bioRxiv preprint doi: https://doi.org/10.1101/004762; this version posted May 6, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Sequence co-evolution gives 3D contacts and structures of protein complexes

Thomas A. Hopf[1,2]*, Charlotta P.I. Schärfe[1,3]*, João P.G.L.M. Rodrigues[4]*, Anna G. Green[1], Chris Sander[5#], Alexandre M.J.J. Bonvin[4#], Debora S. Marks[1#]

[1] Department of Systems Biology, Harvard University, Boston, Massachusetts, USA
[2] Department for Bioinformatics and Computational Biology, Technische Universität München, Garching, Germany
[3] Applied Bioinformatics, Center for Bioinformatics, Quantitative Biology Center and Department of Compute Science, University of Tübingen, Germany
[4] Computational Structural Biology Group, Bijvoet Center for Biomolecular Research, Utrecht University, The Netherlands
[5] Computational Biology Center, Memorial Sloan-Kettering Cancer Center, New York, New York, USA

* Contributed equally to this work
# Corresponding authors email: EVcomplex@gmail.com

Abstract

High-throughput experiments in bacteria and eukaryotic cells have identified tens of thousands of possible interactions between proteins. This genome-wide view of the protein interaction universe is coarse-grained, whilst fine-grained detail of macro-molecular interactions critically depends on lower throughput, labor-intensive experiments. Computational approaches using measures of residue co-evolution across proteins show promise, but have been limited to specific interactions. Here we present a new generalized method showing that patterns of evolutionary sequence changes across proteins reflect residues that are close in space, and with sufficient accuracy to determine the three-dimensional structure of the protein complexes. We demonstrate that the inferred evolutionary coupling scores distinguish between interacting and non-interacting proteins and the accurate prediction of residue interactions. To illustrate the utility of the method, we predict unknown 3D interactions between subunits of ATP synthase and find results consistent with detailed experimental data. We expect that the method can be generalized to genome-wide interaction predictions at residue resolution.

bioRxiv preprint doi: https://doi.org/10.1101/004762; this version posted May 6, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Introduction

A large body of biological research is concerned with the identity, dynamics and specificity of protein interactions. There have also been impressive advances in the three-dimensional (3D) structure determination of protein complexes at residue resolution, which is significantly extended by homology-inferred 3D information[1,2,3,4]. However, there is little or no 3D information for ~80% of known protein interactions in bacteria, yeast and humans, amounting to ~30,000/~6000 uncharacterized interactions[2] (human and *E. coli,* respectively) and experimental methods alone cannot match the rapid increase in the demand for residue–level information of these interactions.

One way to address the knowledge gap has been the use of hybrid, computational-experimental approaches that typically combine 3D structural information at varying resolutions, homology models and other methods[5], with biophysical force fields such as Rosetta Dock, residue cross-linking and data-driven approaches that incorporate various sources of biological information[1,6-15]. However, most of these approaches depend on the availability of prior knowledge and many biologically relevant systems remain out of reach, as additional experimental information is sparse (e.g., membrane proteins, transient interactions, large complexes).

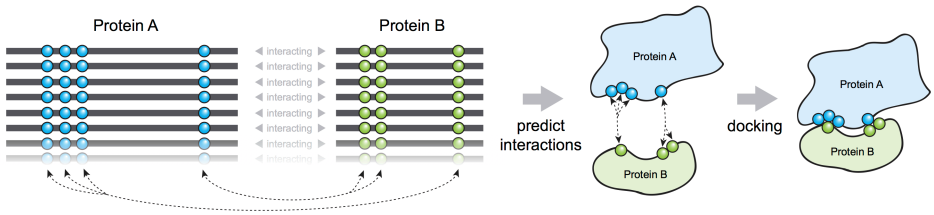

Figure 1. Co-evolution of residues across protein complexes from the evolutionary sequence record. Evolutionary pressure to maintain protein-protein interactions leads to the coevolution of residues between interacting proteins in a complex. By analyzing patterns of amino acid co-variation in an alignment of putatively interacting homologous proteins (left), evolutionary couplings between coevolving *inter*-protein residue pairs can be identified (middle). By defining distance restraints on these pairs, docking software can be used to calculate the 3D structure of the complex from its monomer constituents (right).

Evolutionary analysis of amino acid co-variation was used 20 years ago to identify close residue contacts across protein interactions[16,17] A more advanced method was applied specifically in the case of the interaction between histidine kinases and response regulators[18-

bioRxiv preprint doi: https://doi.org/10.1101/004762; this version posted May 6, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

[20], but this approach has yet to be generalized and used to predict contacts between proteins in complexes of unknown structure. In principle, just a small number of key reside-residue contacts across a protein interface would allow fast computation for 3D models and provide a powerful, orthogonal approach to experiments. Since the recent and successful demonstration of the use of inter-residue evolutionary couplings (ECs) to determine the 3D structure of individual proteins[21-25], including integral membrane proteins[26,27], we reasoned that such an evolutionary statistical approach (such as EVcouplings[21], web site: http://evfold.org) could be used to determine co-evolved residues between proteins. Here we calculate co-evolved residues across protein interactions, and find that the majority of top ranked *inter*-protein EC pairs (inter-ECs) are close in known 3D structures of the complexes, and that these constraints are sufficient to calculate accurate 3D complexes (using HADDOCK software[13]), (Fig. 1). The method, called EVcomplex, also provides a measure for the confidence of interaction by the relative strength of the inter-ECs versus the *intra*-protein ECs (intra-ECs). This approach allowed us to predict the currently *unknown* 3D interaction between the a and b subunits of the E. *coli* ATP synthase complex, to assess its consistency with published cross-linking experiments, and to highlight residues and interactions that may be crucial for ATP synthase function.

Results

Methodological approach

We investigated whether co-evolving residues between proteins are close in three dimensions and can be used to calculate unknown 3D structures of complexes by assessing blinded predictions of experimentally determined 3D complex structures.

