Structure determination of macromolecular complexes by experiment and computation

Frank Alber*, Narayanan Eswar*, Andrej Sali[$]

Departments of Biopharmaceutical Sciences and Pharmaceutical Chemistry, and California Institute for Quantitative Biomedical Research, Mission Bay Genentech Hall, Suite N472D, 600 16th Street, University of California at San Francisco, San Francisco, CA 94143-2240, USA.
tel +1 (415) 514-4227, fax +1 (415) 514-4231, email: sali@salilab.org, web: http://salilab.org, [$]corresponding author, * both authors contributed equally.

Summary

Structures of macromolecular complexes are necessary for a mechanistic description of biochemical and cellular processes. These structures can be defined by experimental methods such as x-ray crystallography, nuclear magnetic resonance spectroscopy, electron microscopy, cross-linking, and footprinting, as well as computational approaches such as protein structure prediction, docking, and bioinformatics analysis of subunit sequences and structures. Technical advances in instrumentation and computer hardware have expanded their applicability and helped close the resolution gaps between different methods. We argue for computing three-dimensional models of a given protein assembly that are consistent with all available information about its composition and structure, which may vary greatly in terms of its source, accuracy, and resolution. For illustration, we describe an approach that combines comparative protein structure modeling of assembly subunits with their docking into low-resolution electron density maps obtained by electron microscopy and its application to the yeast and *E. coli* ribosomes.

1. Introduction

The function of a protein is defined by its interactions with other molecules in its environment. The interactions can be either transient, such as protein-protein interactions involved in intra-cellular signaling, or relatively stable, such as the protein-protein and protein-RNA interactions in ribosomes. A structural description of these interactions is an important step toward understanding the mecha-

nisms of biochemical, cellular, and higher order biological processes. There is a need to integrate structural information gathered at multiple levels of the biological hierarchy – from atoms to cells – into a common framework. Recent developments in several experimental and computational techniques allow structural biology to shift its focus from the structures of the individual proteins to larger assemblies (Sali et al., 2003; Baumeister, 2002).

Macromolecular assemblies vary widely in their functions and sizes (Alberts, 1998; Goto et al., 2002; Grakoui et al., 1999; Courey, 2001; Noji and Yoshida, 2001). They play crucial roles in most cellular processes, and are often depicted as molecular machines (Alberts, 1998). This metaphor accurately captures many of their characteristic features, such as modularity, complexity, cyclic functions, and energy consumption (Nogales and Grigorieff, 2001). For instance, the nuclear pore complex, a 50-100 MD protein assembly, regulates and controls the traffic of macromolecules through the nuclear envelope (Rout et al., 2000); the ribosome is responsible for protein biosynthesis; the RNA polymerase catalyzes the formation of RNA (Murakami and Darst, 2003); and the ATP synthase catalyzes the formation of ATP (Noji and Yoshida, 2001). Macromolecular assemblies are also involved in transcription control (*ie*, IFNβ enhanceosome) (Courey, 2001; Nogales, 2000), regulation of cellular transport (*ie*, microtubulines in complex with molecular motors myosin or kinesin) (Vale, 2003; Goldstein and Yang, 2000; Vale and Milligan, 2000), and are crucial components in neuronal signaling (*eg*, the postsynaptic density complexes) (Gomperts, 1996).

The estimation of the total number of macromolecular complexes in a proteome is a non-trivial task. This difficulty can be partly ascribed to the multitude of component types (*eg*, proteins, nucleic acids, nucleotides, metal ions), and the varying lifespan of the complexes (*eg*, transient complexes such as those involved in signaling and stable complexes such as the ribosome).

The Protein Quaternary Structure Database (PQSD; Nov 2002) contains ~10,000 structurally defined protein assemblies of presumed biological significance, derived from a variety of organisms (http://pqs.ebi.ac.uk/pqs-doc.shtml). Each assembly consists of at least two protein chains. These assemblies can be organized into ~3,000 groups that contain chains with more than 30% sequence identity to at least one other member of the group (Figure 1) (Sali et al., 2003).

(Figure 1)

The most comprehensive information about protein-protein interactions is available for the yeast proteome consisting of ~6,200 proteins. The lower bound on binary protein-protein interactions and functional links in yeast has been estimated to be in the range of ~30,000 (Kumar and Snyder, 2002; von Mering et al., 2002); this number corresponds to ~9 protein partners per protein, though not necessarily all at the same time. The human proteome may have an order of magnitude more complexes than the yeast cell; and the number of different complexes across all relevant genomes may be several times larger still. Therefore, there may be thousands of biologically relevant macromolecular complexes whose structures are yet to be characterized (Abbott, 2002).

In contrast to structure determination of the individual macromolecules, structural characterization of macromolecular assemblies usually poses a more difficult challenge. A comprehensive description of large complexes generally requires the use of several experimental methods, underpinned by a variety of theoretical approaches to maximize efficiency, completeness, accuracy, and resolution of the determination of assembly composition and structure.

X-ray crystallography has been the most prolific technique for the structural analysis of proteins and protein complexes, and is still the 'gold standard' in terms of accuracy. Structures of several macromolecular assemblies have recently been solved: the RNA polymerase (Cramer et al., 2001), the ribosomal subunits (Ban et al., 2000; Harms et al., 2001; Wimberly et al., 2000); the complete ribosome and its functional complexes (Yusupov et al., 2001); the proteasome (Lowe et al., 1995); GroEl (Braig et al., 1994); the cellular transport machinery (Goldstein and Yang, 2000; Vale, 2003), and various viral capsid and virion structures (Grimes et al., 1995; Oda et al., 2000). However, the number of structures of macromolecular assemblies solved by x-ray crystallography is still quite small compared to that of individual proteins (Figure 1). This discrepancy is due mainly to the difficult production of sufficient quantities of the sample and its crystallization.

There are several variants of electron microscopy, including single-particle electron microscopy (EM) (Frank, 1996), electron tomography (Baumeister, 2002), and electron crystallography of regular two-dimensional arrays of the sample (Nogales et al., 1998). For large particles with molecular weights larger than 250 to 500 kD, single particle cryo-EM can reveal the shape and symmetry of an assembly at resolutions of 1-2 nm. Although the electron microscope produces images that represent only 2D projections of the specimen, the full 3D structure of the object can be reconstructed from many such projections, each showing the object from a different angle (Frank, 1996). More importantly, imaging by cryo-EM at these resolutions requires neither large quantities of the sample nor the sample in a crystalline form.

In the absence of high-resolution assembly crystal structures, approximate atomic models of assemblies can still be derived by combining low-resolution cryo-EM data of whole protein assemblies with computational docking of atomic resolution structures of their subunits (Nogales et al., 1998; Volkmann et al., 2000; Spahn et al., 2001; Beckmann et al., 2001; Chiu et al., 2002; Chacon and Wriggers, 2002). Recent developments in the methods for interpretation of low-resolution cryo-EM maps have suggested that docking and fitting of atomic resolution subunit structures can enhance the structural information content of the maps to a large extent. It has been estimated that using fitting techniques improves the accuracy up to one tenth the resolution of the original EM reconstruction (Volkmann and Hanein, 1999; Roseman, 2000; Wriggers et al., 2000; Rossmann et al., 2001; Wriggers and Birmanns, 2001).

Unfortunately, atomic resolution crystal structures of the isolated subunits are frequently not available. Alternatively, the induced fit may severely limit their utility in the reconstruction of the whole assembly. In such cases, it might frequently be possible to get useful comparative protein structure models of the subunits (Blundell et al., 1987; Greer, 1990; Sali and Blundell, 1993; Marti-Renom et

al., 2000; Sauder and Dunbrack, Jr., 2000; Murzin and Bateman, 2001). This approach is increasingly more applicable because of the structural genomics initiative. One of the main goals of structural genomics is to determine a sufficient number of appropriately selected structures from each domain family, such that all sequences are within modeling distance of at least one known protein structure (Baker and Sali, 2001). It has also been shown that the number of models that can be constructed with useful accuracy is already two orders of magnitude higher than the number of available experimental structures (Pieper et al., 2002).

We begin by introducing the need for a multi-scale description of macromolecular assemblies that integrates information derived from multiple sources and variable resolution into a common computational framework (Section 2). Next, we review the role comparative modeling may play in the determination of atomic structures by EM (Section 3). In particular, we introduce automated comparative protein structure modeling (Section 3.1), its errors (Section 3.2), ways to predict errors (Section 3.3), and utility of comparative models in docking of assembly subunits into EM maps. Finally, we illustrate combined comparative modeling and map fitting with two applications, the determination of partial atomic models of the 80S ribosome from *Saccharomyces cerevisiae* and the 70S ribosome of *Escherichia coli* (Sections 3.5 and 3.6).

2. Hybrid approaches to determination of assembly structures

Although x-ray crystallography and EM in combination with atomic structure docking have been successfully employed to solve structures of protein assemblies, they are not capable of efficiently characterizing the myriad of complexes that exist in a cell. For example, most of the transient complexes cannot be addressed at all with these approaches. Therefore, there is a great need for hybrid methods where accuracy, high throughput, and/or high resolution are improved by integrating information from all available sources (Figure 2) (Malhotra et al., 1990; Aloy et al., 2002). Information about the structure of an assembly can be provided by a number of experimental and theoretical methods (Figure 2). For instance, the shape, density and symmetry of a complex or its subunits may be derived from x-ray crystallography (Ban et al., 2000; Zhang et al., 1999) and electron microscopy(Frank, 2002); upper distance bounds on residues from different proteins may be obtained from NMR spectroscopy(Fiaux et al., 2002) and chemical cross-linking (Rappsilber et al., 2000; Young et al., 2000); information that two proteins bind to each other may be discovered by yeast two-hybrid (Phizicky et al., 2003; Uetz et al., 2000) or micro-calorimetry (Lakey and Raggett, 1998) experiments; two proteins can be assigned to be close to each other (relative to the size of the assembly) if they are part of an isolated sub-complex, characterized, for example, by an immuno-purification experiment (Rout et al., 2000; Aebersold and Mann, 2003; Phizicky et al., 2003).

