Structure-based assembly of protein complexes in yeast

Patrick Aloy[1], Bettina Böttcher[1], Hugo Ceulemans[1], Christina Leutwein[2], Christian Mellwig[1], Susanne Fischer[1], Anne-Claude Gavin[2], Peer Bork[1], Giulio Superti-Furga[2], Luis Serrano[1] and Robert B. Russell[1]*

1. EMBL and 2. Cellzome AG
Meyerhofstrasse 1, 69117 Heidelberg, Germany

*Corresponding author:
Tel: +49 6221 387 473; FAX: +49 6221 387 517; Email: russell@embl.de

Supplementary Information

Complex purification, selection and classification

The protein complexes used in this structural analysis were selected from a set of 232 yeast complexes characterized by Tandem Affinity Purification (TAP)(*1*). We selected those complexes that yielded more than 0.1 mgs of protein from 2L of culture (typically between 0.1 to 3 mgs) and for which all components showed a similar relative intensity in coomassie-stained gels (i.e. similar stoichiometry). This lead to a subset of 102 complexes purified from 126 tagged proteins. For these we compiled information from the literature and Internet resources (*2, 3*) and manually classified them according to broad functional categories. The full list of complexes, and associated data are available at http://complexes.embl.de

EM collection and processing

We prepared fresh TAP purified complexes for electron microcopy by negative staining within 6 hours after the final purification step. After making carbon coated copper grids hydrophilic by glow discharge, we applied undiluted complex solution. After incubation for 30-60s we removed the complex solution with filter paper and washed the grid twice with water and 3

times with 2% uranyl acetate. We incubated the last drop of washing solution (2% uranyl acetate) on the grid for 30-90s before removal. To make the preparation of grids consistent all samples were prepared by the same person.

After drying, we examined grids by electron microscopy usually within 1 week using two CM120 Biotwin microscopes (operated at 100kV) equipped with either a LAB6- or tungsten-filament. We examined different areas of the grid at three different magnifications: an overview showing a complete mesh of the grid, a medium magnification showing the spread of the particles and a high magnification showing fine details of individual particles. For documentation we recorded micrographs on CCD camera (either with a wide- or small-angle, depending on the microscope). Two people performed a total of 126 analyses (i.e. sometimes more than one purification per complex). CCD images of the complexes were printed and evaluated by a single person to ensure that similar evaluation criteria were used. These were: homogeneity of the sample, aggregation of particles, recognizable and recurring shapes and common contaminations.

We grouped complexes into one of four classes according to probable suitability for image processing. Particles from class 1 samples were homogeneous in shape and size and suitable for immediate processing. Class 2 complexes had certain recurring particles with recognizable shapes and sizes, meaning that processing might be possible if a single type of particle could be selected manually. Class 3 complexes had distinct shapes but were too polymorphic for further processing, and those of class 4 showed particles with no distinctive shape.

For some complexes preliminary we performed image processing using IMAGIC V (*4*) by following standard procedures.

Sequence searches and modeling

Figure S1 illustrates the complete modeling procedure.

Modeling of individual components

For each component protein we searched for homologous proteins of known 3D structure. Sequences were masked for regions of low complexity (*5*) and

then used in PSI-Blast (*6*) searches (5 interactions with an E-value inclusion threshold of 0.001) against NRdb. We used the resulting profile to search against sequences of known 3D structure (PDB (*7*)) and from the Pfam database (*8*). For sequences with statistically significant matches to known 3D structures (E≤0.001) we aligned sequences using the same profile, and used modeler (*9*) to build homology models (based on the template with the highest sequence identity).

Modeling of interactions and complex assembly

We compared every pair of components in each complex to our database of interacting domains of known 3D structure (*10*). We first sought instances of interacting proteins belonging to the same sequences families (pfam) as those in the component pairs (inferred by sequence). Here we have the greatest confidence that the interaction will be similar, since comparison of different instances of the same interacting family pairs rarely differ (*11*). We ignored those families that have specifically evolved to bind different proteins as we found that the mode of interaction is seldom conserved (i.e. Pfam families: AAA, DEAD, LRR, HEAT, TPR, ank and rrm).

