ARTICLES

Facilities and Methods for the High-Throughput Crystal Structural Analysis of Human Proteins

UDO HEINEMANN,*,[†,‡] KONRAD BÜSSOW,[§,||]
UWE MUELLER,[‡,⊥] AND PATRICK UMBACH[‡,||]
Forschungsgruppe Kristallographie, Max-Delbrück-Centrum für Molekulare Medizin, Robert-Rössle-Strasse 10, D-13125 Berlin, Germany, Institut für Chemie/Kristallographie, Freie Universität, Takustrasse 6, D-14195 Berlin, Germany, Max-Planck-Institut für Molekulare Genetik, Ihnestrasse 73, D-14195 Berlin, Germany, Protein Structure Factory, Heubnerweg 6, Haus D, D-14059 Berlin, Germany, and Protein Structure Factory, c/o BESSY GmbH, Albert-Einstein-Strasse 15, D-12489 Berlin, Germany

Received June 10, 2002

ABSTRACT
Facilities and methods for the high-throughput crystal structure analysis of human proteins are described as recently established in the Protein Structure Factory, a Berlin-area structural genomics project. Genes encoding human proteins are expressed in either recombinant *Escherichia coli* or yeast (*Saccharomyces cerevisiae* or *Pichia pastoris*). To facilitate and standardize protein purification, the target proteins are produced with various tags for affinity chromatography. For high-throughput crystallization, a robotic station is being set up that has the capacity to handle 960 000 experiments simultaneously. The resulting protein crystals will be subjected to X-ray diffraction experiments at the third-generation synchrotron storage ring BESSY where protein crystallography beamlines are currently under construction. The Protein Structure Factory's strategy for high-throughput production and structure analysis of human proteins is evaluated based on first results.

Structural Genomics and the Protein Structure Factory

The Protein Structure Factory (www.proteinstrukturfabrik.de) is part of the international structural genomics initiative[1] that aims at determining the three-dimensional structures of all proteins encoded by the genomes of key organisms. A specific aim of this common endeavor is to achieve a coverage of "protein sequence space" with crystal or NMR structures such that for every protein there will be a template protein of known 3D structure sharing at least 30% of identical residues, on which a homology modeling could be based. It has been estimated[2] that this goal requires to determine ∼16 000 new protein structures that have been selected to evenly cover protein space and serve as modeling templates. If other criteria for the selection of target proteins are also applied, the number of experimental structures to be solved will increase.[2] Considering these numbers it seems obvious that structural genomics will fail without the development and implementation of high-throughput protein production and analysis methods and facilities.[3] The Protein Structure Factory is committed to establishing high-throughput techniques[4] for the NMR and crystal structure analysis of proteins from human origin.

The Protein Structure Factory comprises a complete structural genomics pipeline starting from expression cloning and protein production and extending to protein structure analysis and ligand identification (Figure 1). The consortium focuses on human proteins which hold promise to deepen our understanding of health and disease. These proteins pose certain technical challenges, since they often have a multidomain structure or are of little stability, making their preparation and structure analysis difficult. To exclude the most problematic molecules, the Protein Structure Factory selects small, water-soluble, monomeric or homo-oligomeric proteins for structure analysis. The size limit for crystal structure analysis is currently set to 500 amino acid residues per protein chain, since larger polypeptides are more likely to cause difficulties in recombinant production, especially in a high-throughput regime. Small proteins are alternatively stud-

* To whom correspondence should be addressed. Tel: +49 30 9406 3420. Fax: +49 30 9406 2548. E-mail: Heinemann@MDC-Berlin.de.

[†] Max-Delbrück-Centrum für Molekulare Medizin.

[‡] Freie Universität.

[§] Max-Planck-Institut für Molekulare Genetik.

[||] Protein Structure Factory, Heubnerweg 6.

[⊥] Protein Structure Factory, c/o BESSY GmbH.

Udo Heinemann studied chemistry in Göttingen, where he obtained a doctorate degree for work in protein crystallography carried out at the Max-Planck-Institut für experimentelle Medizin in 1982. After postdoctoral research at the University of California at Los Angeles, he joined the Freie Universität Berlin in 1985. He spent a year at Stuttgart University before being appointed as a research group leader for crystallography at the Max-Delbrück-Centrum für Molekulare Medizin Berlin-Buch in 1993. Since 1995 he has also been a Full Professor for protein crystallography at the Freie Universität Berlin. He is interested in structure−function relations of proteins and nucleic acids and has been actively engaged in structural genomics recently.

Konrad Büssow studied Biochemistry at the Freie Universität Berlin. After spending half a year at the Wellcome Trust Centre for Human Genetics in Oxford, U.K., he joined the department of Prof. Hans Lehrach at the Berlin Max-Planck-Institut für Molekulare Genetik as a graduate student. He obtained a doctoral degree for work on arrayed cDNA expression libraries and high-throughput protein expression and purification methods. Since then he has been head of the *E. coli* expression group in the Protein Structure Factory, where he coordinates the production and characterization of expression clones for protein structure analysis.

