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 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, with sufficient accuracy to determine the three-dimensional structure of the protein complexes. We demonstrate that the inferred evolutionary coupling scores accurately predict inter-protein residue interactions and can distinguish between interacting and non-interacting proteins. To illustrate the utility of the method, we predict co-evolved contacts between 50 E. *coli* complexes (of unknown structure), including the 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.

Introduction

A large part 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 which has been significantly extended by homology-inferred 3D models (Mosca, Ceol et al. 2012, Webb, Lasker et al. 2014),(Hart, Ramani et al. 2006, Zhang, Petrey et al. 2012). However, there is still little, or no, 3D information for ~80% of the known protein interactions in bacteria, yeast or human, amounting to ~30,000/~6000 uncharacterized interactions in human and *E. coli,* respectively (Mosca, Ceol et al. 2012, Rajagopala, Sikorski et al. 2014). Current experimental methods cannot match the rapid increase in the demand for residue–level information of these interactions.

One way to address the knowledge gap of protein interactions has been the use of hybrid, computational-experimental approaches that typically combine 3D structural information at varying resolutions, homology models and other methods (de Juan, Pazos et al. 2013), with biophysical force fields such as Rosetta Dock, residue cross-linking and data-driven approaches that incorporate various sources of biological information (Kortemme and Baker 2002, Dominguez, Boelens et al. 2003, Kortemme and Baker 2004, Kortemme, Joachimiak et al. 2004, Kortemme, Kim et al. 2004, Chaudhury, Berrondo et al. 2011, Schneidman-Duhovny, Rossi et al. 2012, Velazquez-Muriel, Lasker et al. 2012, Karaca and Bonvin 2013, Rodrigues, Melquiond et al. 2013, Webb, Lasker et al. 2014). 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 and large complexes).

One promising computational approach is to use evolutionary analysis of amino acid co-variation to identify close residue contacts across protein interactions, first used 20 years ago (Gobel, Sander et al. 1994, Pazos and Valencia 2001), though with limited success. More recent approaches have been more successful in identifying residue interactions, for instance between histidine kinases and response regulators (Burger and van Nimwegen 2008, Skerker, Perchuk et al. 2008, Weigt, White et al. 2009), 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 residue-residue contacts across a protein interface would allow computation of 3D models and provide a powerful, orthogonal approach to experiments.

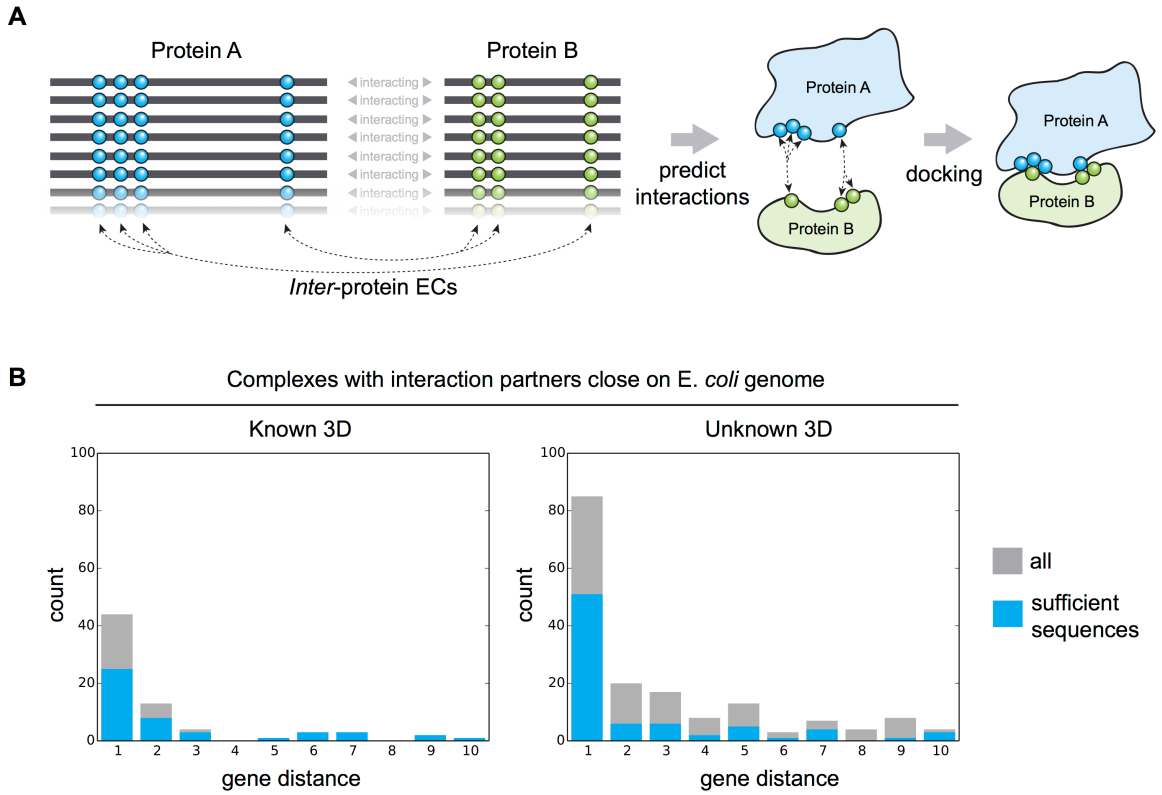


Figure 1. Co-evolution of residues across protein complexes from the evolutionary sequence record. (A) Evolutionary pressure to maintain protein-protein interactions leads to the coevolution of residues between interacting proteins in a complex. By analyzing patterns of amino acid co-variation in an 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 (right). **(B)** Distribution of *E. coli* protein complexes of known and unknown 3D structure where both subunits are close on the bacterial genome (grey bars), allowing sequence pair matching by genomic distance. For a subset of these complexes, sufficient sequence information is available for evolutionary couplings analysis (blue bars). See Figure 1 - figure supplement 1.

