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] Robert B. Russell[1]*

Images of entire cells are preceding atomic structures of the separate molecular machines that they contain. The resulting gap in knowledge can be partly bridged by protein-protein interactions, bioinformatics, and electron microscopy. Here we use interactions of known three-dimensional structure to model a large set of yeast complexes, which we also screen by electron microscopy. For 54 of 102 complexes, we obtain at least partial models of interacting subunits. For 29, including the exosome, the chaperonin containing TCP-1, a 3′-messenger RNA degradation complex, and RNA polymerase II, the process suggests atomic details not easily seen by homology, involving the combination of two or more known structures. We also consider interactions between complexes (cross-talk) and use these to construct a structure-based network of molecular machines in the cell.

Cell and structural biology share the common goal of understanding large, complex biological entities at the highest possible detail. Although different in outlook, the distinction diminishes as techniques improve. Cell biologists can now see structures like the nuclear pore (*1*), or even whole cells (*2*), at resolutions approaching 3 nm. Structural biologists, once restricted for technical reasons to small macromolecules, are now solving atomic resolution structures for large molecular machines like the ribosome (*3*) or RNA polymerases (*4*). In spite of this blurring distinction, technical problems will delay atomic resolution structures of large cellular entities for several years. Microscopy has difficulties reaching this resolution, and expression and crystallization problems still slow x-ray crystallography for large complexes. The result is an information gap between low-resolution images of large cellular entities and atomic structures for the macromolecules that they contain.

Recent developments in functional genomics provide possibilities for bridging this gap. Sequenced genomes give a complete list of macromolecules contained in the cell. Genome-scale interaction discovery approaches, like the two-hybrid system (*5*) or affinity purifications (*6*), have uncovered details of the cell network, although without critical molecular details: what interacts with what, but not how. These details can sometimes come from similarities to interacting proteins of known three-dimensional (3D) structure. Here, we investigate a large set of yeast interactions using structures to give the most complete view currently possible of complexes and their interrelationships. We screen complexes using electron microscopy (EM) and use low-resolution images to help assemble and validate models. We also predict links between complexes and provide a higher order, structure-based network of connected molecular machines within the cell.

We began with a large set of yeast protein complexes identified by tandem affinity purification (TAP) (*6*). From an initial 232 complexes, we selected 102 (126 purifications in total) that yielded samples most promising for EM from analysis of gels and protein concentrations. We prepared negatively stained (uranyl acetate) grids and collected EM data with a charge-coupled device. We used literature and Internet resources to classify the complexes into broad functional groups; for those known previously, we noted missing or unexpected components. [Further experimental details are given in (*7*)].

We built interactions between complex components by their similarity to interacting proteins of known structure. We assigned domains and built 3D models for as many components as possible (*7*). We then searched for suitable structures (templates) on which to model interactions between complex components (Fig. 1). We first inferred interaction models by finding component pairs similar in sequence to interacting regions from a single known structure. We inferred further models from pairs of components showing a similarity to structural domains in the structural classification of proteins (SCOP) (*8*). Knowledge of structure allows domains to be grouped in the absence of sequence similarity, meaning that we could consider interac-

