Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

A large-scale, high-efficiency and low-cost platform for structural genomics studies

Xiao-Dong Su,[a]* Yuhe Liang,[a] Lanfen Li,[a] Jie Nan,[a] Erik Bröstromer,[a] Peng Liu,[b] Yuhui Dong[b] and Dingchang Xian[b]

[a]National Laboratory of Protein Engineering and Plant Genetic Engineering, Peking University, Beijing 100871, People's Republic of China, and [b]Beijing Synchrotron Radiation Facility, Institute of High Energy Physics, Chinese Academy of Sciences, Beijing 100049, People's Republic of China

Correspondence e-mail: su-xd@pku.edu.cn

Received 23 February 2006
Accepted 25 June 2006

A large-scale, high-efficiency and low-cost platform based on a Beckman Coulter Biomek FX and custom-made automation systems for structural genomics has been set up at Peking University, Beijing, People's Republic of China. This platform has the capacity to process up to 2000 genes per year for structural and functional analyses. *Bacillus subtilis*, a model organism for Gram-positive bacteria, and *Streptococcus mutans*, a major pathogen of dental caries, were selected as the main targets. To date, more than 470 *B. subtilis* and 1200 *S. mutans* proteins and hundreds of proteins from other sources, including human liver proteins, have been selected as targets for this platform. The selected genes are mainly related to important metabolism pathways and/or have potential relevance for drug design. To date, 40 independent structures have been determined; of these 11 are in the category of novel structures by the criterion of having less than 30% sequence identity to known structures. More than 13 structures were determined by SAD/MAD phasing. The macromolecular crystallography beamline at the Beijing Synchrotron Radiation Facility and modern phasing programs have been crucial components of the operation of the platform. The idea and practice of the genomic approach have been successfully adopted in a moderately funded structural biology program and it is believed this adaptation will greatly improve the production of protein structures. The goal is to be able to solve a protein structure of moderate difficulty at a cost about US $10 000.

1. Introduction

With the completion of hundreds of genome-sequencing projects, the next step (and, in our view, also the last step) in the analysis of this detailed fundamental information for understanding biological functions is to determine the three-dimensional structures of all the biological molecules encoded by the genomes. Therefore, so-called structural genomics (SG) initiatives have been initiated worldwide since the turn of the new millennium (Brenner, 2001; Burley, 2000; Stevens *et al.*, 2001) and have made a large impact on structural biology, as reviewed recently by Chandonia & Brenner (2006).

In China, SG projects were initiated in early 2001, as proposed by Professors Dongcai Liang, Yunyu Shi, Xiaocheng Gu and Zihe Rao *etc*. In contrast to many other countries, Chinese SG projects have been supported by diverse sources. The pilot funding came from the National Natural Science Foundation of China (NSFC) and, to a smaller extent, from

the Ministry of Education (MOE). Larger funding came later from the Ministry of Science and Technology of China (MOST) and the Chinese Academy of Sciences (CAS) (Gong *et al.*, 2003).

The SG projects pursued at Peking University (PKU) were started in late 2001; the initial approaches included targeting genes related to human diseases or of functional importance (Ding *et al.*, 2002). In the earlier phase (2001–2004), the Gateway cloning system (Invitrogen, The Netherlands) was chosen for high-throughput (HTP) gene cloning and protein expression. Although the Gateway system has many advantages, such as enabling rapid and efficient transfer of DNA fragments produced by polymerase chain reaction (PCR) into multiple vectors for protein expression and suitable adaptability to automation, it is too expensive for large-scale operations in an average-funded university laboratory.

Since 2003, we have been seeking alternative approaches to set up large-scale inexpensive ways to pursue SG projects in a university laboratory environment, as shown in Fig. 1. We have been attempting to use conventional restriction-enzyme digestion-based cloning methods by carefully selecting the cloning sites and at the same time considering how to clone the same DNA fragment into multiple vectors. Eventually, we found that it is quite possible to obtain a similar flexibility and adaptability to that of the Gateway system while keeping the costs at a level of one-quarter to one-fifth. In order to increase the throughput and output of the SG platforms, automation is a must, despite limited funding. We have therefore purchased an automatic liquid handler from Beckman Coulter (Fig. 2) and converted it into a multiple-purpose robot to perform large-scale operations such as PCR, DNA purification, ligation, protein-expression checking and even crystallization screening (as shown in Fig. 1). It has been very cost-effective for a laboratory-based SG platform to use one automatic system for many tasks instead of having many robots as are found in larger SG consortia. Furthermore, we have designed and built our own automatic imaging system for recording the crystallization screens at a cost of one-fifth to one-tenth of the commercial price of such systems.

