT between 50 and 70 are low confidence and should be treated with caution. Finally, regions with pLDDT <50 are probably disordered.

In CASP14, AlphaFold2 was the top-ranked protein structure prediction method, producing predictions with high accuracy [24].

The source code of AlphaFold2 is available on GitHub (https://github.com/deepmind/alphafold). It is also possible to use AlphaFold2 via the Google ColabFold notebooks [25], a free

platform for protein folding that does not require any installation or expensive hardware. Several ColabFold notebooks are available on GitHub (https://github.com/sokrypton/ColabFold).

DeepMind and EMBL's European Bioinformatics Institute (EMBL-EBI) created the AlphaFold database (https://alphafold.ebi.ac.uk) to provide open access to protein structure predictions generated by the AlphaFold2 method. At the moment, the predictions cover almost the entire human proteome [26] and the proteomes of several other key organisms such as *E. coli*, fruit fly, mouse, and zebrafish, among others, totaling over 350,000 protein structures. The database provides three outputs from AlphaFold2: the three-dimensional coordinates, the per-residue confidence metric pLDDT, and the Predicted Aligned Error, which is necessary to assess confidence in the domain packing and large-scale topology of the protein.

3 Protein-Protein Interaction Prediction Using Coevolution

We refer to molecular coevolution when a change in one locus affects the selection pressure at another locus, and this change is reciprocal [27, 28]. In other words, when a mutation occurs in a particular position, another mutation may occur to compensate for the change or restore the protein function. As coevolving residues tend to be close in the tridimensional structure, coevolution has been successfully applied to predict intra- and inter-protein residue contacts [29–32]. When coevolution methods were applied at whole-proteome scale combined with structure modeling to predict protein-protein interactions, the accuracy of interaction prediction is higher than the proteome-wide two-hybrid and mass spectrometry screens [33]. A large panel of methods exists to predict molecular coevolution; all of them use a multiple sequence alignment (MSA) as input. In general, a large number of diverse sequences are required to obtain reliable results. To predict inter-protein coevolution between two proteins A and B, the real input of the coevolution algorithm is the concatenated alignment; protein A and protein B for each organism must be properly paired (Fig. 1). Building the concatenated alignment is not straightforward, because each row of the MSA should contain a pair of interacting proteins out of two protein families. That means that it is desirable to concatenate orthologous proteins, as they are likely to perform an equivalent function, rather than other types of homologs.

The I-COMS web server (http://i-coms.leloir.org.ar) allows computing inter-protein contact prediction using four different covariation methods [34]. The server gives the option to provide the concatenated alignment or build it automatically. The server includes four covariation methods: corrected mutual information,

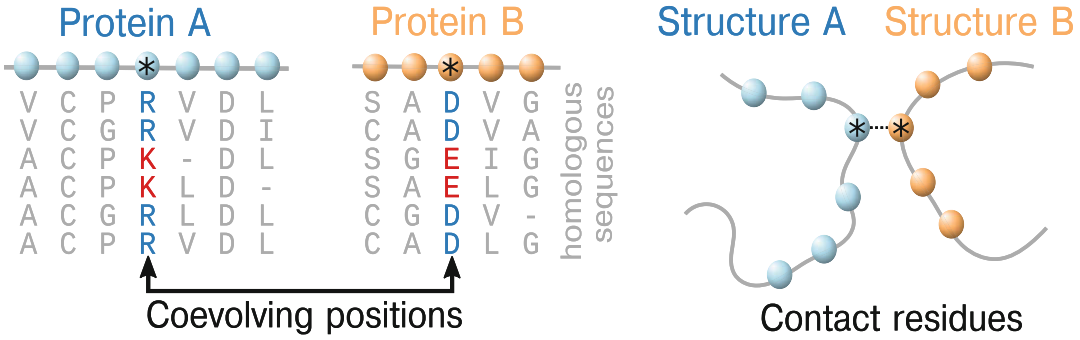

Fig. 1 Inter-protein coevolution. In the concatenated alignment between two interacting proteins A and B, two positions coevolve (indicated with an arrow) to maintain favorable interactions between physically interacting amino acid residues (indicated as *) in the three-dimensional structure

mfDCA, PSICOV, and CCMpred. Intra- and inter-protein results are provided in an interactive visualization allowing the comparison between methods as well as the concordance between results. Covariation positions can be calculated for up to five proteins.

4 Protein Assembly Prediction and Analysis

4.1 Protein-Protein Docking: Principles and Methods

When the structural information of different protein partners is available through experimental data or modeling, the docking approach is used as a standard method to predict the potential interactions. The aim of docking is to find the best matched 3D structure of the protein complex among several protein models. To do so, a fast search algorithm is used to sample all possible spatial conformations, and a scoring function is needed to rank the solutions. Due to the large number of possibilities for the position and angle of protein residues, spatial search algorithms in protein-protein docking can be divided into three main categories: (a) exhaustive global search including fast Fourier transform (FFT)-based search implemented [35, 36] and spherical Fourier transform-based search [37–39], (b) randomized search using Monte Carlo [40, 41], and (c) local shape feature matching including geometric hashing [40]. It is important to notice that all FFT-based approaches perform rigid-body docking because the related grid cannot be updated, unlike randomized search algorithms.

Protein-protein docking methods typically generate thousands of potential solutions for a particular complex. To discriminate near-native solutions, the development of a scoring function is needed and is still challenging. These scoring functions can be divided into several categories, sometimes combined: (a) physics-based scoring function capturing the determinants related to the stability of protein-protein complexes, e.g., shape complementary, van der Waals, electrostatics, and desolvation potential [41–47],

(b) knowledge-based functions taking advantage of the information from available structures [48–51], (c) scoring functions combining physical terms with knowledge-based terms [52–55], (d) evolutionary scoring function based on the protein sequence evolution [56, 57], and (e) consensus-based scoring functions seeking to identify solutions with high occurrence features, independently of any physics-based or evolutionary evaluation, such as conservation of interface contacts [58–62]. Along the same line as the CASP contest for protein structure prediction, the CAPRI competition allows a blind assessment of the most recent methods, offering an updated view of progress in the field [63–65].