When interactions could not be inferred by sequence, we sought further interaction templates by searching for interacting partners sharing a similar structure despite no sequence similarity as known in the SCOP database (inferred by structure). We first identified SCOP classifications for domains in each component by Blast (E-value ≤0.01), and then checked if interactions between proteins with similar folds had ever been seen regardless of whether or not sequence similarity is apparent. However, we ignored interactions inferred by only fold similarities (i.e. different superfamilies in the same fold) since these are rarely associated with a similarity in interaction (*11*).

There is a clear correlation between sequence and interaction conservation (*11*), so where possible we used interactions inferred by sequence similarity in preference to those inferred by structure. Our previous study also shows that there is an interaction similarity twilight zone akin to that known for sequence & structure relationships (*12*). Below a minimum of about 25% between domains in different chains, interacting pairs can differ in

orientation. However, the vast majority of interacting pairs above this value interact in the same way, suggesting that many thousands of interactions can be modeled with confidence (see Figure S2). We are also able to consider even lower sequence identities if we restrict sequences to those found within the Pfam database, since we observed that interaction pairs involving the same Pfam domains rarely different in orientation (*12*). Obviously, whether any particular pair can be modeled depends on preservation of the interaction interface (*13*) or other details as to the biological reality of the interaction, such as a common cellular location or time of expression. We used our method for assessing interaction interface preservation (InterPReTS) (*13*) to score the interface similarity for any interaction inferred by sequence.

We built 3D models by superimposing the individual components to their equivalents in the template via STAMP (*14*). We applied the same approach for components lying in different complexes to model the instances of cross-talk discussed in the text.

Summary of the 3SOM algorithm for fitting X-ray to EM densities

3SOM is an approach for finding the best fit of atomic resolution structures into lower-resolution density maps through surface overlap maximization. It was inspired by the need to fit partial structures or homology models into very low resolution EM density maps, and our observation that manual fitting done by experts typically optimised surface, rather than full density overlap. EM densities are first filtered using and approximate Gaussian filter, and converted to boolean grids based on a threshold. Coordinate data (X-ray or homology models) are converted into grids of the same dimension, and both grids are converted to surface matrices. From these we then seek the transformation that best optimises the overlap of surface voxels. To optimise speed of searching, we first perform a fast-fitting procedure where we consider only key vectors that capture local surface information in the EM density. The best transformations found in this stage are then re-scored with a finer surface overlap measure: the ratio of the number of superimposed surface voxels to the total number of evaluated surface voxels of the EM map. For the fits discussed in the text, we always refer to the highest scoring transformation, unless otherwise stated.

(Hugo Ceulemans and Robert B. Russell, Fast fitting of atomic structures to low resolution electron density maps by surface overlap maximization. Manuscript submitted.)

Assessing confidence in predicted interactions

An important issue whenever a new prediction method appears is to assess the accuracy of its results. A difficulty with assessing the accuracy of any interaction identification method (either experimental or computational) is the lack of an established "gold standard" containing both positive examples, or proteins that definitely interact, and more importantly negative examples, or proteins definitely known not to interact. Predictions based on 3D structures suffer from an additional problem that the current database contains few examples to test the possibility of predicting one complex three or more proteins using binary interactions known from other 3D structures. In the absence of a benchmark set, we have decided to score the predictions as to their confidence in different ways, and to classify them accordingly in the text, and on the accompanying web site.

Predictions of interactions within complexes

Here interactions between homologues of the components in the complexes have been seen before, and thus represent the best hypothesis for the molecular details of the interactions. However, since homology is not always associated with a similarity in function or interaction, it is important to know the degree of homology required to be confident that interactions will be similar. As discussed above, we have greatest confidence in interactions predicted with sequence identities above 25%, or lying in the same Pfam family (see Figure S2). For more remote similarities, there is a "twilight zone" where interactions may or may not be similar (akin to sequence/structure similarity). When there is only one interaction of a particular type in a complex, and sequence similarities are above 25% (or in the same Pfam family), then we feel our predictions are most likely to be correct. We have given a breakdown of the interactions we have made as to how many fit into the category above. Of 196 predictions, 127 meet these criteria (high confidence set). It is important to emphasise that this does *not* mean that the

other interactions are all incorrect; instead they lie in a twilight zone containing correct predictions and artefacts.