Protein complexes were chosen based on the availability of sufficient sequences, (roughly L sequences after removal of redundancy, where L is the length of the protein sequence) and the ability to match pairs of interacting proteins (Methods). The current algorithm uses protein partners encoded close in the bacterial genomes, or requires the single occurrence of the pair of interacting proteins in each genome (Table 1, Supplementary Table 1). We estimate that ~40 of the roughly 3000 non-redundant protein interactions in *E. coli* [2] have

bioRxiv preprint doi: https://doi.org/10.1101/004762; this version posted May 6, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

enough sequences in other species (in the current UniProt database) to allow modeling with the current methods. We selected 10 diverse examples from this set of 40.

For each complex, we retrieved sequence alignments for each monomer and paired the maximum number of proteins from the two alignments in other species using a genome proximity algorithm, in the same spirit as earlier operon-based methods [18]. The paired sequences are concatenated for the EC analysis using EVcouplings[21,22,24] that implements a pseudolikelihood maximization (PLM) approximation to determine the interaction parameters in the maximum entropy equation[25,28] simultaneously generating both intra- and inter-EC scores for all pairs of residues within and across the protein pairs (Fig. 1). After the EC scores have been calculated for all intra- and inter-residue pairs, we use the strength of the top inter-EC to predict the likelihood of accurate inter-ECs. This approach is consistent with the observation for hundreds of single proteins that the lower strength ECs are gradually more likely to be false positives for EC ranks below L[21-23,26], where L is the length of the sequence. We use both genome distance and rank of the inter-ECs scores to estimate the likely accuracy of inter-ECs (Supplementary Fig. 1, Supplementary Data).

Benchmark overview

For the benchmark complexes, the majority of the top 5 ECs between proteins are correct to within 8Å (Table 1). In two cases there are inter-ECs in the ranked positions between 5 and 10, that are distant in the crystal structure (>10Å). In one case, the false positives may be due to large conformational changes of the complex (e.g. in the BtuCDF complex[29]) and another case, due to the low number of sequences in the concatenated alignment (<1 sequence per residue in ClpS-ClpA).

For all the complexes, can the ECs be used for generating high confidence 3D models? Docking with the top 5 or 10 high ranked EC pairs (using HADDOCK[13,30]) over 70% of the generated models were close to the crystal structures of the complexes (<4Å backbone iRMSD) (Fig. 2 and Supplementary Fig. 2).

bioRxiv preprint doi: https://doi.org/10.1101/004762; this version posted May 6, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