Uwe Mueller studied chemistry at the Humboldt-Universität zu Berlin and obtained a doctoral degree for work in macromolecular crystallography performed at the Max-Delbrück-Centrum für Molekulare Medizin Berlin-Buch. He is head of the protein crystallography outstation of the Protein Structure Factory at BESSY, Berlin. His current research interests circle around synchrotron X-ray diffraction and high-throughput protein crystallization techniques.

Patrick Umbach studied chemistry at the Technische Universität Berlin and obtained a doctoral degree from the Freie Universität for work in protein crystallization. After one and a half years as a postdoctoral fellow in the Department of Structural Biology at Abbott Laboratories, Chicago, he joined the Protein Structure Factory, where he was first involved in high-throughput protein crystallization and now acts as adminstrative coordinator and head of the central laboratory unit of the project.

10.1021/ar010129t CCC: $25.00
Published on Web 01/22/2003

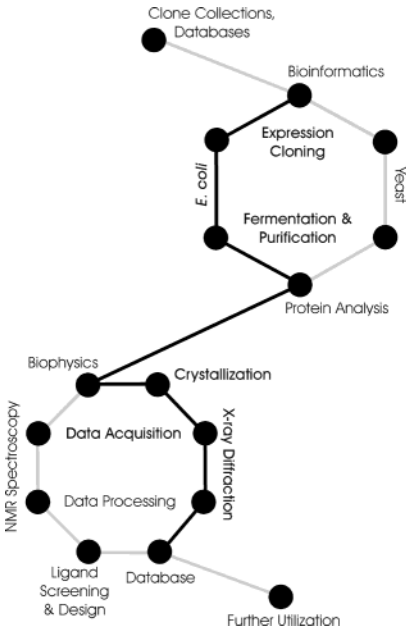

FIGURE 1. Workflow within the Protein Structure Factory. Material and information flows from top to bottom. Bifurcations permit the choice of either *E. coli* or yeast for expression cloning and protein production and either NMR spectroscopy or X-ray crystallography for structure analysis. This Account focuses on the parts of the project highlighted by bold-face print, following the path indicated by the black line.

ied by NMR structure analysis. Proteins containing predicted transmembrane regions, coiled-coil structures, or extended low-complexity sequences (often with runs of like amino acid residues) are likely to be insoluble in aqueous buffers or unstructured. These proteins are not considered for structure analysis within the Protein Structure Factory.

The genes encoding the selected proteins are expressed in recombinant bacterial or yeast cells and the gene products purified by affinity chromatography. Their purity and chemical integrity is examined by SDS−PAGE, and for each protein a biophysical fingerprint is generated. Here, circular dichroism and FT-infrared spectra provide information regarding the protein's secondary structure, spectroscopically or calorimetrically monitored unfolding curves, or calorimetrically characterize its conformational stability, and dynamic light scattering examines the particle distribution in the protein solution. These studies permit to predict whether a protein lends itself to 3D structure analysis and thus to plan the structure solving process. Proteins proving amenable to structure analysis in initial experiments are then produced either with ^{13}C and/or ^{15}N isotope labels for NMR analysis or with selenomethionine to permit anomalous-diffraction phasing[5] in crystal structure analysis. In addition to structure determination, NMR techniques[6] are also employed to characterize the binding of small ligands to a subset of the studied proteins. Proteins to by analyzed by X-ray diffraction methods are crystallized in a high-throughput robotic station.[7] Diffraction experiments are performed at the BESSY II synchrotron storage ring in Berlin. The NMR and X-ray structures, along with all other relevant data, are stored in the project database, before they are released to the Protein Data Bank.[8]

Space does not permit an in-depth discussion of all work done at the Protein Structure Factory. Below, we shall thus concentrate on concepts and facilities employed in the high-throughput crystal structure analysis of proteins (see Figure 1), leaving aside other aspects of the Protein Structure Factory.

Production of Recombinant Proteins for Structure Analysis

Structural genomics requires methods for high-throughput cloning and expression of the genes encoding the target proteins. The strategy of cloning and expression depends on the donor organism, the available expression systems, and the number of targets. If the target proteins originate from a narrow focus, more optimization effort is applied to each protein than in a genome-driven approach where losses of individual polypeptides can be tolerated, because the structure of a homologous protein will often serve the same purpose.

Cloning Strategy. Cloning of open reading frames is straightforward for microorganisms lacking introns, especially if the complete genome sequence is available. For higher organisms, the cloning of open reading frames can be challenging. If suitable cDNA clones are unavailable, open reading frames have to be cloned from RNA extracts of the target tissues. Genetic heterogeneity and alternative splicing have to be taken into account, and the cloned open reading frames have to be verified carefully on the sequence level. On the other hand, clone resources are being generated for man, mice, and other organisms. These clone resources, available from public and commercial sources, can represent an excellent starting material for structural genomics projects, especially if vector systems are used that allow for easy subcloning into various expression vectors.[9]

The expression strategy also depends on the donor organism. Proteins of bacterial origin, especially the stable proteins from thermophilic organisms, are often well suited for production in a bacterial expression system. Proteins of mammalian origin, on the other hand, are often difficult to express in bacterial systems. For them, expression systems such as yeast or cell culture may be more appropriate.

