A structural genomics initiative on yeast proteins

Sophie Quevillon-Cheruel,[a] Bruno Collinet,[a] Cong-Zhao Zhou,[a] Philippe Minard,[a] Karine Blondeau,[b] Gilles Henkes,[b] Robert Aufrère,[b] Jérôme Coutant,[c] Eric Guittet,[c] Anita Lewit-Bentley,[d] Nicolas Leulliot,[d] Isabella Ascone,[d] Isabelle Sorel,[e] Philippe Savarin,[e] Ines Li de La Sierra Gallay,[e] Françoise de la Torre,[e] Anne Poupon,[e] Roger Fourme,[f]* Joël Janin[e] and Herman van Tilbeurgh[a]

[a]Institut de Biochimie et de Biophysique Moléculaire et Cellulaire (UMR 8619), Université Paris-Sud, Bâtiment 430, 91405 Orsay, France, [b]Institut de Génétique et Microbiologie (UMR 8621), Université Paris-Sud, Bâtiment 360, 91405 Orsay, France, [c]Institut de Chimie des Substances Naturelles (UPR 2301), 1 Avenue de la Terrasse, 91198 Gif sur Yvette, France, [d]LURE (UMR 130), Bâtiment 209D, Université Paris-Sud, 91898 Orsay, France, [e]Laboratoire d'Enzymologie et Biochimie Structurale (UPR 9063), Bâtiment 34, 1 Avenue de la Terrasse, 91198 Gif sur Yvette, France, and [f]Synchrotron SOLEIL, Bâtiment 209H, Université Paris-Sud, 91898 Orsay, France.
E-mail: roger.fourme@soleil.u-psud.fr

A canonical structural genomics programme is being conducted at the Paris-Sud campus area on baker's yeast proteins. Experimental strategies, first results and identified bottlenecks are presented. The actual or potential contributions to the structural genomics of several experimental structure-determination methods are discussed.

Keywords: structural genomics; yeast; protein; structure.

1. Introduction

Genomics is the science of genomes. Born at the end of 1995 with the publication of the first complete DNA sequence of a bacterium, it is now in the process of causing a profound change in biology and medicine. Systematic sequencing gives access to all genes of a given organism and thus to all the proteins it is able to produce. The objective of structural genomics is to determine the three-dimensional structure of these proteins, and thus structural genomics lies downstream from sequencing. For many of the proteins, even the function is unknown. A known structure and function can then be exploited, as a target for drug design, for example.

A main objective of structural genomics is to have representative models of all families of homologous proteins and to establish a catalogue of all folds (Kim, 1998; Baker & Sali, 2001). To fill the pages of this catalogue is in itself a scientific task, because the knowledge of the structure of one protein can be extended to all of its homologues, *i.e.* those whose genes evolved from a common ancestor.

According to present estimates, about 10000 families of homologous proteins exist in nature at the level of 30% sequence identity, with perhaps 1000 really distinct folds [see Sali (1998) and Fischer & Eisenberg (1999) for recent reviews]. In contrast, the number of proteins coded for by the human genome is estimated to be 38000, yet today we have representative structures for only 20–30% of the families and 30–50% of the folds. The unknown folds are still numerous and most will correspond to new functions. In the genomes we know today, it is possible to assign a fold to less than half of the proteins on the basis of their homology with a known protein structure. It was estimated that about 16000 carefully selected structures will be needed to construct useful models for the vast majority of proteins (Vitkup *et al.*, 2001).

A second objective is to help determine and understand the biochemical function of these proteins. In several cases already, X-ray structure determinations have hinted at what the function of an unknown protein could be (Zarembinski *et al.*, 1998). In a pilot project on archeon proteins, it was observed that five out of ten structures contained either a bound ligand or a ligand binding site that could be inferred from a structural homologue (Christendat *et al.*, 2000). In all cases, structural studies have complemented other approaches and brought essential information. In that sense, structural genomics is part of functional genomics as exemplified by Hegyi & Gerstein (1999) for the yeast genome.