4.2 ZDOCK

ZDOCK is a protein-protein docking method available through an online web server (https://zdock.umassmed.edu/) [66]. It uses the fast Fourier transform algorithm to enable an efficient docking search. It is a user-friendly server to predict complexes that proceed in three steps. The first step is to provide two input structures (by PDB code or PDB file) and choose the ZDOCK version. The second step is the selection of blocking or contacting residues for each protein submitted. The last step is the result analysis and visualization, including the top ten docking models.

4.3 InterEvDock3

InterEvDock3 (https://bioserv.rpbs.univ-paris-diderot.fr/services/InterEvDock3/) is a server designed for predicting protein pairwise assemblies, based on sequence or on structure, and possibly combined with coevolution data [67]. Three protocols are implemented to use at best the available information.

The first method is template-based docking; it uses sequences to search the protein assembly with already known structures. Template-based docking protocols need two or more sequences and a protocol search among a list of interacting proteins if the structure of protein homologs is available in complex with partners, based on HHsearch. The structural assembly is built with threading for the main parts, and the missing parts are built with the DaReUS-Loop program [68].

The two other methods perform free docking using the FRODOCK software. Then, generated models are ranked according to the coevolution information given by the user or computed by the server.

5 Case Study: Modeling the Succinate-Quinone Oxidoreductase Heterocomplex

We propose to build a structural model of the supramolecular complex succinate-quinone oxidoreductase (SQR). SQR is a key enzyme in the Krebs cycle, oxidizing succinate to fumarate and reducing quinone to quinol, acting as a link between the Krebs cycle and the respiratory chain. *Escherichia coli* SQR has four subunits, two hydrophilic subunits exposed to the cytoplasm (SdhA

and SdhB), which interact with two hydrophobic membrane-intrinsic subunits (SdhC and SdhD) [69]. Interestingly, SdhA and SdhB have already been shown to coevolve together. This information enabled to predict the proper interacting interface [29–32] compared to the crystallographic protein structure of *E.coli* SQR [70, 71] (PDB code: 1NEK, 2WDQ).

For pedagogical purposes, we provide step-by-step instructions to generate the structural models of the heterotetramer subunits and their assembly (Fig. 2). First, we shall build a structural model for all subunits (SdhA, SdhB, SdhC, and SdhD) using either the AlphaFold2 method without template or I-TASSER without using close templates. This choice will mimic cases where no crystallographic information is available. Second, we will use inter-protein coevolution detection to predict residue contacts between the subunits. The dataset for coevolution comes from the available data reported in reference [30] and is provided in supplementary information (SI1). Third, the predicted residue contacts will be used to guide the protein-protein docking. Fourth, a docking was carried out between the dimers SdhA-SdhB and SdhC-SdhD without using coevolution information.

5.1 Building a 3D Model Using AlphaFold2: SQR Subunits, SdhA, and SdhC

To avoid setting up AlphaFold2 on your local computer, we will use an online version to build the 3D models of SdhA and SdhC. The following steps are the same for SdhA (UniProt ID: P0AC4) and SdhC (UniProt ID P69054):

1. Download the amino acid sequence of the target in FASTA format from UniProt.
2. Go to ColabFold repository (https://github.com/sokrypton/ColabFold).
3. Choose the Notebook *AlphaFold2 (from DeepMind)*.
4. Execute the first two cells by clicking the play button. It will install the required programs in the cloud, and not on your computer.
5. Wait until the task is completed, a green tick mark will appear at the left of the play button. You can also visualize the progression of each task in the progress bar (Fig. 3a).
6. Paste the protein sequence without the FASTA header in the text box.
7. Select *Runtime* -> *Run After* in the toolbar at top of screen.
8. Unzip the file downloaded automatically with the results.
9. It's done! Now, we are ready to analyze the results.

To make sure that you can reproduce the result, it is recommended to save a copy of the notebook on your computer. You can find several options to save the notebook in the *File* menu in the top bar.