(Figure 2 in landscape format)
(Figure 3)

To develop a framework for computing the 3D models of a given protein assembly that are consistent with all available information about its composition and structure, we express structure determination of assemblies as an optimization problem (Figure 3). This approach consists of three components (Figure 4): (i) a representation of the modeled assembly (Figure 4a); (ii) a scoring function consisting of the individual spatial restraints (Figure 4b); and (iii) optimization of the scoring function to obtain the models (Figure 4c). The most important aspect of this approach is to accurately capture all available information about the structure of the complex, whether it is high- or low-resolution, experimental, or theoretical. The method should also be capable of calculating all the models that satisfy the input spatial restraints. We illustrate this method by a description of its application to the low-resolution modeling of the configuration of proteins in a given assembly.

(Figure 4)

2.1. Modeling the low-resolution structures of assemblies

Some large assemblies, such as the nuclear pore complex, consist predominantly of subunits whose structures have not yet been defined. If comparative modeling attempts cannot provide atomic structures, such assemblies may be characterized only by low-resolution information about their overall shape and protein-protein proximity. In other words, we can expect to be able to model only the configuration of the proteins in the assembly, not their individual conformations. The following sections outline the three essential aspects of modeling by satisfaction of spatial restraints, introduced above. It has been applied to the low-resolution modeling of the configuration of proteins in the yeast nuclear complex (Alber et al., 2004, in prep).

Representation of molecular assemblies

The system is represented by points that are restrained by spatial restraints. In the absence of any atomic structures, we need to represent each of the assembly proteins as a point. A slightly higher resolution may be achieved by parsing the protein into individual domains, using either bioinformatics tools or biochemical experiments, such as limited proteolysis followed by mass spectroscopy.

Scoring function consisting of individual spatial restraints

The most important aspect of low-resolution modeling is to accurately capture all of the experimental and theoretical information about the structure of the modeled assembly. This aim may be achieved by defining the scoring function as a sum of individual spatial restraints.

The restrained spatial features may include distances, angles, and dihedral angles defined by points and gravity centers of sets of points, as well as symmetry between sets of points. The distance restraints are defined based on the available information about the modeled complex. Typical examples include:
Excluded volume restraints: Lower bounds on protein-protein distances are the sum of the corresponding estimated protein radii (Russel et al., 1997). The radius can be estimated from the number of amino acid residues or from the experimentally determined Stokes radius (Harding and Colfen, 1995).
Symmetry restraints: If EM images and stoichiometry considerations indicate symmetry (Yang et al., 1998), the appropriate result can be achieved by imposing a distance root-mean-square term on the parts of the model that need to have similar conformation or configuration.
Protein localization restraints: Immuno-labeling experiments (Rout et al., 2000) can be readily expressed as distance restraints on the labeled protein, relative to a reference point such as another labeled protein or the gravity center of the complex, upon superimposition of the individual electron microscopy or tomography images containing the labeled proteins.
Protein proximity restraints: "Pullout" experiments (Rout et al., 2000; Aebersold and Mann, 2003; Phizicky et al., 2003), chemical cross-linking (Rappsilber et al., 2000; Young et al., 2000), foot-printing (Kiselar et al., 2002), or yeast two hybrid system assays (Uetz et al., 2000) can be translated into weak upper bounds on the protein-protein distances. Such restraints may also be inferred from a bioinformatics analysis of protein sequences (*eg*, an analysis of correlated mutations (Pazos and Valencia, 2002)).
Shape restraints: EM (Frank, 1996) and tomography images (Baumeister, 2002) may allow defining the volume density map for the complex. The configuration of the proteins in the complex can then be restrained by maximizing the correlation coefficient between the EM map and that implied by a model, similarly to the fitting of higher-resolution atomic models into the EM maps (Roseman, 2000; Wriggers et al., 2000; Rossmann et al., 2001; Wriggers and Birmanns, 2001).

Optimization of the scoring function

An "ensemble" of models that minimize violations of the input restraints may be obtained by optimization of the scoring function. For example, it is possible to start with a random configuration of the proteins, and then apply a combination of the conjugate gradients minimization and simulated annealing with molecular dynamics to the Cartesian coordinates of the points representing the system. Since the optimization is stochastic, a large number of models are generally calculated by starting from a large number of independently generated random configurations (*eg*, 100,000). The aim of this sampling is to find all possible models that satisfy the input restraints.

Analysis of the models

Depending on the resolution of the modeling, a variety of geometrical criteria for comparing two given configurations of points can be used. Examples include the distance root-mean-square deviation that focuses on the protein-protein con-

tacts and a root-mean-square deviation that focuses on the positions of the individual proteins.

Assessing the accuracy of the results is an important and highly non-trivial part of the modeling. There are three conceivable ways of estimating the accuracy of the models, in the absence of a directly determined structure.

First, similarity among the well scoring models is a necessary, but not sufficient condition for their accuracy. If the well scoring models are not similar to each other, there is not sufficient information in the input restraints to define the configuration of the whole complex.

Second, the consistency between the model and the data not used in the model calculation also measures the accuracy of the model. For example, a criterion similar to the crystallographic free R-factor could be used to assess both the model accuracy and the harmony among the input restraints.

Third, the number and properties of the restraints can be correlated with the expected accuracy of the resulting models. Such correlations can be estimated by the use of "toy" models where the native structure of an assembly is known, the restraints are simulated, and their information content is estimated by exhaustive simulation.

3. Comparative modeling for structure determination of macromolecular complexes

Comparative modeling can play an important role in the structure determination of large protein assemblies. Due to the progress in structural biology and structural genomics, the structures of the individual subunits of larger assemblies are frequently already known. Additionally, the structures of large assemblies and their constituent parts also tend to be conserved in evolution. It is then possible to calculate relatively accurate comparative models of the individual subunits that have no available experimental structure. While only ~2% of known protein sequences have had their structures determined by experiment, comparative modeling can currently be used to predict at least the folds for approximately 30% of all domains in the known sequences. Therefore, there is a growing need to improve the use of homologous subunit structures in the modeling of protein assemblies. We now review the comparative modeling method and its limitations, and then continue with its application to the docking of subunit structures into EM maps.

3.1. Automated comparative protein structure modeling

Comparative modeling consists of four main steps (Marti-Renom et al., 2000): (i) fold assignment that identifies similarity between the target sequence of interest and at least one known protein structure (the template); (ii) alignment of the target sequence and the template(s); (iii) building a model based on the chosen tem-

plate(s); and (iv) assessing the model for its accuracy. These steps were assembled into a completely automated pipeline (Sanchez and Sali, 1998; Eswar et al., 2003). Manual intervention is usually required only in difficult cases. Automation of the procedure makes comparative modeling accessible to both experts and the non-specialists alike and enables the calculation of models for more sequences than is practical by hand. There are a number of servers for automated comparative modeling (http://salilab.org/bioinformatics_resources.shtml). Many of these servers are tested at the bi-annual CAFASP meetings (Fischer et al., 2001) and continually by the LiveBench (Bujnicki et al., 2001) and EVA (Eyrich et al., 2001; Koh et al., 2003) web servers for assessment of automated protein structure prediction methods. We now describe MODPIPE, which is our version of an automated scheme for large-scale comparative modeling (Sanchez and Sali, 1998; Eswar et al., 2003).

MODPIPE is an automated software pipeline for comparative protein structure modeling that can calculate comparative models for a large number of protein sequences, using many different template structures and sequence-structure alignments (Figure 5) (Sanchez and Sali, 1998; Marti-Renom et al., 2000; Pieper et al., 2002; Eswar et al., 2003). Sequence-structure matches are established by aligning the PSI-BLAST sequence profile (Altschul et al., 1997) of the target sequence against each of the template sequences extracted from Protein Data Bank (PDB) (Berman et al., 2002), as well as by scanning the target sequence against a database of the template profiles (Schaffer et al., 1999). Significant alignments covering distinct regions of the target sequence are chosen for modeling. Models are calculated for each of the sequence-structure matches using MODELLER, which implements comparative protein structure modeling by satisfaction of spatial restraints (Sali and Blundell, 1993). The resulting models are then evaluated by a composite model quality criterion that depends on the compactness of a model, the sequence identity of the sequence-structure match, and statistical energy Z-scores (Melo et al., 2002).

(Figure 5)

The thoroughness of a search for the best model is modulated by a number of parameters, including the E-value thresholds for identifying useful sequence-structure relationships and the degree of conformational sampling given a sequence-structure alignment. The validity of sequence-structure relationships is not pre-judged at the detection of the fold, but is obtained after the construction of the model and its subsequent evaluation. This approach enables a thorough exploration of fold assignments, sequence-structure alignments, and conformations, with the aim of finding the model with the best model quality score.