The first eukaryotic genome to be sequenced was that of the baker's yeast *Saccharomyces cerevisiae* (Goffeau *et al.*, 1996; Mewes *et al.*, 1997). The choice of the yeast genome for a structural genomics project is justified by the fact that it is one of the laboratory micro-organisms whose genetics have been studied most extensively (Ross-MacDonald *et al.*, 1999; Winzeler *et al.*, 1999). Only 5% of the genes in yeast contain (mostly small) introns, and therefore most genes can be cloned into expression vectors straight from genomic DNA. A number of fascinating post-genomic studies on yeast have been published in recent years (Hughes *et al.*, 2000). Systematic proteomics studies are being conducted on this organism that will define a global protein-interaction map (Ito *et al.*, 2000; Uetz *et al.*, 2000).

The same reasons that have guided our choice of yeast for sequencing apply to a systematic analysis of the three-dimensional structures of the resulting proteins. Yeast is a self-sufficient organism and, in spite of its small size, its genome probably contains representatives of most of the structural families that exist in nature. At a stage where we know only a few percent of human genes, close to one half of them (and already more than 2000) have a homologue in yeast.

In this paper, the general outlines of a relatively small-scale structural genomics project on yeast proteins conducted at the Paris-Sud campus area are presented. Experimental strategies, first results and identified bottlenecks are discussed. Some aspects of the potential of solution X-ray scattering and X-ray absorption spectroscopy as complementary tools to the key structure-determination methods [protein crystallography (PX) and NMR] are discussed.

2. High-throughput production of yeast proteins

The yeast genome contains around 6200 ORFs (open reading frames, or DNA sequences that code for a protein). The yeast structural genomics project was subdivided into several phases, during which the best experimental approaches were tested at the same time as the first results were being obtained, allowing the efficiency of the system to improve during its execution. We are now engaged in the pilot-project phase, with the goal of achieving the preparation of a few hundred proteins from 300 yeast ORFs and the resolution of about 20 three-dimensional structures.

Here we first describe the general strategy used to proceed from genomic DNA to proteins in test tubes ready for structural analysis. The goal is to obtain 10–50 mg of purified protein by a limited series of well standardized steps. The general strategy is close to those adopted in other structural genomics programmes (Christendat *et al.*, 2000; Yee *et al.*, 2002).

Table 1
Selection of ORFs in the Paris-Sud initiative.

Total number of yeast ORFs	6213
Proteins having two or more transmembrane segments	864
Proteins containing coiled-coil regions of more than 40 residues	157
Proteins with low complexity	54
Proteins having more than 20% sequence identity with a protein of known three-dimensional structure	1949
Proteins having more than 500 residues	942
Proteins with more than one domain	877
Proteins paralogous to proteins already in the list	880
Proteins eliminated after manual inspection	112
Targets of other structural genomics projects	65
Final number of potential candidates	313

2.1. Selection of an ORF

Not all proteins are equally well adapted for a high-throughput structure-determination approach. In order to test technologies and to establish protocols, a subset of yeast proteins was selected (Table 1). Membrane and multiple-domain proteins were discarded, as well as proteins containing low-complexity regions and coiled-coil domains. The second filter is the search for homologies using sequence comparisons (*DARWIN*; Gonnet *et al.*, 1992; *FASTA* and *BLAST*). This allowed us to divide the ORFs into three categories: those homologous to a known structure, those homologous to proteins whose structures are not known and those that do not have a clearly identified homologue. This filter also allowed us to define the breakdown of multi-domain proteins, especially if each domain represents homology with a different protein (identified by scanning through databases such as *PRODOM*). The third filter uses the technique of motif search (*ProfileScan*, *PFAM*). It also allowed us to define the domain organization of the protein and sometimes even the definition of its function. The fourth filter is the search for homologies using multiple and/or iterative alignments (*HMMER*, *PFAM*, *PSI-BLAST*), which are more sensitive than pairwise sequence alignments.

The last filter uses fold-recognition techniques [3*DPSSM* (Kelley *et al.*, 2000) and *FROST* (Marin *et al.*, 2001)], which align a sequence based on three-dimensional structure and which can take over from standard sequence-comparison methods when the percentage of identity falls below 20% and hence detection of homologies by sequence comparison is no longer adequate.