Since the recent and successful demonstration of the use of evolutionary couplings (ECs) between residues to determine the 3D structure of individual proteins (Marks, Colwell et al. 2011, Morcos, Pagnani et al. 2011, Aurell and Ekeberg 2012, Jones, Buchan et al. 2012, Kamisetty, Ovchinnikov et al. 2013), including integral membrane proteins (Hopf, Colwell et al. 2012, Nugent and Jones 2012), we reasoned that such an evolutionary statistical approach (such as EVcouplings (Marks, Colwell et al. 2011), web site: http://evfold.org) could be used to determine co-evolved residues *between* proteins. Here we calculate evolutionary couplings across protein interactions resulting in the majority of top ranked *inter*-protein EC pairs (inter-ECs) being in close proximity when compared to known 3D structures and show that these constraints are sufficient to calculate accurate 3D models of the complexes in the subset we docked (Figure 1A). We present the results for evolutionary couplings for 50 complexes of unknown 3D structures that have sufficient number of sequences and high inter-protein

scores, showing experimental support for our predicted interactions between the a-, b- and c-subunit of ATP synthase.

Results

Methodological approach. We investigated whether co-evolving residues between proteins are close in three dimensions by assessing blinded predictions of residue co-evolution against experimentally determined 3D complex structures. Beginning with a published dataset of ~3500 high-confidence protein interactions in *E. coli (Rajagopala, Sikorski et al. 2014)*, we removed redundancy and conditioned their inclusion based on genome distance between the pairs of proteins in the *E. coli* genome resulting in 236 interactions (Figure 1B, Supplementary file 1), see Materials and methods. The current algorithm leverages the fact that some interacting proteins are encoded close in the bacterial genomes, or that there are only single occurrences of the pair of interacting proteins in each genome (Table 1, Supplementary file 1).

The paired sequences are concatenated and co-evolutionary analysis is performed using EVcouplings (Marks, Colwell et al. 2011, Morcos, Pagnani et al. 2011, Aurell and Ekeberg 2012) that implements a pseudolikelihood maximization (PLM) approximation to determine the interaction parameters in the maximum entropy equation (Ekeberg, Lovkvist et al. 2013, Kamisetty, Ovchinnikov et al. 2013), simultaneously generating both intra- and inter-EC scores for all pairs of residues within and across the protein pairs (Figure 1A).

Benchmark calculations here and previous work (Marks, Colwell et al. 2011, Morcos, Pagnani et al. 2011, Hopf, Colwell et al. 2012, Jones, Buchan et al. 2012, Kamisetty, Ovchinnikov et al. 2013) indicate that the number of sequences in the alignment is critical (> 0.5 non-redundant sequences per residue, see also Materials and methods) and that the accuracy of ECs that rank lower than L (the number of residues in the sequence) decays rapidly.

Based on these observations we use the strength of the top inter-protein EC (relative to all intra-protein EC residue pairs) to estimate the likelihood of the predicted interactions to be accurate. We compared our predicted inter-protein ECs against residue distances in the 30 known crystal structures of complexes that had sufficient sequences together with high inter-EC scores (< 0.3) (Supplementary file 1, Figure 2). For the remaining 159 interactions that have no known structural information, 50 complexes had sufficient inter-EC strength and number of sequences indicating high-confidence ECs.

Top ECs are mostly inter-protein contacts. For the top ranked benchmark complexes, the majority of the top 5 ECs between proteins is correct to within 8Å (Table 1 and Supplementary file 3).

