Deciphering the Structure of Condensin Protein Complex –

Supplementary Information:

Materials and Methods:

Obtaining Protein Sequence Data for Condensin Subunits

In order to predict coevolved contacts between the various subunits, the databases of multiple sequence alignments (MSA) for each protein were extracted from Pfam (1) (SMC hinge domain (PF06470), SMC head N_{domain} (PF02463), ScpA protein (PF02616) and ScpB protein (PF04079)). All MSAs that are composed of 25% or more gaps were discarded from the analysis due to poor coverage. In investigating the interaction between the various subunits of the condensin protein complex, we take advantage of the tendency for interacting proteins in bacteria to be encoded under the same operon. Hence, we assume that an ScpA and ScpB protein interacts if they are encoded adjacent to one another on the bacterial genome. We employ DCA to predict the contacts between SMC-ScpA and SMC-ScpB protein pairs using same organism classification. For all systems studied, Pfam datasets contain more than 3000 sequences ensuring the statistical significance of the results and a substantial level of accuracy.

Experimental Protein Structural Data

All experimental crystal structures in this study were retrieved from Protein Data Bank (PDB) (2).

Direct Coupling Analysis

To quantify the amino acid coevolution between residue sites in a MSA, $\{\vec{\sigma}^{(s)}\}_{s=1...M}$, we construct a probabilistic model of a sequence, $\vec{\sigma}=(\sigma_1,\sigma_2,...,\sigma_L)$ that is most consistent with the MSA data. The probability distribution resulting from our model, $P(\vec{\sigma})$, must reproduce the single-site and pairwise frequencies of the dataset, i.e., it must satisfy the marginalization conditions $\sum_{\substack{\sigma_k \\ k\neq i}} P(\vec{\sigma}) = f_i(\sigma_i)$ and $\sum_{\substack{\sigma_k \\ k\neq i,j}} P(\vec{\sigma}) = f_{ij}(\sigma_i,\sigma_j)$, where $f_i(\sigma_i)$ and $f_{ij}(\sigma_i,\sigma_j)$ denote the single-site and pairwise frequency, respectively, and *i* and *j* are indices of the MSA. The most general form of such a probabilistic model is given by the principle of maximum entropy as a Boltzmann distribution:

$$P(\vec{\sigma}) = \frac{1}{Z}\exp(-H(\vec{\sigma}))$$

where

$$H(\vec{\sigma}) = -\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} J_{ij}(\sigma_i,\sigma_j) - \sum_{i=1}^{N} h_i(\sigma_i).$$

www.pnas.org/cgi/doi/10.1073/pnas.1812770115

where the J_{ij} parameters capture the correlations between pairs of sites in a MSA and the h_i parameters are the single-site fields related to the amino acid composition at a site. We adopted the pseudolikelihood maximization approach of Ekeberg et al (3) to estimate the parameters of a probabilistic model that is most consistent with the sequence data. The coevolution between a pair of residue sites *i* and *j* is calculated using an average product corrected Frobenius norm of the couplings matrix J_{ij} (3). The Frobenius norm of J_{ij} is expressed as

$$F_{ij} = \sqrt{\sum_{\sigma_i}\sum_{\sigma_j} J_{ij}(\sigma_i, \sigma_j)^2}$$

where the double summation is taken over all amino acid identities (plus gap) at sites *i* and *j*, respectively. The average product correction is applied as

$$F_{ij}^{(APC)} = F_{ij} - \frac{F_{\bullet j} F_{i\bullet}}{F_{\bullet\bullet}}$$

Where $F_{\bullet j}$, $F_{i\bullet}$ and $F_{\bullet\bullet}$ are averaged between all sites *i*, *j*, and both *i* as well as *j*, respectively.

MD Simulation for Prediction of Unknown Protein Complex

Because no complete structural data exist for the SMC-ScpA and SMC-ScpB complexes, we predict their 3D structure by combining the DCA-derived contacts with structure-based models (SBM) (4-5). Similar approaches have been used to investigate both the folding and the docking of the protein subunits (7). In order to reproduce docking of the selected subunits, the structure of each subunit was initially separated by 40 Å and was processed by the SMOG server (7), generating topology files that contain the SBM coarse-grained C_α potentials. DCA-derived constraints were incorporated in an SBM using a pairwise potential energy function that combines a repulsive, excluded volume interaction with an attractive Gaussian potential at long range (8-9):

$$U_G(r_{ij}) = \left(1 + \left(\frac{\sigma_{NC}}{r_{ij}}\right)^{12}\right)\left(1 + G(r_{ij})\right) - 1$$

where

$$G(r_{ij}) = -A\exp\left(-\frac{(r_{ij} - r_0)^2}{2\sigma^2}\right)$$

Here, σ is the width of the Gaussian well, *A* is the amplitude, r_{ij} is the distance between C_α atoms *i* and *j,* and r^0 is the equilibrium distance..