After these automated filters, each protein was examined individually. At this stage, about one candidate in four was eliminated for some reason: for example, because of the presence of a transmembrane segment that was not automatically detected or of a low-complexity segment (using hydrophobic cluster analysis; Callebaut *et al.*, 1997), or because the protein is already targeted in another structural genomics project. To reduce the chances of targeting multi-domain proteins, only proteins with size lower than 500 residues were selected.

2.2. The choice of an expression system

We have chosen to express most yeast proteins in a bacterial host, *Escherichia coli*. Expression is carried out by inserting the ORF into an expression vector with the following characteristics:

(i) the presence of a selection tag, needed to follow the presence of the vector in the host cell (resistance to kanamycin, to chloramphenicol or to tetracycline);

(ii) a plasmid present in a large number of copies, favouring the production of large quantities of protein;

(iii) the presence of an extremely efficient promoter allowing optimal induction of protein production; vectors carrying T7 polymerase and inducible by IPTG were selected.

Given the eukaryotic origin of the proteins, we foresaw a second host organism for their expression, *P. Pastoris*. This yeast is commonly used for the production of large quantities of proteins and provides a good system for eukaryotic expression, allowing important post-translational modifications such as glycosylations. It is as easy to manipulate as *E. coli*, and several comparative studies have shown that the expression of heterologous proteins is ten to 100 times more efficient in *P. pastoris* than in *S. cerevisiae*.

2.3. Cloning and expressing yeast ORFs

In *E. coli*, a commonly used expression system is based on the phage T7 polymerase promoter with the product fused to a polyhistidine tag. A procedure to check the efficiency of expression and the solubility of the expressed protein in parallel was set up. Our preliminary results show that a large fraction of the cloned ORFs are expressed with a good yield but as insoluble material in inclusion bodies. We therefore developed a simple renaturation procedure, which can be performed systematically on a test-tube culture using the solubility of the His-tagged material as a criterion (Colinet, 2003). Expression, resolubilization and refolding are carried out in parallel on several ORFs in a standardized procedure. Those ORFs that cannot be solubilized at this stage are inserted into an alternative expression system, which makes use of cotransfection with chaperone-carrying plasmids. Fusion with a solubility-promoting partner such as maltose binding protein did not work well in our hands, because of the very low yields in the final proteolytic cleavage and concentration steps.

2.4. Extraction and purification of proteins

The introduction of a polyhistidine tag, allowing the selective affinity purification of the overexpressed proteins on a Ni^{2+}-grafted column in a single step, proved to be a very robust and efficient method for obtaining sufficiently purified material. We opted for the shortest possible linker between the protein and the His-tag, avoiding as much as possible any interference with the crystallization process. The choice of a short linker was coupled to the decision not to cleave the His-tag in the final preparation steps. Proteolytic cleavage on a preparative scale is a costly and time-consuming step, which has often to be optimized when changing the target protein. In view of the relatively high success in obtaining crystals from soluble proteins, this choice seems to be justified. The final step of the purification procedure was gel filtration, mainly to remove the high imidazole concentrations needed for elution from the Ni matrix. This step also allowed us to estimate the monodispersity of the concentrated protein sample.

Mass spectroscopy and SDS-PAGE (sodium dodecyl sulfate-polyacrylamide gel electrophoresis) were used systematically for checking the purity of the samples. Systematic isoelectric focusing will be added at a later stage. This will show the presence of possible microheterogeneities and will give information on the experimental isoelectric point of the protein, which could help in setting up crystallization strategies.

3. Experimental structure determination

3.1. High-resolution methods: protein crystallography and NMR

Protein crystallography (PX) is the most prominent method for structure determination at near-atomic resolution. In recent years, long-awaited structures have been solved with milestones such as the nucleosome core particle, GroEL-GroES, F1 ATPase, photosystem I and the ribosome. The case of ribosome is especially noteworthy. It took a long time to master technical difficulties (*e.g.* crystallization, cryocooling, heavy-atom substitution), yet the final crystal structure determination proper has been impressively fast (Ban *et al.*, 2000; Wimberley *et al.*, 2000; Yusupov *et al.*, 2001). Instrumental progress (synchrotron radiation, detectors, cryocooling methods) coupled to theoretical and computing advances have led to faster data collection and structure solution. Structural genomics programmes are a strong incentive to go further towards the high-throughput stage.