MODPIPE has been used to calculate models for all sequences in the SwissProt database (Boeckmann et al., 2003) with detectable similarity to a known protein structure. The results are available through MODBASE, a relational database that allows flexible and efficient querying of its contents (http://salilab.org/modbase) (Pieper et al., 2002). Currently, MODBASE contains models for domains in 415,937 out of 733,239 (~57%) unique protein sequences found in SwissProt (March 2002). Most of the models are based on less than 30% sequence identity to

the closest structure and cover only a single domain in the protein sequence, corresponding on average to one third of the whole protein. The automation and archival of such comparative models reflect the ultimate goal of the structural genomics initiative (Sali, 1998; Sanchez et al., 2000; Vitkup et al., 2001; Burley and Bonanno, 2002).

3.2. Accuracy of comparative models

The accuracy of comparative models is most easily quantified by the extent of sequence similarity between the sequence and the known structure (Chothia and Lesk, 1986; Sanchez and Sali, 1998; Marti-Renom et al., 2000; Baker and Sali, 2001). Accuracy of a model tends to increase with the target-template sequence identity (Figure 6). In general, models based on alignments with more than 40% sequence identity frequently tend to have close to 80% of their backbone atoms superposable with their actual structures with an RMS error less than 3.5Å (Sanchez and Sali, 1998).

(Figure 6)

High accuracy comparative models are based on more than 50% sequence identity to their templates (Marti-Renom et al., 2000; Fiser and Sali, 2001). They tend to have approximately 1Å RMS error for the main-chain atoms, which is comparable to the accuracy of a medium resolution nuclear magnetic resonance (NMR) spectroscopy structure or a low-resolution x-ray structure. The errors are mostly mistakes in side-chain packing, small shifts or distortions of the core main-chain regions, and occasionally larger errors in loops. Medium accuracy comparative models are usually based on 30-50% sequence identity. They tend to have approximately 90% of the main-chain modeled with 1.5Å RMS error. There are more frequent side-chain packing, core distortion, and loop modeling errors, and there are occasional alignment mistakes. And finally, low accuracy comparative models are generally based on less than 30% sequence identity. The alignment errors increase rapidly below 30% sequence identity and become the most significant origin of errors in comparative models. In addition, when a model is based on an almost insignificant alignment to a known structure, it may also have an entirely incorrect fold.

3.3. Prediction of model accuracy

The folds of the comparative models in MODPIPE are evaluated by a composite scoring function (Melo et al., 2002; John and Sali, 2003):

$$\mathrm{GA341} = 1 - \left[\cos\left(\mathrm{sequence\_identity}\right)\right]^{(\mathrm{compactness+sequence\_identity})/\exp(\mathrm{z\text{-}score})}$$

Sequence identity is the fraction of positions with identical residues in the target-template alignment. Structural compactness is the ratio between the sum of the standard volumes of the amino acid residues in the protein and the volume of the sphere with the diameter equal to the largest dimension of the model. The Z-score is calculated for the combined statistical potential energy of a model, using the mean and standard deviation of the 200 random sequences with the same composition and structure as the model (Melo et al., 2002). The combined statistical potential energy of a model is the sum of the solvent accessibility terms for all C^{β} atoms and distance-dependent terms for all pairs of C^{α} and C^{β} atoms. The solvent accessibility term for a C^{β} atom depends on its residue type and the number of other C^{β} atoms within 10Å; the non-bonded terms depend on the atom and residue types spanning the distance, the distance itself, and the number of residues separating the distance-spanning atoms in sequence. These potential terms reflect the statistical preferences observed in 760 non-redundant proteins of known structure. The GA341 scoring function was evolved by a genetic algorithm that explored many combinations of a variety of mathematical functions and model features, to optimize the discrimination between good and bad models in a training set of models. The GA341 score ranges from 0 for models that tend to have an incorrect fold to 1 for models that tend to be comparable to at least low-resolution x-ray structures. GA341 scores greater than 0.7 indicate a correct fold with more than 35% of the backbone atoms superposable to better than 3.5Å.

3.4. Docking of comparative models into low-resolution cryo-EM maps

The usefulness of comparative models is limited by their accuracy and the resolution of the density map; similar limitations may also apply to the experimentally determined subunit structures, due to the induced fit. It is usually possible to generate a set of comparative models that are based on alternate alignments, templates, and domain orientations; some of these models may be more accurate than others. The best subunit models and their positions in the complex may then be identified by manual or automated docking of the alternate models into the electron density data from electron microscopy or low-resolution x-ray crystallography. Ultimately, the best protein assembly model may be obtained by satisfying simultaneously the homology-derived restraints on the individual subunits and shape restraints on the whole complex.

The useful accuracy of comparative models for docking into the EM density map varies with the resolution of the map (Figure 7). At resolutions worse than 10Å, only the shape and size of a subunit can be identified and models based on different but related template structures could be chosen for the docking without loss of accuracy. The different template structures could account for variable conformations of the subunit (*eg,* open/closed forms) or different orientations of the constituent rigid bodies. At medium resolutions, between 5 and 10Å, it is usually possible to discern the positions of secondary structural elements and the domain

structure of the components. In these cases, models calculated with one or more templates but with several variations in the alignments to reposition secondary structures and loops could be useful for identifying the optimal fit of the structure in the density map. Additionally, loop regions can be independently optimized to account for differences in conformations between the model and the observed density. The backbone trace as well as the positions and boundaries of the secondary structure elements can be identified more accurately at even higher resolutions (~5Å). Models of at least medium accuracy (Section 3.2) are required for docking into maps at this resolution. In addition to the use of multiple templates, multiple models could also be sampled by an optimization scheme that explores the conformational degrees of freedom for the backbone and side-chains based on a single target-template alignment.

(Figure 7 in landscape)

3.5. Example 1: A partial molecular model of the 80S ribosome from *Saccharomyces cerevisiae*

As an illustration of the integrated strategies introduced earlier, we now describe the fitting of comparative protein structure models into an electron density maps of the whole yeast (Spahn et al., 2001) and *E. coli* ribosomes (Spahn et al., 2001; Gao et al., 2003). Partial or complete molecular models of the ribosomes are obtained by the use of information from two sources, experimental low-resolution (≤ 10Å) cryo-EM maps and all-atom comparative models for the individual RNA and protein components of the ribosomes.

Ribosomes are macromolecular machines responsible for protein biosynthesis in the cell and consist of ribosomal RNA (rRNA) molecules and 50-80 ribosomal proteins. They are made up of two subunits, a small subunit responsible for decoding in protein translation (*ie*, selection of cognate tRNA) (Carter et al., 2000) and a large subunit, primarily responsible for the catalytic activity (*ie*, peptidyl transferase) (Nissen et al., 2000). Atomic resolution x-ray structures are available for the small 30S subunit from the thermophile bacterium *Thermus thermophilus* (Schluenzen et al., 2000; Wimberly et al., 2000) as well as the large 50S subunit from the halophile archaebacterium *Haloarcula marismortui* (Ban et al., 2000) and mesophilic eubacterium *Deinococus radiodurans* (Harms et al., 2001). While a relatively large amount of high-resolution structural information is available for prokaryotic ribosomes or their individual subunits, there is only sparse data for their eukaryotic counterparts. Fortunately, the eukaryotic ribosomal RNA and proteins are evolutionarily related to their prokaryotic homologs. Despite the different sizes of the rRNA, additional proteins, and more complex functions of the eukaryotic ribosome, it can be anticipated that the overall spatial arrangement of the subunits and the fundamental process of protein biosynthesis are similar to those in the prokaryotes.

To gain structural insights into the machinery of eukaryotic ribosomes, we combined a low-resolution cryo-EM map (~15Å) of the *Saccharomyces cerevisiae*

ribosome with comparative modeling and docking (Spahn et al., 2001). The yeast ribosomal complex is made up of a 40S small subunit, composed of a 1798 nucleotide (nt) long 18S rRNA and 32 ribosomal proteins, and a large 60S subunit composed of a 25S rRNA (3392 nt), 5.8S rRNA (158 nt), 5S rRNA (121 nt), and 45 ribosomal proteins (Spahn et al., 2001). To facilitate the docking, the map of the 80S ribosome was computationally separated into the protein and RNA parts, using a method that takes into account the differences in the density distribution of RNA and proteins, as well as the molecular masses and contiguity constraints (Spahn et al., 2000). rRNA models from the crystal structures of the 30S subunit from *T. thermophilus* (Wimberly et al., 2000) and the 50S subunit from *H. marismortui* (Ban et al., 2000) were fitted into the resulting maps for the small subunit rRNA and large subunit rRNA of yeast, respectively. Where necessary, the x-ray models were modified by moving the non-fitting parts (*eg*, helices) as rigid bodies relative to the rest of the model.

Comparative models for the yeast ribosomal proteins were constructed using MODPIPE (Sanchez and Sali, 1998; Spahn et al., 2001) and are available through MODBASE (http://salilab.org/modbase). Structural templates used to calculate the models consisted of all the individual chains from structures in PDB (as of September 2000), clustered such that the sequences of no two chains from any two clusters were more than 95% identical. In addition, the structures of the small subunit from *T. thermophilus* (PDB code: 1FJF) and the large subunit from *H. marismortui* (PDB code: 1FKF) were considered as separate sets of templates. In total, comparative models were obtained for 43 yeast ribosomal proteins; 15 for the 40S subunit (Figure 8a) and 28 for the 60S subunit (Figure 8b). The models were derived from alignments with sequence identities in the range 20%–56% (with an average of 32%) and E-values better than 0.0001. The coverage of the models (fraction of the yeast ribosomal sequence modeled) ranges between 34%–99% (with an average of 75%). Docking of atomic models into the cryo-EM density map was done manually using program O (Jones et al., 1991).