We have been trying to use a wide variety of methods and programs to solve our crystal structures and an in-house X-ray system has been used for all preliminary crystal screening and diffraction data collection. The Beijing Synchrotron Radiation Facility (BSRF) shown in Fig. 5 is our major source of synchrotron radiation (SR) for diffraction data collection, although other international SR light sources such as DESY in Hamburg, Germany and MAX-Lab in Lund, Sweden have also been used during the last few years. A flowchart of our SG platform is shown in Fig. 1 and a more detailed description of the steps developed for this platform is presented below.

2. Materials and methods

2.1. Bioinformatics for target selection and LIMS (laboratory information-management system)

The first step in our platform or the pipeline of structure genomics is target selection, as shown in Fig. 1. In our laboratory, in addition to working on several human proteins, we have been focusing on two model bacteria as SG targets, the non-pathogen *Bacillus subilis* and the dental pathogen *Streptococcus mutans* that causes dental caries. As in most of the SG laboratories in the world, our targets are mostly selected from gene families of unknown structure or/and function, which are more likely to represent new three-dimensional structures.

The large numbers of data and images produced by our SG platform are far beyond the traditional methods of manual

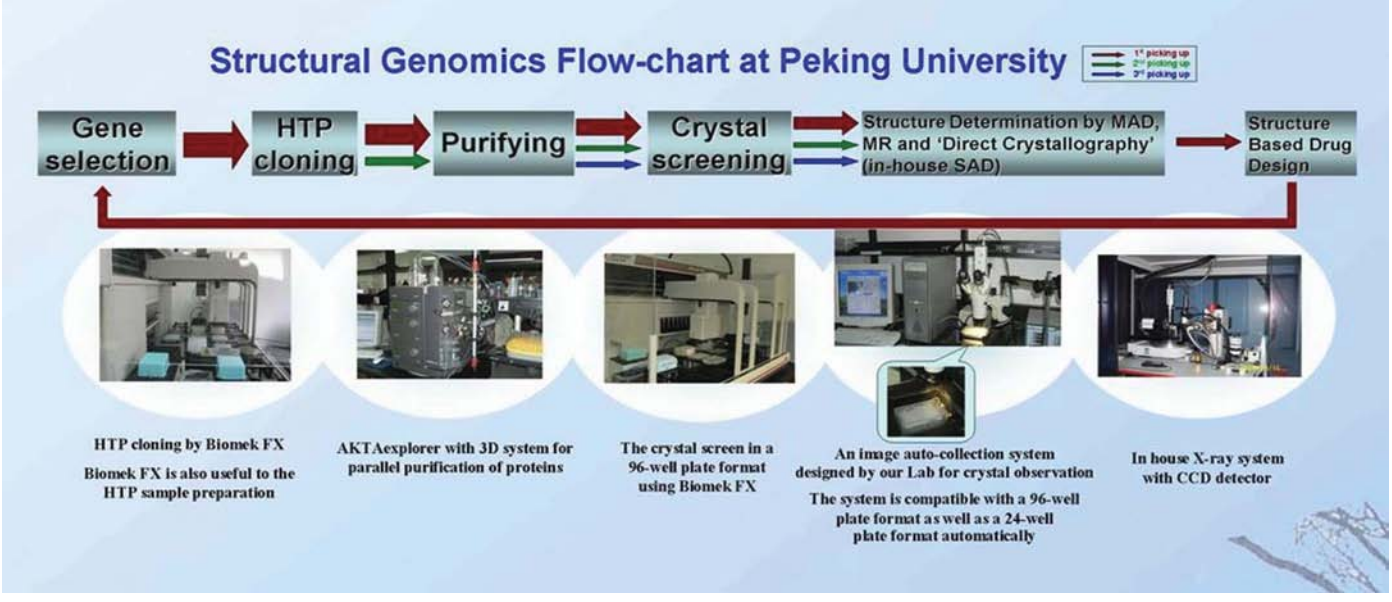

Figure 1
Flowchart of the SG platform at PKU. The five steps are described in more detail in the text, but the final step 'structure-based drug design' will not be elaborated since it is only one of the many applications of protein structures.

recording and management. A laboratory information-management system (LIMS) could help to solve this problem. The commercially available LIMS software is not only expensive, but also is not suitable for our specific needs. Therefore, we designed and developed our own LIMS.