In the Protein Structure Factory, human open reading frames are mainly obtained from partially sequenced clones of the IMAGE consortium,[10] which are distributed by the German Resource Center (www.rzpd.de). Sequence analysis of the IMAGE clones identified a subset of clones containing complete open reading frames. The available open reading frames were compared with a list of suitable target proteins for structural analysis, chosen as explained above.

Expression Strategy. Subject to the expression system and the protein sequence, the results of gene expression and protein purification experiments are highly diverse.

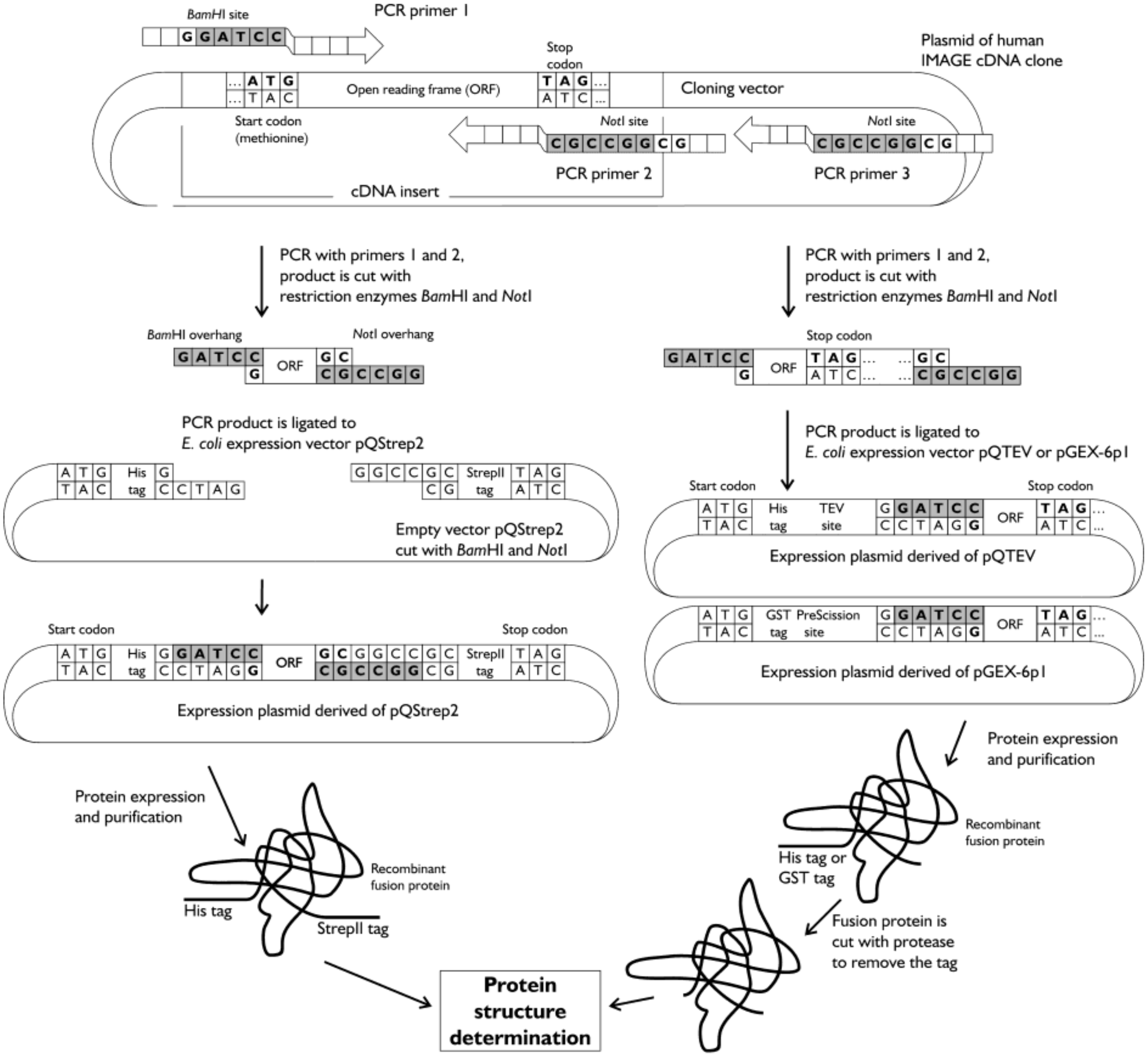
BamHI site
G G A T C C
PCR primer 1
Stop codon
Plasmid of human IMAGE cDNA clone
... A T G
... T A C
Open reading frame (ORF)
T A G ...
A T C ...
Cloning vector
Start codon (methionine)
NotI site
C G C C G G C G
NotI site
C G C C G G C G
PCR primer 2
PCR primer 3
cDNA insert
PCR with primers 1 and 2, product is cut with restriction enzymes BamHI and NotI
PCR with primers 1 and 2, product is cut with restriction enzymes BamHI and NotI
BamHI overhang
NotI overhang
G A T C C
G
ORF
G C
C G C C G G
Stop codon
G A T C C
G
ORF
T A G ... ... G C
A T C ... ... C G C C G G
PCR product is ligated to E. coli expression vector pQStrep2
PCR product is ligated to E. coli expression vector pQTEV or pGEX-6p1
A T G His G
T A C tag C C T A G
G G C C G C StrepII T A G
C G tag A T C
Empty vector pQStrep2 cut with BamHI and NotI
Start codon
Stop codon
A T G His TEV G G A T C C ORF T A G ...
T A C tag site C C T A G G A T C ...
Expression plasmid derived of pQTEV
A T G GST PreScission G G A T C C ORF T A G ...
T A C tag site C C T A G G A T C ...
Expression plasmid derived of pGEX-6p1
Start codon
Stop codon
A T G His G G A T C C ORF G C G G C C G C StrepII T A G
T A C tag C C T A G G C G C C G G C G tag A T C
Expression plasmid derived of pQStrep2
Protein expression and purification
Protein expression and purification
Recombinant fusion protein
Recombinant fusion protein
His tag
StrepII tag
His tag or GST tag
Fusion protein is cut with protease to remove the tag
Protein structure determination