(Figure 8)

The composite map, consisting of docked RNA and comparative models of proteins into the 15.4Å cryo-EM map, provides for the structural interpretation of the eukaryotic ribosome complex. The common core of the eukaryotic ribosome was found to agree well with x-ray structures of the bacterial and archaebacterial subunits. It reinforces the notion that the fundamental mechanism of protein synthesis is highly conserved throughout all kingdoms. The differences in the structures of the prokaryotic and eukaryotic ribosomes could be localized to regions in the density map corresponding to either yeast proteins without homologous counterparts or those with additional domains. These differences occur mainly on the solvent exposed faces of the subunits, conserving the core of the ribosome. It was also found that the inter-subunit interactions, important for communication between the subunits, and the ribosome-tRNA interactions were largely conserved. Additionally, the structure enabled the identification of four new protein-protein

contacts. For more information, see references (Spahn et al., 2001; Beckmann et al., 2001).

3.6. Example 2: A molecular model of the *E. coli* 70S ribosome

The aim of this study was to capture the dynamic features of the ribosome, the 'ratchet-like' inter-subunit motion, by trapping functionally meaningful states by cryo-EM (Gao et al., 2003). The limited resolution of the cryo-EM maps was overcome by docking comparative models of rRNA and proteins into the maps of the different states of the ribosome: (i) a 11.5-Å map (Gabashvili et al., 2000) of the control, an initiation-like complex with fMet-tRNA$_f^{Met}$ at the P site (Malhotra et al., 1998); and (ii) a 12.3-Å map of the EF-G·GTP-bound complex (*ie*, a ribosome complex with EF-G in the presence of a non-hydrolysable GTP analog) (Frank and Agrawal, 2000). The *E. coli* 70S ribosome consists of two subunits: the 30S subunit, comprising 16S rRNA (1542 nt) and 21 proteins, and the 50S subunit, comprising 23S rRNA (2904 nt), 5S rRNA (120 nt), and 36 proteins. The models of *E. coli* 23S rRNA and 5S rRNA were generated from the crystal structure of *H. marismortui* (PDB code 1FFK) (Ban et al., 2000), while the model of *E. coli* 16S rRNA was generated from the crystal structure of *T. thermophilus* (PDB code 1IBL) (Ogle et al., 2001) using the molecular modeling package Insight II (Accelrys Inc. Insight II 2003).

Models for the *E. coli* ribosomal proteins were calculated by MODPIPE as described earlier. The crystal structures of the proteins from the small subunit of *T. thermophilus* (PDB code 1FJG) (Carter et al., 2000) were chosen as the structural templates to model 19 proteins of the 30S small subunit (S2-S20) of *E. coli*. For proteins of the 50S subunit, 29 out of the 36 *E. coli* proteins were modeled based on the crystal structures of *H. marismortui* (PDB code 1JJ2) (Klein et al., 2001), *D. radiodurans* (1LNR) (Harms et al., 2001), and *T. thermophilus* (1GIY; L9, L25) (Yusupov et al., 2001).

The starting models of the whole ribosome were built by manually docking the individual rRNA and protein models as rigid bodies into the cryo-EM density maps using the interactive program O (Jones et al., 1991). The initial positions of each of the rRNA structures and those of the proteins were taken from the corresponding positions of the template crystal structures. The program RSRef (Chapman, 1995), a real-space refinement module for the TNT program (Tronrud, 1997), was then employed for automatically and simultaneously refining both the stereochemistry and the fit of the atomic structures to the density map. Since the resolutions of the experimental density maps are not suitable for refinement of independent atoms, a multi-rigid-body refinement was employed.

Comparison of the two resulting atomic models revealed that the ribosome changes from a compact structure in the initiation-like form to a looser one in the EF-G bound form. This change is coupled with the rearrangement of many of the proteins. Furthermore, it could be seen that in contrast to the unchanged inter-subunit bridges formed wholly by RNA, the bridges involving ribosomal proteins undergo large conformational changes following the ratchet-like motion. Observa-

tions suggested an important role of ribosomal proteins in facilitating the dynamics of translation.

4. Conclusions

We are now poised to integrate structural information gathered at multiple levels of the biological hierarchy — from atoms to cells — into a common framework. The goal is a comprehensive description of the multitude of interactions between molecular entities, which in turn is a prerequisite for the discovery of general structural principles that underlie all cellular processes. In contrast to structure determination of individual proteins, structural characterization of macromolecular assemblies usually requires diverse sources of information (Sali et al., 2003). This information may vary greatly in terms of its accuracy and resolution, and includes data from both experimental and computational methods, such as X-ray crystallography, NMR spectroscopy, electron microscopy, chemical cross-linking, affinity purification, yeast two-hybrid system experiments, calorimetry, computational docking, and bioinformatics analysis of protein sequences and structures. Structural genomics will bring us closer to a comprehensive dictionary of proteins in the foreseeable future, while electron microscopy techniques and other approaches will allow us to assemble proteins into complexes. A comprehensive description of large complexes will generally require the use of a number of experimental methods, underpinned by a variety of theoretical approaches to maximize efficiency, completeness, accuracy, and resolution of the experimental determination of assembly composition and structure. In conjunction with the non-invasive 3D imaging of whole cells, these approaches might ultimately enable us to read the molecular book of the cell.

Acknowledgments

We are grateful to Fred Davies, Damien Devos, Dimitry Korkin, and Maya Topf for discussions about the modeling of assembly structures. This review is based on the following publications: (Sali et al., 2003), (Spahn et al., 2001), (Gao et al., 2003), (Eswar et al., 2003), and (Marti-Renom et al., 2000).

LEGENDS TO FIGURES

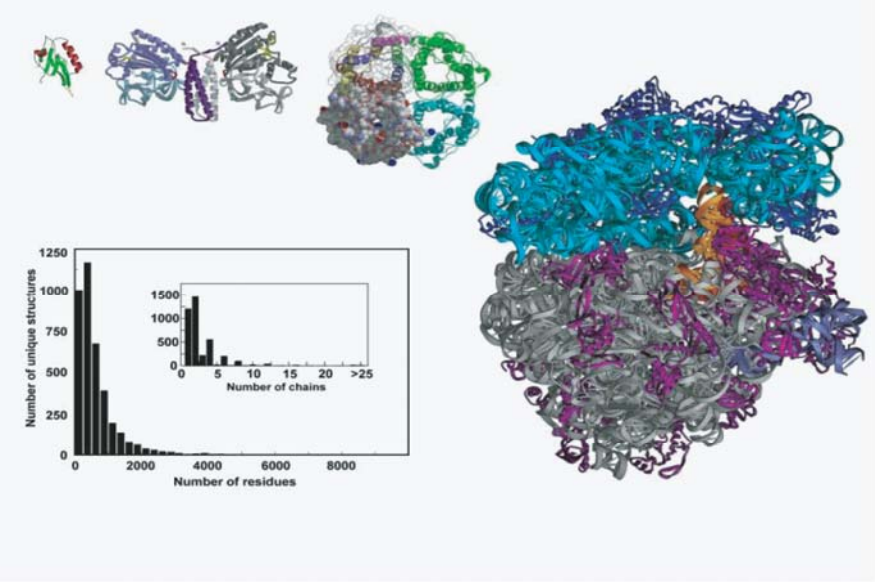

Fig. 1: Illustration of the size range of biomolecular structures solved by x-ray crystallography and the size distribution of structures contained in the Protein Quaternary Structure database. Structures of (top left to right) the PDZ domain, a molecular recognition domain that leads to protein-protein interactions; CheA, a dimeric multidomain bacterial signaling molecule; aquaporin, which serves as a transmembrane water channel; and 70S ribosome, which is the molecular machine for protein biosynthesis. The histogram shows the distribution of the size of the entries in the Protein Quaternary Structure (PQS) database (http://pqs.ebi.ac.uk). The 15,190 entries with at least one protein chain of at least 30 residues, when compared with each other, produced 3,876 clusters with more than 30% sequence identity and less than 30 residue length difference among the members within the same cluster. The distributions of the numbers of residues and chains (inset) in the representative structures for each group are shown. As expected, the structures of large complexes are under-represented, given an estimated average size of a yeast complex of 7.5 proteins (see text).

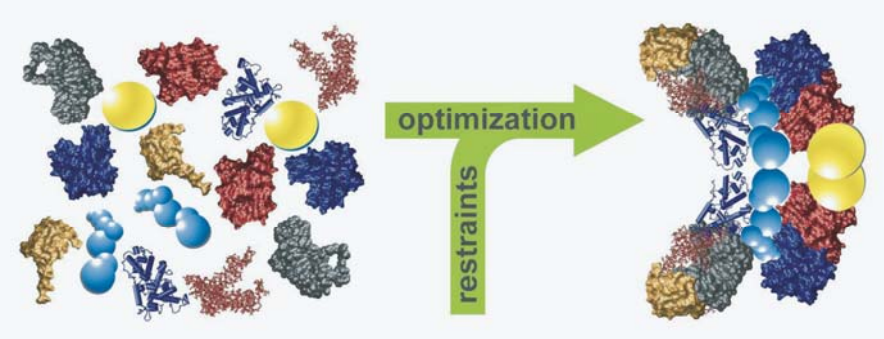

Fig. 2. Experimental and theoretical methods that can provide information about a macromolecular assembly structure. The annotations below each of the panels list the aspects of an assembly that might be obtained by the corresponding method. Subunit and assembly structure indicate an atomic or near atomic resolution at 3Å or better. Subunit and assembly shape indicate the density or surface envelope at a low-resolution of worse than 3Å. Subunit-subunit contact indicates knowledge about protein pairs that are in contact with each other, and in some cases about the face that is involved in the contact. Subunit proximity indicates whether two proteins are close to each other relative to the size of the assembly, but not necessarily in direct contact. Subunit stoichiometry indicates the number of subunits of a given type that occur in the assembly. Assembly symmetry indicates the symmetry of the arrangement of the subunits in the assembly. Gray boxes indicate extreme difficulty in obtaining the corresponding information by a given method (Sali et al., 2003).