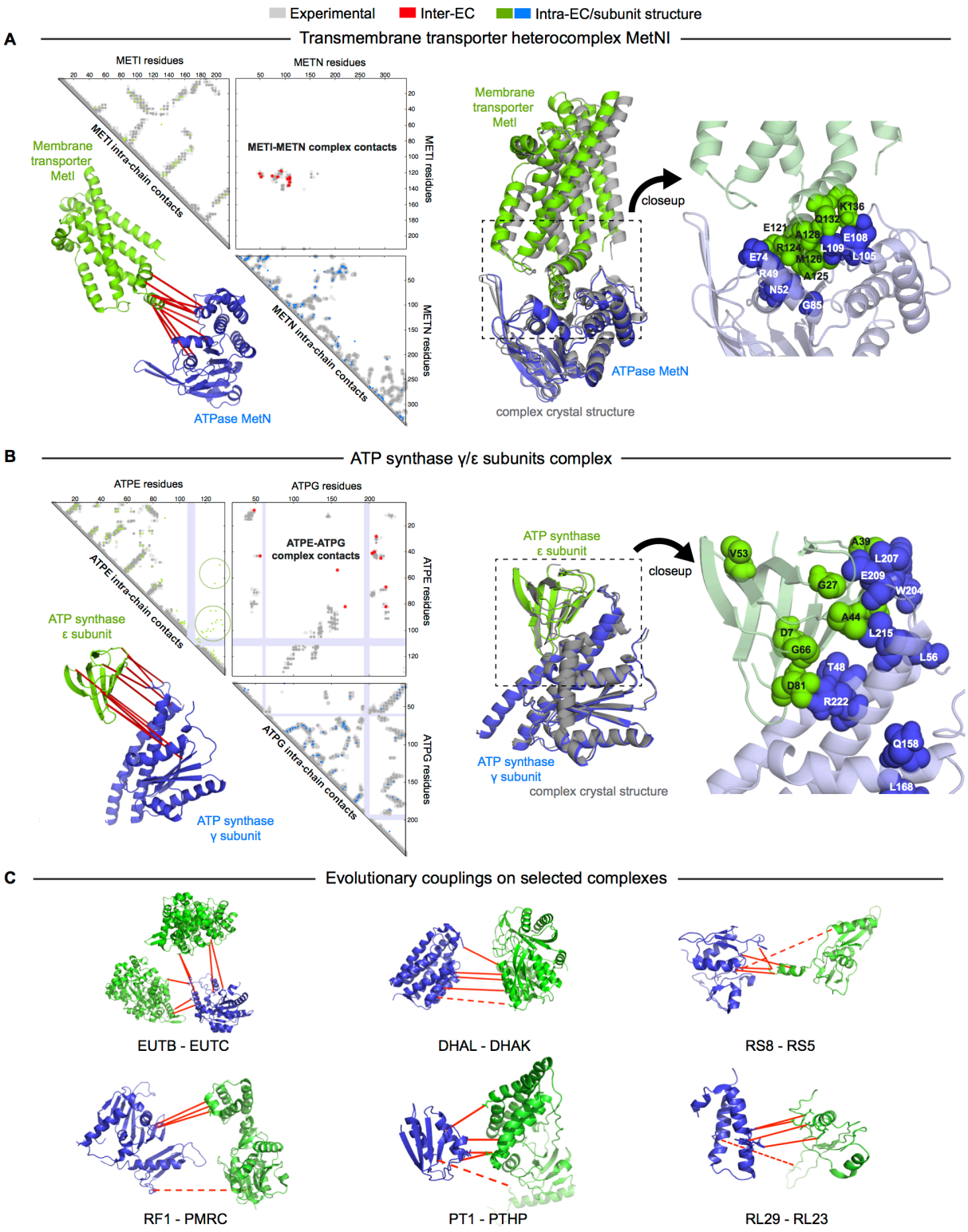


Figure 2. Evolutionary couplings give accurate 3D structures of complexes. EVcomplex predictions and comparison to crystal structure for **(A)** the methionine-importing transmembrane transporter heterocomplex MetNI from E. coli (PDB: 3tui) and **(B)** the gamma/epsilon subunit interaction of E. coli ATP synthase (PDB: 1fs0). **(A,B)** ***Left panels:*** 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 panels:*** 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)** High-ranking inter-ECs on a selection of benchmark complexes (hetero-complex subunits: blue, green cartoons, multiple copies of the same subunit in identical color; correct inter-ECs: solid red lines, incorrect inter-ECs: dashed red lines). See Figure 2 - figure supplement 1.

The normalized inter EC-score demarks those complexes that have accurate EC predictions; inter-EC scores < 0.3 distinguishes mostly accurate ECs in the top 5, and those with inter-EC scores > 0.6 have mostly inaccurate ECs (Table 1, Figure 3 and Supplementary files 2 and 3). Since the co-evolution score indicates incorrect ECs, we reasoned that the scores can also distinguish interaction from non-interaction of protein pairs. We tested this on the a-, b-, c- and α-subunit interactions of ATP synthase and correctly paired the interacting subunits using the inter-protein EC scores (Figure 4A and B).

Some of the high confidence inter-protein ECs are not close in 3D when compared to crystal structures. These false positives may be a result of a number of reasons, including incorrect assumptions about interacting pairs across the alignments, oligomerization, as well as genuine co-evolution 'at a distance'. In addition, the complexes may also exist in alternative conformations not necessarily captured by a single crystal structure, for instance in the case of the large conformational changes of the BtuCDF complex (Hvorup, Goetz et al. 2007). To test whether the inter-protein ECs are sufficient for computing accurate 3D structures, we selected 7 diverse examples for docking. Using high ranked inter EC pairs (5 and 10) as distance constraints between the individual proteins we generated 100 3D models of the complexes (HADDOCK (Dominguez, Boelens et al. 2003, de Vries, van Dijk et al. 2007)), see Materials and methods. Over 70% of the generated models were close to the crystal structures of the complexes (< 4Å backbone iRMSD) (Figure 2, Supplementary file 4, Supplementary data).

Evolutionary conserved residue networks across proteins

The top 10 inter-EC pairs between MetI and MetN are accurate to within 8Å in the MetNI complex (PDB: 3tui (Johnson, Nguyen et al. 2012)), resulting in an average of 1.6 Å iRMSD from the crystal structure (standard deviation 0.06) for all 100 computed 3D models (Table 1, Supplementary file 4 and 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 (Johnson, Nguyen et al. 2012) (Figure 2A).

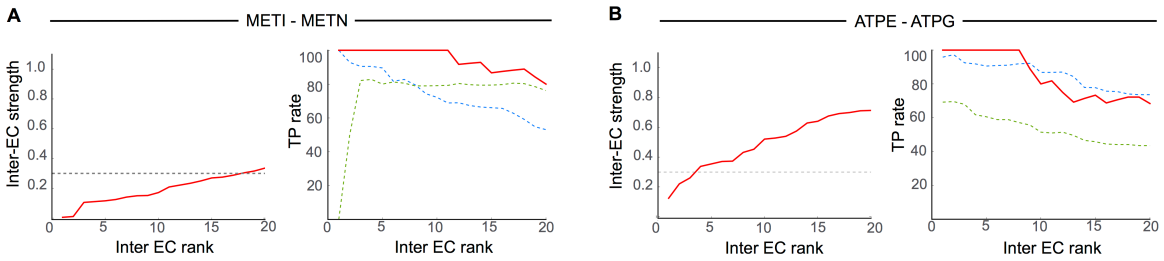


Figure 3. The accuracy of inter-protein ECs is predicted by their relative rank. See Supplemental files 2 and 3. (A) Evolutionary couplings between MetI and MetN are high relative to all intra- and inter-protein ECs (left plot, red). 16 inter-ECs have a normalized inter-EC score <0.3, of which 85% are true positives (right plot, red line) **(B)** In contrast, compared to MetI-MetN, fewer ECs between AtpE and AtpG have high inter-EC scores (left plot, red line), but all 4 are true positives (right plot, red line). The normalized inter-EC strength is the rank percentile of the inter-EC in the list of top L intra- and inter- ECs, where L is the concatenated length of the proteins in the complex. **(A, B)** The true positive rates for the inter-protein ECs weakly corresponds the accuracy of the intra-protein ECs for the individual proteins (right plots, green and blue dashed lines).

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

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

The vitamin B12 transporter (BtuC) belongs to a different structural class of ABC transporters, but also uses ATP hydrolysis via an interacting ATPase (BtuD). EVcomplex identifies sufficient inter-protein residue contacts to calculate accurate complex 3D structures, the top ranked model is 0.8 Å backbone iRMSD compared to the crystal structure, 1l7v (Locher, Lee et al. 2002), (Table 1, Supplementary file 4 and Supplementary data). 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. 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 (Kadaba, Kaiser et al. 2008, Johnson, Nguyen et al. 2012).