Figure 2
A multi-purpose liquid handler based on the Beckman Coulter Biomek FX automation system. It has been used for PCR, different kinds of DNA purification, ligation, expression checking, small-scale protein expression and some initial crystallization screening.

Our LIMS adopts Browser/Server (B/S) mode, so that the users (normally students) can reach the server through an internet connection. Another advantage of B/S mode is that the operating system is more independent and much safer in comparison with the Client/Server mode. The server was built on PHP5 (The PHP Group; http://www.php.net/) and Apache HTTP Server v.2.0 (The Apache Software Foundation; http://httpd.apache.org/) running on a Linux Fedora Core 1 system. MySQL 5.0 (MySQL AB, Uppsala, Sweden; http://www.mysql.com/) was also installed on the same server, which integrates the database server. All the software and the computational environment are available completely free of charge and thus could easily be adapted for other academic users.

2.2. Highly efficient ways of gene cloning and expression using conventional technology

After target selection, an HTP and automated method for the parallel cloning of hundreds or thousands of genes is essential for any SG platform. We initially took advantage of the Gateway system, which is flexible and adaptable to many different vectors and can be easily implemented in a robotic environment. However, the cost of the Gateway system and the PCR primers is too high since the BP reaction of the pENTR cloning for the Gateway system requires the primer to contain about 30 extra nucleotides at both the 5′ and 3′ ends

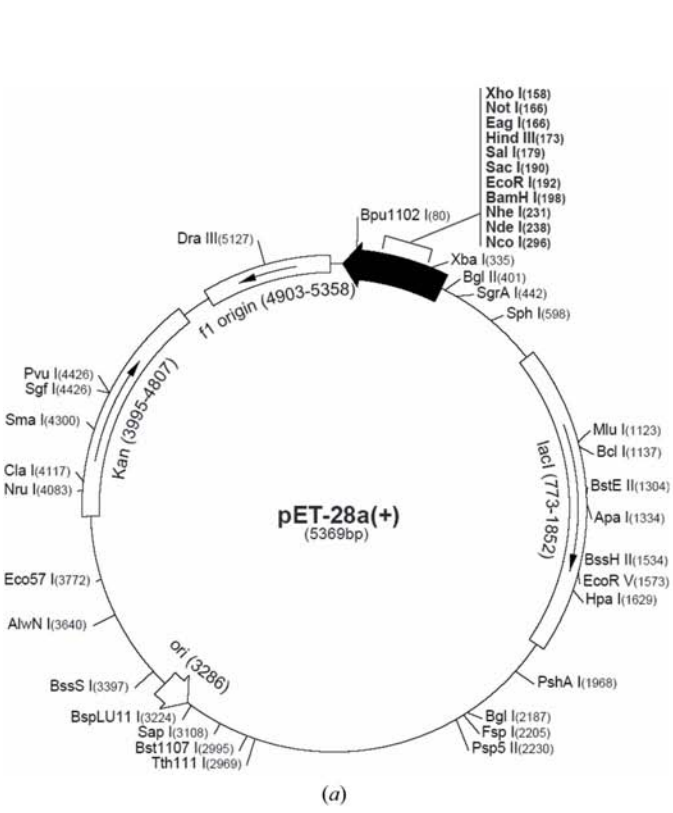

Figure 3
(*a*) The most widely used expression vector pET28a in our laboratory (other pET vectors have also been used); *Bam*HI and *Xho*I restriction enzymes have been chosen for the majority of cloning work for the reasons described in the text. (*b*) By choosing *Bam*HI and *Xho*I restriction sites, pFastBac vectors can easily be coupled to pET vectors.

for each gene. The problem of extra nucleotides could be partly overcome by using the Gateway/TOPO Cloning Technology (TOPO and Gateway Cloning Technology are both from Invitrogen, The Netherlands). In the case of the TOPO cloning system, only the forward primer needs to contain four extra nucleotides CACC in front of the start codon ATG and no extra nucleotide is needed for the reverse primer in order to clone the target gene directionally into the pENTR vector. However, the total cost of the Gateway/TOPO cloning is not significantly lower owing to the price of the cloning kits. It is estimated that in our SG platfrom the overall material cost is about US $75 for each gene to be cloned into an expression vector by either Gateway or Gateway/TOPO cloning systems, which is too expensive for a university laboratory-based SG platform.