The potentials were added into the topology generated by SMOG, alongside with a Gaussian potential (8). For each system, 100 simulations were performed, with a total of $1x10^9$ timesteps for hinge simulations and $4x10^9$ timesteps for ScpAB and ScpA-SMC head docking. The binding simulations consist of 3 stages in which the equilibrium distance (r_0) was modulated, along with the Gaussian parameters of amplitude and decay A and σ, respectively. Similarly to (9) from the first stage to the final stage, the equilibrium distance was decreased, the amplitude was increased, and the width of the Gaussian well was decreased. Each simulation stage was carried until the observation of a stabilized conformation, observed by a reduced variance on each stage's RMSD. The parameters used for each simulation stage and the variances at the last stage of the simulation are summarized in the following table:

MD Stage	σ (Å)	A (Å)	r_0(Å)
1	4	3	15
2	4	5	8
3	1	5	8

The obtained structures of our MD simulations (Figure 1b, Figure 3c, Figure 3d, Figure 5e and Figure 5f) were reconstructed to all-atom representations using an in-house code.

Data deposition

The atomic coordinates of both open and closed condensin protein complex have been deposited in the Model Archive database, modelarchive.org, (Model Archive ID code ma-abc6s and ma-ac7jp).

In main text (Figures 3a, 3b, 4a and 5a) we demonstrate the overlap between our direct coupling analysis (DCA) maps and exemplary PDB experimental results. Here, we provide correlation between DCA of each condensin subunit with full PDB structural data, corresponding to its Pfam database:

SMC head

PDB code: 1E69

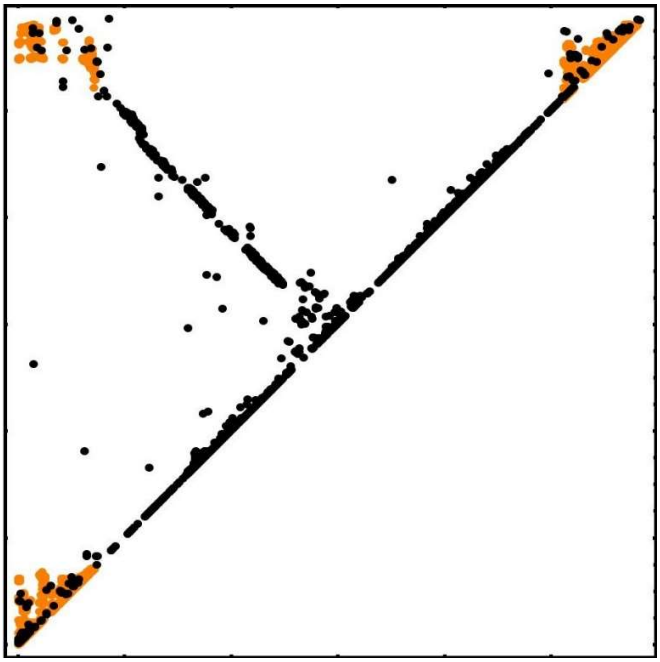

Fig. S1A. Overlap of DCA contact prediction (top contacts appear in black) with experimental results (PDB code: 1E69) obtained for SMC head domain of *Themotoga maritima* bacteria (orange color) (10).