The ATP synthase ε and γ subunit complex 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 (Cingolani and Duncan 2011). 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 1fs0 (Rodgers and Wilce 2000), with the top inter-EC, εA40-γL207, providing contact between the subunits at the end of an inter-protein beta sheet. The C-terminal helices of the ε subunit are significantly different across 3 crystal structures (PDB IDs: 1fs0, 1aqt (Uhlin, Cox et al. 1997), 3oaa (Cingolani and Duncan 2011)). The top ranked intra-ECs support the conformation seen in 1aqt, with the C-terminal helices tucked against the N-terminal beta barrel (Figure 2B, green circles) and do not contain a high ranked evolutionary trace for the extended helical contact to the γ subunit seen in 1fs0 or 3oaa. Docking with the top inter-ECs results in models with 1.5/1.3 Å backbone iRMSDs to the crystal structure, for the interface between the N-terminal domain of the ε subunit and the γ subunit (Table 1, Figure 2B and Supplementary file 4). εD81 and γR222 connect the ε- subunit via a network of 3 high intra ECs between the N and C terminal helices to the core of the F1 ATP synthase. In summary, these benchmarks show the power of evolutionary information to infer protein complex 3D structure phenotypes, and demonstrate the criteria needed for successful prediction of unknown interactions. 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 predictions of unknown complexes

The 50 *E. coli* complexes that had sufficient sequences and high inter-EC scores form a non-redundant set of diverse functions and stochiometries (Supplementary files 1, 2 and 6), and include the remaining unsolved interactions of the ATP synthase complex. We highlight the EVcomplex predictions of the a-, b- and c-subunit interactions since the structure is of wide

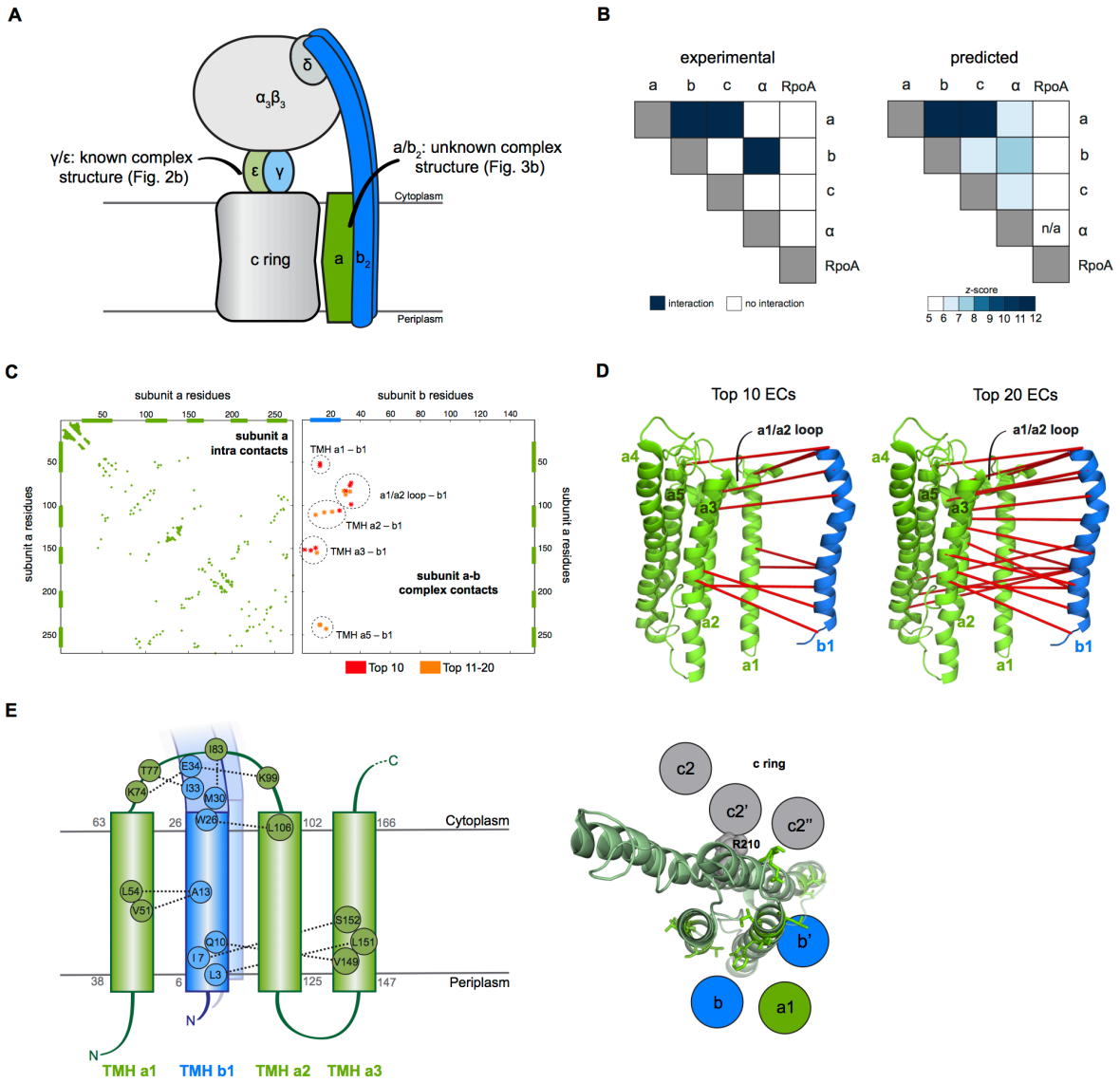


Figure 4. Predicted interactions between the a-, b- and c- subunits of ATP synthase. (A) The a- and b-subunits of E. coli ATP synthase interact, but the monomer structure of both the individual subunits and the complex is unknown. **(B)** Experimental evidence for binary interactions between the ATP 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, 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).

biological interest (reviewed in (Walker 2013)), and has proven experimentally challenging (Baker, Watt et al. 2012) (Figure 4A). Since the 3D structure of the membrane-integral penta-helical a-subunit is unknown, we computed a *de novo* 3D model using evolutionary couplings as described previously (Hopf, Colwell et al. 2012). The resulting model of the a-subunit is consistent with topologies that have been inferred from crosslinking studies (Long, DeLeon-

Rangel et al. 2002, DeLeon-Rangel, Zhang et al. 2003, Fillingame and Steed 2014). The 10 highest-ranked inter-ECs between the a- and b- subunits (Figure 4C, D and E) are completely consistent with experimental crosslinking studies (Supplementary file 6) and the proposed geometry of the interaction (DeLeon-Rangel, Ishmukhametov et al. 2013) (Figure 4E). The top ranked coupling between the a- and b-subunit (aK74 - bE34) coincides with experimental evidence of the interaction of aK74 with the b-subunit (Long, DeLeon-Rangel et al. 2002, DeLeon-Rangel, Zhang et al. 2003).