FIGURE 2. Construction of *E. coli* expression clones. Open reading frames are amplified by the polymerase chain reaction (PCR) from IMAGE cDNA clones. PCR products are equipped with overhangs that can be cut with the restriction enzymes *Bam*HI and *Not*I. Alternative enzymes are used if open reading frames contain recognition sites of these enzymes (*Bgl*II, *Eco*31I, *Esp*3I). PCR products of primers 1 and 2 do not contain a stop codon and are inserted into the plasmid vector pQStrep2. This vector allows for expression of fusion proteins with N-terminal histidine tag and C-terminal StrepII tag. PCR products of primers 1 and 3 are inserted into vectors pQTEV and pGEX-6p1 for expression of cleavable fusion proteins with N-terminal GST or histidine tag, respectively.

The different protein biosynthesis systems of the expression host and the nonphysiological expression levels can lead to a lack of expression of the heterologous protein, the production of partially misfolded protein, or the production of incomplete polypeptide chains.

The most widely used and best established expression systems use the bacterium *Escherichia coli*. It is especially well suited to produce the milligram amounts of highly pure protein needed for structure determination. *E. coli* cells are the system of choice for bacterial proteins and also work with proteins from eukaryotes. However, many human proteins, which are the targets of the Protein Structure Factory, fail to be expressed in a suitable form. Alternatives to *E. coli* are baculovirus-infected insect cells[11] or recombinant yeasts such as *Saccharomyces cerevisiae* or *Pichia pastoris*, which are used in the Protein Structure Factory to complement the *E. coli* system (see below).

Since only a fraction of human proteins can be expressed in *E. coli* and because the structure of only a small fraction of the human proteins has been determined, we used a screening approach to identify human proteins which can be produced through an *E. coli* expression system. A human expression cDNA library enriched for expression clones[12] (hex clones) was screened by small-scale expression and purification for clones producing human proteins in soluble form. These clones were subsequently assigned to human proteins by sequencing their insert. Clones expressing proteins of unknown structure are now being used to produce these proteins for experimental structure determination.

In the Protein Structure Factory, an expression vector is used that equips the target proteins with short affinity peptide tags on C- and N-terminus (pQStrep2 and pQStrep4, Figure 2). These affinity tags allow for purifica-

Table 1. Numbers of Target Proteins in Different Experimental Steps from Template cDNA Clone to Expression of Soluble Protein in *E. coli*[a]

experimental step	expression vector	target proteins (IMAGE clones)	target proteins (hex1 clones)	total target proteins
IMAGE/hex1 clones corresponding to target proteins	n.a.	566	33	599
PCR amplification of open reading frames from IMAGE clones	n.a.	397	29	426
E. coli expression constructs	any	332	27	359
E. coli expression constructs yielding soluble protein	any	139	24	163
	pQStrep	81	17	98
	pGEX-6p1	68	11	81
	pQTEV	30	10	40
	other vector	29	11	40
soluble protein with high yield	any	92	21	113

[a] Hex1 clones were preselected for expression as described in the text.

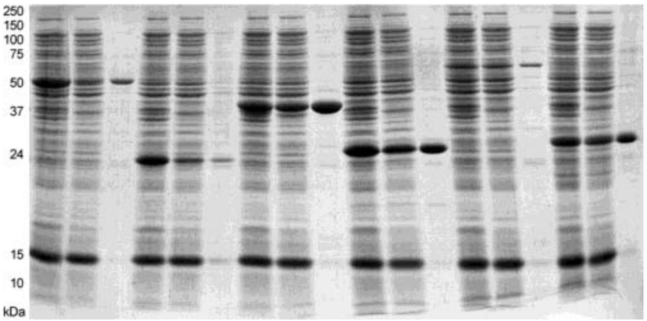

FIGURE 3. Characterization of *E. coli* expression clones by small-scale protein expression and purification, 15% SDS-polyacrylamide gel electrophoresis, Coomassie-stained. Six clones expressing soluble protein are shown. Each clone is characterized by three lanes on the gel (from left to right): whole cellular protein, soluble cellular protein, and protein purified by affinity chromatography.