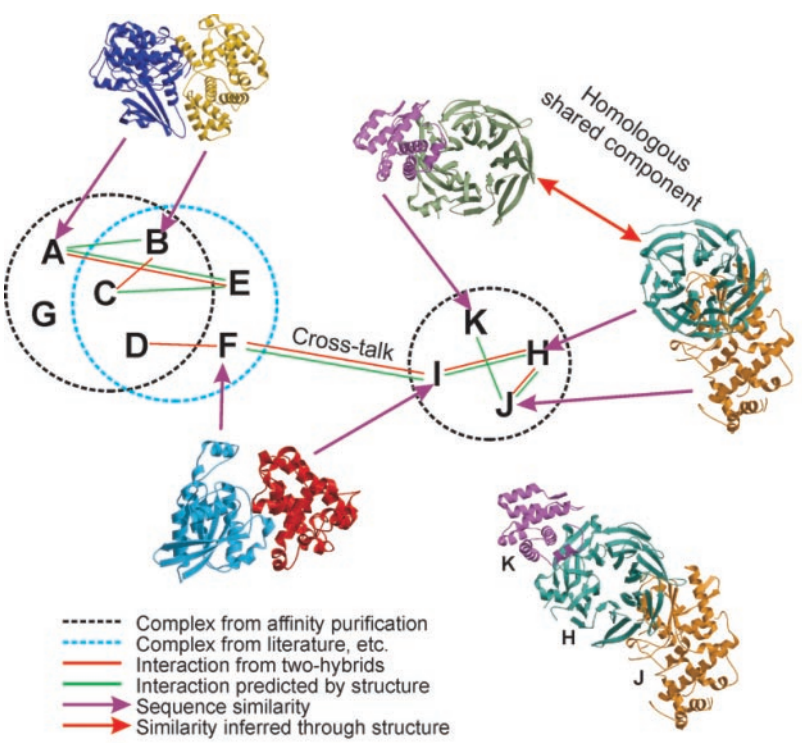

Fig. 1. Illustration of the methods and concepts used. How predictions are made within complexes (circles) and between them (cross-talk). Bottom right shows two binary interactions combined into a three-component model.

[1]European Molecular Biology Laboratory, Structural and Computational Biology Programme, [2]Cellzome AG Meyerhofstrasse 1, 69117 Heidelberg, Germany.

*To whom correspondence should be addressed. E-mail: russell@embl.de

tions between proteins adopting folds like those found in the complex, despite no direct sequence similarity (i.e., inferred by structure) (Fig. 1). We removed pairs of domains likely to interact differently: those lying in different superfamilies within the same fold or those known to be promiscuous with interaction partners (e.g., armadillo repeats or ankyrin repeats) (*9*). We used interactions inferred by sequence in preference to those inferred by structure, because these are most likely to be similar in interaction orientation (*9*). We repeated this process for components in different complexes to identify instances of cross-talk, and cross-referenced interaction data from other experiments (*10*). We define high-confidence models as those occurring between the same homologous families and/or having greater than 25% sequence identity. Others lie within a twilight zone where interaction orientations may or may not be similar (*9*).

We then assembled the modeled interaction pairs into the most complete model for each complex. The best models come when all components in a TAP-purified complex are similar to proteins in a single structure. This is comparatively rare: Even complexes such as RNA polymerase II contained additional components absent from known structures, and many had only separate modeled pairs. To build larger complexes, it was often necessary to combine different interaction templates, for which we used homologous proteins present in multiple structures (homologous shared components) as links to construct a single chimeric model. When more than one predicted interaction of the same type is present in a single complex, there is an additional problem related to selecting the best of several alternative models, although we did not attempt to address this here. When we had both an EM reconstruction and models of sufficient size, we searched for best fits using a surface overlap maximization procedure (*7*, *11*).

Assessing the accuracy of genome-scale interaction discovery approaches is a critical issue. The lack of a test set of adequate size makes standard assessment measures like specificity and sensitivity difficult to compute. There is currently only one complex of known 3D structure to benchmark our approach, i.e., involving three or more proteins and two or more interaction types, where the interactions have been seen in other structures: CDK6 in complex with p18(INK4c) and a viral cyclin (*12*) (which would indeed be modeled correctly).

However, confidence in models and predicted interactions can come from other sources. The greatest comes when the proteins are known to be physically associated: those within the same TAP-identified complex or involved in cross-talk interactions validated by other experiments. Here, the model is the best currently possible inference of molecular details. Confidence also comes from a high degree of sequence similarity (*9*), preservation of residues at the interaction interface (*13*), or similarities in broad functional class. With or without support, it is likely that some interactions are probably incorrect, related to artifacts of the experimental and computational techniques or to physically possible interactions between proteins that never meet in the cell.

A majority of components (408 of 634) contain at least one domain for which 3D structures are known or modelable. A total of 196 interactions within the complexes could also be modeled, leading to quaternary structures with varying degrees of completeness (Table 1). We obtained nearly complete models for 42 complexes and partial models for another 12, but could only model separate components for most of the remainder. For 27 complexes, our procedure assembled more components than was previously possible with any single structure (e.g., RNA polymerase II and Ski).

There are several complexes containing multiple components that each shows good homology to a known structure, despite no suitable template on which to model any interaction. There are 30 pairs among these where the interaction is also supported by yeast two-hybrid experiments. These are excellent candidates for techniques to study interaction details, such as nuclear magnetic resonance (*14*) or docking (*15*).

We classified EM grids into three categories on the basis of visual inspection: 6 were judged suitable for reconstruction, 9 had minor impurities that might be improved by alterations to the purification procedure, and 111 were deemed to require more extensive purifications. Four are discussed below.

A detailed summary of the exosome was published previously (*16*). Similarities among nine components with domains of polyribonucleotide phosphorylase (PNPase) (*17*) produced a model that fit well into the EM density (Fig. 2A), but that left a lump of extra density that might correspond to components lacking equivalents in PNPase (Rrp44, Rrp6, or Ski7).