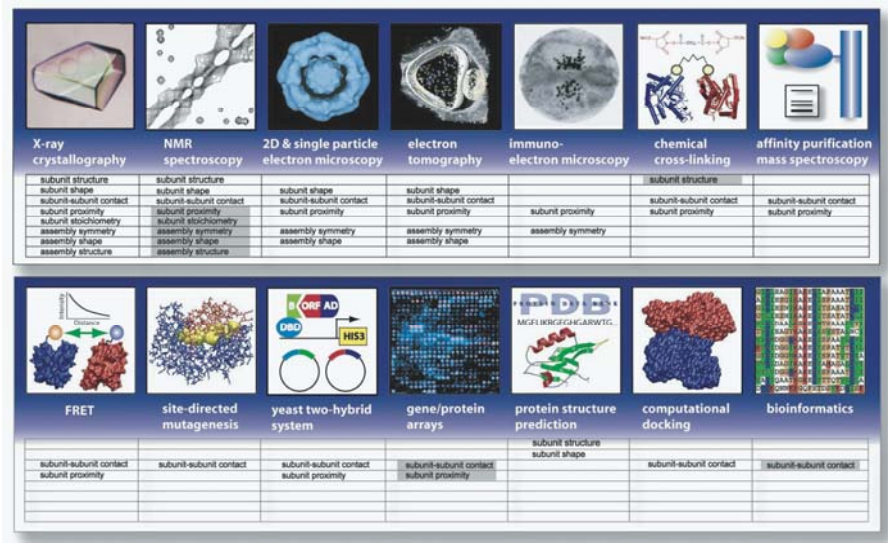

Fig. 3. The scheme that illustrates how the subunits of a hypothetical complex (left) may be assembled through optimization with respect to restraints from a variety of methods to obtain the final assembly model (right) (Sali et al., 2003).

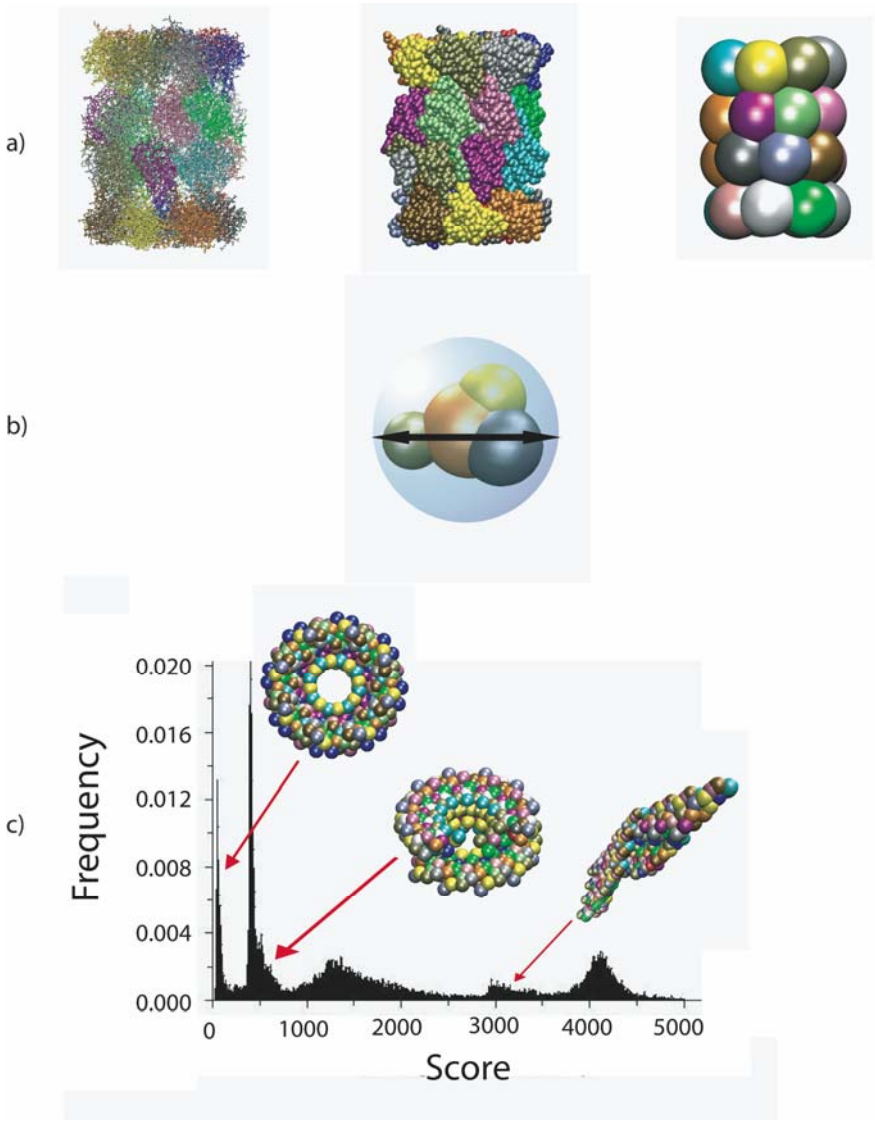

Fig. 4. Modeling of the configuration of proteins in an assembly by satisfaction of spatial restraints. (a) From left to right, representations of the proteasome assembly of 28 proteins with points per atom, residue and protein, respectively. (b) Derivation of upper distance bounds on all pairs of proteins that have been shown to be a part of the same subcomplex by an affinity chromatography experiment. An estimate for the diameter of the whole subcomplex is needed and can be obtained, for example, from the measured Stokes radius or the total number of residues in the subcomplex. (c) The distribution of an objective function score for many optimized configurations. A desired ring structure is indicated on the left, but stochastic optimization that starts from random configurations also results in a variety of other distorted solutions that do not satisfy input restraints.

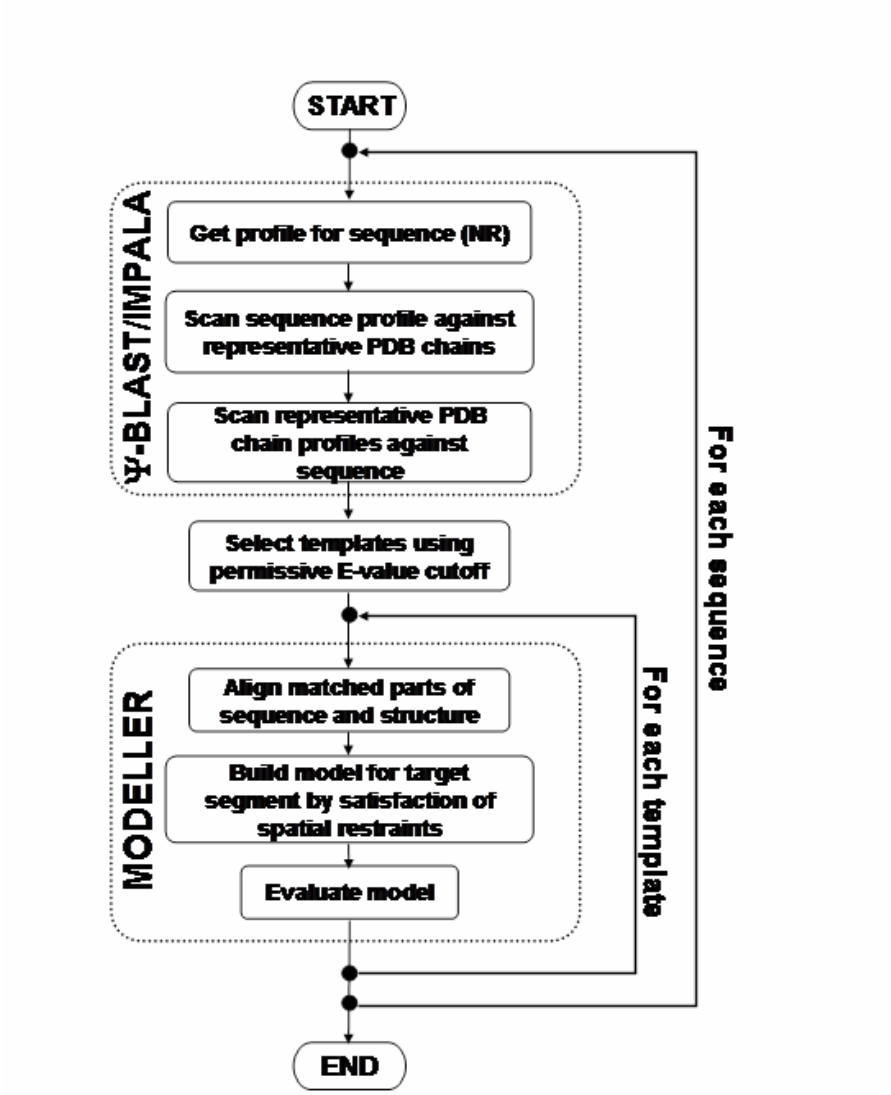

Fig. 5. Flowchart of MODPIPE, a large-scale protein structure modeling pipeline (Eswar et al., 2003). See text for details.