Stringent conditions are required for PX. First, a concentrated solution of pure and structurally homogeneous macromolecules must be prepared. Then, X-ray-grade crystals must be grown. The efficient exploration of the space of crystallization parameters is facilitated by statistical experimental approaches and the use of commercial screening kits. Automation of the crystallization steps will certainly be one of the major outcomes of structural genomics projects. It will dramatically increase the speed and the number of experiments that can be carried out. At the present stage of our project, we are making use of a pipetting robot to set up the crystallization trays. Micro- to X-ray-grade crystals were obtained from an unexpectedly high number of ORFs that made it to the stage of pure, concentrated and soluble sample. The optical verification of the crystallization experiments, however, quickly becomes cumbersome and very time-consuming. Once micro-crystals are obtained, the crystallization conditions have to be repeated and optimized. This step has not been automated yet and requires a lot of manual intervention. At this stage, any biochemical information on the target protein is very helpful. This knowledge can be exploited, for instance, by adding small molecule inhibitors or stabilizing metal ions to the initial screening crystallization conditions. Going from small crystals in the initial robot assay to well behaved reproducible X-ray-grade crystals is a far from trivial aspect of the whole process and the most time-consuming step at the moment. Once crystals have been obtained, diffraction data have to be collected. The high brilliance of synchrotron sources, in our case LURE at Orsay and ESRF at Grenoble, reduces the data-collection times by orders of magnitude, and the continuous spectrum of radiation gives a choice of the wavelength that optimizes anomalous scattering. One can expect that further improvements in data collection will not come primarily from an increase in beam intensities available at third-generation facilities but rather in speeding up the changing of samples at the beamlines. Indeed, during a synchrotron data collection, much if not most of the time is lost with sample mounting, testing and actually making the trip to the synchrotron source. A great effort must therefore be made in the optimization of sample handling and automatic data collection and reduction. Some prototypes of automatic sample changers are in their testing phase. Expert systems being developed at present will considerably reduce the data-reduction process and should also speed up data collection. Solving the phase problem in protein crystallography now relies heavily on multiple anomalous dispersion (MAD) on selenomethionine substituted crystals. This procedure works very well routinely, but a number of our crystallized ORFs did not contain (enough) methionines. Searches for heavy-metal binding has to be banned from high-throughput procedures and other solutions have to be devised. One possibility is the incorporation of methionines using site-directed mutagenesis. *In vitro* expression of proteins will in principle enlarge the scope of labels that can be incorporated.

NMR spectroscopy has long been limited to the study of small proteins. Recent developments of the technique and the availability of high-field spectrometers have greatly extended the field of application. The feasibility limit is being continually pushed further and it is now possible to study proteins of 30–40 kDa. Over the past few years, NMR has seen spectacular success in the study of protein structure and dynamics, as well as protein interactions.

NMR has the advantage of avoiding the cumbersome need for crystals, but the technique does sometimes impose constraints on other aspects of the samples under study (pH and temperature stability, monodispersity *etc.*). NMR studies require conditions where the protein is correctly folded and does not aggregate. Experience shows that the search for appropriate conditions (salt, pH, temperature, concentration *etc.*) is crucial to the success of the study. It is often necessary to work with low protein concentrations (100 µ*M*) in order to reduce aggregation. This requires very sensitive spectrometers together with recently developed cryoprobes and very high magnetic fields (Chou *et al.*, 2001). The protein has to be stable throughout the period of study (about two weeks for data collection overall).

The assignment is the identification of the individual resonance frequencies for all atoms in the protein. The availability of samples labelled with ^{15}N and ^{13}C offers the only possibility of solving this problem for larger proteins (above 100 amino acids). Multidimensional and multinuclear NMR techniques make assignment a relatively easy and rapid task. We must mention that, in the case of proteins of 25 kDa and over, (total or partial) deuterium labelling may help simplifying the spectra and reduce spin relaxation. This labelling has proved to be feasible in a high-throughput manner.