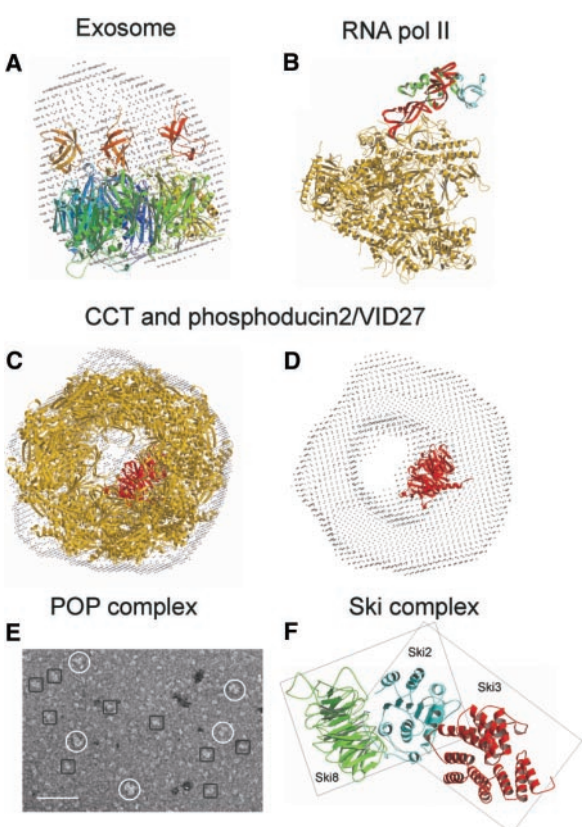

Fig. 2. Models of yeast complexes. (**A**) Exosome model on PNPase fit into EM map. (**B**) RNA polymerase II with RPB4 (green)/RPB7 (red) built on *Methanococcus jannaschii* equivalents, and SPT5/pol II (cyan) built with IF5A. (**C** and **D**) Views of CCT (gold) and phosphoducin 2/VID27 (red) fit into EM map. (**E**) Micrograph of POP complex, with particle types highlighted. (**F**) Ski complex built by combination of two complexes.

Table 1. "Nearly complete model" means that for two-thirds of the complex components, at least one domain and one interaction per component can be modeled. "Most components" means that models are possible for two-thirds of separate components.

Overview of structural information	
Nearly complete model	42
Most components and some interactions	12
Most components	20
Some components	25
No structure	3

The 4.1 Å resolution yeast structure of the RNA polymerase II primary transcription complex contains 12 subunits (*18*), although affinity purifications revealed an additional two: TFG2 and SPT5, both involved in transcription initiation, which we suspect are genuine because pol II recognizes the preinitiation complex (*19*). TFG2 could be modeled with confidence, although there was no suitable template to model an interaction with pol II. We built a model of SPT5 on the SH3 domainlike structure of translation initiator factor 5A. In this structure, the SH3 domain is in contact with an OB fold containing protein homologous to pol II component RPB7, thus providing a template to join SPT5 (inferred by structure) (Fig. 2B). The combined model shows some clashes easily accounted for by minor conformational changes or discrepancies in the model (or, of course, the model could be wrong). The sample was also of sufficient quality to give a preliminary EM reconstruction, which confirms the approximate size and shape of RNA pol II, though without apparent density for SPT5/RPB7, possibly due to the more transient nature of this interaction or disruption owing to the use of uranyl acetate at low pH.

Affinity purifications identified a complex consisting of the chaperonin CCT (*20*), together with a phosducin (PLP2) and G protein γ homolog (VID27), which are thought to act as inhibitors (*21*). All 10 components are similar to known structures and can be assembled into two subcomplexes (Fig. 2, C and D). The eight CCT components are similar to each other, and to subunits of the thermosome (*22*), which has been used to construct a hexadecameric model [two of each CCT component (*20*)]. A structure of a phosducin interacting with a G-γ homolog is also known (*23*), yielding a high-confidence interaction model, although there is currently no structure on which to model any interaction of these with proteins similar to CCT. The homogenous sample gave an EM reconstruction into which the CCT model fit best, as shown in Fig. 2C, with several alternative solutions owing to internal symmetry. All fits left extra density in the interior, into which the small PLP2/VID27 model fit best even when the entire map was considered (i.e., including that accounted for by CCT) (Fig. 2D). Although it is not possible to place the structures precisely or to know which of the symmetric fits of CCT is correct, the overall model supports the stoichiometry of one PLP2/VID27 molecule per CCT hexadecamer (*21*). It also suggests coarse details of the interaction: e.g., that PLP2/VID27 probably interacts with the N terminus of the CCT subunits.