Assembly of multiple components of the same type in one complex

We found several examples where there are many possible interactions of the same type in one complex (e.g. the exosome and the CCT complex). Here one is faced with an *assembly* problem. For example, in CCT there are eight *different* homologous copies the CCT subunits that are known to form a complex that probably resembles the thermosome. We confidently predict that the CCT subunit interactions will resemble those of the thermosome subunits (i.e. the interfaces are correct), but we do not know the precise ordering. We do not attempt to address this difficult problem here. It is important to understand that this is different from general predictions of interaction.

Cross-talk: interactions between complexes

These interactions are more difficult to validate. We have the most confidence for those where our prediction is confirmed by another interaction discovery method (such as the two-hybrid system or TAP). This is analogous to why we believe interactions predicted within complexes to be more accurate (above). For the others, we can measure the strength of the prediction by our method to assess protein-interface similarity (*13*), the degree of sequence similarity, or the similarity in functional classification. Note that here all the predicted interactions lie in the same Pfam family, thus putting them in the high-confidence set above.

In summary, of 70 cross-talk instances (Figure 3):

4 are validated by experiment and interface similarity
14 are validated by experiment only
6 are validated by interface similarity only
30 occur between complexes of the same broad functional class
(42 if one allows those between transcription and RNA synthesis)

There are only 27 (or 19) where we have little confidence in the interaction apart from homology to a known structure. Some of these might be artefacts, such as that between mannosyl transferase and the ribosome, and probably require further analysis. Note that this is also the case for the genome-scale interaction discovery experiments: they find many tantalizing, new interactions that are readily believable, but they also find others that look more dubious, and which probably will not stand up under closer experimental scrutiny.

Structural templates used to build the models presented

Details such as sequence identities, domains in Pfam and many others are available on the associated web site (http://complexes.embl.de).

Exosome (Fig 2a)

Polynucleotide Phosphorylase (PNPase); 1e3p (*15*)

RNA polymerase II (Fig 2b)

RNA polymerase II; 1i3q (*16*)

Full coordinates for RPB4 and RPB7; 1go3 (*17*)

Interaction between SPT5/RNA pol II; 2eif (*18*)

CCT and phosphoducin2/VID27 (Figs 2c,d)

CCT; 1a6d (*19*)

phosphoducin2/VID27; 1b9x (*20*)

Ski complex (Fig 2f)

Ski2 and Ski3; 1gp2 (*21*)

Ski2 and Ski8; 1e96 (*22*)

Cross talk between translation initiation complexes

TOA1, TOA2, TFIID and DNA ;1ytf(*23*)

TFIID and SUA7; 1c9b(*24*)