Table 2. Number of Different Proteins Obtained in Soluble Form from Different Host Systems

E. coli (only/total)	115/163
S. cerevisiae (only/total)	39/56
P. pastoris (only/total)	39/70
E. coli/*S. cerevisiae*	17
E. coli/*P. pastoris*	31
large-scale purification	43

tion of proteins via two consecutive steps of affinity purification. Six consecutive histidines fused to a target protein (histidine tag) allow for purification by metal-chelate affinity chromatography. The StrepII tag consists of eight amino acids which bind to streptavidin and permit affinity chromatography with immobilized variants of this protein.[13]

Protein preparations resulting from these two affinity chromatography purifications are homogeneous and devoid of incomplete polypeptide chains (Figure 3). Affinity chromatography does not require optimization such as ion-exchange chromatography, and the purification of different proteins can be performed under essentially identical conditions.

While affinity tags can streamline the protein purification, there are concerns that they might lead to problems during structure determination. Affinity tags will add complexity to NMR spectra of proteins. For crystallization, their effects depend on the target protein. While affinity tags might be beneficial to the crystallization of some proteins, they are generally believed to be either neutral or detrimental to the success of crystallization experiments.[14] We could recently show that a small protein domain could be subjected to both NMR and crystal structure analysis in the presence of fused histidine and StrepII tags comprising more than 25% of the total protein mass (Mueller et al., in preparation). Thus, peptide tags are in principle compatible with protein structure analysis even in extreme cases. However, protein crystallization remains an empirical procedure. If a protein carrying an affinity tag fails to yield crystals, the experiment should be repeated with a protein variant devoid of tags.

Alternatives to the crystallization of proteins carrying tags is to purify native proteins, or to use specific proteases to remove tags. Both options will increase the effort of protein purification. In the Protein Structure Factory, the vectors pGEX-6p1 and pQTEV are currently used to produce tag-free proteins (see Figure 2). With pGEX-6p1 and pQTEV, fusion proteins of glutathione S-transferase (GST)[15] or a histidine tag are produced. The vectors encode recognition sequences for specific proteases, the protease 3C from a picornavirus ("PreScission")[16] and a protease from tobacco etch virus (TEV),[17] that allow for removal of affinity tags from the synthesized proteins. The use of additional expression strategies significantly increases the yield of soluble protein from *E. coli* compared to the exclusive use of the pQStrep vector (Table 1). However, of the 359 expression constructs generated, less than half produce soluble protein suitable for structure analysis, and less than one-third give a high yield of soluble protein that will allow easy purification. We are uncertain at present why the remaining proteins are insoluble in aqueous buffers although the bioinformatics analysis of their sequences predicted soluble and globular products. One obvious problem may be that some of these chains are subunits of hetero-oligomeric proteins or depend in other ways on the presence of tightly bound partner molecules for their solubility.

As an alternative host to *E. coli*, we have chosen yeast over the baculovirus/Sf9 system[11] because of easy handling, inexpensive media, short incubation times, and high protein yields, in particular from high-density *Pichia* strains. Gene expression in eukaryotic cells indeed allows to purify a subset of proteins that could not be produced in *E. coli*. Table 2 summarizes the success achieved so far in the production of soluble human proteins using the different host systems. Out of 241 different proteins so far obtained in soluble form, roughly one-third are from *S. cerevisiae* or *P. pastoris*. Milligram amounts of 43 different, pure proteins have been passed on to protein crystallization.

Laboratory Information Management System. Handling of large numbers of expression constructs for a list of target proteins requires an electronic Laboratory Information Management System (LIMS). Storage of laboratory data in electronic form is most often done by spreadsheet applications or desktop database programs. These programs are convenient for users lacking programming knowledge but have limitations when it comes to sharing data between different laboratories. Web-based information retrieval is most suitable for this purpose. The data of interest can be retrieved from any computer with a web browser, without the need of installing a specific software or exchanging files. Web-based information retrieval requires a central database and customized software for data entry and display. The implementation of such a system will therefore usually require a dedicated software developer.

Here, a LIMS has been designed to hold a characteristic set of data for each expression clone and protein preparation that is in the production pipeline. This enables, for example, a crystallographer to retrieve the sequence of a clone that was used to make the protein he is currently studying. The local LIMS, organizing data concerning the expression cloning, is linked to the general database of the Protein Structure Factory.

High-Throughput Protein Crystallization

The Protein Structure Factory has set up a robotic station for protein crystallization that handles screen mixing, protein drop and precipitant setting, plate storage, and database evaluation.[7] To speed up protein crystallization, conditions are screened in 96-well microtiter plates developed in collaboration with Greiner BioOne (Frickenhausen, Germany). Both sitting-drop and hanging-drop vapor-diffusion experiments are performed in a semi-automated manner.