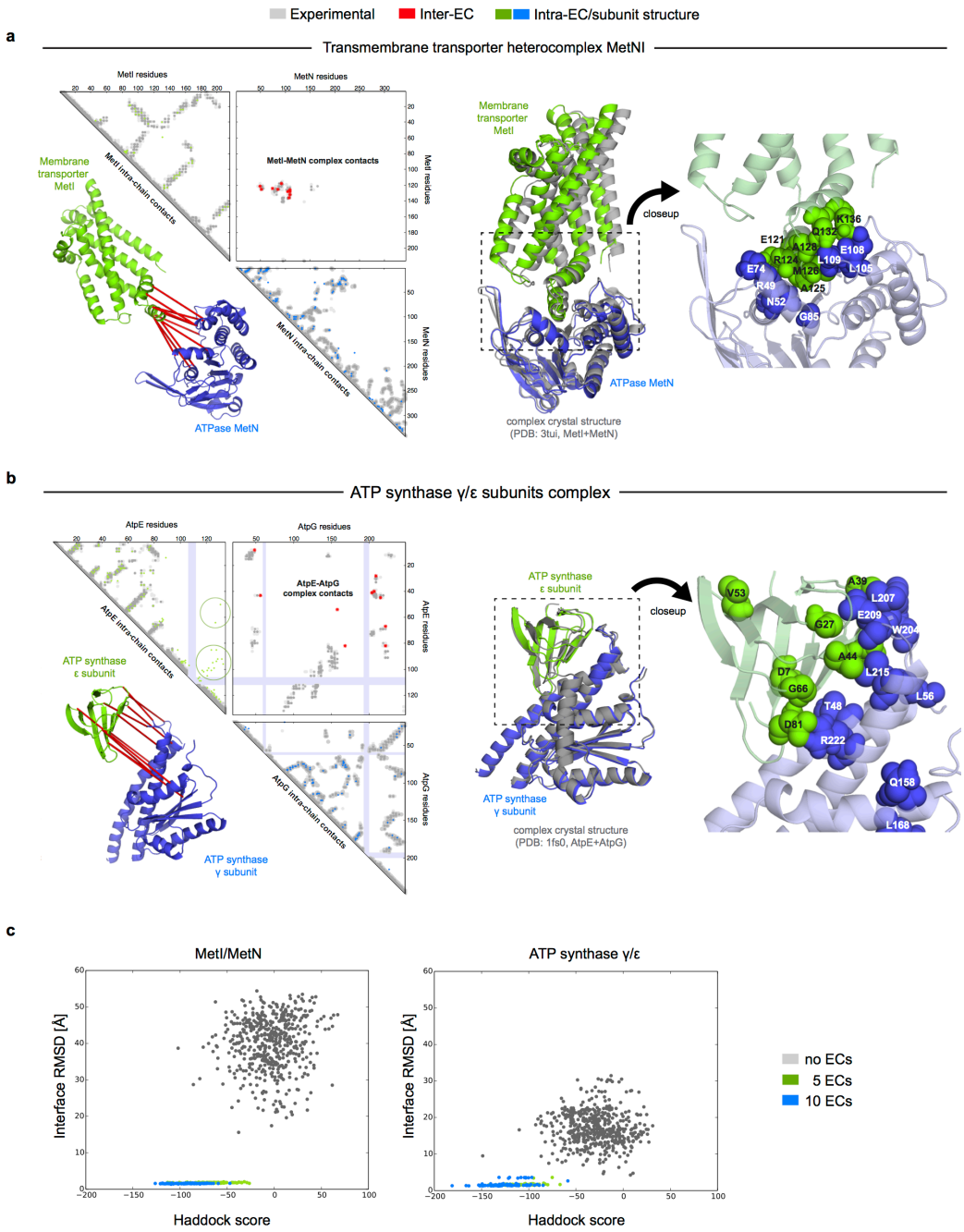

Figure 2. Accurate prediction of protein complex 3D structures using evolutionary couplings. EVcomplex predictions and comparison to crystal structure information for **(a)** the methionine-importing transmembrane transporter heterocomplex MetNI from E. *coli* (PDB: 3tui[31]) and **(b)** the gamma/epsilon subunit interaction of E. *coli* ATP synthase (PDB: 1fs0[32]). **(a,b)** Items shown for both protein complexes: *Left panel*: Complex contact map comparing predicted top ranked inter-ECs (red stars, upper right quadrant) and intra-ECs (to the occurrence of the 10th inter-EC; green and blue stars, top left and lower right triangles) to close pairs in the complex crystal (dark/mid/light grey points for minimum atom distance cutoffs of 5/6/7 Å; missing crystal data: shaded blue rectangles). The top 10 inter-ECs are also displayed on the spatially separated subunits of the complex (red lines on green and blue cartoons, lower left). *Right panel*: Superimposition of the top ranked model from 3D docking (green/blue cartoon, left) onto the complex crystal structure (grey cartoon), and close-up of the interface region with highly coupled residues (green/blue spheres, residues involved in top 10 inter-ECs). **(c)** HADDOCK scores of docked models versus their iRMSDs to the complex crystal structure (outliers with HADDOCK scores > 100 excluded from display).

bioRxiv preprint doi: https://doi.org/10.1101/004762; this version posted May 6, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

After calculating inter- and intra-ECs for 3 of the complexes (Table 1), we predict that the inter-ECs are likely to be incorrect. This prediction is based on the relative rank of the inter-ECs to high confidence intra-ECs (Supplementary Figs. 3, 4 & 5). Analysis of the scores suggests that relative ranks of <0.5 will clearly distinguish correct from incorrect predictions with unknown complexes (Table 1).

Table 1. Accuracy of complex residue interactions

Protein complex[a]	PDB ID	Effective # sequences/ length[c]	Relative inter EC rank[d]	False Positive ECs[e]	Top ranked iRMSD[f,g]	Best iRMSD[f,g]
ATP synthase γ and ε subunits[b]	1FS0	1.9	0.12	0 / 2	1.4 / 1.3	1.4 / 1.2
Vitamin B12 uptake system permease & ATP-binding domain	1L7V	4.4	0.02	1 / 1	0.8 / 0.8	0.7 / 0.8
Vitamin B12 uptake system SBP & permease	2QI9	3.8	0.17	1 / 5	2.2 /1.4	2.0 / 1.3
Methionine transporter complex	3TUI	1.1	0.02	0 / 0	1.7 / 1.5	1.6 / 1.4
Molybdopterin synthase	1FM0	2.0	0.05	0 / 0	5.2 / 2.9	5.0 / 2.7
Histidine kinase - response regulator complex[b]	3DGE	38.4	0.09	0 / 0	1.6 / 1.8	1.4 / 1.8
ClpAS chaperone-protease complex[b]	1R6Q	0.7	0.03	1 / 5	2.0 / 2.0	1.5 / 2.0
Ferrodoxin Reductase complex	1EWY	4.7	0.72	5 / 10	- / -	- / -
Thioredoxin reductase complex	1F6M	1.9	1.37	5 / 10	- / -	- / -
Imidazoleglycerol phosphate synthase Bienzyme complex	1GPW	1.4	0.60	5 / 10	- / -	- / -

a. UniProt IDs for complex pairs: ATPE_ECOLI/ATPG_ECOLI, BTUC_ECOLI/BTUD_ECOLI, BTUC_ECOLI/BTUF_ECOLI, METI_ECOLI/METN_ECOLI, MOAD_ECOLI/MOAE_ECOLI, Q9WZV7_THEMA/Q9WYT9_THEMA, CLPS_ECOLI/CLPA_ECOLI FENR_ANASO/FER1_ANASO, TRXB_ECOLI/THIO_ECOLI, HIS6_THEMA/HIS5_THEMA.
b. docked with unbound structures
c. reweighted number of sequences per residue of concatenated length
d. rank of first inter-EC normalized by concatenated alignment length
e. False Positive Contacts (dist > 8Å in Xtal) in top 5/top 10 ECs
f. iRMSD interface residues, see main paper for definition.
g. Top 5/top10 inter-protein ECs used for docking, all model IDs in Supplementary Table 2 and coordinates in Supplementary Data.

Lastly, we show that EVcomplex can predict the subunit-subunit interactions of the ATP synthase complex. Interactions between tested monomer subunits of ATP synthase, namely the a-, b-, c- and α-subunits, are correctly paired using a z-score of the inter-protein EC scores

bioRxiv preprint doi: https://doi.org/10.1101/004762; this version posted May 6, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

(Fig. 3a & b). Taken together, the benchmark complexes indicate that EVcomplex can be used to identify accurate inter-ECs, generate accurate 3D models, and distinguish interactions from non-interactions in complexes.

Inter-protein functional residue networks

All top 10 inter-EC pairs between MetI and MetN are accurate (<8Å in the MetNI complex (PDB:3tui[31]), resulting in an average of 1.6 Å iRMSD from the crystal structure (standard deviation 0.06, PDB: 3tui[31]) for all 100 computed 3D models (Supplementary Data). The top 3 inter-EC residue pairs (K136-E108, A128-L105, and E121-R49, MetI-MetN respectively) constitute a residue network coupling the ATP binding pocket of MetN to the membrane transporter MetI. This network calculated from the alignment corresponds to residues identified experimentally that couple ATP hydrolysis to the open and closed conformations of the MetI dimer[31] (Fig. 2a). The vitamin B12 transporter (BtuC) belongs to a different class of ABC transporters, but also uses ATP hydrolysis via an interacting ATPase (BtuD), and EVcomplex identifies sufficient inter-protein residue contacts to calculate accurate complex 3D structures (top ranked model is 0.8 Å backbone iRMSD compared to PDB: 1l7v[33], Table 1). The top 5 inter-ECs co-locate the L-loop of BtuC close to the Q-loop ATP-binding domain of the ATPase, hence coupling the transporter with the ATP hydrolysis state in an analogous way to MetI-MetN (Supplementary Data). The identification of these coupled residues across the different subunits suggests that EVcomplex identifies not only residues close in space, but also particular pairs that are constrained by the transporter function of these complexes[31,34].