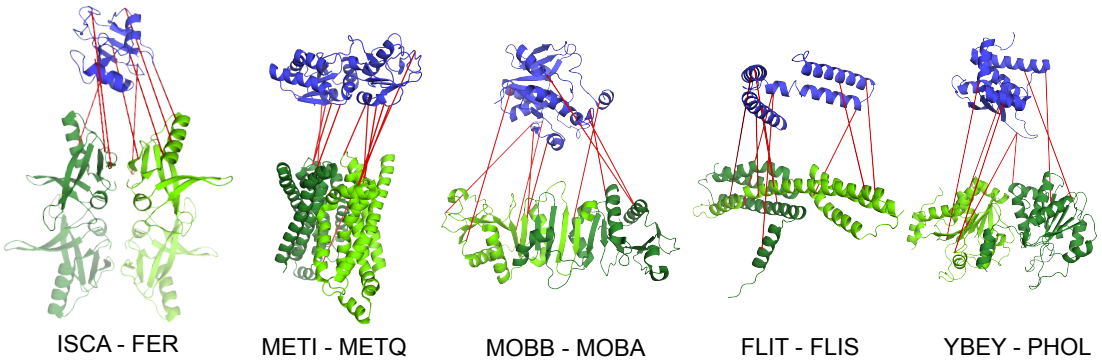


Figure 5. Evolutionary couplings in complexes of unknown 3D structure. Top ranked inter-ECs for five *de novo* prediction candidates without E*. coli* or interolog complex 3D structure (Subunits: blue/green cartoons; top ranked inter-ECs: red lines). For complex subunits which homomultimerize (light/dark green cartoon), inter-ECs are shown to the monomer partner to enable the identification of multiple interaction sites. Left to right: *(1)* the iron binding protein IscA (PDB: 1r95) and ferredoxin (PDB: 1i7h); *(2)* the membrane subunit of methionine-importing transporter heterocomplex MetI (PDB: 3tui) together with its periplasmic binding protein MetQ (Swissmodel: P28635); *(3)* molybdenum cofactor biosynthesis adapter protein MobB (PDB: 1np6) and molybdenum cofactor guanylyltransferase MobA (PDB: 1e5k); *(4)* the flagella proteins FliT and FliS (comparative models based on PDB 3a7m and 1vh6); *(5)* and the endoribonuclease yebY (PDB: 1xm5) and the PhoH-like protein PhoL (Swissmodel: P0A9K3).

Most other high-ranking co-evolved pairs are between the cytoplasmic loop connecting transmembrane helices TMHa1 and TMHa2 in the a-subunit, with residues 30-34 in the b-subunit (Figure 4C, D and E). The top EC between the a- and c-subunits (aG213 cM64) lies close to the functionally critical aR210, cD61 interaction (Dmitriev, Jones et al. 1999) on the same helical faces of the respective subunit (Figure 4E).

Other high confidence predictions include the flagella FliT-FliS, MobB-MobC, IscA-Fer and YbeY-PhoL complexes that show distinct interaction faces consistent with dimer interactions with the monomer (Figure 5). The top 9 inter-ECs between MetQ and MetI, are from one interface of MetQ to the MetI periplasmic loops, consistent with the known binding of MetQ binding to MetI in the periplasm. In general, 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 1000 sequences for the concatenated proteins, each a 100 residues in length, after filtering redundancy and down-weighting highly similar proteins in the alignments (0.5 sequences per residue) (Table 1, Supplementary file 1). Although nearly 3000 bacterial genomes have now been sequenced, the number of sequenced genomes remains a bottleneck for the current method, as many protein complexes are not conserved across all bacteria. If there are multiple paralogs of the interacting proteins, as in the case of the HK-RR interaction, then the method can successfully partner proteins and hence construct correctly concatenated alignments. However, with the rapidly expanding number of sequenced genomes it is plausible that we will be able to explore far higher percentage of protein interactions in the near future.