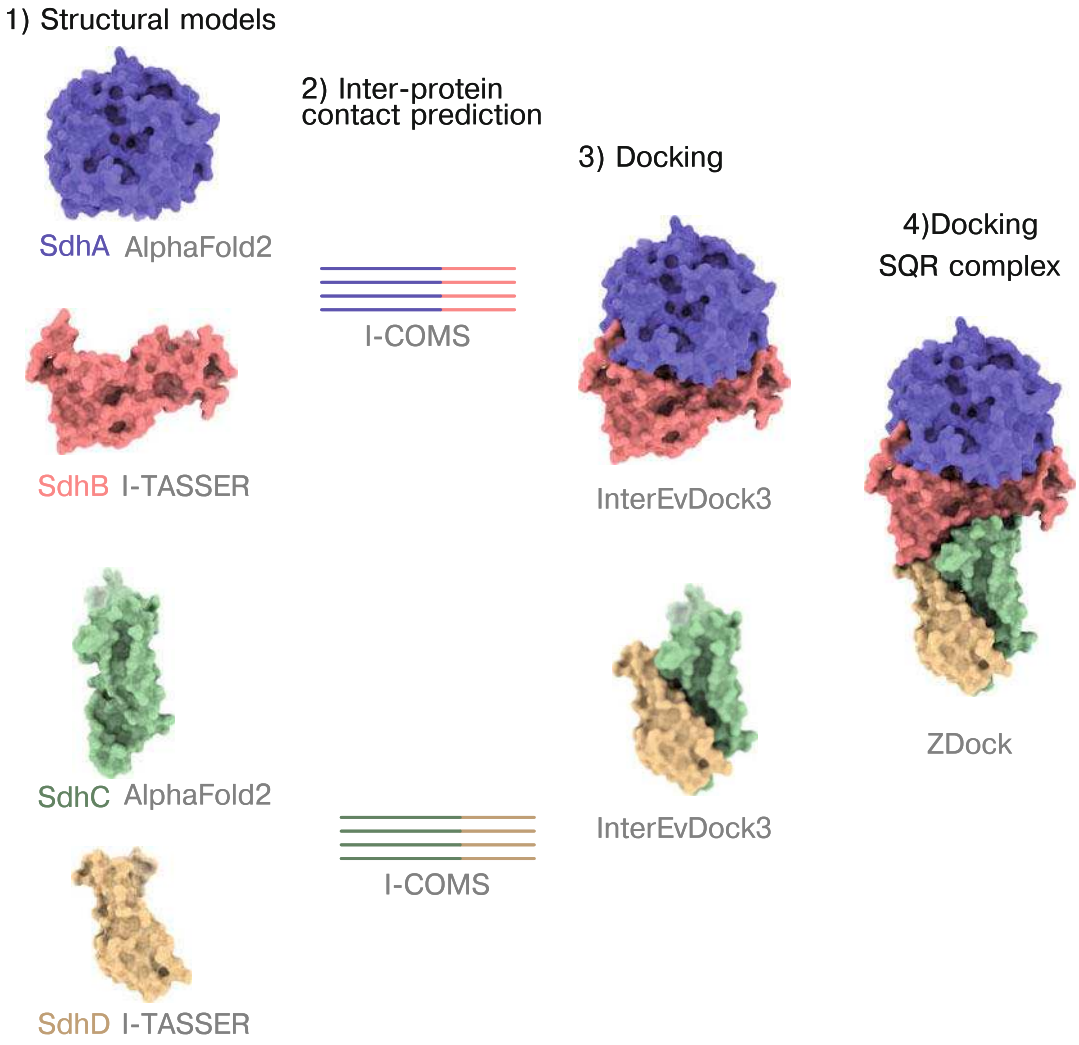

Fig. 2 Strategy to model the heterocomplex succinate-quinone oxidoreductase (SQR). The complex model was built as follows. First, we shall build a structural model for all subunits (SdhA, SdhB, SdhC, and SdhD) using either the AlphaFold2 method without template or I-TASSER without using close templates. Second, we will use inter-protein coevolution detection to predict residue contacts between the subunits. The dataset for coevolution comes from the available data reported in reference [32] and is provided in supplementary information (SI1). The inter-protein contact prediction was carried out using I-COMS. Third, the two subunits were docked using InterEvDock3 with coevolution information, and in the fourth step a docking was carried out between the dimers using ZDOCK without coevolution information

To analyze the results, we will visualize two parameters: (a) the number of sequences and gaps for contact prediction (Fig. 3b) and (b) the AlphaFold per-residue confidence score (pLDDT) that is found in the B-factor fields of the coordinate files (Fig. 3c). Both sequence information and pLDDT score per residue provided on average a good confidence about the quality of 3D models

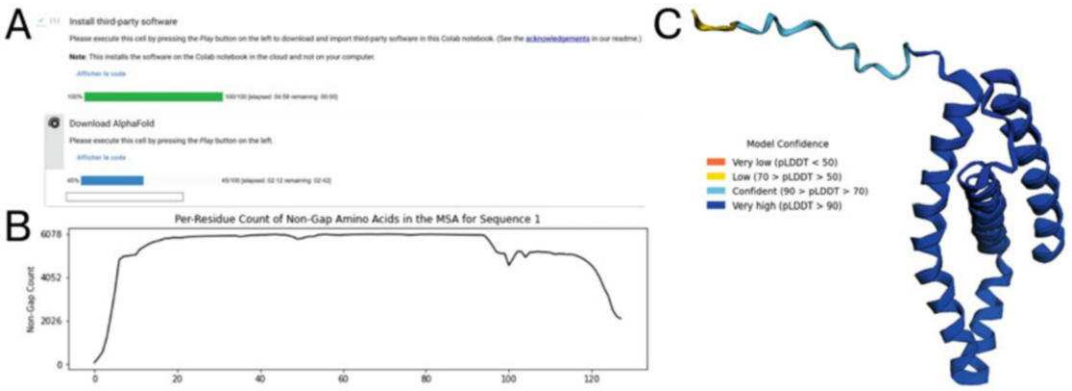

Fig. 3 Building a 3D model of SdhC from *E. coli* using AlphaFold2. Following the ColabFold notebook running process (**a**). Coverage of the multiple sequence alignment used by AlphaFold2 (**b**). Structural model colored by pLDDT (**c**). The AlphaFold2 method predicts a bundle of transmembrane helices and a disordered/coil region in N-term. In this latter, a low confidence is determined due to the lack of information in this region (N-term region in B)

(SdhA and SdhC). To confirm this result, both 3D models were compared with the corresponding X-ray structures (PDB code: 1NEK chain A and C). Using TM-align server (https://zhanggroup.org/TM-align/), structural alignments between models and solved structures gave RMSD values of 0.73 Å and 1.33 Å for SdhA and SdhC, respectively. It is worth noting that these RMSD values correspond to aligned residues; thus these latter can increase when considering the whole structure as the loop/coil/disordered regions highlighted in Fig. 4.

5.2 Building a 3D Model Using I-TASSER: SQR Subunits, SdhB, and SdhD

To avoid installation and set up programs on your computer, we will use the widely used I-TASSER webserver to build the 3D models of SdhB and SdhD.

1. Register yourself (https://zhanggroup.org/I-TASSER/registration.html).
2. Download the amino acid sequence of the target in FASTA format from UniProt (UniProt ID: P07014 and P0AC44 for SdhB and SdhD, respectively).
3. Go to I-TASSER webserver (https://zhanggroup.org/I-TASSER/).
4. Paste the protein sequence in FASTA format in the text box (Fig. 5a).
5. Type 60% to exclude homologous templates in the Option II section.
6. Identify you with email and password.
7. Click on the "Run I-TASSER" box.
8. Wait for results sent by email.