The complex of the ε and γ subunits from *E. coli* ATP synthase provides a challenge to our approach, since the ε subunit can take different positions relative to the γ subunit, executing the auto-inhibition of the enzyme by dramatic conformational changes[35]. In a real-world scenario, where we might not know this *a priori*, there may be conflicting constraints in the evolutionary record corresponding to the different positions of the flexible portion of ε subunit. EVcomplex accurately predicts 8 of the top 10 inter-EC pairs (within 8Å in the crystal structure; PDB: 1fs0[32]), with the top inter-EC, εA40-γL207 providing contact between the subunits at the at the end of an inter-protein beta sheet. Docking with these inter-ECs results top-ranked models having 1.5/1.3 Å backbone iRMSDs to the crystal structure (Table 1, Fig.

bioRxiv preprint doi: https://doi.org/10.1101/004762; this version posted May 6, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

2b, Supplementary Table 2) for the interface between the N-terminal domain of the ε subunit and the γ subunit. The C-terminal helices of the ε subunit are significantly different across 3 crystal structures (PDB: 1fs0, 1aqt[36], 3oaa[35]). The top ranked intra-ECs support the conformation seen in 1aqt, with the C-terminal helices tucked against the N-terminal beta barrel (Fig. 2b, green circles), and do not seem to leave a high ranked trace for the extended helical contact to the γ subunit seen in 1fs0 or 3oaa.

In summary, these benchmarks show the power of evolutionary information to infer protein complex 3D structure phenotypes, and demonstrate the unbiased assessment of the likelihood of successful predictions. The benchmarks also show that ECs provide precise relationships across the proteins that could be critical for the identification of functional coupling pathways in addition to the 3D models.

De novo **prediction** ***E. coli*** **ATP synthase a and b subunits**

The mechanism of production of ATP from ADP through a transmembrane protomotive force and rotary mechanism has been beautifully exposed in a body of work over the last 50 years (reviewed in [37]), and the mammalian complex in 3D most recently described using cryo-EM at a resolution of 18 Å[38]. However, the a-subunit remains structurally unsolved at atomic resolution and the precise interaction of the a and b subunits is unknown (Fig. 3a). Using the method described for the benchmark proteins, we predicted the interacting residues of the a and b subunits of E. *coli* ATP synthase (Fig. 3, Supplementary Data). Since the 3D structure of the membrane-integral pentahelical a-subunit is unknown, we started by using EVfold-membrane[26] to compute an all-atom 3D model based on an alignment of 17237 homologous sequences. The resulting model of a-subunit is consistent with topologies that have been inferred from crosslinking studies[39-41].

bioRxiv preprint doi: https://doi.org/10.1101/004762; this version posted May 6, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