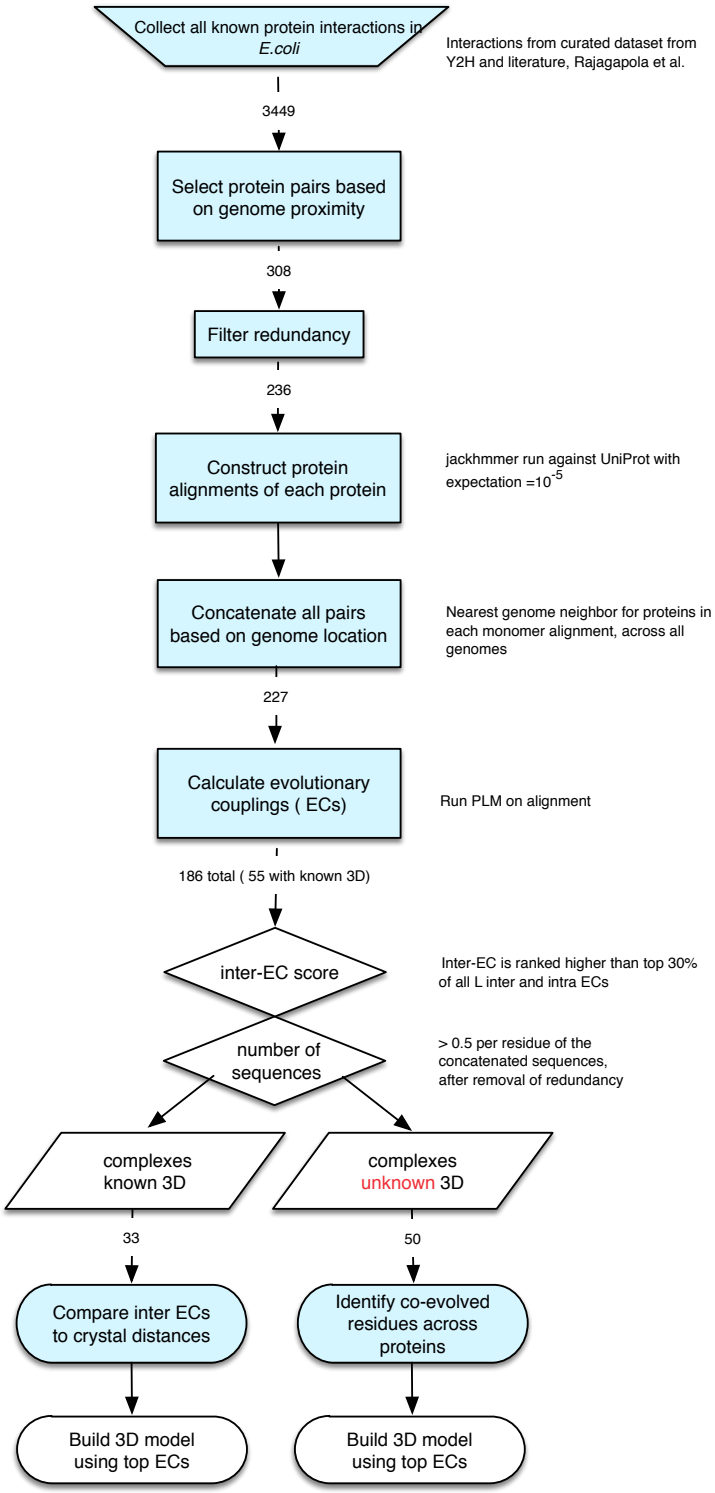
A second limitation is the dependence of the approach on genome proximity of the interacting pairs of proteins such that the sequences can be correctly concatenated for the evolutionary couplings calculation. Since only ~10% of interacting protein pairs in *E. coli* are distributed in this way, the challenge is now to develop an algorithm that can calculate the probability that any pair are co-evolved. However, as with the first limitation, the meaningful pairing of interacting proteins from paralogous families will become unnecessary as many more thousands of genomes are sequenced.

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. As with single protein (intra-EC) predictions, evolutionary conserved conformational flexibility and oligomerization can result in more than one set of contacts that must be de-convoluted. 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. 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.

Materials and methods

Workflow



Selection of interacting protein pairs for co-evolution calculation.

The candidate set of complexes for benchmark and *de novo* prediction was derived starting from a dataset of binary protein-protein interactions in E. *coli* including yeast two-hybrid experiments, literature-curated interactions and 3D complex structures in the PDB (Rajagopala, Sikorski et al. 2014). Since our algorithm for concatenating multiple sequence pairs per species assumes the proximity of the interacting proteins on the respective genomes (see below), we exclude any complex with a gene distance > 10 from further analysis. The gene distance is calculated as the number of genes between the interacting partners based on an ordered list of genes in the E. *coli* genome obtained from the UniProt database. The resulting list of pairs (~ 300) was then filtered for redundant homologous complexes and heterodimers based on the identification of Pfam domains in the interacting proteins (236). All remaining complexes with a known 3D structure or an interlog 3D structure (71) (identified by intersecting the results of HMMER searches against the PDB for both monomers) were used for benchmarking the method, while complexes without known structure (159) were assigned to the *de novo* prediction set, see Workflow box above.

Multiple sequence alignments.

Each protein from all pairs in our dataset was used to seed a multiple sequence alignment (MSA) using jackhmmer (Johnson, Eddy et al. 2010) and the alignment depth for the calculations was chosen to optimize the number of sequences retrieved and the coverage of the protein length, as in previous work (Hopf, Colwell et al. 2012) (Supplementary files 1 and 4). 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 (Skerker, Perchuk et al. 2008, Weigt, White et al. 2009). 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 (Pakseresht, Alako et al. 2014) for mapping), and (ii) each pair must be the closest to one another (on genome), when compared to all other possible pairs. The concatenated sequence pairs were filtered based on the distribution of genomic distances to exclude outlier pairs with high genomic distance (Figure 1 Figure supplement 1, Supplementary data). Alignment members were clustered together and reweighted if 80% or more of their residues were identical (and thus implicitly removing duplicate sequences from the alignment). Supplementary file 1 reports the total number of concatenated sequences, the lengths, and the effective number of sequences remaining after down-weighting highly similar proteins and removing fragments as has been done in previous work (Marks, Colwell et al. 2011)

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 (Marks, Colwell et al. 2011, Morcos, Pagnani et al. 2011, Hopf, Colwell et al. 2012) using a PLM (Ekeberg, Lovkvist et al. 2013) rather than mean field approximation to calculate the coupling parameters. Columns in the alignment that contain more than 80% gaps 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. 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 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 and intra EC list. Since ECs ranked in the top $L/2$ are likely to be true positives when there are sufficient sequences (Marks, Colwell et al. 2011, Morcos, Pagnani et al. 2011, Hopf, Colwell et al. 2012, Jones, Buchan et al. 2012), this ranking 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). Using these observations we defined an EVcomplex inter-EC score based on the relative rank of the first inter-protein residue pair in the list of all intra and inter pairs of ECs calculated from the multiple sequence alignment. Based on previous work on single proteins, we filter for sequence sufficiency requiring > 0.5 residues per residue in the concatenated sequence by counting the effective number of sequences after redundancy (Table 1, Supplementary file 2).

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 (Krivov, Shapovalov et al. 2009) resulting in ~ 1.5 Å RMSD over the side chains relative to the original 3D structure. We used HADDOCK (Dominguez, Boelens et al. 2003), a widely used docking program based on ARIA (Linge, Habeck et al. 2003) and the CNS software (Brunger 2007) (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) (de Vries, van Dijk et al. 2007) alone and in the case of the controls 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})(Fernandez-Recio, Totrov et al. 2004). 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.