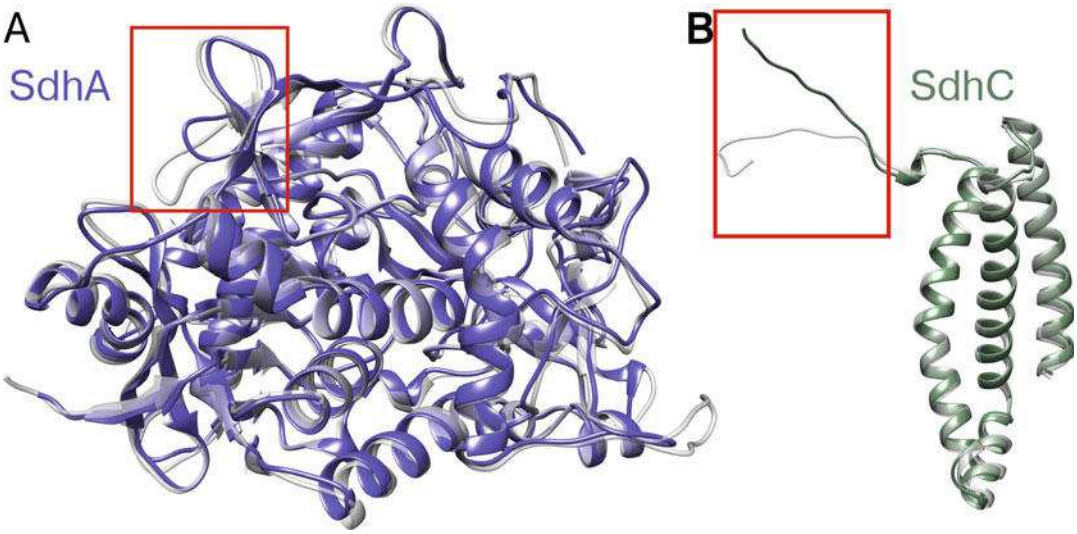

Fig. 4 Structural comparison between X-ray structure (1nek) and 3D models from AlphaFold2. SdhA (**a**) and SdhC (**b**) structures are shown in cartoon and colored as in Fig. 2. X-ray structures are displayed in transparent gray cartoon representation. Red squares highlight the main regions where Alphafold2 differs from the X-ray structure

To analyze the results, we will visualize two parameters: (a) the threading templates used by I-TASSER and the alignment quality against the target sequence (Norm Z-score) (Fig. 5b) and (b) the I-TASSER score (c-score) that gives the confidence of each model based on the significance of threading template alignments and the convergence parameters of the structure assembly simulations (Fig. 5c). This score is comprised between −5 and 2, with higher values (close to 2) indicating a higher confidence on the 3D model and vice-versa. Both templates and C-score (1.23 and 0.53 for SdhB and SdhD, respectively) provided good confidence about the quality of 3D models. Indeed, the best C-score was obtained using the templates chain B and C from 1YQ3 for SdhB and SdhD, respectively. Even if the sequences from 1YQ3 share ~50% and 20% of identity with SdhB and SdhD, respectively, the selected template corresponds to the same functional complex from another organism (*Gallus gallus*). To confirm this result, both 3D models were compared with the corresponding X-ray structures (PDB code: 1NEK chain B and D). Using the TM-align server, structural alignments between models and solved structures gave RMSD values of 2.01 Å and 2.19 Å for SdhB and SdhD, respectively.

5.3 Modeling SdhA-SdhB and SdhC-SdhD Using Protein-Protein Docking and Coevolution Information

Among the six possible protein pairs composing the heterotetramer, we focused on the prediction of SdhA-SdhB and SdhC-SdhD, the first pair corresponding to the cytosolic subunits and the second one to the membrane domains. We will use inter-protein coevolution to predict contacts between these two subunit pairs using I-COMS server. The input will be the alignments taken from a

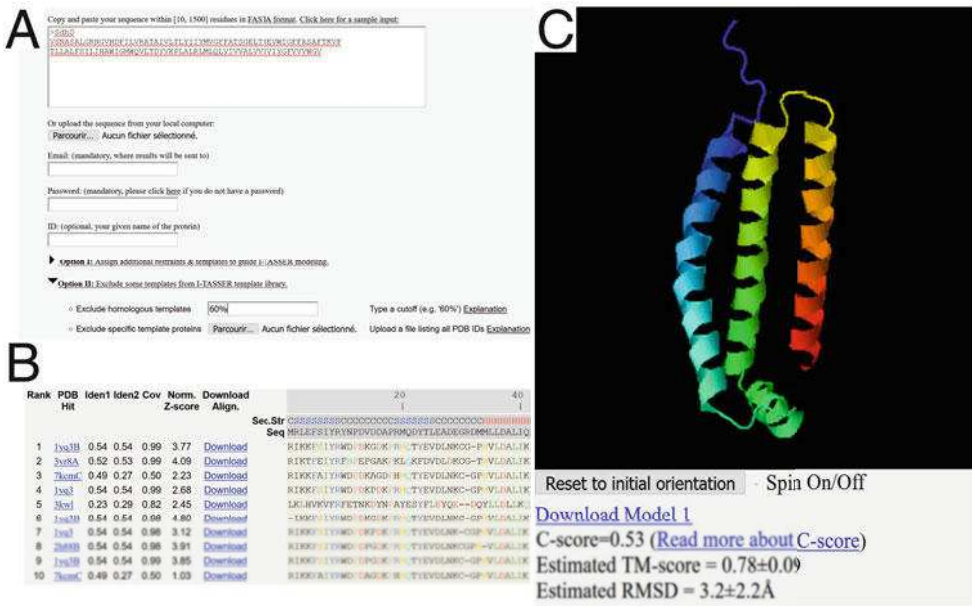

Fig. 5 Building a 3D model of SdhD from *E. coli* using I-TASSER. Following the submission process described between steps 4 and 7 (**a**). Top ten of threading templates (**b**). Best 3D model out of the top five final models (**c**)

previously published and publicly available dataset and provided in supplementary information (SI1).

1. Download the alignments from SI1.
2. Go to the I-COMS server (http://i-coms.leloir.org.ar/index.php).
3. Select the option "Upload your own alignments."
4. Optionally, you can describe the uploaded dataset.
5. Upload the two alignments using the "Browse. . ." button.
6. Click on "Upload and submit."
7. Choose the method for coevolution: plmDCA.
8. Optionally you can indicate the job description and your email address.

Results include information about the alignment used, such as the number of sequences and clusters. If the number of clusters is low (<400), it means that there is little diversity in the MSA and the results should be interpreted with caution. Results are shown in a circos representation of the covariation scores of each of the selected methods, and protein pairs are displayed (Fig. 6).