Eventually, we developed a robot-based HTP cloning method based on conventional cloning with pET (Novagen, Madison, USA), pFastBac (Invitrogen, the Netherlands) and modified pGEX (GE Healthcare Life Sciences, USA) systems on a Beckman Coulter Biomek FX robotic system (Beckman Coulter, Inc., Fullerton, CA, USA) as shown in Fig. 2.

The two vectors most commonly used in our SG platform are shown in Figs. 3(*a*) and 3(*b*); other vectors with suitable restriction sites have also been used. Apart from the economic factors, the chosen restriction sites are readily adaptable to different expression systems. We are also trying to modify existing vectors to increase the adaptability or to develop new vectors for protein expression. In particular, yeast systems have been under development for HTP protein-expression needs.

2.3. Automated parallel methods for protein purification using ÄKTA systems

All proteins produced in our platform are His_6-tagged at either the N-terminus or the C-terminus, and nickel-column-based affinity purification on ÄKTA chromatography systems (GE Healthcare Life Sciences, USA) is the most commonly used method for protein purification. After an initial solubility check at various temperatures (310, 298 or 291 K), proteins with good expression and solubility were identified and cultured in large flasks with 1 l LB medium and induced with 0.5 m*M* isopropyl-β-D-thiogalactoside (IPTG). The bacteria were then harvested by centrifugation and lysed by sonication. Purification of the His_6-tagged proteins using a nickel column was often first performed on an ÄKTA Purifier or FPLC (GE Healthcare Life Sciences, USA); parallel purification methods were then applied to those proteins that behaved well.

Of the several different types of ÄKTA purification system available in our laboratory, ÄKTA Explorer 100 with 3D kit as shown in Fig. 4 is the most powerful, since up to six His_6-tagged proteins can be automatically purified in a single run using protocols containing two chromatography steps. The 3D Kit has been developed to facilitate automatic purification of tagged recombinant proteins from clarified cell cultures.

Figure 4
ÄKTA Explorer 100 with 3D kit for parallel purification of His_6-tagged proteins.

2.4. High-throughput methods for protein crystal screening

In our SG platform, we have also been trying to adapt and develop HTP methods for crystallization screening and optimization that do not rely on specialized crystallization robots. For the initial crystallization screenings, ANSI (American National Standards Institute) SBS (Society for Biomolecular Sciences) standard plates with 96 wells are used to screen sparse-matrix conditions under oil (Ding *et al.*, 2002). The promising conditions containing microcrystals are normally optimized by the hanging-drop vapour-diffusion method using Hampton Research 24-well VDX plates (Hampton Research, USA). We have been trying to implement methods and protocols on the Beckman Coulter Biomek FX for automatic crystallization screens since the Biomek FX does not contain standard procedures for such operations.

We are currently moving from the classic hanging-drop methods of crystallization screening to an HTP approach that is suitable for both manual setup and automation. As a first step, we have started to use SBS standard microplates for crystallization under oil and for sitting-drop experiments. The crystallizations under oil can be set up by the Biomek FX system using 96-well microplates. However, the sitting-drop plates are currently set up manually using eight-channel pipettes and transparent tape as plate sealant; the time and work used to set up a screen plate has been greatly reduced. Laboratory members collaborate during the crystallization experiments, thus maximizing the number of proteins screened at a given time. We have been trying to design and manufacture our own sitting-drop crystallization plates with SBS standard plates, with the aim of such plates being automatically handled by the Biomek FX robot.

To monitor the crystallization experiments, an automated imaging system has been developed in-house by our graduate students. The system accepts SBS standard microplates and the recorded images are transferred to an image and plate database system that is accessible through a web interface from any internet-connected computers around our SG platform. While stand-alone, the database can be reached through the LIMS interface using existing user identifiers. This system

not only speeds up the crystallization validation process, but also helps to preserve and analyze information and generate crystallization statistics.

So far, the screening kits used in our SG platform have been the standard commercially available kits such as Hampton kits (Hampton Research, USA). We have been investigating our own statistics for successful crystallization conditions and hope to develop our own crystallization screening kits in the near future.