The RNase P (POP) complex is involved in processing RNA and tRNA in the nucleus (*24*) and consists of eight apparently nonhomologous proteins. EM suggested that the sample contained at least two complexes of different sizes (Fig. 2E), the larger being about 100 times bigger than expected from a 1:1 stoichiometry of the components (white circles). Some components contain ribonuclease domains, although only one domain from one protein is similar to any known structure. Here, a preliminary view of the structure could well be possible, but virtually no structural information comes from homology.

The Ski complex is involved in cytoplasmic 3′-mRNA degradation (*25*) and includes proteins containing two P-loop adenosine triphosphatase (ATPase) domains (Ski2), Leucine-rich repeats (Ski3),

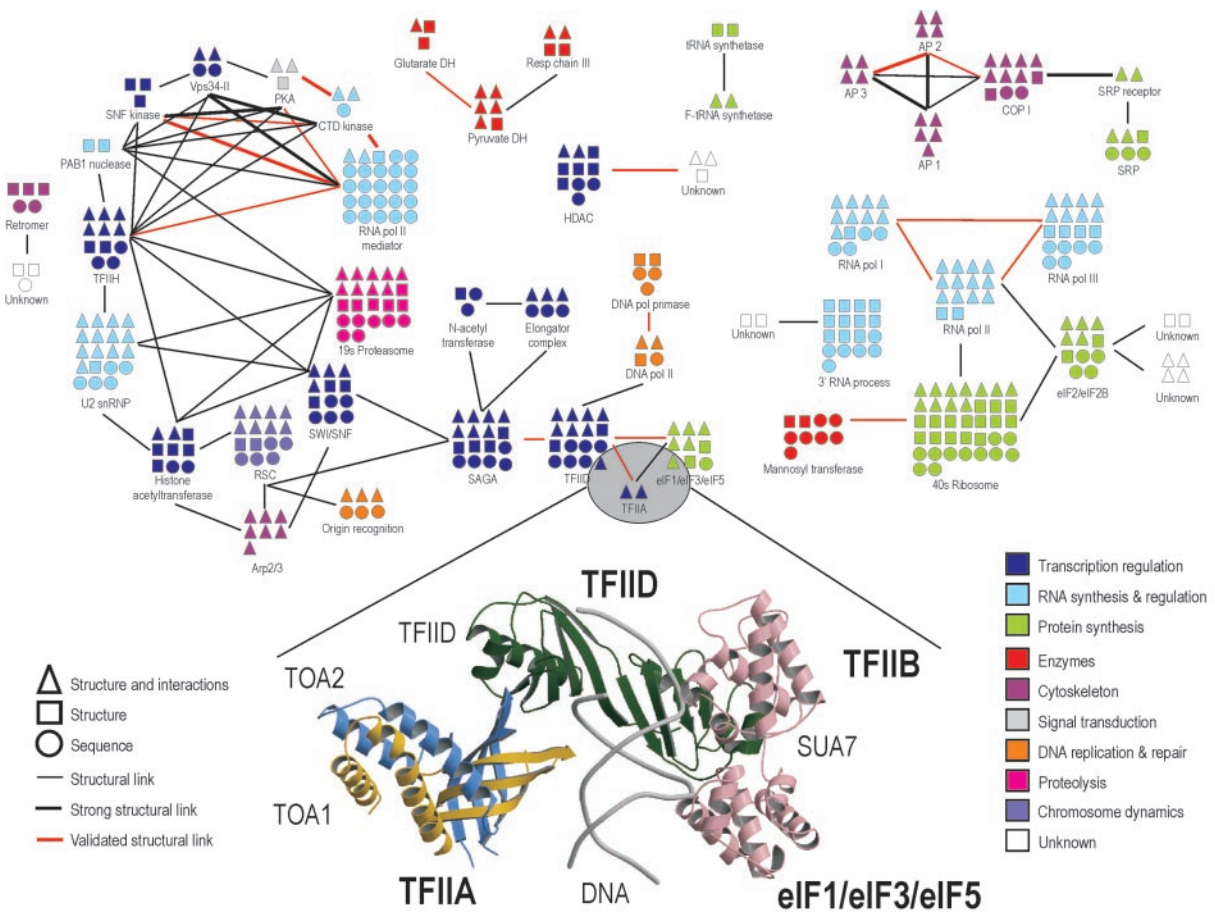

Fig. 3. Cross-talk between complexes. (**Top**) Triangles show components with at least one modelable structure and interaction; squares, structure only; circles, others. Lines show predicted interactions: thick lines imply a conserved interaction interface (*12*); red, those supported by experiment. (**Bottom**) Expanded view of cross-talk between transcription complexes built on by a combination of two complexes.