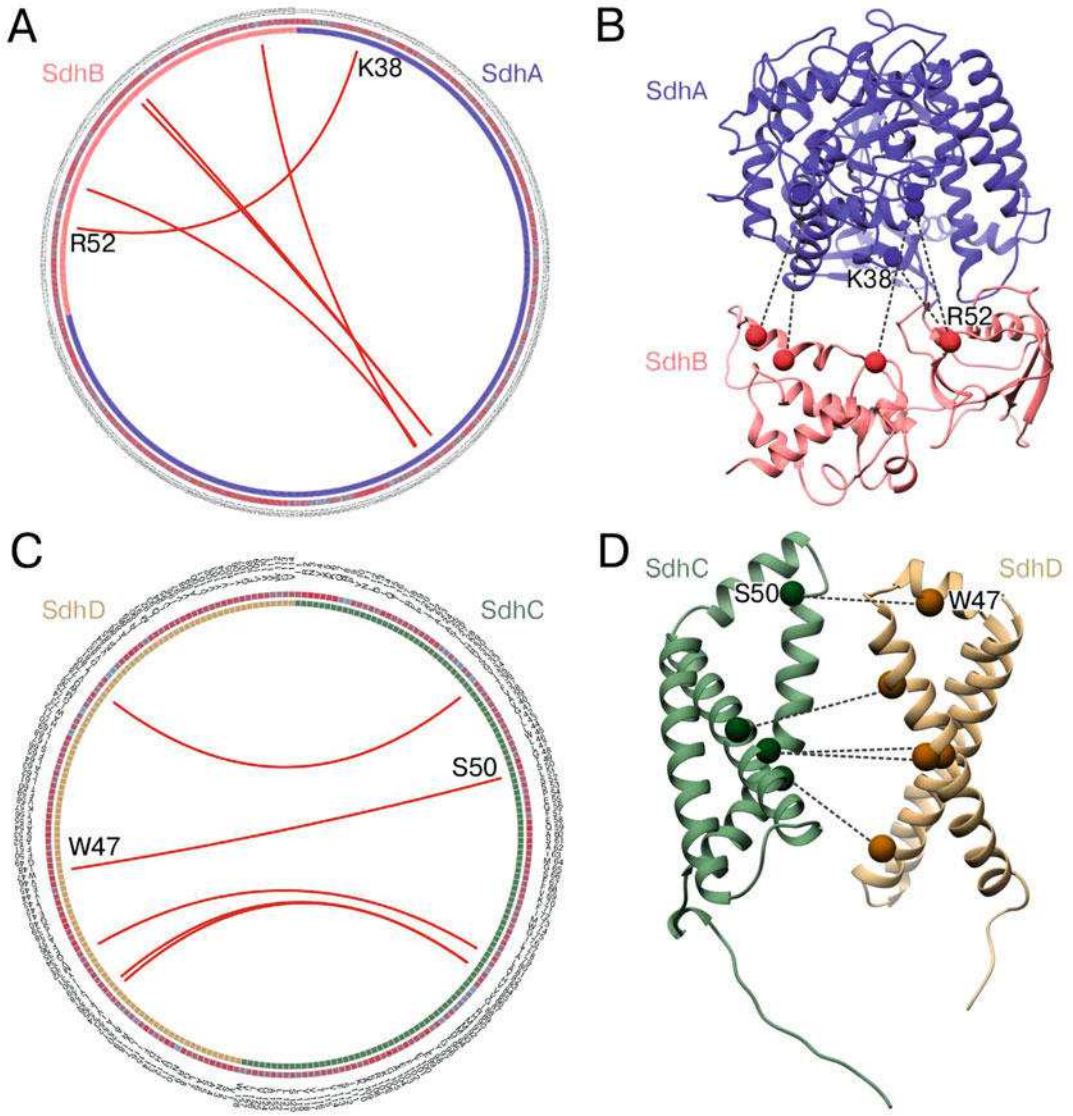

Fig. 6 Docking of subunits using coevolution information. Top five inter-protein coevolution results from I-COMS server. The inner circle represents the sequence positions in boxes colored according to the sequence they belong to (SdhA or SdhB). The correlated mutation scores are represented as lines between positions in the center of the circle. Given as example, the coevolving positions K38 and R52 from SdhA and SdhB, respectively, are indicated (**a**). Top five inter-protein coevolving positions are shown in the modeled subunits; the Cα of coevolving positions are shown in sphere representation (**b**). Analogous results are given for subunits SdhC and SdhD, the top five coevolution results (**c**) and the same coevolving pairs mapped on the models (**d**)

To visualize the inter-protein results:

1. Choose the pair of proteins (SdhA vs SdhB) or (SdhC vs SdhD).
2. Select the method.
3. Click on "Draw Circos."

4. Click on “Inter-protein” links.
5. You can select the number of edges to visualize.
6. Download covariation raw data, it will be used in the next steps.

Protein docking of SdhA-SdhB and SdhC-SdhD will be performed using InterEvdock3 server and residue contact predictions from I-COMS as described previously. The inputs will be the pdb files of the two partners to dock and a list of residue pair contacts.

7. Go to InterEvdock3 server (https://mobyle.rpbs.univ-paris-diderot.fr/cgi-bin/portal.py#forms::InterEvDock3).
8. Upload Partner A and Click on “Browse...” to browse and select pdb file.
9. Upload Partner B and Click on “Browse...” to browse and select pdb file.
10. Click on “Advanced Options.”
11. Go to “Use of co-evolution or deep-learning maps.”
12. Upload the coevolution map (Top 100) from I-COMS given in SI2.
13. Select “Yes” in “Minimize the output models using gromacs.”
14. Click on *run*.

InterEvdock3 web portal enables to follow the job progress at any time without any specific link. The https link associated with the job can be stored locally for caution.

Main InterEvdock3 output provides two kinds of rankings limited to the top ten poses: (a) based on the number of structural contacts matching the predicted coevolution pairs and (b) based on the scoring function related to the sum of the best predicted coevolution pairs.

In this study, the best docking poses for both heterodimers are selected from the second type of ranking, which leads to favor the most probable pairs related to their coevolution score. The resulting models of the heterodimers are provided in SI3.