The crystallization of a new protein usually starts with a multifactorial buffer and precipitant screening, where the protein solution is brought to supersaturation in a controlled way and crystals form under favorable conditions. The mixing the screening solutions, one of the most time-consuming steps in a crystallization experiment, has been delegated to a modified pipetting robot (Zinsser Analytic, Frankfurt/M., Germany). The protocol for the crystallization screen is obtained from a remote database after recognizing the microtiter plate's unique barcode. The following step includes the dispensing of 96 different screening solutions in parallel onto the screening plate using a 96-syringe robot and adding the appropriate volume of protein solution to the precipitant by a non-contact droplet dispenser based on micro-solenoid-valve technology. The preparation of a complete microtiter plate takes less than 5 min. Thus, close to 100 plates can be prepared over an 8 h working day and stored in the robotic station. Solutions of 1 μL can be handled in the broadest range of viscosities, the dispensation of smaller volumes down to 400 nL protein plus 400 nL of precipitant solution is under development and testing. The plates are stored at 20 °C in an automated storage system with a capacity of over 10 000 plates and an attached microscope equipped with CCD camera and computer interface. During the following 100 days, the screening plate is being checked according to a pre-defined protocol for crystals and other features appearing in the droplet. When small crystals are observed, the crystallization conditions are improved by subsequent fine screens. Optimized crystals are harvested, shock-frozen and sent on liquid N_2 to the synchrotron beamline for diffraction experiments. At present, the throughput and storage capacity of the robotic station exceeds our capabilities of producing homogeneous protein samples for crystallization considerably. The facility is therefore offered to interested laboratories for external crystallization trials.

All steps are recorded by database clients developed by the Protein Structure Factory which are linked to an Oracle database. This archive stores all relevant information about the protein from earlier studies such as isoelectric focusing or dynamic light scattering experiments and all information related to the crystallization experiment itself. This includes all positive and negative experimental results so that the operator is able to trace back any experiment even to vendors and distributors of the chemicals used. We believe the in-house developed LIMS system to give a higher flexibility than commercial alternatives.

High-Throughput X-Ray Diffraction Techniques and Facilities

High-throughput crystal structure analysis of proteins is inconceivable without unrestricted access to synchrotron radiation. In contrast to laboratory sources, the synchrotron X-ray beam may be tuned to energies at or around the absorption edge of specific atoms in the crystal. It thus permits to make optimal use of anomalous diffraction for phasing through the MAD or SAD approaches[5] and eliminates the need to collect datasets from different, isomorphous crystals (native and derivative). The single, necessary heavy-atom marker may be introduced into a protein systematically, e.g. as selenomethionine, in a process that can be standardized and automated. For these reasons, the use of synchrotron radiation must be an integral part of every high-throughput structural genomics effort.

To prevent crystal decay caused by the brilliant synchrotron X-ray beam at room temperature, diffraction experiments must be carried out under cryogenic cooling.[18] Using data measured at a suitable synchrotron beamline on a cryo-cooled crystal and applying modern crystallographic software[19] may reduce the time from the diffraction experiment to the fully refined structure to a few days under favorable circumstances.

Following the above considerations, the Protein Structure Factory is currently engaged in setting up protein crystallography facilities at the 1.7 GeV third-generation electron synchrotron source BESSY (www.bessy.de) in Berlin. This facility comprises three independently operat-

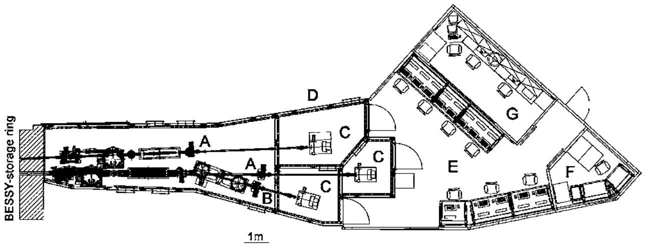

FIGURE 4. Protein crystallography facilities of the Protein Structure Factory in the experimental hall of the synchrotron storage ring BESSY in Berlin. Individual parts of the setup are explained in the text.

ing beamlines and experimental end stations. It will be open to external users.

The synchrotron radiation is produced by a superconducting 7 T wavelength shifter[20] with a critical energy of $E_{\text{crit}} = 13.5$ keV manufactured at the Budker Institut (Novosibirsk, Russia), which is inserted into a straight section of the BESSY ring. Beamline components outside the BESSY beam tunnel are shown in Figure 4. Optical beamline components (ACCEL Instruments, Bergisch-Gladbach, Germany) include two double-crystal monochromators permitting energy tuning (A). A third beamline (B) will operate at a constant wavelength near E_{crit}. First experimental results from the commissioning of the beamlines indicate a high brilliance at the central, tunable-energy station where 7.3×10^{10} photons per s impinge on the sample through 200×200 μm slits at 100 mA beam current, an energy of 13 keV (wavelength of 0.95 Å), and an energy resolution ($\Delta\lambda/\lambda$) of 1.5×10^{-4}. This is comparable with the performance of a bending-magnet beamline at the European Synchrotron Radiation Facility (ESRF, Grenoble, France). The beamline is tunable from 4.5 to 17.5 keV (2.75−0.71 Å). The other two beamlines, which are currently being commissioned, are desgined to meet similar specifications. Experimental end stations are equipped with a variety of position-sensitive detectors (C). Optical beamline elements and end stations are shown inside their lead enclosures (D) that protect the user from radiation and allow the independent use of the experimental end stations. Access to the experimental stations is through the operator room (E), a noise-reduced and climatized area. Within this room, users will control their experiments and the beamlines from PC terminals. User-friendly interfaces are under development to allow crystallographers with various degrees of experience the use of the station and protect the systems against operator errors.