The usefulness of NMR in structural proteomics was recently illustrated by the results of a study on 513 proteins from five different microorganisms (Yee *et al.*, 2002). The proteins were distributed from a central production unit to several NMR laboratories. Protein aggregation and poor solubility are equally recognized as major bottlenecks. NMR also struggles with long data-collection times (three to four weeks per sample). Much improvement should come from (i) central sample preparation facilities that could rapidly test solubilization conditions and (ii) development of automatic resonance and nuclear Overhauser effect-assignment programs.

PX and NMR tend to require partnerships with other structural techniques such as cryoelectron microscopy (van den Ent *et al.*, 2001) and atomic force microscopy (Smith *et al.*, 2001). The whole field of structural genomics will benefit from an integration of a number of other structural methods. In particular, solution X-ray scattering (possibly complemented by neutron scattering) and X-ray absorption spectroscopy are clearly complementary to PX and NMR in structural genomics programmes.

3.2. Solution X-ray scattering

The *ab initio* shape-determination methods (Svergun *et al.*, 1996; Chacon *et al.*, 1998; Svergun, 1999; Walther *et al.*, 1999) allow the restoration of the low-resolution (20–30 Å) shape of macromolecules. A recently developed method, in which a protein is represented by an ensemble of dummy residues, allows reliable reconstruction of the domain structure of proteins at a resolution of about 10 Å, based on small- and medium-angle X-ray scattering data (Svergun *et al.*, 2001). The above methods can be employed when high-quality crystals are not available and/or the molecular weight precludes structure determination by NMR. The *ab initio* models, especially those provided by the dummy residue method, yield a description of the general architecture of the macromolecule. The domain structure of a 50 kDa protein can be routinely determined within half a day, including 4 h for a full-scale data collection at small and medium angles on a synchrotron radiation beamline plus 8 h for an *ab initio* calculation using a standard personal computer. Such models can be used for molecular replacement in protein crystallography, and they can be incorporated as constraints in protein-folding predictions (Hao *et al.*, 1999; Ockwell *et al.*, 2000).

Protein crystallography often yields models of macromolecules in which some parts are missing. This is the case in particular when flexible loops or domains in proteins are disordered, or when portions of the structure have been genetically removed to facilitate crystallization. To add missing parts, the known part of the structure is fixed and the rest is built around it to fit the experimental scattering data from the entire particle. Moreover, solution scattering patterns of multi-domain proteins and macromolecular complexes can also be fitted using rigid-body modelling if high-resolution structures of individual domains or subunits are available.

Solution X-ray scattering, as a specific probe of quaternary structure, can also be used to monitor the structural response of a macromolecule to a perturbation, such as ligand binding. Here, one is not primarily concerned with the structure, which, in the most favourable case, has been determined at high resolution, but with its modifications in a direct study of the structure-function relationship. Structural transitions can be studied at equilibrium, but the kinetics can also be investigated using a combination of fast-mixing techniques and time-resolved recording of scattering patterns (*e.g.* Segel *et al.*, 1999).

Finally, X-ray solution scattering can be used as a tool in the crystallization stage, especially in difficult cases that do not yield crystals after a reasonable number of trials. The coarse-grain characterization of the presence of aggregates or large-molecular-weight oligomers in protein solution provided by elastic or quasi-elastic light scattering can be refined by using X-ray scattering on solutions of the protein of interest, providing a factorial exploration of the parameter subspace already determined (*e.g.* Mourey *et al.*, 1997). Within the framework of a structural genomics initiative, this contribution would be of interest if a robot is available to prepare small-volume samples to be automatically loaded in the beam for the few-second exposure required on a third-generation synchrotron radiation source.