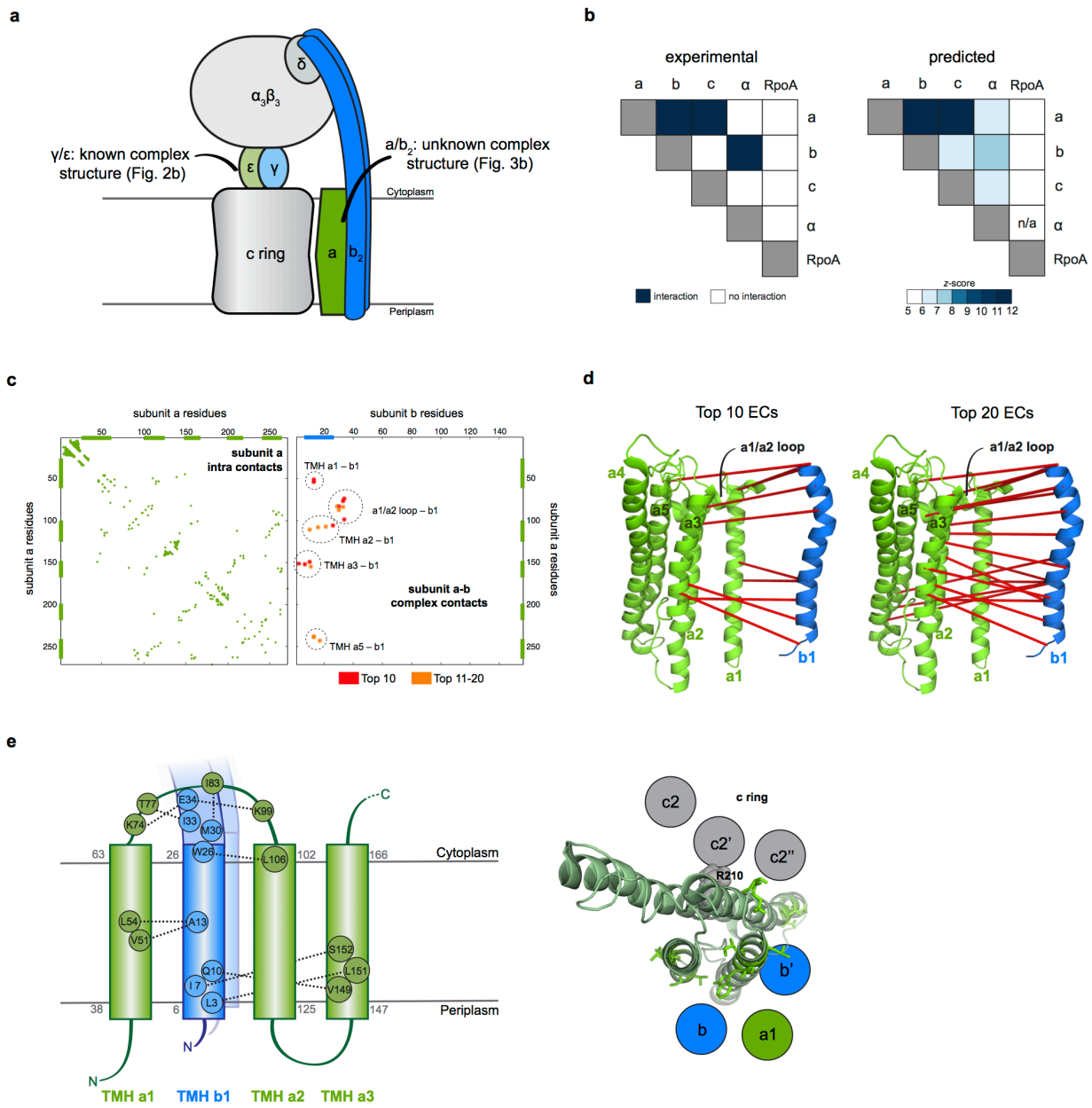

Figure 3. De novo prediction of ATP synthase subunit a and b interaction. (a) The a and b subunits of E. *coli* ATP synthase are known to interact, but the monomer structure of both the individual subunits and the complex is unknown. **(b)** Experimental evidence for binary interactions between the ATP synthase subunits a, b, c, and α and RNA polymerase subunit α (negative control) agrees with EVcomplex predictions. **(c)** Complex contact map of 10 top ranked inter-ECs and corresponding top ranked subunit a intra-ECs (subunit b intra-ECs not shown). **(d)** Inter-ECs (red lines) between subunit a (model predicted with EVfold-membrane, green) and subunit b (PDB: 1b9u[42], blue). **(e)** *Left panel:* Residue detail of predicted interaction between subunit a and b (dotted lines, predicted transmembrane helices as grey numbers). *Right panel:* Proposed helix-helix interactions between ATP synthase subunits a (green), b (blue, homodimer), and the c ring (grey).

The 10 highest-ranked inter-ECs are consistent with published experimental crosslinking studies on the a and b subunits (Supplementary Table 3) and the proposed geometry of the interaction[43]. K74 (subunit a) to E34 (subunit b) is the top-ranked inter-EC, coinciding with

bioRxiv preprint doi: https://doi.org/10.1101/004762; this version posted May 6, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

experimental evidence of the interaction of K74 with the b-subunit[39,40]. Most of the other high-ranking co-evolved pairs are between the cytoplasmic loop connecting transmembrane helices TMH a1 and TMH a2 in subunit a with residues 30-34 in subunit b, between the mid-membrane positions of TMH a1 and TMH b1, and also TMH a3 with the N terminal part of the b subunit (Fig. 3c-e). The 12th and 13th highest ranked inter-ECs are between TMH a5 and the b subunit mid-membrane. Since subunit b is thought to homo-dimerize in the complex[43], these interactions could be either with the same or the second b subunit (Fig. 3e). Similarly, the top ranked inter-EC between the a-subunit and c-subunit between aG213 cM64 lies one the same interaction interface, one turn away, of the functionally critical aR210, cD61 interaction[42] (Fig. 3e). The agreement between our *de novo* predicted *inter*-protein ECs with available experimental data serves as a measure of confidence for the predicted residue pair interactions, and suggests that EVcomplex can be used to reveal the 3D structural details of yet unsolved protein complexes given sufficient evolutionary information.

Discussion

A primary limitation of our approach is its dependence on the availability of a large number of evolutionarily related sequences. We estimate that one needs more than one sequence per residue (after reduction for redundancy) of the concatenated sequence length to produce reliable predictions (Table 1, Supplementary Table 1). Since only ~2000 bacterial genomes have been sequenced to date, this constitutes a bottleneck for the current method, unless there are multiple paralogs of the complex per species. In the latter case of multiple paralogs per species EVcomplex currently relies on the genome proximity of interacting pairs for concatenation, a condition that is violated in many cases. However, with the rapidly expanding number of sequenced genomes it is plausible that we will be able to explore genome-wide interactions of bacterial proteins in the near future. The work presented here is in anticipation of this genome-wide exploration and, as a proof of principle, shows the accurate prediction of inter-protein contacts and their ability to determine 3D structures across diverse complex interfaces.

In this work we chose benchmark protein pairs that were relatively easy to concatenate (Fig. 2, Online Methods). However, outside of multi-domain proteins, this approach will not work for

bioRxiv preprint doi: https://doi.org/10.1101/004762; this version posted May 6, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

most eukaryotic complexes with no bacterial homologs, or for proteins that interact but are not proximal on bacterial genomes. Clearly this is an important hurdle to overcome for the full potential of this evolutionary approach.

Finally, there is the question of conformational flexibility; as with intra-EC predictions, false positive contacts between proteins may be true positives in another, perhaps yet unseen conformation[26]. At least 3 of the complexes in this report have conformational flexibility critical to their function and it is plausible that large proportion of protein interactions exert their functions through dynamic changes in 3D relationships to greater or lesser extents. Can evolutionary information help to predict the details and extent for each complex? The challenge will certainly involve the development of algorithms that can disentangle evolutionary signals caused by alternative conformations of single complexes, alternative conformations of homologous complexes, or simply false positive signals. Taken together, these limitations highlight fruitful areas for future development of the methodology.

Despite strong conditions for the successful *de novo* calculation of co-evolved residues, the power of the method illustrated here may hugely accelerate the exploration of the protein-protein interaction world and the determination of protein complexes on a genome-wide scale. ECs between proteins may also be used to rank the likelihood of interaction as well as the interacting residues. The use of co-evolutionary analysis towards computational models to determine protein specificity and promiscuity, co-evolutionary dynamics and functional drift will be exciting future research questions.

Methods

Choice of prediction benchmark set. Protein complexes for testing were chosen based on sufficient sequences and the ability to match pairs of interacting proteins, for all sequences.