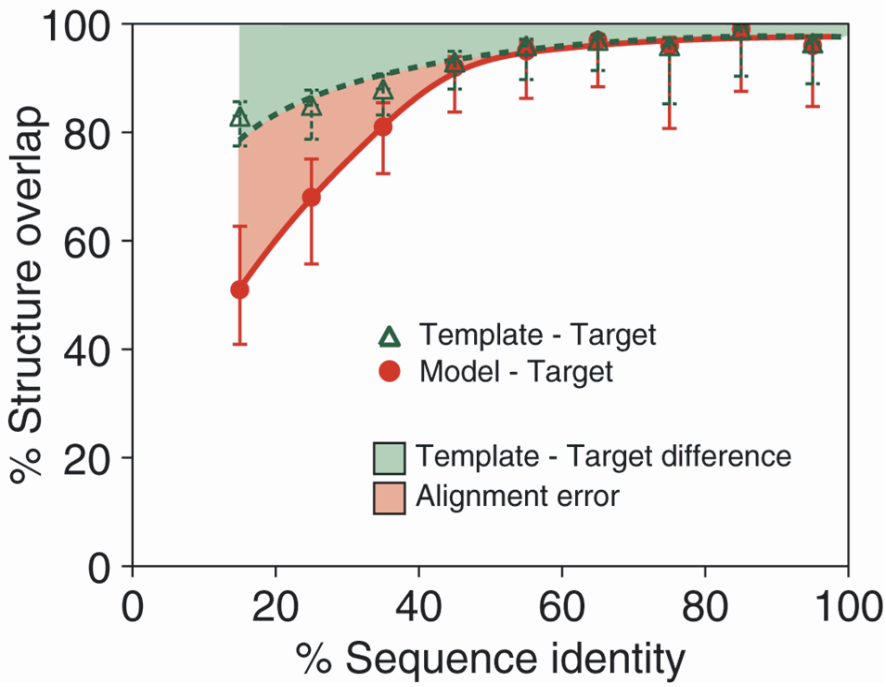

Fig. 6. The relationship between the accuracy of a reliable model and the percentage sequence identity to the template. The overlaps of an experimentally determined protein structure with its model (red continuous line) and with a template on which the model was based (green dashed line) are shown as a function of the target-template sequence identity. The structure overlap is defined as the fraction of the equivalent C^{α} atoms. For comparison of the model with the actual structure (filled circles), two C^{α} atoms were considered equivalent if they were within 3.5 Å of each other and belonged to the same residue. For comparison of the template structure with the actual target structure (open triangles), two C^{α} atoms were considered equivalent if they were within 3.5 Å after alignment and rigid-body superposition. The points correspond to the median values, and the error bars in the positive and negative directions correspond to the average positive and negative differences from the median, respectively (Sanchez and Sali, 1998).

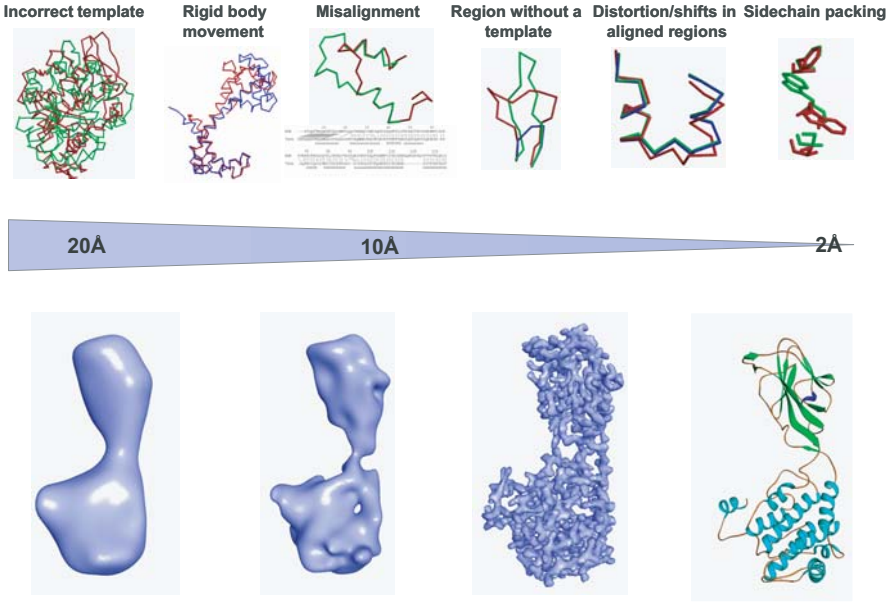

Fig. 7. Usefulness of comparative models for docking into EM electron density maps (the maps are courtesy of Dr. Wah Chiu). Examples of errors in comparative models that can be identified at various resolutions of the density maps are indicated. See text for details.

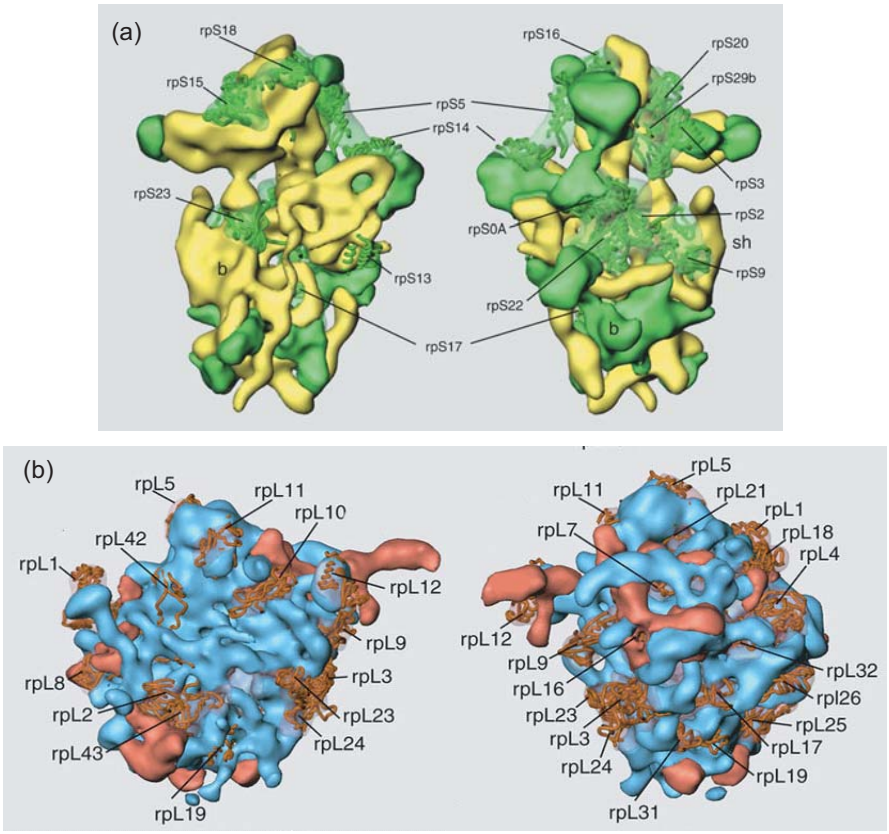

Fig. 8. Structures of the (a) 40S small subunit and (b) 60S large subunit of the yeast ribosome. The RNA and protein partitions are shown in yellow and turquoise respectively for the small subunit; they are depicted as blue and orange respectively for the large subunit. Wherever comparative models could be docked into the map, the protein partition is shown transparently. Therefore, solid parts of the protein partition predict the position of additional proteins with no homologous counterparts in prokaryotes (Spahn et al., 2000).

References

Abbott,A. (2002). The society of proteins. Nature *417*, 894-896.

Aebersold,R. and Mann,M. (2003). Mass spectrometry-based proteomics. Nature *422*, 198-207.

Alber, F, Dokudovskaya, S, Kipper, J., Supraptro, A., Veenhoff, L. M., Zhang, W., Chait, B. T., Rout, M. P., and Sali, A. Modeling the yeast nuclear pore complex. 2003 (in preparation).

Alberts,B. (1998). The cell as a collection of protein machines: preparing the next generation of molecular biologists. Cell *92*, 291-294.

Aloy,P., Ciccarelli,F.D., Leutwein,C., Gavin,A.C., Superti-Furga,G., Bork,P., Bottcher,B., and Russell,R.B. (2002). A complex prediction: three-dimensional model of the yeast exosome. EMBO Rep. *3*, 628-635.

Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W., and Lipman,D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. *25*, 3389-3402.

Baker,D. and Sali,A. (2001). Protein structure prediction and structural genomics. Science *294*, 93-96.

Ban,N., Nissen,P., Hansen,J., Moore,P.B., and Steitz,T.A. (2000). The complete atomic structure of the large ribosomal subunit at 2.4 A resolution. Science *289*, 905-920.

Baumeister,W. (2002). Electron tomography: towards visualizing the molecular organization of the cytoplasm. Curr. Opin. Struct. Biol. *12*, 679-684.