During an experiment which may run for approximately 2−5 h, it is necessary to have access to the collected data, using fast experimental data processing and highly available raw data storages. A high-performance computing system using an 8-CPU high-performance computer with an attached SAN storage array (fiber-channel-based storage array network) of 3 TB net capacity (F) will serve this purpose. These CPU and storage resources are designed to accept up to 100 GB of raw data per day per experiment. The data will be automatically saved on a daily basis and archived after evaluation using a managed enterprise backup solution which is a member of the SAN.

Since simple notebook tracking of the experiments is not possible with a planned throughput of up to four diffraction experiments per beamline per day, a mySQL (www.mysql.com) based database, together with a specific web-form frontend is under development that will store all relevant experimental data (Manjasetty et al., in preparation). This local database is also interfaced with the general database of the Protein Structure Factory to exchange relevant information in both directions on demand via a secure shell tunnel. A sample preparation laboratory (G) completes the outstation of the Protein Structure Factory at BESSY. There, crystals are prepared for data collection and stored for further use.

Conclusions and Outlook

The Protein Structure Factory has completed the major part of its first project phase dedicated to establishing a technical infrastructure for the high-throughput structure analysis of human proteins. In this phase, a central laboratory for expression cloning, protein purification and characterization, and crystallization has been set up in parallel to two outstations for NMR spectroscopy and X-ray crystallography, respectively. The facilities now available have to be further upgraded, especially regarding the automation of NMR structure analysis and synchrotron-based X-ray diffraction, as well as the development of pattern-recognition software for crystal detection in the robotic station. At the same time, the high-throughput protein structure analysis is beginning to take shape. A substantial number of human proteins have been produced in soluble form or purified for NMR or crystal structure analysis. Because of the emphasis on infrastructure development early in the project, less than 10 proteins have been crystallized so far, and only a few structures were solved. In part, the small number of completed structure analyses is also due to the choice of human proteins that are harder to deal with than the bacterial proteins targeted by many other structural genomics consortia. However, the structure analysis pipeline is now filled with expression clones and purified soluble proteins awaiting crystallization. It is planned to complete work on the currently targeted set of small to medium-sized, monomeric or homooligomeric soluble human proteins before addressing problems regarding longer protein chains, hetero-oligomers or membrane-associated proteins.

The Leitprojektverbund "Proteinstrukturfabrik" is funded by the German Federal Ministry of Education and Research (BMBF). Support of our work by the Berlin Senate, the European Fund for Regional Development (EFRE) and the Fonds der Chemischen Industrie is also acknowledged.

References

(1) (a) Burley, S. K. An overview of structural genomics. *Nat. Struct. Biol.* **2000**, *7*, 932−934. (b) Stevens, R. C.; Yokoyama, S.; Wilson, I. A. Global efforts in structural genomics. *Science* **2001**, *294*, 89−92. (c) Terwilliger, T. C. Structural genomics in North America.

Nat. Struct. Biol. **2000**, *7*, 935−939. (d) Heinemann, U. Structural genomics in Europe: Slow start, strong finish? *Nat. Struct. Biol.* **2000**, *7*, 940−942. (e) Yokoyama, S.; Hirota, H.; Kigawa, T.; Yabuki, T.; Shirouzo, M.; Terada, T.; Ito, Y.; Matsuo, Y.; Kuroda, Y.; Nishimura, Y.; Kyogoku, Y.; Miki, K.; Masui, R.; Kuramitsu, S. Structural genomics projects in Japan. *Nat. Struct. Biol.* **2000**, *7*, 943−945.

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

(3) (a) Abola, E.; Kuhn, P.; Earnest, T.; Stevens, R. C. Automation of X-ray crystallography. *Nat. Struct. Biol.* **2000**, *7*, 973−977. (b) Blundell, T. L.; Jhoti, H.; Abell, C. High-throughput crystallography for lead discovery in drug design. *Nat. Rev. Drug Discov.* **2002**, *1*, 45−54. (c) Heinemann, U.; Illing, G.; Oschkinat, H. High-throughput three-dimensional protein structure determination. *Curr. Opin. Biotechnol.* **2001**, *12*, 348−354. (d) Jhoti, H. High-throughput structural proteomics using x-rays. *Trends Biotechnol.* **2001**, *19*, S67−S71.

(4) (a) Heinemann, U.; Frevert, J.; Hofmann, K. P.; Illing, G.; Maurer, C.; Oschkinat, H.; Saenger, W. An integrated approach to structural genomics. *Prog. Biophys. Mol. Biol.* **2000**, *73*, 347−362. (b) Heinemann, U. Establishing a structural genomics platform: The Berlin-based Protein Structure Factory. *Gene Funct. Dis.* **2002**, *3*, 25−32.

(5) (a) Hendrickson, W. A.; Horton, J. R.; LeMaster, D. M. Selenomethionyl proteins produced for analysis by multiwavelength anomalous diffraction (MAD): a vehicle for direct determination of three-dimensional structure. *EMBO J.* **1990**, *9*, 1665−1672. (b) Hendrickson, W. A.; Ogata, C. M. Phase determination from multiwavelength anomalous diffraction measurements. *Methods Enzymol.* **1997**, *276*, 494−523. (c) Rice, L. M.; Earnest, T. N.; Brunger, A. T. Single-wavelength anomalous diffraction phasing revisited. *Acta Crystallogr., Sect. D* **2000**, *56*, 1413−1420.