Pipeline Overview

(i) Construct alignments for pairs of individual proteins in a complex or interacting pair
(ii) Assess whether the number of sequences and coverage is sufficient.
(iii) Run EVcomplex (a generalization from EVfold[21,26]) using pseudolikelihood maximization[28] (PLM) to get intra- and inter-ECs
(iv) Discard complexes if the rank of the top inter EC is lower than L/2

bioRxiv preprint doi: https://doi.org/10.1101/004762; this version posted May 6, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

(v) Prepare monomers for docking with ECs, using either unbound conformations or randomized side-chains
(vi) Use top 5, 10 inter-ECs as distance restraints for docking to predict 3D complex using HADDOCK[44]

Multiple sequence alignments. Each protein was used to build a series of multiple sequence alignments (MSA) using jackhmmer[45] at different expectation value thresholds. The MSA was chosen to optimize the trade-off between the number of sequences retrieved and the coverage of the protein length, as in previous work[26]. In order to calculate co-evolved residues across different proteins we need to match the pairs of interacting protein sequences. To match the pairs, we assume interacting proteins are located in close proximity on their respective genomes, often on the same operon, as in the methods used previously matching histidine kinase and response regulator interacting pairs[18,19]. We retrieved the genomic locations of proteins in the alignments and concatenated pairs following 2 rules. (i) the CDS of each concatenated protein pair must be located on the same genomic contig (using ENA[46] for mapping), and (ii) each pair must be the closest to one another (on genome), when compared to all other possible pairs. Additionally, the concatenated sequence pairs were filtered blindly based on the distribution of genomic distances to exclude outlier pairs with high genomic distance (Supplementary Fig. 1, Supplementary Data). Furthermore, alignment members were clustered together and reweighted if 70% (theta 0.3) or more of their residues were identical (and thus implicitly removing duplicate sequences from the alignment). The total number of pairs lost from individual monomer counts after concatenation as well as the number of non-redundant sequences in the alignments is reported in Supplementary Table 1.

Evolutionary couplings calculation. Inter- and intra-ECs were calculated on the alignment of concatenated sequences using a global probability model of sequence co-evolution, adapted from the method for single proteins[21,22,26] using a PLM[28] rather than mean field approximation to calculate the coupling parameters. Columns in the alignment that contain more then 80% gaps (m80) were excluded and the weight of each sequence was adjusted to represent its cluster size in the alignment thus reducing the influence of identical or near-identical sequences in the calculation. All sequences >70% identity was down-weighted in the calculation in proportion to the number of sequences in that cluster. We can then compare the predicted ECs for both within and between the protein/domains to the crystal structures of the complexes (for contact maps and all EC scores, see Supplementary Fig. 3, Supplementary Data).

To predict the accuracy of the calculated inter-EC, we examined the rank of the first inter-EC for each complex, relative to *all* intra- and inter-ECs, calculated on the concatenated alignment. For instance the length of the concatenated MoaD/MoaE alignment is 231, resulting in 26,565 pairs of EC scores, of which all 12150 *inter*-protein EC scores and a subset of 13290 intra-EC scores remain when excluding intra-EC pairs with a primary sequence distance of up to 5. The highest rank inter-EC is 11th in the combined inter+intra EC list. Since ECs ranked in the top $L/2$ are likely to be true positives when there are sufficient sequences[21-23,26], this ranking

bioRxiv preprint doi: https://doi.org/10.1101/004762; this version posted May 6, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

provides an important metric to assess how likely the inter-EC contacts are accurate (see *_ECs_mapped.txt and *_evaluation_inter.txt files in Supplementary Data). The relative score of the first inter-EC also correctly reflects whether or not subunits of the ATP synthase complex interact (Fig. 3).

Docking. Monomer structures for each of the proteins in the HK-RR and CLPS-CLPA complexes and ATPE were taken from crystallized unbound conformations. For the other benchmark complexes we randomize the side chains of the monomers before docking because subunits that have been crystallized together in a complex will be biased due to the complementary positions of the surface side chains, and hence docking these proteins with no restraints between them could artificially produce high-ranking correct structures. Therefore, starting monomers (i.e. those extracted from complex structures) were subjected to side chain replacement using SCWRL4[47] resulting in ~ 1.5 Å RMSD over the side chains relative to the original 3D structure. We used HADDOCK[13], a widely used docking program based on ARIA[48] and the CNS software[49] (Crystallography and NMR System), to dock the monomers for each protein pair with 5, 10 inter-ECs as distance restraints on the α–carbon atoms of the backbone. (For interest only we also provide results from using 15 and 20 constraints.)

Each docking calculation starts with a rigid-body energy minimization, followed by semi-flexible refinement in torsion angle space, and ends with further refinement of the models in explicit solvent (water). 500/100/100 models generated for each of the 3 steps, respectively. All other parameters were left as the default values in the HADDOCK protocol. Each protein complex was run using predicted ECs as unambiguous distance restraints on the Cα atoms (d_{eff} 5Å, upper bound 2Å, lower bound 2Å; input files available in Supplementary Data). As a negative control, each protein complex was also docked using center of mass restraints (*ab initio* docking mode of HADDOCK)[30] alone and generating 10000/500/500 models.

Each of the generated models is scored using a weighted sum of electrostatic (E_{elec}) and van der Waals (E_{vdw}) energies complemented by an empirical desolvation energy term (E_{desolv})[50]. The distance restraint energy term was explicitly removed from the equation in the last iteration (Edist3 = 0.0) to enable comparison of the scores between the runs that used a different number of ECs as distance restraints.

$$\text{HADDOCKscore} = 0.2\ E_{elec} + 1.0\ E_{vdw} + 1.0\ E_{desol}$$

Comparison of models to crystal structures. All computed models in the benchmark were compared to the cognate crystal structures by the RMSD of all backbone atoms at the interface of the complex using ProFit v.3.1 (http://www.bioinf.org.uk/software/profit/). The interface is defined as the set of all residues that contain an atom < 6 Å away from any atom of the complex partner. For the ATPE-ATPG complex we excluded the 2 C-terminal helices of ATPE as these helices are mobile and take many different positions relative to

bioRxiv preprint doi: https://doi.org/10.1101/004762; this version posted May 6, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

other ATP synthase subunits[35]. Similarly, since the DHp domain of histidine kinases can take different positions relative to the CA domain, the HK-RR complex was compared over the interface between the DHp domain alone and the response regulator partner. Accuracy of the computed models with EC restraints were compared with computed models with center of mass restraints alone (negative controls), (Supplementary Fig.5, Supplementary Table 2).