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

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 other ATP synthase subunits (Cingolani and Duncan 2011). 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), (Figure 2 supplemental figure 2, Supplementary file 5).

<u>Note added during submission</u>

As this manuscript was completed, an 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 (Ovchinnikov, Kamisetty et al. 2014).

References

Aurell, E. and M. Ekeberg (2012). "Inverse Ising inference using all the data." Phys Rev Lett **108**(9): 090201.

Baker, L. A., I. N. Watt, M. J. Runswick, J. E. Walker and J. L. Rubinstein (2012). "Arrangement of subunits in intact mammalian mitochondrial ATP synthase determined by cryo-EM." Proc Natl Acad Sci U S A **109**(29): 11675-11680.

Brunger, A. T. (2007). "Version 1.2 of the Crystallography and NMR system." Nature protocols **2**(11): 2728-2733.

Burger, L. and E. van Nimwegen (2008). "Accurate prediction of protein-protein interactions from sequence alignments using a Bayesian method." Mol Syst Biol **4**: 165.

Chaudhury, S., M. Berrondo, B. D. Weitzner, P. Muthu, H. Bergman and J. J. Gray (2011). "Benchmarking and analysis of protein docking performance in Rosetta v3.2." PLoS One **6**(8): e22477.

Cingolani, G. and T. M. Duncan (2011). "Structure of the ATP synthase catalytic complex (F(1)) from Escherichia coli in an autoinhibited conformation." Nat Struct Mol Biol **18**(6): 701-707.

de Juan, D., F. Pazos and A. Valencia (2013). "Emerging methods in protein co-evolution." Nat Rev Genet **14**(4): 249-261.

de Vries, S. J., A. D. van Dijk, M. Krzeminski, M. van Dijk, A. Thureau, V. Hsu, T. Wassenaar and A. M. Bonvin (2007). "HADDOCK versus HADDOCK: new features and performance of HADDOCK2.0 on the CAPRI targets." Proteins **69**(4): 726-733.

DeLeon-Rangel, J., R. R. Ishmukhametov, W. Jiang, R. H. Fillingame and S. B. Vik (2013). "Interactions between subunits a and b in the rotary ATP synthase as determined by cross-linking." FEBS letters **587**(7): 892-897.

DeLeon-Rangel, J., D. Zhang and S. B. Vik (2003). "The role of transmembrane span 2 in the structure and function of subunit a of the ATP synthase from Escherichia coli." Arch Biochem Biophys **418**(1): 55-62.

Dmitriev, O. Y., P. C. Jones and R. H. Fillingame (1999). "Structure of the subunit c oligomer in the F1Fo ATP synthase: model derived from solution structure of the monomer and cross-linking in the native enzyme." Proc Natl Acad Sci U S A **96**(14): 7785-7790.

Dominguez, C., R. Boelens and A. M. Bonvin (2003). "HADDOCK: a protein-protein docking approach based on biochemical or biophysical information." Journal of the American Chemical Society **125**(7): 1731-1737.

Ekeberg, M., C. Lovkvist, Y. Lan, M. Weigt and E. Aurell (2013). "Improved contact prediction in proteins: using pseudolikelihoods to infer Potts models." Phys Rev E Stat Nonlin Soft Matter Phys **87**(1): 012707.

Fernandez-Recio, J., M. Totrov and R. Abagyan (2004). "Identification of protein-protein interaction sites from docking energy landscapes." J Mol Biol **335**(3): 843-865.

Fillingame, R. H. and P. R. Steed (2014). "Half channels mediating H transport and the mechanism of gating in the F sector of Escherichia coli FF ATP synthase." Biochim Biophys Acta.

Gobel, U., C. Sander, R. Schneider and A. Valencia (1994). "Correlated mutations and residue contacts in proteins." Proteins **18**(4): 309-317.
Hart, G. T., A. K. Ramani and E. M. Marcotte (2006). "How complete are current yeast and human protein-interaction networks?" Genome biology **7**(11): 120.
Hopf, T. A., L. J. Colwell, R. Sheridan, B. Rost, C. Sander and D. S. Marks (2012). "Three-dimensional structures of membrane proteins from genomic sequencing." Cell **149**(7): 1607-1621.
Hvorup, R. N., B. A. Goetz, M. Niederer, K. Hollenstein, E. Perozo and K. P. Locher (2007). "Asymmetry in the structure of the ABC transporter-binding protein complex BtuCD-BtuF." Science **317**(5843): 1387-1390.
Johnson, E., P. T. Nguyen, T. O. Yeates and D. C. Rees (2012). "Inward facing conformations of the MetNI methionine ABC transporter: Implications for the mechanism of transinhibition." Protein science : a publication of the Protein Society **21**(1): 84-96.
Johnson, L. S., S. R. Eddy and E. Portugaly (2010). "Hidden Markov model speed heuristic and iterative HMM search procedure." BMC Bioinformatics **11**: 431.
Jones, D. T., D. W. Buchan, D. Cozzetto and M. Pontil (2012). "PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments." Bioinformatics **28**(2): 184-190.
Kadaba, N. S., J. T. Kaiser, E. Johnson, A. Lee and D. C. Rees (2008). "The high-affinity E. coli methionine ABC transporter: structure and allosteric regulation." Science **321**(5886): 250-253.
Kamisetty, H., S. Ovchinnikov and D. Baker (2013). "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**(39): 15674-15679.
Karaca, E. and A. M. Bonvin (2013). "Advances in integrative modeling of biomolecular complexes." Methods **59**(3): 372-381.
Kortemme, T. and D. Baker (2002). "A simple physical model for binding energy hot spots in protein-protein complexes." Proc Natl Acad Sci U S A **99**(22): 14116-14121.
Kortemme, T. and D. Baker (2004). "Computational design of protein-protein interactions." Curr Opin Chem Biol **8**(1): 91-97.
Kortemme, T., L. A. Joachimiak, A. N. Bullock, A. D. Schuler, B. L. Stoddard and D. Baker (2004). "Computational redesign of protein-protein interaction specificity." Nat Struct Mol Biol **11**(4): 371-379.
Kortemme, T., D. E. Kim and D. Baker (2004). "Computational alanine scanning of protein-protein interfaces." Sci STKE **2004**(219): pl2.
Krivov, G. G., M. V. Shapovalov and R. L. Dunbrack, Jr. (2009). "Improved prediction of protein side-chain conformations with SCWRL4." Proteins **77**(4): 778-795.
Linge, J. P., M. Habeck, W. Rieping and M. Nilges (2003). "ARIA: automated NOE assignment and NMR structure calculation." Bioinformatics **19**(2): 315-316.
Locher, K. P., A. T. Lee and D. C. Rees (2002). "The E. coli BtuCD structure: a framework for ABC transporter architecture and mechanism." Science **296**(5570): 1091-1098.