(6) Shuker, S. B.; Hajduk, P. J.; Meadows, R. P.; Fesik, S. W. Discovering high-affinity ligands for proteins: SAR by NMR. *Science* **1996**, *274*, 1531−1534.

(7) Mueller, U.; Nyarsik, L.; Horn, M.; Rauth, H.; Przewieslik, T.; Saenger, W.; Lehrach, H.; Eickhoff, H. Development of a technology for automation and miniaturisation of protein crystallisation. *J. Biotechnol.* **2001**, *85*, 7−14.

(8) Berman, H. M.; Bhat, T. N.; Bourne, P. E.; Feng, Z.; Gilliland, G.; Weissig, H.; Westbrook, J. The Protein Data Bank and the challenge of structural genomics. *Nat. Struct. Biol.* **2000**, *7*, 957−959.

(9) (a) Strausberg, R. L.; Feingold, E. A.; Klausner, R. D.; Collins, F. S. The mammalian gene collection. *Science* **1999**, *286*, 455−457. (b) Wiemann, S.; Weil, B.; Wellenreuther, R.; Gassenhuber, J.; Glassl, S.; Ansorge, W.; Böcher, M.; Blöcker, H.; Bauersachs, S.; Blum, H.; Lauber, J.; Düsterhöft, A.; Beyer, A.; Köhrer, K.; Strack, N.; Mewes, H.-W.; Ottenwälder, B.; Obermaier, B.; Tampe, J.; Heubner, D.; Wambutt, R.; Korn, B.; Klein, M.; Poustka, A. Towards a catalog of human genes and proteins: Sequencing and analysis of 500 novel complete protein coding human cDNAs. *Genome Res.* **2001**, *11*, 422−435.

(10) Lennon, G.; Auffray, C.; Polymeropoulos, M.; Soares, M. B. The I.M.A.G.E. consortium: An integrated molecular analysis of genomes and their expression. *Genomics* **1996**, *33*, 151−152.

(11) Merrington, C. L.; King, L. A.; Possee, R. D. Baculovirus expression systems. In *Protein Expression*; Higgins, S. J., Hames, B. D., Eds.; Oxford University Press: Oxford, 1999; pp 101−127.

(12) Büssow, K.; Nordhoff, E.; Lubbert, C.; Lehrach, H.; Walter, G. A human cDNA library for high-throughput protein expression screening. *Genomics* **2000**, *65*, 1−8.

(13) (a) Hochuli, E.; Bannwarth, W.; Dobeli, H.; Gentz, R.; Stüber, D. Genetic approach to facilitate purification of recombinant proteins with a novel metal chelate adsorbent. *Biotechnology* **1988**, *6*, 1321−1325. (b) Schmidt, T. G.; Skerra, A. One-step affinity purification of bacterially produced proteins by means of the "Strep-tag" and immobilized recombinant core streptavidin. *J. Chromatogr. A* **1994**, *676*, 337−345.

(14) Bucher, M. H.; Evdokimov, A. G.; Waugh, D. S. Differential effects of short affinity tags on the crystallization of *Pyrococcus furiosus* maltodextrin-binding protein. *Acta Crystallogr., Sect. D* **2002**, *58*, 392−397.

(15) Smith, D. B.; Johnson, K. S. Single-step purification of polypeptides expressed in *Escherichia coli* as fusions with glutathione-S-transferase. *Gene* **1988**, *67*, 31−40.

(16) Walker, P. A.; Leong, L. E.; Ng, P. W.; Tan, S. H.; Waller, S.; Murphy, D.; Porter, A. G. Efficient and rapid affinity purification of proteins using recombinant fusion proteases. *Bio/Technol.* **1994**, *12*, 601−605.

(17) Parks, T. D.; Leuther, K. K.; Howard, E. D.; Johnston, S. A.; Dougherty, W. G. Release of proteins and peptides from fusion proteins using a recombinant plant virus proteinase. *Anal. Biochem.* **1994**, *216*, 413−417.

(18) (a) Hope H. Cryocrystallography of biological macromolecules. *Acta Crystallogr., Sect. B* **1988**, *44*, 22−26. (b) Garman, E. Cool data: quantity and quality. *Acta Crystallogr., Sect. D* **1999**, *55*, 1641−1653.

(19) (a) Terwilliger, T. C.; Berendzen, J. Automated structure solution for MIR and MAD. *Acta Crystallogr., Sect. D* **1999**, *55*, 849−861. (b) Perrakis, A.; Morris, R.; Lamzin, V. S. Automated protein model building combined with iterative structure refinement. *Nat. Struct. Biol.* **1999**, *6*, 458−463. (c) Lamzin, V. S.; Perrakis, A. Current state of automated crystallographic data analysis. *Nat. Struct. Biol.* **2000**, *7*, 978−981.

(20) Borovikov, V. M.; Djurba, V. K.; Fedurin, M. G.; Repkov, V. V.; Karpov, G. V.; Kulipanov, G. N.; Kuzin, M. V.; Mezentsev, N. A.; Shkaruba, V. A.; Kraemer, D.; Richter, D. Superconducting 7T wave length shifter for BESSY-II. *Nucl. Instrum. Methods Phys. Res. A* **2001**, *467−468*, 181−184.

AR010129T