Note added in proof

As this manuscript was completed, an interesting similar report with high-quality predicted protein complexes, also based on a global probability model approach, was published by Sergey Ovchinnikov, Hetunandan Kamisetty, and David Baker[51].

bioRxiv preprint doi: https://doi.org/10.1101/004762; this version posted May 6, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

References

1 Webb, B. *et al.* Modeling of proteins and their assemblies with the Integrative Modeling Platform. *Methods Mol Biol* **1091**, 277-295, doi:10.1007/978-1-62703-691-7_20 (2014).

2 Mosca, R., Ceol, A. & Aloy, P. Interactome3D: adding structural details to protein networks. *Nat Methods* **10**, 47-53, doi:10.1038/nmeth.2289 (2012).

3 Hart, G. T., Ramani, A. K. & Marcotte, E. M. How complete are current yeast and human protein-interaction networks? *Genome Biol* **7**, 120, doi:10.1186/gb-2006-7-11-120 (2006).

4 Zhang, Q. C. *et al.* Structure-based prediction of protein-protein interactions on a genome-wide scale. *Nature* **490**, 556-560, doi:10.1038/nature11503 (2012).

5 de Juan, D., Pazos, F. & Valencia, A. Emerging methods in protein co-evolution. *Nature reviews. Genetics* **14**, 249-261, doi:10.1038/nrg3414 (2013).

6 Chaudhury, S. *et al.* Benchmarking and analysis of protein docking performance in Rosetta v3.2. *PLoS One* **6**, e22477, doi:10.1371/journal.pone.0022477 (2011).

7 Kortemme, T., Kim, D. E. & Baker, D. Computational alanine scanning of protein-protein interfaces. *Science's STKE : signal transduction knowledge environment* **2004**, pl2, doi:10.1126/stke.2192004pl2 (2004).

8 Kortemme, T. *et al.* Computational redesign of protein-protein interaction specificity. *Nature structural & molecular biology* **11**, 371-379, doi:10.1038/nsmb749 (2004).

9 Kortemme, T. & Baker, D. Computational design of protein-protein interactions. *Current opinion in chemical biology* **8**, 91-97, doi:10.1016/j.cbpa.2003.12.008 (2004).

10 Kortemme, T. & Baker, D. A simple physical model for binding energy hot spots in protein-protein complexes. *Proc Natl Acad Sci U S A* **99**, 14116-14121, doi:10.1073/pnas.202485799 (2002).

11 Schneidman-Duhovny, D. *et al.* A method for integrative structure determination of protein-protein complexes. *Bioinformatics* **28**, 3282-3289, doi:10.1093/bioinformatics/bts628 (2012).

12 Velazquez-Muriel, J. *et al.* Assembly of macromolecular complexes by satisfaction of spatial restraints from electron microscopy images. *Proc Natl Acad Sci U S A* **109**, 18821-18826, doi:10.1073/pnas.1216549109 (2012).

13 Dominguez, C., Boelens, R. & Bonvin, A. M. HADDOCK: a protein-protein docking approach based on biochemical or biophysical information. *J Am Chem Soc* **125**, 1731-1737, doi:10.1021/ja026939x (2003).

14 Karaca, E. & Bonvin, A. M. Advances in integrative modeling of biomolecular complexes. *Methods* **59**, 372-381, doi:10.1016/j.ymeth.2012.12.004 (2013).

15 Rodrigues, J. P. *et al.* Defining the limits of homology modelling in information-driven protein docking. *Proteins*, doi:10.1002/prot.24382 (2013).

16 Gobel, U., Sander, C., Schneider, R. & Valencia, A. Correlated mutations and residue contacts in proteins. *Proteins* **18**, 309-317, doi:10.1002/prot.340180402 (1994).

17 Pazos, F., Helmer-Citterich, M., Ausiello, G. & Valencia, A. Correlated mutations contain information about protein-protein interaction. *J Mol Biol* **271**, 511-523, doi:10.1006/jmbi.1997.1198 (1997).

18 Skerker, J. M. *et al.* Rewiring the specificity of two-component signal transduction systems. *Cell* **133**, 1043-1054, doi:10.1016/j.cell.2008.04.040 (2008).

bioRxiv preprint doi: https://doi.org/10.1101/004762; this version posted May 6, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

19 Weigt, M., White, R. A., Szurmant, H., Hoch, J. A. & Hwa, T. Identification of direct residue contacts in protein-protein interaction by message passing. *Proc Natl Acad Sci U S A* **106**, 67-72, doi:10.1073/pnas.0805923106 (2009).

20 Burger, L. & van Nimwegen, E. Accurate prediction of protein-protein interactions from sequence alignments using a Bayesian method. *Molecular systems biology* **4**, 165, doi:10.1038/msb4100203 (2008).

21 Marks, D. S. *et al.* Protein 3D structure computed from evolutionary sequence variation. *PLoS One* **6**, e28766, doi:10.1371/journal.pone.0028766 (2011).

22 Morcos, F. *et al.* Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proc Natl Acad Sci U S A* **108**, E1293-1301, doi:10.1073/pnas.1111471108 (2011).

23 Jones, D. T., Buchan, D. W., Cozzetto, D. & Pontil, M. PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. *Bioinformatics* **28**, 184-190, doi:10.1093/bioinformatics/btr638 (2012).

24 Aurell, E. & Ekeberg, M. Inverse Ising inference using all the data. *Physical review letters* **108**, 090201 (2012).

25 Kamisetty, H., Ovchinnikov, S. & Baker, D. Assessing the utility of coevolution-based residue-residue contact predictions in a sequence- and structure-rich era. *Proc Natl Acad Sci U S A* **110**, 15674-15679, doi:10.1073/pnas.1314045110 (2013).

26 Hopf, T. A. *et al.* Three-dimensional structures of membrane proteins from genomic sequencing. *Cell* **149**, 1607-1621, doi:10.1016/j.cell.2012.04.012 (2012).