3.3. X-ray absorption spectroscopy (XAS)

XAS focuses investigation on an absorbing atomic species and the first coordination spheres of relevant atoms. Accessible species range from relatively light atoms (Mg, P, S, Cl, . . .) with *K*-edges at long wavelengths, which require special instrumentation, to many metallic atoms (from K and Ca to heavier atoms) with edges accessible on conventional XAS synchrotron radiation beamlines. The occurrence of metal centres in proteins is high, as about 30% of proteins coded by genomes are metalloproteins. 30–50% of all enzymes carry protein-bound metal centre(s) located mostly at the catalytic site. XAS does not require crystals and is applicable to insoluble as well as soluble proteins (Hasnain & Hodgson, 1999). As a local method focused on a limited and structurally quite stable portion of the protein structure, it is likely to be more tolerant than PX or NMR to sample inhomogeneity and aggregation. In order to develop XAS studies on metalloproteins in the frame of structural genomics programmes, two problems will have to be carefully considered: the identification of potential metal-coordinating sites in the genetic information and the expression of a potential metalloprotein with incorporation of the biologically relevant metal, which is an important post-transcriptional event.

The structural genomics initiative presented in this article was one of the first to incorporate explicitly a methodological programme on the development of XAS for structural genomics. A more thorough discussion on how XAS could contribute to structural genomics is given in the introductory article in this issue (Ascone *et al.*, 2002).

4. Current status of the initiative and perspectives

A global overview of the state of the Paris-Sud initiative on yeast proteins is represented in Fig. 1. 86% of the target proteins have been expressed and 61% are in a soluble state. The remaining proteins were usually produced as inclusion bodies. Today, 53% of the soluble proteins have been purified. NMR analysis, very useful at the initial stages of protein characterization, revealed that about one-third of the proteins are non-structured. 40% of the purified structured proteins have yielded crystals of some sort. Structures of three proteins have been solved by the MAD method. Good MAD data sets have been collected for two other proteins. High-quality crystals have been obtained for five other proteins. An NMR study is in progress.

The high-throughput structure determination of proteins seems to be a realistic objective even for a small-scale budget, as was the case for this project. The standardization and semi-automatization of the cloning, purification, expression and crystallization will certainly allow us to obtain tens of protein crystals and protein structures. The optimization of the production conditions compensates in many cases for the lack of biochemical knowledge of the protein targets. In a classical approach, one usually gathers a wealth of biochemical and functional information on a protein target before starting structural studies. This information helps in determining the strategy of structure determination. For instance, the choice of an appropriate inhibitor may be a decisive factor in obtaining high-quality crystals, the knowledge of domain organization may help in designing more workable protein constructions *etc.* Even if this information were present, it cannot always be included in a general production approach, yet these considerations may become crucial during the optimization process necessary to obtain well diffracting crystals. This step remains a bottleneck in the whole process, because it is difficult to automate since it depends heavily on the nature of the protein. Ongoing developments such as the systematic screening of refolding buffers and the improvement of protein solubilization during expression (co-expression with chaperones, *in vitro* expression or expression in strains resistant to toxic proteins) should further improve the yields of proteins amenable to structural studies.

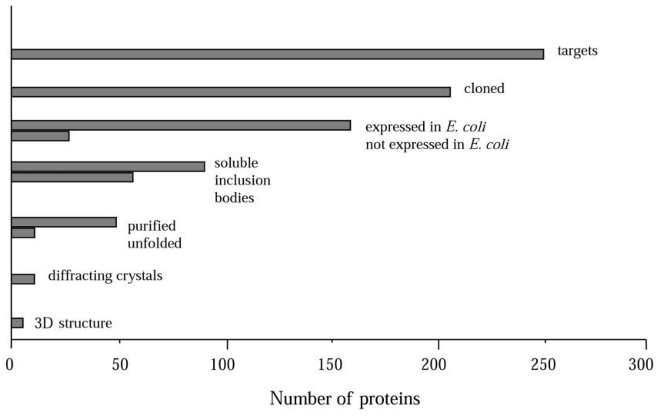

Figure 1
Global overview of the Paris-Sud initiative by May 2002.

Dmitri Svergun (EMBL, Hamburg) and Patrice Vachette (LURE, Orsay) are gratefully acknowledged for their suggestions and comments about the use of solution scattering for structural genomics.

References

Ascone, I., Fourme, R. & Hasnain, S. S. (2003). **10**, 1–3.

Baker, D. & Sali, A. (2001). *Science*, **294**, 93–96.

Ban, N., Nissen, P., Moore, P. B. & Steitz, T. A. (2000). *Science*, **289**, 905–920.