PDB code: 5H66

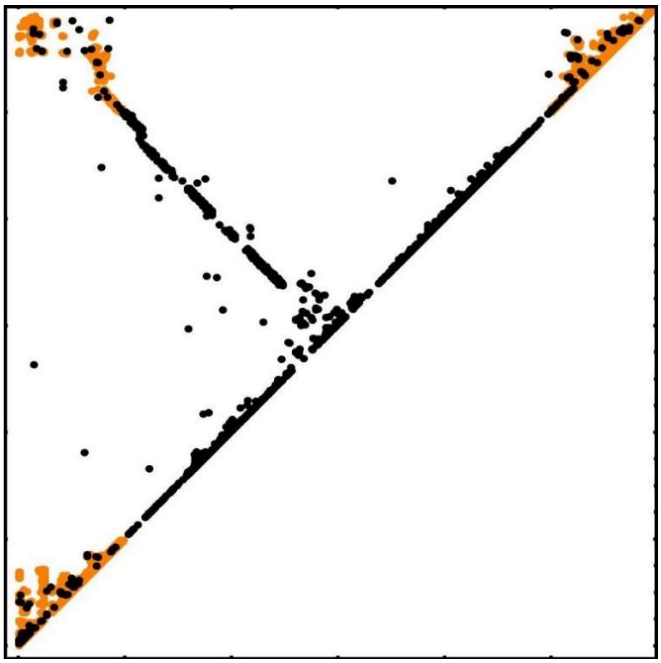

Fig. S1B. Overlap of DCA contact prediction (top contacts appear in black) with experimental results (PDB code: 5H66) obtained for SMC head domain of *Bacillus subtilis* bacteria (orange color) (11).

PDB code: 3KTA

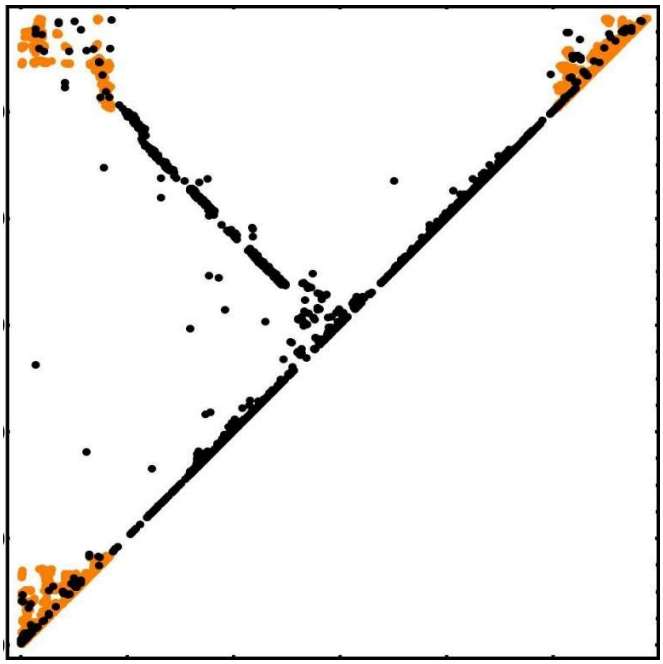

Fig. S1C. Overlap of DCA contact prediction (top contacts appear in black) with experimental results (PDB code: 3KTA) obtained for the structural basis for adenylate kinase activity in SMC head ATPases of the *Pyrococcus furiosus* bacteria (orange color) (12).

PDB code: 1XEX

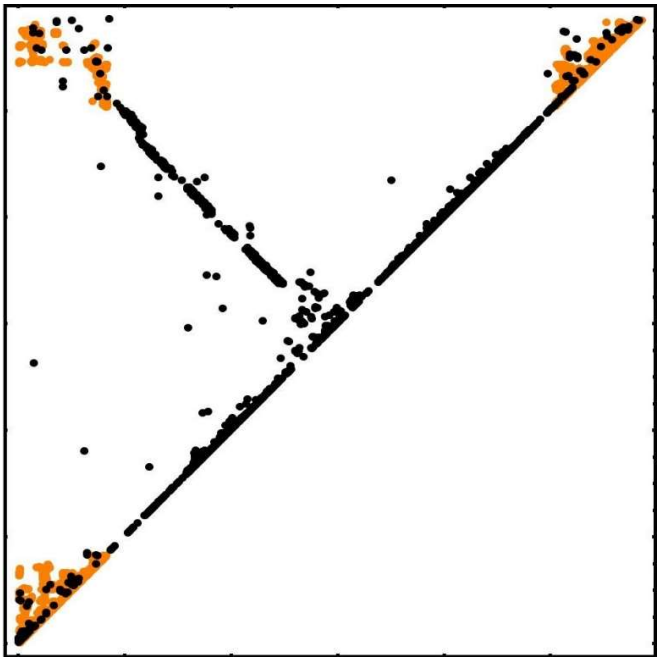

Fig. S1D. Overlap of DCA contact prediction (top contacts appear in black) with experimental results (PDB code: 1XEX) obtained for the ATP-driven dimerization and DNA stimulated activation of SMC head ATPases for the *Pyrococcus furiosus* bacteria (orange color) (13).