and a G protein γ homolog (Ski8). TAP did not give a sample suitable for an EM reconstruction, but comparison to known structures yielded a model based on two separate templates containing P-loop ATPases in contact with leucine-rich repeats and G protein γ subunits (Fig. 2F). The interactions were modeled on templates with little or no sequence similarity (inferred by structure). Nevertheless, the model contains only minor clashes and provides a plausible mode of interaction. Ski2 and Ski3 are long proteins (1287 and 1432 amino acids, respectively) and contain regions lacking homology to other sequences or structures. The model suggests shorter regions mediating the interaction.

Most of the complexes we considered consist of proteins that are bound together for most of their existence and that have little functional significance in isolation (*26*). However, the complexity of the cellular process requires a still higher level of organization involving transient interactions (cross-talk) between complexes (*27*). Instances in which it is possible to model structures for the interactions are good candidates to study how such processes are mediated.

There are 246 potential pairs of interacting components that can be modeled on known 3D structures, leading to 70 instances of cross-talk involving 46 complexes. Evidence from numerous additional sources gives further confidence to several of these: 18 are supported by experiments, 30 are between complexes of the same functional class, and 10 show preservation of the interaction interface (*13*). Some 27 lack additional evidence and require further study to confirm or dismiss.

These interactions lead to a network of complexes (Fig. 3, top) where connections signify a possible means to model the molecular basis of the interaction. It provides a more realistic picture of cell structures than those derived only from interaction data. For example, many of the individual molecular machines shown in our network as single functional units appear as highly connected subnetworks in those described previously (*28*).

The cross-talk places complexes into a higher order network, and they appear as static entities. However, complexes are often dynamic and can include different components depending on cellular conditions. The same components can also be used repeatedly as modules to perform related functions (*27*). Several of the cross-talk interactions are the result of these phenomena in which proteins (or their close homologs) are present in multiple complexes. For example, the RNA polymerases are variations on a common theme, with different proteins complementing a common core. There are 33 components present in more than one TAP-identified complex, either singly (e.g., dihydrolipoamide dehydrogenase in the 2-oxoglutarate and pyruvate dehydrogenase complexes) or together as functional modules (e.g., ARP9/ARP7 in SWI/SNF and RSC).

Other cross-talk occurs between protein families that are unique to the separate complexes and provide possible structural models for their interactions. Transcription complexes TFIIA, B, and D are particularly illustrative examples. The structure of yeast TFIIA bound to the TATA box–binding protein (TBP), the principal component of the TFIID complex, is known (*29*). The structure of human TBP-TFIIB complex (*30*) is also known, and the yeast equivalents, TBP and SUA7, share sufficient sequence identity with the two human proteins (81 and 28%) to build a high-confidence interaction model by homology. This leads to a good model for TFIIA, B, and D with DNA (Fig. 3, bottom). SUA7 is also an unexpected component of the cytoplasmic translation initiation eIF1/eIF3/eIF5 complex involved in ribosome assembly (*31*), allowing (if true) an intriguing link to be modeled between the transcription and translation machinery.

We have shown that a combination of 3D structure and protein-interaction data can already provide a partial view of complex cellular structures. The predicted details of how proteins interact and assemble into complexes generate many hypotheses to be tested further. For example, the predicted domains and interface residues in the Ski complex or the proposed site of CCT-inhibitor interaction are readily testable by mutagenesis. Even when structures are not available, density that is unaccounted for in an EM reconstruction can suggest an approximate location, as for the exosome. Such models might also be used to probe larger EM images to locate complexes in the cell (*32*). The structure-based network derived from cross-talk between complexes provides a more realistic picture to complement lower resolution images of cell structures than those derived blindly from interaction data, because it suggests molecular details for how they are mediated.