5.4 Modeling the Succinate-Quinone Oxidoreductase Heterocomplex Using Protein-Protein Docking and Restraints

As there is not enough information when merging concatenated MSA from SdhA, SdhB, SdhC, and SdhD, coevolution cannot be used to predict residue contacts. Therefore, the docking between the predicted partners will be done using “classical” docking. Free docking and docking with restraints will be performed using ZDOCK server. To avoid clashes and improve docking prediction, N-term disordered regions for SdhC and SdhD are removed, corresponding to the first 13 residues and the 10 residues, respectively.

1. Go to https://zdock.umassmed.edu/.
2. Choose "PDB file" in the scrolling list close to "Input Protein 1" keyword.
3. Click on "Browse ..." to select PDB file corresponding to SdhA-SdhB.
4. Repeat steps 2 and 3 for *Input Protein 2*.
5. Fill up the form "Enter your email."
6. Optionally, for free docking, check the box close to *Skip residue selection*.
7. Click on "Submit" button.
8. If *Skip residue selection* was not checked, select interactively the residues belonging to the binding site for guiding docking.
9. Click on "Submit" button.
10. Wait for results sent by email.
11. Download top ten predictions.
12. Select the first docking poses.

This particular case seems to be difficult for good docking prediction. Indeed, free docking does not provide a good solution compared to the X-ray structure. To get a correct assembly, a list of 17 and 19 residues from SdhB and SdhC-SdhD (given in SI2) had to be provided to guide the docking. The binding residues at the interface can be selected on distance threshold criteria, 3.2 Å on heavy atoms from X-ray structure in this work. Having this kind of information helps to have better predictions as shown in Fig. 7. Superposition of the modeled heterotetramer onto the X-ray structure (PDB code: 1NEK) showed an RMSD of ~0.73 Å based on TM-align server, indicating a very good fit. The coordinate file of the final model is provided in SI3.

6 Conclusions

Overall, this study shows that protein complex prediction is not a trivial question. The first crucial work is to obtain the structure of each protein partner. According to the available data, different approaches can be applied with a new methodology outperforming the others, called Alphafold2. Part of the success in the assembly construction will first depend on the quality of the 3D structural model of each partner. Therefore, assessment such as pLDDT is an important step at this stage. Then, protein-protein interactions can be predicted with reasonable confidence when diverse information, such as coevolution prediction or experimental results, is available to guide toward the most probable assembly. In this study, both cases are exemplified. Two heterodimers were quite well predicted

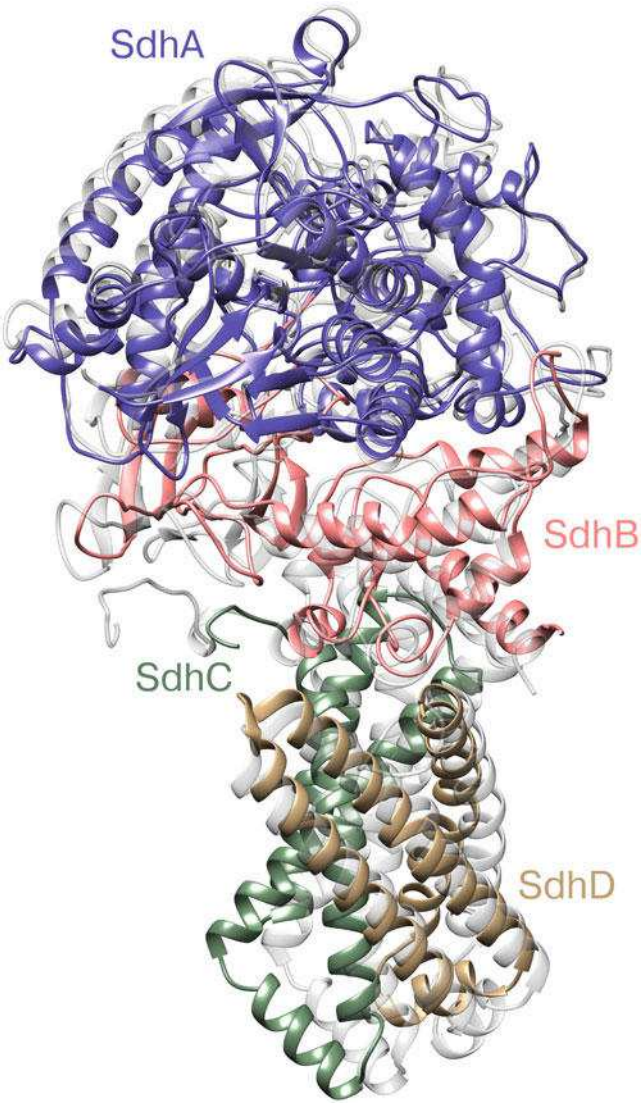

Fig. 7 Superposition of the modeled SDQ heterotetramer onto the reference structure. Each modeled subunits SdhA, SdhB, SdhC, and SdhD is shown in cartoon representation and is colored according to the corresponding label. The heterotetramer is obtained from the docking of the two main units SdhA-SdhB and SdhC-SdhD. The reference corresponds to the X-ray structure (PDB code: 1NEK), which is shown in white cartoon representation for clarity

using coevolution information thanks to the diversity of the data. However, construction of the heterotetramer assembly was quite challenging because the interactions with the membrane are not taken into account in the docking procedure. To circumvent this limitation, a set of amino acid residues from the protein interface identified from experimental data was used to guide the construction of the heterotetramer assembly.