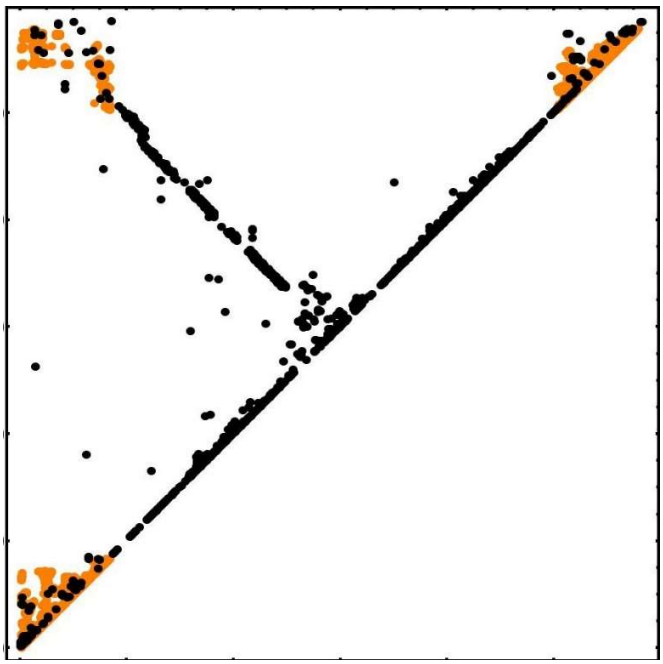

Fig. S1E. Overlap of DCA contact prediction (top contacts appear in black) with experimental results (PDB code: 4I99) obtained for the SMC head domain for the *Pyrococcus furiosus* bacteria (orange color) (14).

SMC hinge

PDB code: 4RSJ

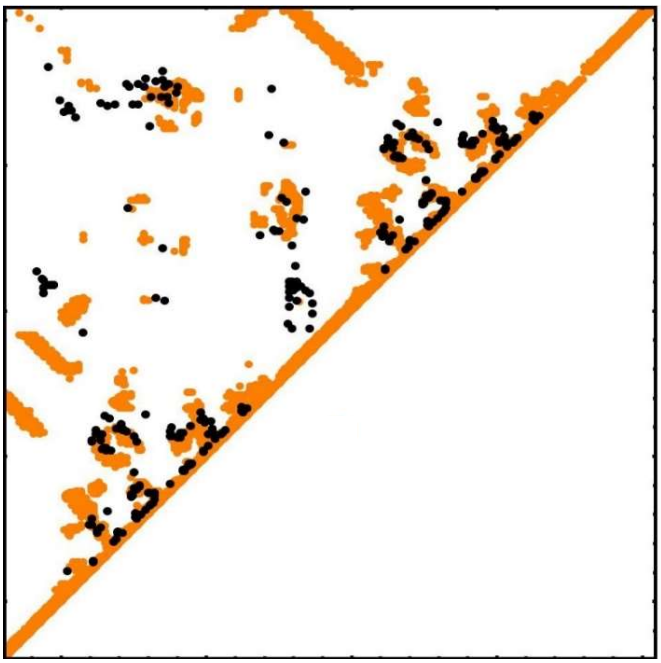

Fig. S1F. Overlap of DCA contact prediction (top contacts appear in black) with experimental results (PDB code: 4RSJ) obtained for the SMC hinge domain with extended coiled coil for the *Pyrococcus furiosus* bacteria (orange color). Monomers are separated with blue lines (15).

ScpA

PDB code: 3W6J

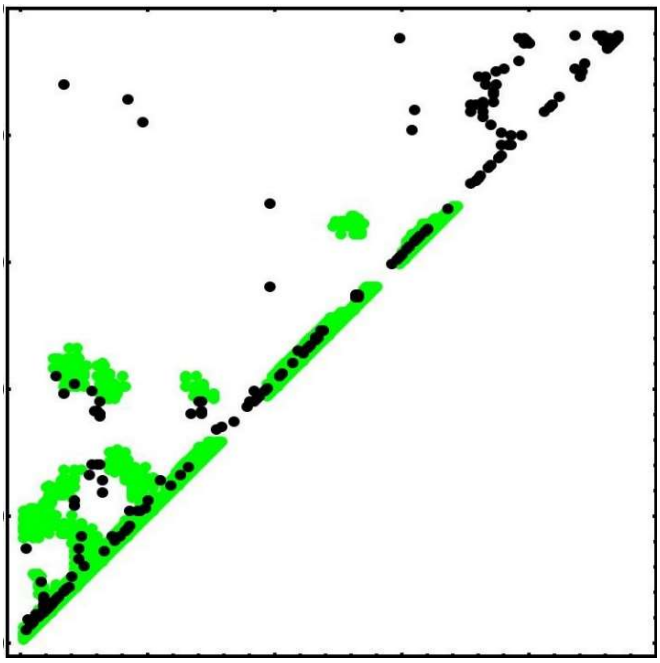

Fig. S1G. Overlap of DCA contact prediction (top contacts appear in black) with experimental results (PDB code: 3W6J) obtained for the ScpA protein for the *Geobacilles sp.* bacteria (green color) (15).