Callebaut, I., Labesse, G., Durand, P., Poupon, A., Canard, L., Chomilier, J., Henrissat, B. & Mornon, J.-P. (1997). *Cell. Mol. Life Sci.* **53**, 621–645.

Chacon, P., Moran, F., Diaz, J. F., Pantos, E. & Andreu, J. M. (1998). *Biophys. J.* **74**, 2760–2775.

Chou, J. J., Li, S., Klee, C. B. & Bax, A. (2001). *Nature Struct. Biol.* **8**, 990–997.

Christendat, D., Yee, A., Dharamsi, A., Kluger, Y., Savchenko, A., Cort, J. R., Booth, V., Mackereth, C. D., Saridakis, V., Ekiel, I., Kozlov, G., Maxwell, K. L., Wu, N., McIntosh, L. P., Gehring, K., Kennedy, M. A., Davidson, A. R., Pai, E. F., Gerstein, M., Edwards, A. M. & Arrowsmith, C. H. (2000). *Nature Struct. Biol.* **7**, 903–909.

Colinet, B. (2003). In preparation.

Ent, F. van den, Amos, L. A. & Lowe, J. (2001). *Nature* (*London*), **413**, 39–44.

Fischer, D. & Eisenberg, D. (1999). *Curr. Opin. Struct. Biol.* **9**, 208–211.

Goffeau, A., Barrell, B. G., Bussey, H., Davis, R. W., Dujon, B., Feldmann, H., Galibert, F., Hoseihel, J. D., Jacq, C., Johnston, M., Louis, E. J., Mewes, H. W., Murakami, Y., Phillippsen, P., Tettelin, H. & Oliver, S. G. (1996). *Science*, **274**, 546–567.

Gonnet, G., Cohen, M. & Benner, S. (1992). *Science*, **256**, 1443–1445.

Hao, Q., Dodd, F. E., Grossmann, J. G. & Hasnain, S. S. (1999). *Acta Cryst.* D**55**, 243–246.

Hasnain, S. S. & Hodgson, K. O. (1999). *J. Synchrotron Rad.* **6**, 852–864.

Hegyi, H. & Gerstein, M. (1999). *J. Mol. Biol.* **288**, 147–164.

Hughes, T. R., Marton, M. J., Jones, A. R., Roberts, C. J., Stoughton, R., Armour, C. D., Bennett, H. A., Coffey, E., Dai, H., He, Y. D., Kidd, M. J., King, A. M., Meyer, M. R., Slade, D., Lum, P. Y., Stepaniants, S. B., Shoemaker, D. D., Gachotte, D., Chakraburtty, K., Simon, J., Bard, M. & Friend, S. H. (2000). *Cell*, **102**, 109–126.

Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M. & Sakaki, Y. (2000). *Proc. Natl Acad. Sci. USA*, **98**, 4569–4574.

Kelley, L. A., MacCallum, R. M. & Sternberg, M. (2000). *J. Mol. Biol.* **299**, 499–520.

Kim, S. H. (1998). *Nature Struct. Biol.* **5**, 643–645.

Marin, A., Pothier, J., Zimmerman, K. & Gibrat, J.-F. (2001). *Protein Structure Prediction: Bioinformatic Approach*, edited by I. Tsigelny. La Jolla, CA: International University Line.

Mewes, H., Albedrman, K., Bähr, M., Frishman, D., Gleissner, A., Hani, J., Heumann, K., Kleine, K., Maierl, A., Olver, S., Pfeiffer, F. & Zollner, A. (1997). *Nature* (*London*), **387**, 7–9.

Mourey, L., Pedelacq, J. D., Fabre, C., Causse, H., Rouge, P. & Samama, J. P. (1997). *Proteins*, **29**, 433–442.

Ockwell, D. M., Hough, M. A., Grossmann, J. G., Hasnain, S. S. & Hao, Q. (2000). *Acta Cryst.* D**56**, 1002–1006.

Ross-MacDonald, P., Coelho, P. S., Roemer, T., Agarwal, S., Kumar, A., Jansen, R., Cheung, K. H., Sheehan, A., Symoniatis, D., Umansky, L., Heidtman, M., Nelson, F. K., Iwasaki, H., Hager, K., Gerstein, M., Miller, P., Roeder, G. S. & Snyder, M. (1999). *Nature* (*London*), **402**, 413–418.