Beckmann,R., Spahn,C.M., Eswar,N., Helmers,J., Penczek,P.A., Sali,A., Frank,J., and Blobel,G. (2001). Architecture of the protein-conducting channel associated with the translating 80S ribosome. Cell *107*, 361-372.

Berman,H.M., Battistuz,T., Bhat,T.N., Bluhm,W.F., Bourne,P.E., Burkhardt,K., Feng,Z., Gilliland,G.L., Iype,L., Jain,S., Fagan,P., Marvin,J., Padilla,D., Ravichandran,V., Schneider,B., Thanki,N., Weissig,H., Westbrook,J.D., and Zardecki,C. (2002). The Protein Data Bank. Acta Crystallogr. D. Biol. Crystallogr. *58*, 899-907.

Blundell,T.L., Sibanda,B.L., Sternberg,M.J., and Thornton,J.M. (1987). Knowledge-based prediction of protein structures and the design of novel molecules. Nature *326*, 347-352.

Boeckmann,B., Bairoch,A., Apweiler,R., Blatter,M.C., Estreicher,A., Gasteiger,E., Martin,M.J., Michoud,K., O'Donovan,C., Phan,I., Pilbout,S., and Schneider,M. (2003). The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Res. *31*, 365-370.

Braig,K., Otwinowski,Z., Hegde,R., Boisvert,D.C., Joachimiak,A., Horwich,A.L., and Sigler,P.B. (1994). The crystal structure of the bacterial chaperonin GroEL at 2.8 A. Nature *371*, 578-586.

Bujnicki,J.M., Elofsson,A., Fischer,D., and Rychlewski,L. (2001). LiveBench-2: large-scale automated evaluation of protein structure prediction servers. Proteins *Suppl 5*, 184-191.

Burley,S.K. and Bonanno,J.B. (2002). Structural genomics of proteins from conserved biochemical pathways and processes. Curr. Opin. Struct. Biol. *12*, 383-391.

Carter,A.P., Clemons,W.M., Brodersen,D.E., Morgan-Warren,R.J., Wimberly,B.T., and Ramakrishnan,V. (2000). Functional insights from the structure of the 30S ribosomal subunit and its interactions with antibiotics. Nature *407*, 340-348.

Chacon,P. and Wriggers,W. (2002). Multi-resolution contour-based fitting of macromolecular structures. J. Mol. Biol. *317*, 375-384.

Chapman,M.S. (1995). Restrained Real-Space Macromolecular Atomic Refinement using a New Resolution-Dependent Electron Density Function. Acta Crystallogr. A *A51*, 69-80.

Chiu,W., Baker,M.L., Jiang,W., and Zhou,Z.H. (2002). Deriving folds of macromolecular complexes through electron cryomicroscopy and bioinformatics approaches. Curr. Opin. Struct. Biol. *12*, 263-269.

Chothia,C. and Lesk,A.M. (1986). The relation between the divergence of sequence and structure in proteins. EMBO J. *5*, 823-826.

Courey,A.J. (2001). Cooperativity in transcriptional control. Curr. Biol. *11*, R250-R252.

Cramer,P., Bushnell,D.A., and Kornberg,R.D. (2001). Structural basis of transcription: RNA polymerase II at 2.8 angstrom resolution. Science *292*, 1863-1876.

Eswar,N., John,B., Mirkovic,N., Fiser,A., Ilyin,V.A., Pieper,U., Stuart,A.C., Marti-Renom,M.A., Madhusudhan,M.S., Yerkovich,B., and Sali,A. (2003). Tools for comparative protein structure modeling and analysis. Nucleic Acids Res. *31*, 3375-3380.

Eyrich,V.A., Marti-Renom,M.A., Przybylski,D., Madhusudhan,M.S., Fiser,A., Pazos,F., Valencia,A., Sali,A., and Rost,B. (2001). EVA: continuous automatic evaluation of protein structure prediction servers. Bioinformatics. *17*, 1242-1243.

Fiaux,J., Bertelsen,E.B., Horwich,A.L., and Wuthrich,K. (2002). NMR analysis of a 900K GroEL GroES complex. Nature *418*, 207-211.

Fischer,D., Elofsson,A., Rychlewski,L., Pazos,F., Valencia,A., Rost,B., Ortiz,A.R., and Dunbrack,R.L., Jr. (2001). CAFASP2: the second critical assessment of fully automated structure prediction methods. Proteins *Suppl 5*, 171-183.

Fiser,A. and Sali,A. (2001). MODELLER: generation and refinement of homology models. In Methods in Enzymology, C.W.Carter and R.M.Sweet, eds. Academic Press).

Frank,J. (1996). Three-dimensional electon microscopy of macromolecular assemblies. (London: Academic).

Frank,J. (2002). Single-particle imaging of macromolecules by cryo-electron microscopy. Annu. Rev. Biophys. Biomol. Struct. *31*, 303-319.

Frank,J. and Agrawal,R.K. (2000). A ratchet-like inter-subunit reorganization of the ribosome during translocation. Nature *406*, 318-322.

Gabashvili,I.S., Agrawal,R.K., Spahn,C.M., Grassucci,R.A., Svergun,D.I., Frank,J., and Penczek,P. (2000). Solution structure of the E. coli 70S ribosome at 11.5 A resolution. Cell *100*, 537-549.

Gao,H., Sengupta,J., Valle,M., Korostelev,A., Eswar,N., Stagg,S.M., Van Roey,P., Agrawal,R.K., Harvey,S.C., Sali,A., Chapman,M.S., and Frank,J. (2003). Study of the structural dynamics of the E coli 70S ribosome using real-space refinement. Cell *113*, 789-801.

Goldstein,L.S. and Yang,Z. (2000). Microtubule-based transport systems in neurons: the roles of kinesins and dyneins. Annu. Rev. Neurosci. *23*, 39-71.

Gomperts,S.N. (1996). Clustering membrane proteins: It's all coming together with the PSD-95/SAP90 protein family. Cell *84*, 659-662.

Goto,N.K., Zor,T., Martinez-Yamout,M., Dyson,H.J., and Wright,P.E. (2002). Cooperativity in transcription factor binding to the coactivator CREB-binding protein (CBP). The mixed lineage leukemia protein (MLL) activation domain binds to an allosteric site on the KIX domain. J. Biol. Chem. *277*, 43168-43174.

Grakoui,A., Bromley,S.K., Sumen,C., Davis,M.M., Shaw,A.S., Allen,P.M., and Dustin,M.L. (1999). The immunological synapse: a molecular machine controlling T cell activation. Science *285*, 221-227.

Greer,J. (1990). Comparative modeling methods: application to the family of the mammalian serine proteases. Proteins *7*, 317-334.

Grimes,J., Basak,A.K., Roy,P., and Stuart,D. (1995). The crystal structure of bluetongue virus VP7. Nature *373*, 167-170.

Harding,S.E. and Colfen,H. (1995). Inversion formulae for ellipsoid of revolution macromolecular shape functions. Anal. Biochem. *228*, 131-142.

Harms,J., Schluenzen,F., Zarivach,R., Bashan,A., Gat,S., Agmon,I., Bartels,H., Franceschi,F., and Yonath,A. (2001). High resolution structure of the large ribosomal subunit from a mesophilic eubacterium. Cell *107*, 679-688.

John,B. and Sali,A. (2003). Comparative protein structure modeling by iterative alignment, model building and model assessment. Nucleic Acids Res. *31*, 3982-3992.

Jones,T.A., Zou,J.Y., Cowan,S.W., and Kjeldgaard (1991). Improved methods for building protein models in electron density maps and the location of errors in these models. Acta Crystallogr. A *47 (Pt 2)*, 110-119.

Kiselar,J.G., Maleknia,S.D., Sullivan,M., Downard,K.M., and Chance,M.R. (2002). Hydroxyl radical probe of protein surfaces using synchrotron X-ray radiolysis and mass spectrometry. Int. J. Radiat. Biol. *78*, 101-114.

Klein,D.J., Schmeing,T.M., Moore,P.B., and Steitz,T.A. (2001). The kink-turn: a new RNA secondary structure motif. EMBO J. *20*, 4214-4221.

Koh,I.Y.Y., Eyrich,V.A., Marti-Renom,M.A., Przybylski,D., Madhusudhan,M.S., Eswar,N., Grana,O., Pazos,F., Valencia,A., Sali,A., and Rost,B. (2003). EVA: evaluation of protein structure prediction servers. Nucleic Acids Res. *in press*.

Kumar,A. and Snyder,M. (2002). Protein complexes take the bait. Nature *415*, 123-124.

Lakey,J.H. and Raggett,E.M. (1998). Measuring protein-protein interactions. Curr. Opin. Struct. Biol. *8*, 119-123.

Lowe,J., Stock,D., Jap,B., Zwickl,P., Baumeister,W., and Huber,R. (1995). Crystal structure of the 20S proteasome from the archaeon T. acidophilum at 3.4 A resolution. Science *268*, 533-539.

Malhotra,A., Penczek,P., Agrawal,R.K., Gabashvili,I.S., Grassucci,R.A., Junemann,R., Burkhardt,N., Nierhaus,K.H., and Frank,J. (1998). Escherichia coli 70 S ribosome at 15 A resolution by cryo-electron microscopy: localization of fMet-tRNAfMet and fitting of L1 protein. J. Mol. Biol. *280*, 103-116.

Malhotra,A., Tan,R.K., and Harvey,S.C. (1990). Prediction of the three-dimensional structure of Escherichia coli 30S ribosomal subunit: a molecular mechanics approach. Proc. Natl. Acad. Sci. U. S. A *87*, 1950-1954.