2.5. X-ray sources and methods for rapid crystal structure determination

We used an in-house Bruker–Nonius FR591 X-ray source and a SMART 6000 CCD detector (Bruker Nonius BV, Delft, The Netherlands) for all crystal tests and screening prior to data collection at synchrotron-radiation (SR) sources. The in-house machine has also been successfully used for structure determination by SAD and SIRAS methods.

The Beijing Synchrotron Radiation Facility (BSRF) shown schematically in Fig. 5 is our primary SR source for diffraction data collection. The detailed design and construction of the MAD beamline at BSRF was described at the Get-Phases Workshop by Yuhui Dong. Other international SR light sources such as DESY, Hamburg, Germany and MAX-Lab, Lund, Sweden have also been used for our SG platform and produced several structures successfully.

2.6. Structure determination and analysis platform

For high-throughput and cost-effective determination of crystal structures, we have been using high-performance medium-priced personal computer (PC) systems with the Linux operating system. About 20 such PCs are connected in three rooms of the wet laboratory to form a local network. These PCs are physically connected in a star topology, but the logical connections are of bus topology. All the PCs have software installed locally and work independently, which means that a crash of any computer would not affect the others. Every student working and sitting in the wet laboratory has access to one of these PCs for daily operation, including interaction with the LIMS and routine use of structure-determination and analysis software.

For extensive structure refinement and analysis, a better equipped network system in physical and logical star topology was built in the so-called dry laboratory, near the X-ray instrument. For this system, in addition to NFS (network file sharing), an NIS (network information system) was used to enhance sharing between different systems and to guarantee that users can access all resources at any one computer after having registered at the server. The important feature is that all software and packages only need to be installed and updated once on the server and the changes will then affect all clients in the network. As clients, high-end PCs with enhanced features such as high-performance graphics cards and stereoglasses were connected to the server and used for both structure determination and refinement.

The software installed on both wet laboratory and dry laboratory computing environments include *HKL*-2000 (Otwinowski & Minor, 1997) for raw data processing, *CCP*4 and related packages (Collaborative Computational Project, Number 4, 1994; Winn *et al.*, 2002), *SOLVE*/*RESOLVE* (Terwilliger, 2003), *PHENIX* (Adams *et al.*, 2002), *AutoSHARP*/*SHARP* (Bricogne *et al.*, 2003) for phasing, and *CNS* (Brünger *et al.*, 1998), *O* (Jones *et al.*, 1991) and *Coot* (Emsley & Cowtan, 2004) for model building and refinement. Software such as *OASIS*2004 (Wang *et al.*, 2004), *Phaser* (Storoni *et al.*, 2004) and *ARP*/*wARP* (Perrakis *et al.*, 1999) were also installed to enhance the phasing and model building.

3. Results and discussions

3.1. Target selection and LIMS

In collaboration with the School of Stomatology, Peking University, we have initiated an SG project to study the functions and structures of proteins from the dental pathogen *S. mutans*, funded by the National Natural Science Foundation of China (NSFC). It is expected that the expression of all the proteins encoded by the *S. mutans* genome and selection of those that are soluble for further detailed functional and structural studies may help in understanding the pathogenic mechanism of *S. mutans* and aid in future drug design and vaccine research.