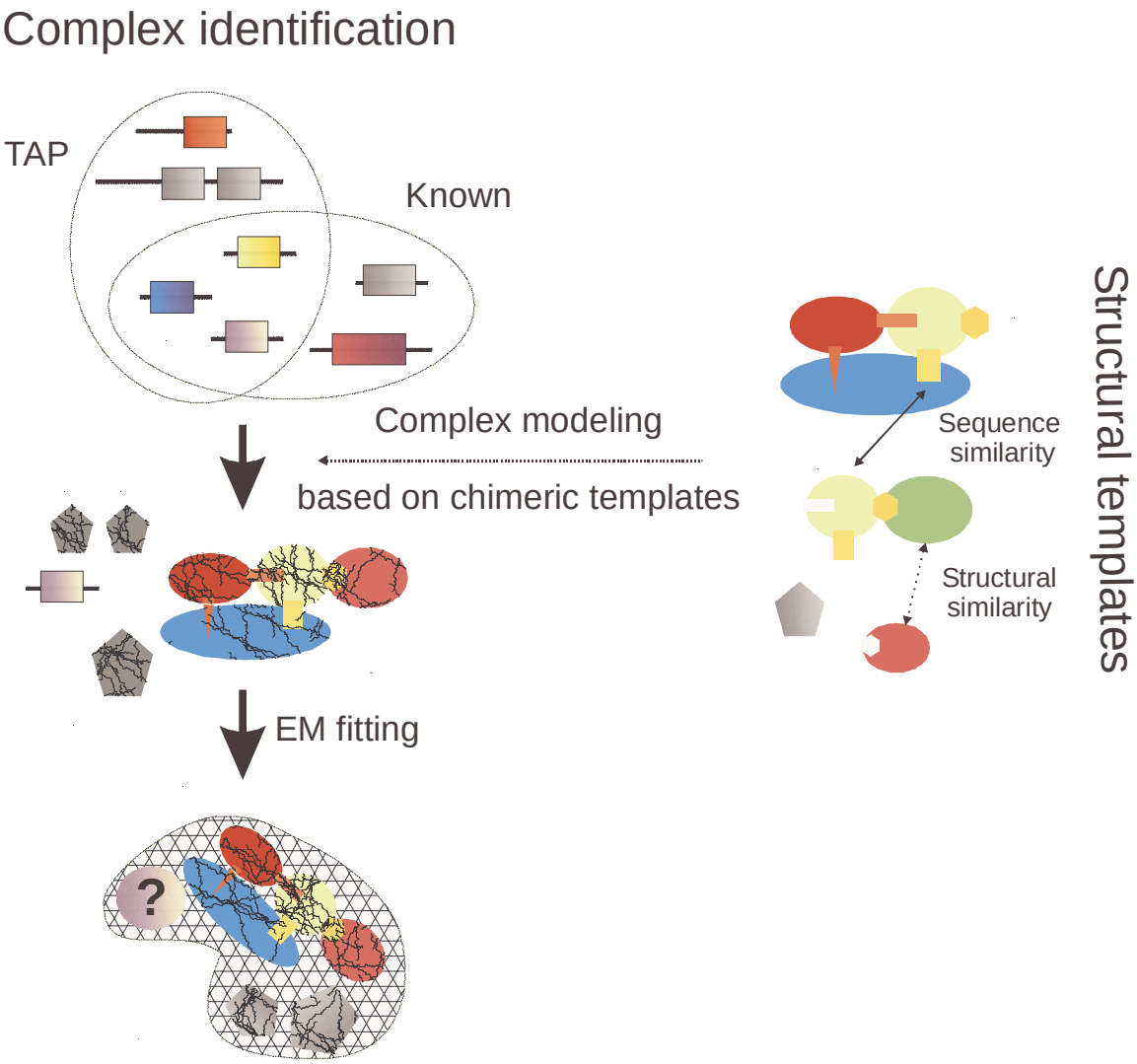

Figure S1 The complex modeling procedure: similar colors and shapes indicate sequence and structure similarity respectively. Cracked shapes show homology models.