References

1. Pieters BJGE, van Eldijk MB, Nolte RJM, Mecinović J (2016) Natural supramolecular protein assemblies. Chem Soc Rev 45:24–39
2. Berggård T, Linse S, James P (2007) Methods for the detection and analysis of protein-protein interactions. Proteomics 7:2833–2842
3. Soni N, Madhusudhan MS (2017) Computational modeling of protein assemblies. Curr Opin Struct Biol 44:179–189
4. Sweetlove LJ, Fernie AR (2018) The role of dynamic enzyme assemblies and substrate channelling in metabolic regulation. Nat Commun 9:2136
5. Chiesa G, Kiriakov S, Khalil AS (2020) Protein assembly systems in natural and synthetic biology. BMC Biol 18:35
6. Zhang Y, Fernie AR (2021) Stable and temporary enzyme complexes and metabolons involved in energy and redox metabolism. Antioxid Redox Signal 35:788–807
7. Rao VS, Srinivas K, Sujini GN, Kumar GNS (2014) Protein-protein interaction detection: methods and analysis. Int J Proteomics 2014: 147648
8. Wu F, Minteer S (2015) Krebs cycle metabolon: structural evidence of substrate channeling revealed by cross-linking and mass spectrometry. Angew Chem Int Ed Engl 54:1851–1854
9. Yang J, Anishchenko I, Park H, Peng Z, Ovchinnikov S, Baker D (2020) Improved protein structure prediction using predicted interresidue orientations. Proc Natl Acad Sci U S A 117:1496–1503
10. Koonin EV, Wolf YI, Karev GP (2002) The structure of the protein universe and genome evolution. Nature 420:218–223
11. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O et al (2021) Highly accurate protein structure prediction with AlphaFold. Nature. https://doi.org/10.1038/s41586-021-03819-2
12. Sander C, Schneider R (1991) Database of homology-derived protein structures and the structural meaning of sequence alignment. Proteins 9:56–68
13. Chung SY, Subbiah S (1996) A structural explanation for the twilight zone of protein sequence homology. Structure 4:1123–1127
14. Heilmann N, Wolf M, Kozlowska M, Sedghamiz E, Setzler J, Brieg M et al (2020) Sampling of the conformational landscape of small proteins with Monte Carlo methods. Sci Rep 10:18211
15. Geng H, Chen F, Ye J, Jiang F (2019) Applications of molecular dynamics simulation in structure prediction of peptides and proteins. Comput Struct Biotechnol J 17:1162–1170
16. Sali A, Blundell TL (1993) Comparative protein modelling by satisfaction of spatial restraints. J Mol Biol 234:779–815
17. Pieper U, Webb BM, Dong GQ, Schneidman-Duhovny D, Fan H, Kim SJ et al (2014) ModBase, a database of annotated comparative protein structure models and associated resources. Nucleic Acids Res 42:D336–D346
18. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W et al (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389–3402
19. Kelley LA, Mezulis S, Yates CM, Wass MN, Sternberg MJE (2015) The Phyre2 web portal for protein modeling, prediction and analysis. Nat Protoc 10:845–858
20. Roy A, Kucukural A, Zhang Y (2010) I-TASSER: a unified platform for automated protein structure and function prediction. Nat Protoc 5:725–738
21. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H et al (2000) The protein data bank. Nucleic Acids Res 28: 235–242
22. Hornak V, Abel R, Okur A, Strockbine B, Roitberg A, Simmerling C (2006) Comparison of multiple Amber force fields and development of improved protein backbone parameters. Proteins 65:712–725
23. Eastman P, Swails J, Chodera JD, McGibbon RT, Zhao Y, Beauchamp KA et al (2017) OpenMM 7: rapid development of high performance algorithms for molecular dynamics. PLoS Comput Biol 13:e1005659
24. Pereira J, Simpkin AJ, Hartmann MD, Rigden DJ, Keegan RM, Lupas AN (2021) High-accuracy protein structure prediction in CASP14. Proteins. https://doi.org/10.1002/prot.26171
25. Mirdita M, Ovchinnikov S, Steinegger M (2021) ColabFold - Making protein folding accessible to all bioRxiv. p. 2021.08.15.456425. https://doi.org/10.1101/2021.08.15.456425
26. Tunyasuvunakool K, Adler J, Wu Z, Green T, Zielinski M, Žídek A et al (2021) Highly accurate protein structure prediction for the human proteome. Nature. https://doi.org/10.1038/s41586-021-03828-1