PDB code: 3ZGX

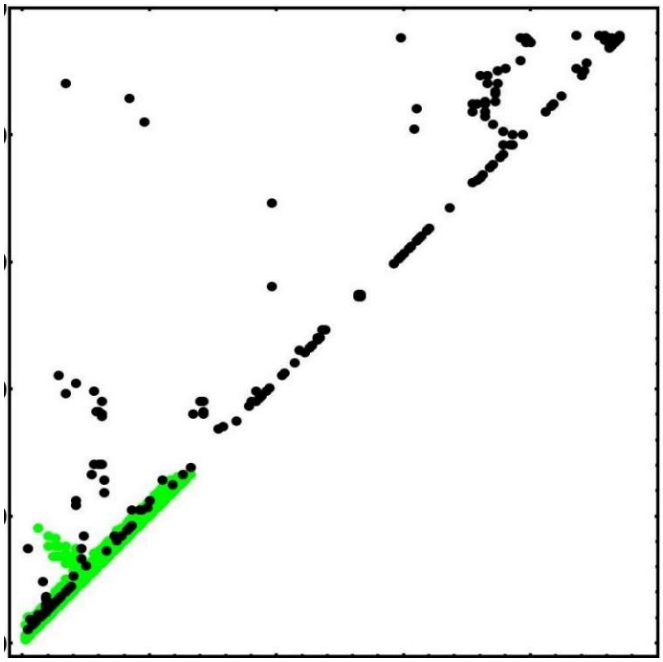

Fig. S1H. Overlap of DCA contact prediction (top contacts appear in black) with experimental results (PDB code: 3ZGX) obtained for the crystal structure of the kleisin N-domain for the *Bacillus subtilis* bacteria (green color) (14).

ScpB

PDB code: 3W6J

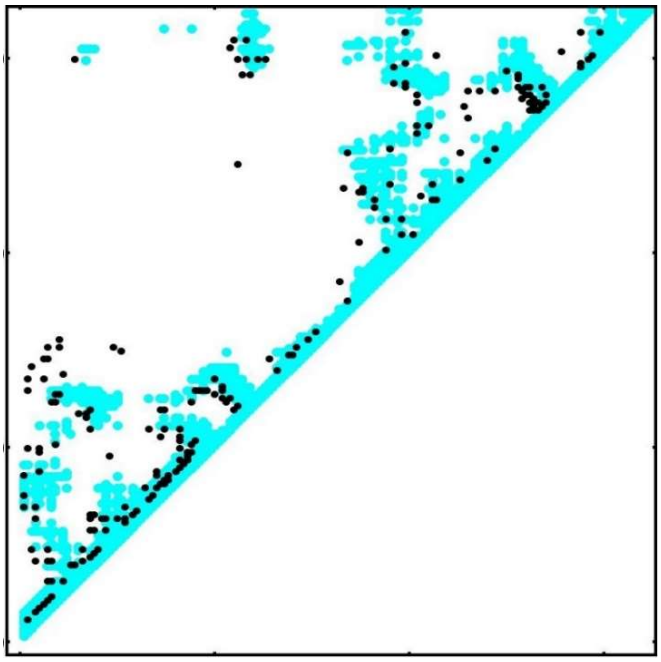

Fig. S1I. Overlap of DCA contact prediction (top contacts appear in black) with experimental results (PDB code: 3W6J) obtained for the ScpB protein for the *Geobacilles sp.* bacteria (cyan color) (15).