Long, J. C., J. DeLeon-Rangel and S. B. Vik (2002). "Characterization of the first cytoplasmic loop of subunit a of the Escherichia coli ATP synthase by surface labeling, cross-linking, and mutagenesis." The Journal of biological chemistry **277**(30): 27288-27293.
Marks, D. S., L. J. Colwell, R. Sheridan, T. A. Hopf, A. Pagnani, R. Zecchina and C. Sander (2011). "Protein 3D structure computed from evolutionary sequence variation." PloS one **6**(12): e28766.
Morcos, F., A. Pagnani, B. Lunt, A. Bertolino, D. S. Marks, C. Sander, R. Zecchina, J. N. Onuchic, T. Hwa and M. Weigt (2011). "Direct-coupling analysis of residue coevolution captures native contacts across many protein families." Proc Natl Acad Sci U S A **108**(49): E1293-1301.
Mosca, R., A. Ceol and P. Aloy (2012). "Interactome3D: adding structural details to protein networks." Nature methods **10**(1): 47-53.
Nugent, T. and D. T. Jones (2012). "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**(24): E1540-1547.
Ovchinnikov, S., H. Kamisetty and D. Baker (2014). "Robust and accurate prediction of residue-residue interactions across protein interfaces using evolutionary information." Elife: e02030.
Pakseresht, N., B. Alako, C. Amid, A. Cerdeno-Tarraga, I. Cleland, R. Gibson, N. Goodgame, T. Gur, M. Jang, S. Kay, R. Leinonen, W. Li, X. Liu, R. Lopez, H. McWilliam, A. Oisel, S. Pallreddy, S. Plaister, R. Radhakrishnan, S. Riviere, M. Rossello, A. Senf, N. Silvester, D. Smirnov, S. Squizzato, P. ten Hoopen, A. L. Toribio, D. Vaughan, V. Zalunin and G. Cochrane (2014). "Assembly information services in the European Nucleotide Archive." Nucleic Acids Res **42**(Database issue): D38-43.
Pazos, F. and A. Valencia (2001). "Similarity of phylogenetic trees as indicator of protein-protein interaction." Protein Eng **14**(9): 609-614.
Rajagopala, S. V., P. Sikorski, A. Kumar, R. Mosca, J. Vlasblom, R. Arnold, J. Franca-Koh, S. B. Pakala, S. Phanse, A. Ceol, R. Hauser, G. Siszler, S. Wuchty, A. Emili, M. Babu, P. Aloy, R. Pieper and P. Uetz (2014). "The binary protein-protein interaction landscape of Escherichia coli." Nat Biotechnol **32**(3): 285-290.
Rodgers, A. J. and M. C. Wilce (2000). "Structure of the gamma-epsilon complex of ATP synthase." Nature structural biology **7**(11): 1051-1054.
Rodrigues, J. P., A. S. Melquiond, E. Karaca, M. Trellet, M. van Dijk, G. C. van Zundert, C. Schmitz, S. J. de Vries, A. Bordogna, L. Bonati, P. L. Kastritis and A. M. Bonvin (2013). "Defining the limits of homology modelling in information-driven protein docking." Proteins.
Schneidman-Duhovny, D., A. Rossi, A. Avila-Sakar, S. J. Kim, J. Velazquez-Muriel, P. Strop, H. Liang, K. A. Krukenberg, M. Liao, H. M. Kim, S. Sobhanifar, V. Dotsch, A. Rajpal, J. Pons, D. A. Agard, Y. Cheng and A. Sali (2012). "A method for integrative structure determination of protein-protein complexes." Bioinformatics **28**(24): 3282-3289.
Skerker, J. M., B. S. Perchuk, A. Siryaporn, E. A. Lubin, O. Ashenberg, M. Goulian and M. T. Laub (2008). "Rewiring the specificity of two-component signal transduction systems." Cell **133**(6): 1043-1054.

Uhlin, U., G. B. Cox and J. M. Guss (1997). "Crystal structure of the epsilon subunit of the proton-translocating ATP synthase from Escherichia coli." Structure **5**(9): 1219-1230.
Velazquez-Muriel, J., K. Lasker, D. Russel, J. Phillips, B. M. Webb, D. Schneidman-Duhovny and A. Sali (2012). "Assembly of macromolecular complexes by satisfaction of spatial restraints from electron microscopy images." Proc Natl Acad Sci U S A **109**(46): 18821-18826.
Walker, J. E. (2013). "The ATP synthase: the understood, the uncertain and the unknown." Biochem Soc Trans **41**(1): 1-16.
Webb, B., K. Lasker, J. Velazquez-Muriel, D. Schneidman-Duhovny, R. Pellarin, M. Bonomi, C. Greenberg, B. Raveh, E. Tjioe, D. Russel and A. Sali (2014). "Modeling of proteins and their assemblies with the Integrative Modeling Platform." Methods Mol Biol **1091**: 277-295.
Weigt, M., R. A. White, H. Szurmant, J. A. Hoch and T. Hwa (2009). "Identification of direct residue contacts in protein-protein interaction by message passing." Proc Natl Acad Sci U S A **106**(1): 67-72.
Zhang, Q. C., D. Petrey, L. Deng, L. Qiang, Y. Shi, C. A. Thu, B. Bisikirska, C. Lefebvre, D. Accili, T. Hunter, T. Maniatis, A. Califano and B. Honig (2012). "Structure-based prediction of protein-protein interactions on a genome-wide scale." Nature **490**(7421): 556-560.