27. Lovell SC, Robertson DL (2010) An integrated view of molecular coevolution in protein--protein interactions. Mol Biol Evol 27:2567–2575
28. Atchley WR, Wollenberg KR, Fitch WM, Terhalle W, Dress AW (2000) Correlations among amino acid sites in bHLH protein domains: an information theoretic analysis. Mol Biol Evol 17:164–178
29. Marks DS, Colwell LJ, Sheridan R, Hopf TA, Pagnani A, Zecchina R et al (2011) Protein 3D structure computed from evolutionary sequence variation. PLoS One 6:e28766
30. Hopf TA, Schärfe CPI, Rodrigues JPGLM, Green AG, Kohlbacher O, Sander C et al (2014) Sequence co-evolution gives 3D contacts and structures of protein complexes. elife 3. https://doi.org/10.7554/eLife.03430
31. Clark GW, Dar V-U-N, Bezginov A, Yang JM, Charlebois RL, Tillier ERM (2011) Using coevolution to predict protein-protein interactions. Methods Mol Biol 781:237–256
32. Green AG, Elhabashy H, Brock KP, Maddamsetti R, Kohlbacher O, Marks DS (2021) Large-scale discovery of protein interactions at residue resolution using co-evolution calculated from genomic sequences. Nat Commun 12:1396
33. Cong Q, Anishchenko I, Ovchinnikov S, Baker D (2019) Protein interaction networks revealed by proteome coevolution. Science 365:185–189
34. Iserte J, Simonetti FL, Zea DJ, Teppa E, Marino-Buslje C (2015) I-COMS: Interprotein-COrrelated mutations server. Nucleic Acids Res 43:W320–W325
35. Chen R, Li L, Weng Z (2003) ZDOCK: an initial-stage protein-docking algorithm. Proteins 52:80–87
36. Kozakov D, Hall DR, Xia B, Porter KA, Padhorny D, Yueh C et al (2017) The ClusPro web server for protein-protein docking. Nat Protoc 12:255–278
37. Ritchie DW, Kozakov D, Vajda S (2008) Accelerating and focusing protein-protein docking correlations using multi-dimensional rotational FFT generating functions. Bioinformatics 24: 1865–1873
38. Ritchie DW, Kemp GJ (2000) Protein docking using spherical polar Fourier correlations. Proteins 39:178–194
39. Garzon JI, Lopéz-Blanco JR, Pons C, Kovacs J, Abagyan R, Fernandez-Recio J et al (2009) FRODOCK: a new approach for fast rotational protein–protein docking. Bioinformatics 25: 2544–2551
40. Christoffer C, Chen S, Bharadwaj V, Aderinwale T, Kumar V, Hormati M et al (2021) LZerD webserver for pairwise and multiple protein–protein docking. Nucleic Acids Res 49:W359–W365
41. Dominguez C, Boelens R, Bonvin AMJ (2003) HADDOCK: a protein−protein docking approach based on biochemical or biophysical information. J Am Chem Soc:1731–1737. https://doi.org/10.1021/ja026939x
42. Cheng TM-K, Blundell TL, Fernandez-Recio J (2007) pyDock: electrostatics and desolvation for effective scoring of rigid-body protein-protein docking. Proteins 68:503–515
43. Pierce B, Weng Z (2007) ZRANK: reranking protein docking predictions with an optimized energy function. Proteins 67:1078–1086
44. Pierce B, Weng Z (2008) A combination of rescoring and refinement significantly improves protein docking performance. Proteins 72: 270–279
45. Moal IH, Bates PA (2010) SwarmDock and the use of normal modes in protein-protein docking. Int J Mol Sci 11:3623–3648
46. Ritchie DW, Venkatraman V (2010) Ultra-fast FFT protein docking on graphics processors. Bioinformatics 26:2398–2405
47. Ohue M, Shimoda T, Suzuki S, Matsuzaki Y, Ishida T, Akiyama Y (2014) MEGADOCK 4.0: an ultra-high-performance protein-protein docking software for heterogeneous supercomputers. Bioinformatics 30:3281–3283
48. Lu H, Lu L, Skolnick J (2003) Development of unified statistical potentials describing protein-protein interactions. Biophys J 84:1895–1901
49. Huang S-Y, Zou X (2008) An iterative knowledge-based scoring function for protein-protein recognition. Proteins 72:557–579
50. Mezei M (2017) Rescore protein-protein docked ensembles with an interface contact statistics. Proteins 85:235–241
51. Khashan R, Zheng W, Tropsha A (2012) Scoring protein interaction decoys using exposed residues (SPIDER): a novel multibody interaction scoring function based on frequent geometric patterns of interfacial residues. Proteins 80:2207–2217
52. Kozakov D, Brenke R, Comeau SR, Vajda S (2006) PIPER: an FFT-based protein docking program with pairwise potentials. Proteins 65: 392–406
53. Liang S, Meroueh SO, Wang G, Qiu C, Zhou Y (2009) Consensus scoring for enriching near-native structures from protein-protein docking decoys. Proteins 75:397–403

54. Feliu E, Aloy P, Oliva B (2011) On the analysis of protein-protein interactions via knowledge-based potentials for the prediction of protein-protein docking. Protein Sci 20:529–541
55. Vreven T, Hwang H, Weng Z (2011) Integrating atom-based and residue-based scoring functions for protein-protein docking. Protein Sci 20:1576–1586
56. Andreani J, Faure G, Guerois R (2013) InterEvScore: a novel coarse-grained interface scoring function using a multi-body statistical potential coupled to evolution. Bioinformatics 29:1742–1749
57. Yu J, Andreani J, Ochsenbein F, Guerois R (2017) Lessons from (co-)evolution in the docking of proteins and peptides for CAPRI Rounds 28-35. Proteins 85:378–390
58. Oliva R, Vangone A, Cavallo L (2013) Ranking multiple docking solutions based on the conservation of inter-residue contacts. Proteins 81: 1571–1584
59. Oliva R, Chermak E, Cavallo L (2015) Analysis and ranking of protein-protein docking models using inter-residue contacts and inter-molecular contact maps. Molecules 20: 12045–12060
60. Vangone A, Cavallo L, Oliva R (2013) Using a consensus approach based on the conservation of inter-residue contacts to rank CAPRI models. Proteins 81:2210–2220
61. Chermak E, Petta A, Serra L, Vangone A, Scarano V, Cavallo L et al (2015) CONSRANK: a server for the analysis, comparison and ranking of docking models based on inter-residue contacts. Bioinformatics 31:1481–1483
62. Chermak E, De Donato R, Lensink MF, Petta A, Serra L, Scarano V et al (2016) Introducing a clustering step in a consensus approach for the scoring of protein-protein docking models. PLoS One 11:e0166460
63. Lensink MF, Méndez R, Wodak SJ (2007) Docking and scoring protein complexes: CAPRI 3rd edition. Proteins 69:704–718
64. Lensink MF, Velankar S, Wodak SJ (2017) Modeling protein-protein and protein-peptide complexes: CAPRI 6th edition. Proteins 85: 359–377
65. Lensink MF, Wodak SJ (2010) Docking and scoring protein interactions: CAPRI 2009. Proteins 78:3073–3084
66. Pierce BG, Wiehe K, Hwang H, Kim B-H, Vreven T, Weng Z (2014) ZDOCK server: interactive docking prediction of protein-protein complexes and symmetric multimers. Bioinformatics 30:1771–1773
67. Quignot C, Postic G, Bret H, Rey J, Granger P, Murail S et al (2021) InterEvDock3: a combined template-based and free docking server with increased performance through explicit modeling of complex homologs and integration of covariation-based contact maps. Nucleic Acids Res 49:W277–W284
68. Karami Y, Guyon F, De Vries S, Tufféry P (2018) DaReUS-Loop: accurate loop modeling using fragments from remote or unrelated proteins. Sci Rep 8:13673
69. Horsefield R, Iwata S, Byrne B (2004) Complex II from a structural perspective. Curr Protein Pept Sci 5:107–118
70. Yankovskaya V, Horsefield R, Törnroth S, Luna-Chavez C, Miyoshi H, Léger C et al (2003) Architecture of succinate dehydrogenase and reactive oxygen species generation. Science 299:700–704
71. Ruprecht J, Yankovskaya V, Maklashina E, Iwata S, Cecchini G (2009) Structure of *Escherichia coli* succinate:quinone oxidoreductase with an occupied and empty quinone-binding site. J Biol Chem 284:29836–29846