PDB code: 2Z99

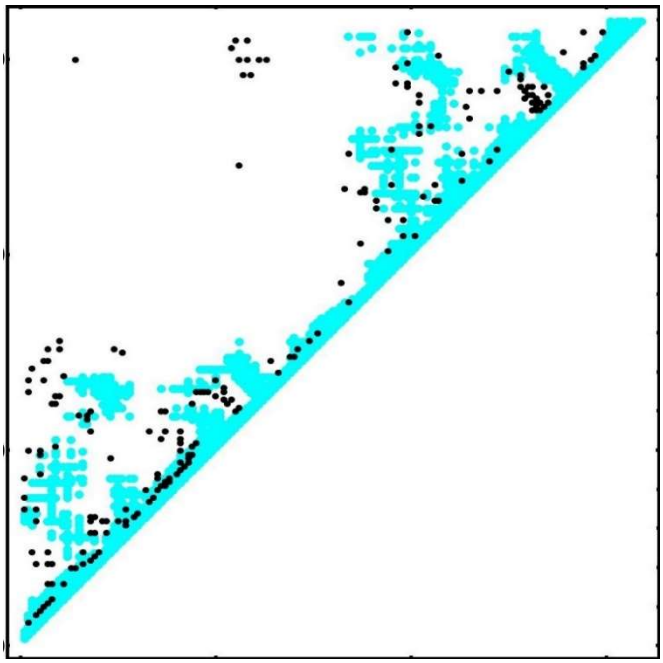

Fig. S1J. Overlap of DCA contact prediction (top contacts appear in black) with experimental results (PDB code: 2Z99) obtained for the crystal structure of the ScpB protein for the *Mycobacterium tuberculosis* bacteria (cyan color) (16).

PDB code: 1T6S

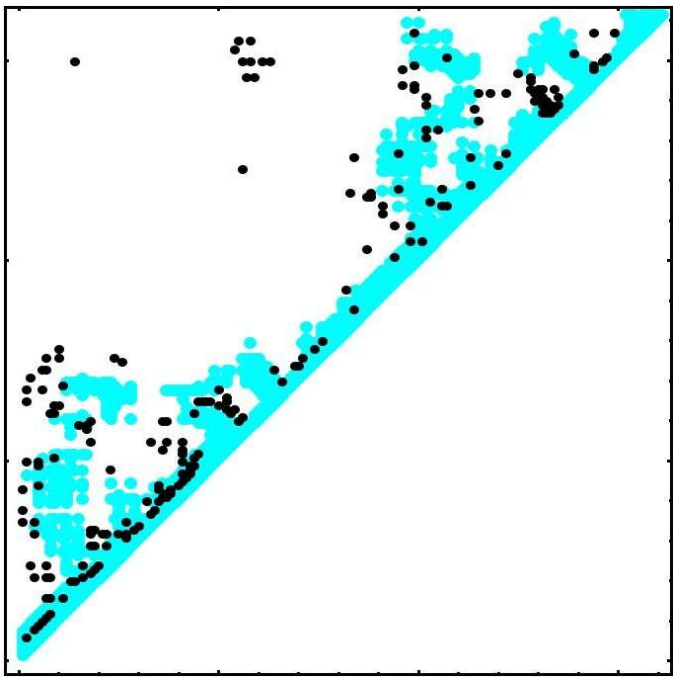

Fig. S1K. Overlap of DCA contact prediction (top contacts appear in black) with experimental results (PDB code: 1T6S) obtained for the crystal structure of the ScpB protein for the *Chlorobium tepidum* bacteria (cyan color) (17).

In figure 2 in main test we present an overview of all experimental structural data for the condensin complex together with co-evolutionary information. In order to predict DCA-derived contacts, we use the sequences under the same organism classification. Obtaining the intra-protein contacts of a single SMC, DCA results in excellent agreement with experimental data (see figure S2, panel a). Here, an overall of over 5000 sequences under the same organism classification were used. Inter-protein contacts between SMC and ScpA (panel b) and between SMC and ScpB (panel c) result in sufficient agreement with experimental data, providing the overall structure of each constituent. For these DCA calculations, a total of 2700 and 2500 same organism sequences were used. These ensure sufficient statistical significance of the results and a substantial level of accuracy in obtaining the top inter-protein contacts for the SMC-ScpA and SMC-ScpB systems.