Of course, the picture is still far from complete and there are numerous new challenges. Complex identification techniques do not always provide samples suitable for structural studies, and efforts to improve sample quality will yield more EM reconstructions or even allow the possibility of x-ray studies. Detailed studies are also required to assess the accuracy of predictions. For example, a carefully derived benchmark set of interacting proteins or domains would provide a guide to interpret the accuracies for all interaction discovery methods. The computational challenges are equally daunting, requiring nothing less than the emergence of a new field for structure prediction that must cope with individual proteins, complexes, and the sophisticated dynamic network that connects them. The structure-based network derived here provides a useful initial framework for further studies. Its beauty is that the whole is greater than the sum of its parts: Each new structure can help to understand multiple interactions. The complex predictions and the associated network will thus improve exponentially as the numbers of structures and interactions increase, providing an ever more complete molecular anatomy of the cell.

References and Notes

1. Q. Yang, M. P. Rout, C. W. Akey, *Mol. Cell* **1**, 223 (1998).
2. O. Medalia *et al.*, *Science* **298**, 1209 (2002).
3. N. Ban, P. Nissen, J. Hansen, P. B. Moore, T. A. Steitz, *Science* **289**, 905 (2000).
4. C. L. Poglitsch *et al.*, *Cell* **98**, 791 (1999).
5. P. Uetz *et al.*, *Nature* **403**, 623 (2000).
6. A. C. Gavin *et al.*, *Nature* **415**, 141 (2002).
7. Materials and methods are available as supporting material on *Science* Online.
8. L. Lo Conte, S. E. Brenner, T. J. Hubbard, C. Chothia, A. G. Murzin, *Nucleic Acids Res.* **30**, 264 (2002).
9. P. Aloy, H. Ceulemans, A. Stark, R. B. Russell, *J. Mol. Biol.* **332**, 989 (2003).
10. H. W. Mewes *et al.*, *Nucleic Acids Res.* **30**, 31 (2002).
11. H. Ceulemans, R. B. Russell, *J. Mol. Biol.*, in press.
12. P. D. Jeffrey, L. Tong, N. P. Pavletich, *Genes Dev.* **14**, 3115 (2000).
13. P. Aloy, R. B. Russell, *Proc. Natl. Acad. Sci. U.S.A.* **99**, 5896 (2002).
14. A. Douangamath *et al.*, *Mol. Cell* **10**, 1007 (2002).
15. J. Janin *et al.*, *Proteins* **52**, 2 (2003).
16. P. Aloy *et al.*, *EMBO Rep.* **3**, 628 (2002).
17. M. F. Symmons, G. H. Jones, B. F. Luisi, *Struct. Fold Des.* **8**, 1215 (2000).
18. D. A. Bushnell, R. D. Kornberg, *Proc. Natl. Acad. Sci. U.S.A.* **100**, 6969 (2003).
19. G. Orphanides, D. Reinberg, *Cell* **108**, 439 (2002).
20. J. M. Valpuesta, J. Martin-Benito, P. Gomez-Puertas, J. L. Carrascosa, K. R. Willison, *FEBS Lett.* **529**, 11 (2002).
21. J. N. McLaughlin *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **99**, 7962 (2002).
22. L. Ditzel *et al.*, *Cell* **93**, 125 (1998).
23. R. Gaudet, J. R. Savage, J. N. McLaughlin, B. M. Willardson, P. B. Sigler, *Mol. Cell* **3**, 649 (1999).
24. J. R. Chamberlain, Y. Lee, W. S. Lane, D. R. Engelke, *Genes Dev.* **12**, 1678 (1998).
25. J. T. Brown, X. Bai, A. W. Johnson, *RNA* **6**, 449 (2000).
26. P. Aloy, R. B. Russell, *Trends Biochem. Sci.* **27**, 633 (2002).
27. A. C. Gavin, G. Superti-Furga, *Curr. Opin. Chem. Biol.* **7**, 21 (2003).
28. S. Maslov, K. Sneppen, *Science* **296**, 910 (2002).
29. S. Tan, Y. Hunziker, D. F. Sargent, T. J. Richmond, *Nature* **381**, 127 (1996).
30. F. T. Tsai, P. B. Sigler, *EMBO J.* **19**, 25 (2000).
31. T. V. Pestova *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **98**, 7029 (2001).
32. A. Sali, R. Glaeser, T. Earnest, W. Baumeister, *Nature* **422**, 216 (2003).

33. We thank A. Musacchio (European Institute of Oncology, Italy) for the functional classification of complexes and K. Leonard and E. Conti (EMBL) for help during the EM screen.

Supporting Online Material

www.sciencemag.org/cgi/content/full/303/5666/2026/DC1
Materials and Methods
Figs. S1 and S2
References

16 October 2003; accepted 26 January 2004