27 Nugent, T. & Jones, D. T. Accurate de novo structure prediction of large transmembrane protein domains using fragment-assembly and correlated mutation analysis. *Proc Natl Acad Sci U S A* **109**, E1540-1547, doi:10.1073/pnas.1120036109 (2012).

28 Ekeberg, M., Lovkvist, C., Lan, Y., Weigt, M. & Aurell, E. Improved contact prediction in proteins: using pseudolikelihoods to infer Potts models. *Physical review. E, Statistical, nonlinear, and soft matter physics* **87**, 012707 (2013).

29 Ovchinnikov, S., Kamisetty, H. & Baker, D. Robust and accurate prediction of residue-residue interactions across protein interfaces using evolutionary information. *Journal Article* **10.7554/eLife.02030**, doi:10.7554/eLife.02030 (2014).

30 de Vries, S. J. *et al.* HADDOCK versus HADDOCK: new features and performance of HADDOCK2.0 on the CAPRI targets. *Proteins* **69**, 726-733, doi:10.1002/prot.21723 (2007).

31 Johnson, E., Nguyen, P. T., Yeates, T. O. & Rees, D. C. Inward facing conformations of the MetNI methionine ABC transporter: Implications for the mechanism of transinhibition. *Protein Sci* **21**, 84-96, doi:10.1002/pro.765 (2012).

32 Rodgers, A. J. & Wilce, M. C. Structure of the gamma-epsilon complex of ATP synthase. *Nat Struct Biol* **7**, 1051-1054, doi:10.1038/80975 (2000).

33 Locher, K. P., Lee, A. T. & Rees, D. C. The E. coli BtuCD structure: a framework for ABC transporter architecture and mechanism. *Science* **296**, 1091-1098, doi:10.1126/science.1071142 (2002).

34 Kadaba, N. S., Kaiser, J. T., Johnson, E., Lee, A. & Rees, D. C. The high-affinity E. coli methionine ABC transporter: structure and allosteric regulation. *Science* **321**, 250-253, doi:10.1126/science.1157987 (2008).

35 Cingolani, G. & Duncan, T. M. Structure of the ATP synthase catalytic complex (F(1)) from Escherichia coli in an autoinhibited conformation. *Nature structural & molecular biology* **18**, 701-707, doi:10.1038/nsmb.2058 (2011).

bioRxiv preprint doi: https://doi.org/10.1101/004762; this version posted May 6, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

36 Uhlin, U., Cox, G. B. & Guss, J. M. Crystal structure of the epsilon subunit of the proton-translocating ATP synthase from Escherichia coli. *Structure* **5**, 1219-1230 (1997).
37 Walker, J. E. The ATP synthase: the understood, the uncertain and the unknown. *Biochemical Society transactions* **41**, 1-16, doi:10.1042/BST20110773 (2013).
38 Baker, L. A., Watt, I. N., Runswick, M. J., Walker, J. E. & Rubinstein, J. L. Arrangement of subunits in intact mammalian mitochondrial ATP synthase determined by cryo-EM. *Proc Natl Acad Sci U S A* **109**, 11675-11680, doi:10.1073/pnas.1204935109 (2012).
39 DeLeon-Rangel, J., Zhang, D. & Vik, S. B. The role of transmembrane span 2 in the structure and function of subunit a of the ATP synthase from Escherichia coli. *Archives of biochemistry and biophysics* **418**, 55-62 (2003).
40 Long, J. C., DeLeon-Rangel, J. & Vik, S. B. Characterization of the first cytoplasmic loop of subunit a of the Escherichia coli ATP synthase by surface labeling, cross-linking, and mutagenesis. *J Biol Chem* **277**, 27288-27293, doi:10.1074/jbc.M202118200 (2002).
41 Fillingame, R. H. & Steed, P. R. Half channels mediating H transport and the mechanism of gating in the F sector of Escherichia coli FF ATP synthase. *Biochim Biophys Acta*, doi:10.1016/j.bbabio.2014.03.005 (2014).
42 Dmitriev, O., Jones, P. C., Jiang, W. & Fillingame, R. H. Structure of the membrane domain of subunit b of the Escherichia coli F0F1 ATP synthase. *J Biol Chem* **274**, 15598-15604 (1999).
43 DeLeon-Rangel, J., Ishmukhametov, R. R., Jiang, W., Fillingame, R. H. & Vik, S. B. Interactions between subunits a and b in the rotary ATP synthase as determined by cross-linking. *FEBS letters* **587**, 892-897, doi:10.1016/j.febslet.2013.02.012 (2013).
44 de Vries, S. J., van Dijk, M. & Bonvin, A. M. The HADDOCK web server for data-driven biomolecular docking. *Nat Protoc* **5**, 883-897, doi:10.1038/nprot.2010.32 (2010).
45 Johnson, L. S., Eddy, S. R. & Portugaly, E. Hidden Markov model speed heuristic and iterative HMM search procedure. *BMC bioinformatics* **11**, 431, doi:10.1186/1471-2105-11-431 (2010).
46 Pakseresht, N. *et al.* Assembly information services in the European Nucleotide Archive. *Nucleic Acids Res* **42**, D38-43, doi:10.1093/nar/gkt1082 (2014).
47 Krivov, G. G., Shapovalov, M. V. & Dunbrack, R. L., Jr. Improved prediction of protein side-chain conformations with SCWRL4. *Proteins* **77**, 778-795, doi:10.1002/prot.22488 (2009).
48 Linge, J. P., Habeck, M., Rieping, W. & Nilges, M. ARIA: automated NOE assignment and NMR structure calculation. *Bioinformatics* **19**, 315-316 (2003).
49 Brunger, A. T. Version 1.2 of the Crystallography and NMR system. *Nat Protoc* **2**, 2728-2733, doi:10.1038/nprot.2007.406 (2007).
50 Fernandez-Recio, J., Totrov, M. & Abagyan, R. Identification of protein-protein interaction sites from docking energy landscapes. *J Mol Biol* **335**, 843-865 (2004).
51 Ovchinnikov, S., Kamisetty, H. & Baker, D. Robust and accurate prediction of residue-residue interactions across protein interfaces using evolutionary information. *eLife* (2014).