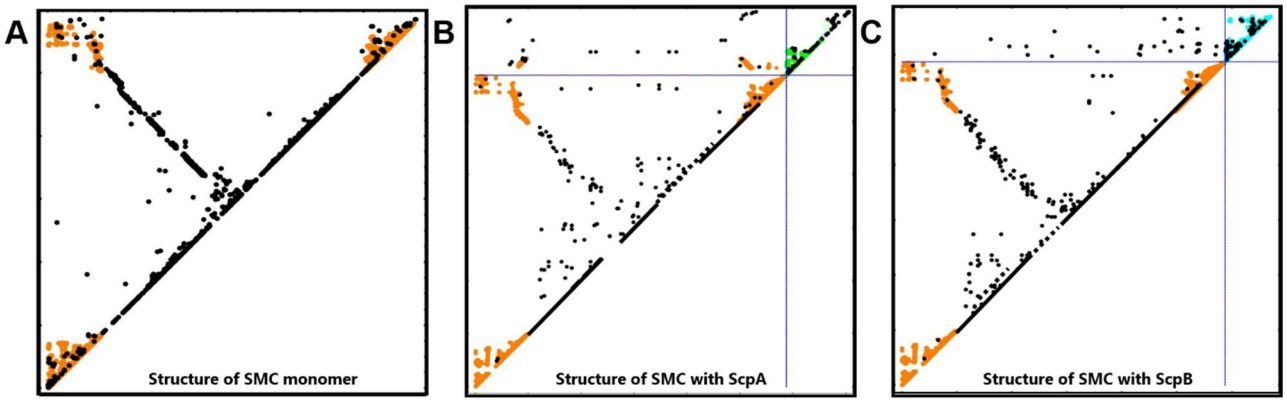

Fig. S2. Same organism classification provides co-evolutionary information for both intra- and inter-protein systems. (A) DCA (black) recapitulates available crystallographic data for a single SMC head domain as well as for (B) a single SMC head domain (orange) with crystallographically determined contacts from a single ScpA (green) protein complex and (C) a single SMC head domain (orange) with crystallographically determined contacts from a single ScpB (cyan) protein complex. Proteins are separated with dash blue line. Overlap of DCA contact prediction (top contacts appear in black) with experimental results obtained for the crystal structure of all systems studied is of sufficient accuracy.

In figure 2 in the main text, we present an overview of all crystallographic structural data for the condensin complex together with our co-evolutionary information. In order to quantify the accuracy of our predicted DCA contacts with crystallographic contacts, we have plotted the Positive Predictive Value (PPV) for the top 200 DCA contacts compared with the corresponding crystallographic contacts at a distance of 8Å. We define PPV in the following manner:

$$PPV = \frac{TP}{TP + FP}$$

where true positive (TP) and false positive (FP) refer to the fraction of contacts that is detected or isn't detected in the corresponding crystallographic contacts. This analysis was done for the subunits which have been used in figure 2: the ScpA subunit, the ScpB subunit and ScpA-ScpB inter-protein domain (shown in green, cyan and grey, respectively). PPV of the SMC head and SMC hinge domains (shown in orange and brown) have also been calculated and resulted in lower values. We note that these results are expected due to the dynamics of both SMC head and hinge domains as predicted by our DCA analysis (see Figures 4 and 5 in main text). For MD simulations, the top 15 DCA contacts have been used throughout the research. For these top DCA contacts a PPV value of 0.8 and higher were obtained."