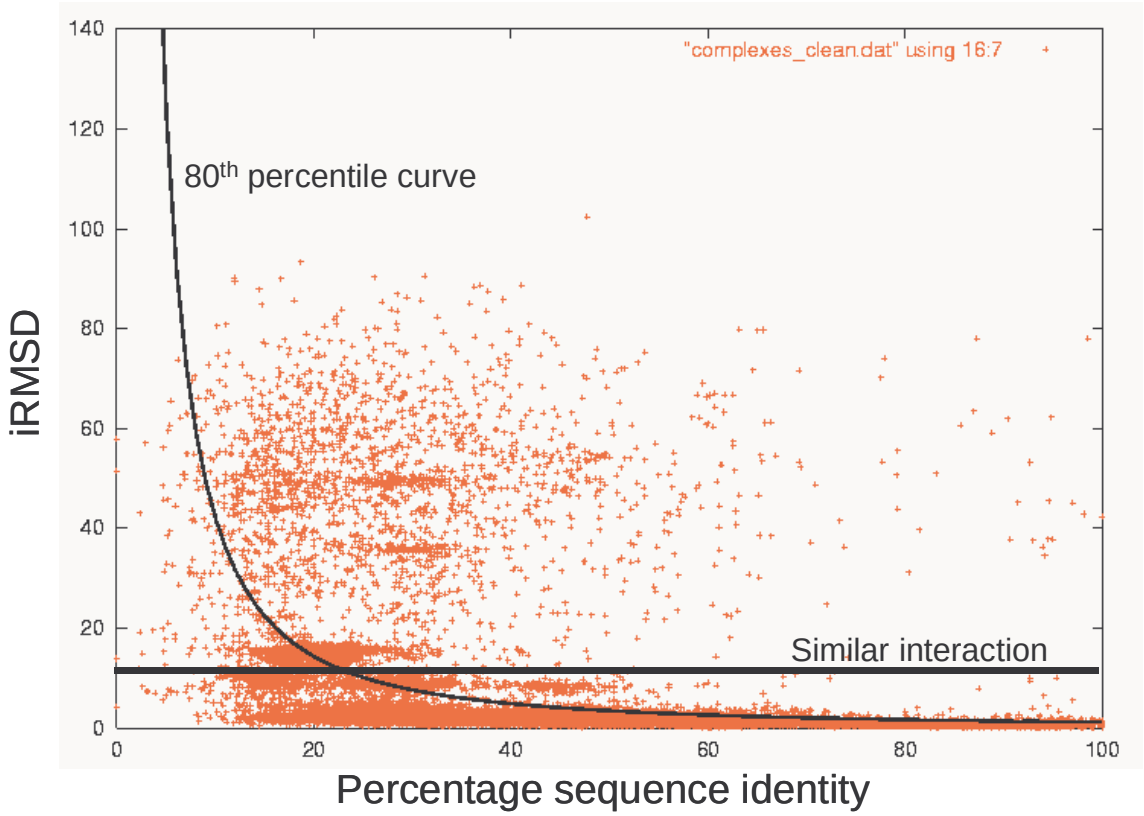

Figure S2 Plot showing interaction similarity (iRMSD) versus percentage sequence identity for all the available pairs of interacting domains with known 3D structure. The curve shows the 80th percentile (i.e. 80% of the data lies below the curve), and points below the line (iRMSD = 10 Å) are similar in interaction.

References

1. A. C. Gavin *et al.*, *Nature* **415**, 141 (2002).
2. http://www.ncbi.nlm.nih.gov/Entrez/.
3. H. W. Mewes *et al.*, *Nucleic Acids Res* **30**, 31 (Jan 1, 2002).
4. M. van Heel, G. Harauz, E. V. Orlova, R. Schmidt, M. Schatz, *J Struct Biol* **116**, 17 (1996).
5. F. S. Wooton JC, *Methods Enzymol* **266**, 554 (1996).
6. S. F. Altschul *et al.*, *Nucleic Acids Res* **25**, 3389 (1997).
7. H. M. Berman *et al.*, *Nucleic Acids Res* **28**, 235 (2000).
8. A. Bateman *et al.*, *Nucleic Acids Res* **30**, 276 (2002).
9. A. Sali, T. L. Blundell, *J Mol Biol* **234**, 779 (1993).
10. http://3did.embl.de.
11. P. Aloy, H. Ceulemans, A. Stark, R. B. Russell, *J Mol Biol* **332**, 989 (2003).
12. R. Schneider, C. Sander, *Nucleic Acids Res* **24**, 201 (Jan 1, 1996).
13. P. Aloy, R. B. Russell, *Proc. Natl. Acad. Sci. USA* **99**, 5896 (2002).
14. R. B. Russell, G. J. Barton, *Proteins* **14**, 309 (1992).
15. M. F. Symmons, G. H. Jones, B. F. Luisi, *Structure Fold Des* **8**, 1215 (2000).
16. P. Cramer *et al.*, *Science* **288**, 640 (Apr 28, 2000).
17. F. Todone, P. Brick, F. Werner, R. O. Weinzierl, S. Onesti, *Mol Cell* **8**, 1137 (Nov, 2001).
18. K. K. Kim, L. W. Hung, H. Yokota, R. Kim, S. H. Kim, *Proc Natl Acad Sci U S A* **95**, 10419 (Sep 1, 1998).
19. L. Ditzel *et al.*, *Cell* **93**, 125 (Apr 3, 1998).
20. R. Gaudet, J. R. Savage, J. N. McLaughlin, B. M. Willardson, P. B. Sigler, *Mol Cell* **3**, 649 (May, 1999).
21. M. A. Wall *et al.*, *Cell* **83**, 1047 (Dec 15, 1995).
22. K. Lapouge *et al.*, *Mol Cell* **6**, 899 (Oct, 2000).
23. S. Tan, Y. Hunziker, D. F. Sargent, T. J. Richmond, *Nature* **381**, 127 (May 9, 1996).
24. F. T. Tsai, P. B. Sigler, *Embo J* **19**, 25 (Jan 4, 2000).