Marti-Renom,M.A., Stuart,A.C., Fiser,A., Sanchez,R., Melo,F., and Sali,A. (2000). Comparative protein structure modeling of genes and genomes. Annu. Rev. Biophys. Biomol. Struct. *29*, 291-325.

Melo,F., Sanchez,R., and Sali,A. (2002). Statistical potentials for fold assessment. Protein Sci. *11*, 430-448.

Murakami,K.S. and Darst,S.A. (2003). Bacterial RNA polymerases: the wholo story. Curr. Opin. Struct. Biol. *13*, 31-39.

Murzin,A.G. and Bateman,A. (2001). CASP2 knowledge-based approach to distant homology recognition and fold prediction in CASP4. Proteins *Suppl 5*, 76-85.

Nissen,P., Hansen,J., Ban,N., Moore,P.B., and Steitz,T.A. (2000). The structural basis of ribosome activity in peptide bond synthesis. Science *289*, 920-930.

Nogales,E. (2000). Recent structural insights into transcription preinitiation complexes. J. Cell Sci. *113 Pt 24*, 4391-4397.

Nogales,E. and Grigorieff,N. (2001). Molecular Machines: putting the pieces together. J. Cell Biol. *152*, F1-10.

Nogales,E., Wolf,S.G., and Downing,K.H. (1998). Structure of the alpha beta tubulin dimer by electron crystallography. Nature *391*, 199-203.

Noji,H. and Yoshida,M. (2001). The rotary machine in the cell, ATP synthase. J. Biol. Chem. *276*, 1665-1668.

Oda,Y., Saeki,K., Takahashi,Y., Maeda,T., Naitow,H., Tsukihara,T., and Fukuyama,K. (2000). Crystal structure of tobacco necrosis virus at 2.25 A resolution. J. Mol. Biol. *300*, 153-169.

Ogle,J.M., Brodersen,D.E., Clemons,W.M., Jr., Tarry,M.J., Carter,A.P., and Ramakrishnan,V. (2001). Recognition of cognate transfer RNA by the 30S ribosomal subunit. Science *292*, 897-902.

Pazos,F. and Valencia,A. (2002). In silico two-hybrid system for the selection of physically interacting protein pairs. Proteins *47*, 219-227.

Phizicky,E., Bastiaens,P.I., Zhu,H., Snyder,M., and Fields,S. (2003). Protein analysis on a proteomic scale. Nature *422*, 208-215.

Pieper,U., Eswar,N., Stuart,A.C., Ilyin,V.A., and Sali,A. (2002). MODBASE, a database of annotated comparative protein structure models. Nucleic Acids Res. *30*, 255-259.

Rappsilber,J., Siniossoglou,S., Hurt,E.C., and Mann,M. (2000). A generic strategy to analyze the spatial organization of multi-protein complexes by cross-linking and mass spectrometry. Anal. Chem. *72*, 267-275.

Roseman,A.M. (2000). Docking structures of domains into maps from cryo-electron microscopy using local correlation. Acta Crystallogr. D. Biol. Crystallogr. *56 (Pt 10)*, 1332-1340.

Rossmann,M.G., Bernal,R., and Pletnev,S.V. (2001). Combining electron microscopic with x-ray crystallographic structures. J. Struct. Biol. *136*, 190-200.

Rout,M.P., Aitchison,J.D., Suprapto,A., Hjertaas,K., Zhao,Y., and Chait,B.T. (2000). The yeast nuclear pore complex: composition, architecture, and transport mechanism. J. Cell Biol. *148*, 635-651.

Russel,M., Linderoth,N.A., and Sali,A. (1997). Filamentous phage assembly: variation on a protein export theme. Gene *192*, 23-32.

Sali,A. (1998). 100,000 protein structures for the biologist. Nat. Struct. Biol. *5*, 1029-1032.

Sali,A. and Blundell,T.L. (1993). Comparative protein modelling by satisfaction of spatial restraints. J. Mol. Biol. *234*, 779-815.

Sali,A., Glaeser,R., Earnest,T., and Baumeister,W. (2003). From words to literature in structural proteomics. Nature *422*, 216-225.

Sanchez,R., Pieper,U., Melo,F., Eswar,N., Marti-Renom,M.A., Madhusudhan,M.S., Mirkovic,N., and Sali,A. (2000). Protein structure modeling for structural genomics. Nat. Struct. Biol. *7 Suppl*, 986-990.

Sanchez,R. and Sali,A. (1998). Large-scale protein structure modeling of the Saccharomyces cerevisiae genome. Proc. Natl. Acad. Sci. U. S. A *95*, 13597-13602.

Sauder,J.M. and Dunbrack,R.L., Jr. (2000). Genomic fold assignment and rational modeling of proteins of biological interest. Proc. Int. Conf. Intell. Syst. Mol. Biol. *8*, 296-306.

Schaffer,A.A., Wolf,Y.I., Ponting,C.P., Koonin,E.V., Aravind,L., and Altschul,S.F. (1999). IMPALA: matching a protein sequence against a collection of PSI-BLAST-constructed position-specific score matrices. Bioinformatics. *15*, 1000-1011.

Schluenzen,F., Tocilj,A., Zarivach,R., Harms,J., Gluehmann,M., Janell,D., Bashan,A., Bartels,H., Agmon,I., Franceschi,F., and Yonath,A. (2000). Structure of functionally activated small ribosomal subunit at 3.3 angstroms resolution. Cell *102*, 615-623.

Spahn,C.M., Beckmann,R., Eswar,N., Penczek,P.A., Sali,A., Blobel,G., and Frank,J. (2001). Structure of the 80S ribosome from Saccharomyces cerevisiae--tRNA-ribosome and subunit-subunit interactions. Cell *107*, 373-386.

Spahn,C.M., Penczek,P.A., Leith,A., and Frank,J. (2000). A method for differentiating proteins from nucleic acids in intermediate-resolution density maps: cryo-electron microscopy defines the quaternary structure of the Escherichia coli 70S ribosome. Structure Fold. Des *8*, 937-948.

Tronrud,D.E. (1997). TNT refinement package. Methods Enzymol. *277*, 306-319.

Uetz,P., Giot,L., Cagney,G., Mansfield,T.A., Judson,R.S., Knight,J.R., Lockshon,D., Narayan,V., Srinivasan,M., Pochart,P., Qureshi-Emili,A., Li,Y., Godwin,B., Conover,D., Kalbfleisch,T., Vijayadamodar,G., Yang,M., Johnston,M., Fields,S., and Rothberg,J.M. (2000). A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. Nature *403*, 623-627.

Vale,R.D. (2003). The molecular motor toolbox for intracellular transport. Cell *112*, 467-480.

Vale,R.D. and Milligan,R.A. (2000). The way things move: looking under the hood of molecular motor proteins. Science *288*, 88-95.

Vitkup,D., Melamud,E., Moult,J., and Sander,C. (2001). Completeness in structural genomics. Nat. Struct. Biol. *8*, 559-566.

Volkmann,N. and Hanein,D. (1999). Quantitative fitting of atomic models into observed densities derived by electron microscopy. J. Struct. Biol. *125*, 176-184.

Volkmann,N., Hanein,D., Ouyang,G., Trybus,K.M., DeRosier,D.J., and Lowey,S. (2000). Evidence for cleft closure in actomyosin upon ADP release. Nat. Struct. Biol. *7*, 1147-1155.

von Mering,C., Krause,R., Snel,B., Cornell,M., Oliver,S.G., Fields,S., and Bork,P. (2002). Comparative assessment of large-scale data sets of protein-protein interactions. Nature *417*, 399-403.

Wimberly,B.T., Brodersen,D.E., Clemons,W.M., Jr., Morgan-Warren,R.J., Carter,A.P., Vonrhein,C., Hartsch,T., and Ramakrishnan,V. (2000). Structure of the 30S ribosomal subunit. Nature *407*, 327-339.

Wriggers,W., Agrawal,R.K., Drew,D.L., McCammon,A., and Frank,J. (2000). Domain motions of EF-G bound to the 70S ribosome: insights from a hand-shaking between multi-resolution structures. Biophys. J. *79*, 1670-1678.

Wriggers,W. and Birmanns,S. (2001). Using situs for flexible and rigid-body fitting of multiresolution single-molecule data. J. Struct. Biol. *133*, 193-202.

Yang,Q., Rout,M.P., and Akey,C.W. (1998). Three-dimensional architecture of the isolated yeast nuclear pore complex: functional and evolutionary implications. Mol. Cell *1*, 223-234.

Young,M.M., Tang,N., Hempel,J.C., Oshiro,C.M., Taylor,E.W., Kuntz,I.D., Gibson,B.W., and Dollinger,G. (2000). High throughput protein fold identification by using experimental constraints derived from intramolecular cross-links and mass spectrometry. Proc. Natl. Acad. Sci. U. S. A *97*, 5802-5806.

Yusupov,M.M., Yusupova,G.Z., Baucom,A., Lieberman,K., Earnest,T.N., Cate,J.H., and Noller,H.F. (2001). Crystal structure of the ribosome at 5.5 A resolution. Science *292*, 883-896.

Zhang,G., Campbell,E.A., Minakhin,L., Richter,C., Severinov,K., and Darst,S.A. (1999). Crystal structure of Thermus aquaticus core RNA polymerase at 3.3 A resolution. Cell *98*, 811-824.