Sali, A. (1998). *Nature Struct. Biol.* **5**, 1029–1032.

Segel, D. J., Bachmann, A., Hofrichter, J., Hodgson, K. O., Doniach, S. & Kiefhaber, T. (1999). *J. Mol. Biol.* **288**, 489–499.

Smith, D. E. *et al.* (2001). *Nature* (*London*), **413**, 748–752.

Svergun, D. I. (1999). *Biophys. J.* **76**, 2879–2886.

Svergun, D. I., Petoukhov, M. V. & Koch, M. H. (2001). *Biophys. J.* **80**, 2946–2953.

Svergun, D. I., Volkov, V. V., Kozin, M. B. & Stuhrmann, H. B. (1996). *Acta Cryst.* A**52**, 419–426.

Uetz, P., Glot, L., Cagney, G., Mansfield, T. A., Judson, R. S., Knight, J. R., Lockshon, D., Narayan, V., Srinlvasan, M., Pochart, P., Qureshi-Emili, A., Ying, L., Godwin, B., Conover, D., Kalbfleish, T., Vijayadamodar, G., Meijia, Y., Johnston, M., Fields, S. & Rothberg, J. M. (2000). *Nature* (*London*), **403**, 623–627.

Vitkup, D., Melamud, E., Moult, J. & Sander, C. (2001). *Nature Struct. Biol.* **8**, 559–566.

Walther, D., Cohen, F. E. & Doniach, S. (1999). *J. Appl. Cryst.* **33**, 350–363.

Wimberley, B. T., Brodersen, D. E., Clemons, W. M. Jr, Morgan-Warren, R. J., Carter, A. P., Vonrhein, C., Hartsch, T. & Ramakrishan, V. (2000). *Nature* (*London*), **407**, 327–339.

Winzeler, E. A., Shoemaker, D. D., Astromoff, A., Liang, H., Anderson, K., Andre, B., Bangham, R., Benito, R., Boeke, J. D., Bussey, H., Chu, A. M., Connelly, C., Davis, K., Dietrich, F., Dow, S. W., El-Bakoury., M., Foury, F., Friend, S. H., Gentalen, E., Giaever, G., Hegemann, J. H., Jones, T., Laub, M., Hong, L., Liebundguth, N., Lockhart, D. J., Lucau-Danila, A., Lussier, M., M'-Rabet, N., Menard, P., Mittmann, M., Chai, P., Rebischung, C., Revuelta, J. L., Riles, L., Roberts, C. J., Ross-Macdonald, P., Scherens, B., Snyder, M., Sookhai-Mahadeo, S., Storms, R. K., Veronneau, S., Voet, M., Volckaert, G., Ward, T. R., Wysocki, R., Yen, G. S., Yu, K., Zimmermann, K., Philipsen, P., Johnston, M. & Davis, R. W. (1999). *Science*, **285**, 901–906.

Yee, A., Chang, X., Pineda-Lucena, A., Wu, B., Semesi, A., Le, B., Ramelot, T., Lee, G. M., Bhattacharyya, S., Gutierrez, P., Denisov, A., Lee, C. H., Cort, J. R., Kozlov, G., Liao, J., Finak, G., Chen, L., Wishart, D., Lee, W., McIntosh, L. P., Gehring, K., Kennedy, M. A., Edwards, A. M. & Arrowsmith, C. H. (2002). *Proc. Natl Acad. Sci. USA*, **99**, 1825–1830.

Yusupov, M. M., Yusupova, G. Z., Baucom, A., Lieberman, K., Earnest, T. N., Cate, J. H. D. & Noller, H. F. (2001). *Science*, **292**, 883–896.

Zarembinski, T. I., Hung, L. W., Mueller-Dieckmann, H. J., Kim, K. K., Yokota, H., Kim, R. & Kim, S. H. (1998). *Proc. Natl Acad. Sci. USA*, **95**, 15189–15913.

Received 25 May 2002 • Accepted 29 August 2002