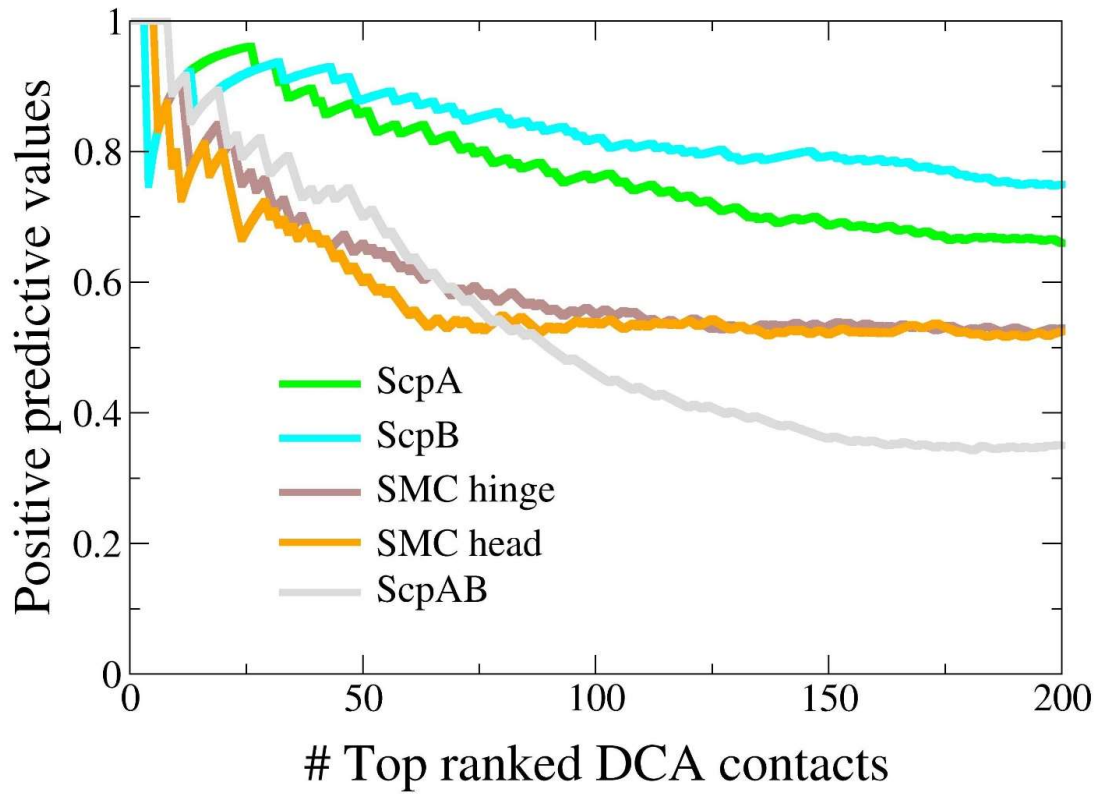

Fig. S3. Plot of the positive predictive value (PPV) as a function of the top 200 DCA contacts for the ScpA subunit (green), the ScpB subunit (cyan), the SMC hinge domain (brown), the SMC head domain (orange) and the ScpA-ScpB inter-protein domain (grey). Our results denote that top 15 DCA contacts used for MD simulations result in PPV value of 0.8 and higher.

References:

1. Finn RD, et al. (2010) The Pfam protein families database. *Nucleic Acids Res*. 38:D211-D222.
2. Berman HM, et al. (2000) The Protein Data Bank Nucleic Acids Research. *Nucleic Acids Res.* 28:235-242.
3. Ekeberg M, et al. (2013) Improved contact prediction in proteins: Using pseudolikelihoods to infer Potts models. *Phys. Rev. E*. 87:012707.
4. Morcos F, et al. (2011) Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proc. Natl. Acad. Sci*. 108(49):E1293-E1301.
5. Van Der Spoel D, et al. (2005) GROMACS: fast, flexible, and free. *J. Comput. Chem*. 26(16):1701-1718.
6. Noel JK, et al. (2010) SMOG@ ctbp: simplified deployment of structure-based models in GROMACS. *Nucleic Acids Res*. 38 (Web issue):W657-W661.
7. Whitford PC et al. (2009) An all-atom structure-based potential for proteins: Bridging minimal models with all-atom empirical forcefields. *Proteins: Struct., Funct., Bioinf.* 75(2):430-441.
8. Lammert H, Schug A, Onuchic JN (2009) Robustness and generalization of structure-based models for protein folding and function. *Proteins Struct Func Bioinf* 77(4):881- 891.
9. Dos Santos RN, et al. (2015) Dimeric interactions and complex formation using direct coevolutionary couplings. *Sci. Rep*. 5:13652.
10. Kim J-S et al. (2008) Crystal structure and domain characterization of ScpB from Mycobacterium tuberculosis. *Proteins* 71:1553-1556.

11. Lowe J, Cordell SC, Van Den Ent F (2001) Crystal Structure of The Smc Head Domain: An Abc ATPase with 900 Residues Antiparallel Coiled-Coil Inserted. *J. Mol. Biol.* 306:25.
12. Kamada K et al. (2017) Overall Shapes of the SMC-ScpAB Complex Are Determined by Balance between Constraint and Relaxation of Its Structural Parts. *Structure* 25:603-616.e4.
13. Lammens A, Hopfner KP (2010) Structural Basis for Adenylate Kinase Activity in ABC ATPases. *J. Mol. Biol.* 401*:*265-273.
14. Lammens A, Schele A, Hopfner K-P (2004) Structural biochemistry of ATP-driven dimerization and DNA-stimulated activation of SMC ATPases. *Curr. Biol.* 14*:*1778-1782.
15. Burmann F. et al. (2013) An asymmetric SMC-kleisin bridge in prokaryotic condensin. *Nat. Struct. Mol. Biol.* 20*:*371-379.
16. Soh YM et al. (2015) Molecular Basis for SMC Rod Formation and Its Dissolution upon DNA Binding. *Mol. Cell* 57*:*290-303.
17. Kamada K, Miyata M, Hirano T (2013) Molecular basis of SMC ATPase activation: role of internal structural changes of the regulatory subcomplex ScpAB. *Structure* 21:581-594.