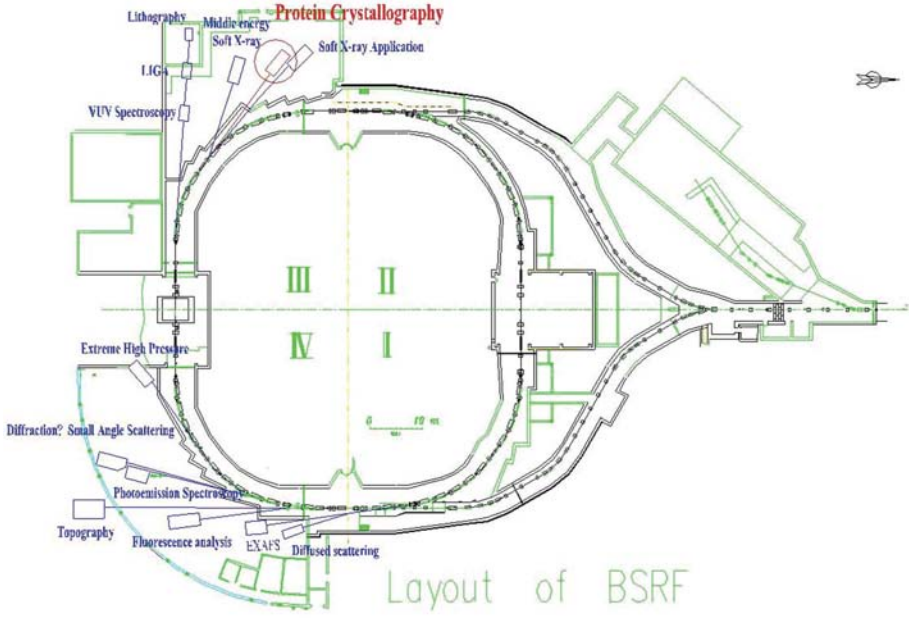

Figure 5
A schematic layout of BSRF, with the macromolecular MAD beamline indicated in red.

Table 1
Summary of the production of the whole platform: first-round, all stages.

Progress in each stage of gene targeting, cloning, protein production, solubility, crystallization and structure determination from early 2003 to December 2005. The two parts of the table show statistics from the projects started before and after September 2004, when the platform switched from TOPO/Gateway to conventional methods and used the Biomek FX robot for operations.

	Targets	PCR	Cloned	Expressed	Soluble	Crystals	Diffracted crystals	Complete data sets	Solved
TOPO/Gateway (from early 2003 to October 2004)									
B. subtilis	173	133	98	72	51	28	26	21	20
Others†	128	104	74	56	36	14	8	8	8
Total	301	237	172	128	87	42	34	29	28‡
Conventional methods (from October 2004 to December 2005)									
Pilot§	54	52	46	41	32	26	23	14	6
96-well (Dec. 2004)	288	275	162	127	75	30	25	16	4
Plate (Jun. 2005)	274	192	137	72	55	19	13	7	2
Batches (Dec. 2005)¶	480	467							
Total	1096	986	345	240	162	75	61	37	12

† Genes selected here include some human secreted proteins (Dai *et al.*, 2005), some SARS CoV genes and other proteins (or gene products) from various collaborative projects. § Pilot targets include both *B. subtilis* and *S. mutans* genes, which were randomly selected to test the HTP concepts in a manual way. ‡ Most of the structures listed here were solved and refined during 2004–2005 and are listed here owing to the origin of the genes, not the actual time the structure were solved. ¶ This batch is still in the process of cloning into expression vectors.

There are about 1960 genes in the genome of *S. mutans* (Ajdic' *et al.*, 2002) and 1532 could be related to 1124 families in Pfam (8183 families to November 2005; Pfam v.19.0), including 15 uncharacterized protein families and 102 domains of unknown function. About 800 gene products have novel structures, as judged by the criterion of having sequence identity below 30%. Thus, the *S. mutans* genome is a good source of SG targets.

We checked the restriction-enzyme site distribution in the *S. mutans* genome with the commonly used enzymes and found that with *Bam*HI or *Eco*RI at the 5′ end and *Xho*I at the 3′ end, 1874 genes (corresponding to 19 96-well plates) could be covered. For non-membrane proteins, the pair *Bam*HI and *Xho*I could cover 1235 genes, corresponding to 12 or almost 13 96-well plates. *Bam*HI, *Eco*RI and *Xho*I are among the cheapest restriction enzymes commercially available. We have therefore chosen the *Bam*HI and *Xho*I pair as the primary digestion sites and the *Eco*I and *Xho*I pair as the secondary digestion sites.

Our LIMS was designed in 2004 and built and tested in 2005. It is now in the process of user input and application. The whole system is composed of four components: Experimental Data-Management Subsystem, Commonly Used Tools, General Laboratory Management Subsystem and Safety Management Subsystem.

The Experimental Data Management system is mainly for protein production and diffraction data collection, processing, storage and analysis. With the large number of data stored and managed by the LIMS, many types of statistical analyses and comparisons become possible. Interfaces of databases are also provided for the addition of new machines and data management.

We have also developed some user-friendly software, such as *PrimerDesign*, which can design primers not only for a single gene but also for large number of genes at once, and *ImgAlign*, which allows users to optimize the results produced by structure-based multiple sequence alignment and generate illustrations with sufficiently good resolution for publication.

The General Laboratory Management Subsystem was developed for the daily management of laboratory affairs, such as seminars, journal clubs, purchase orders, protocols, publications, contact information and so on. These three subsystems are comparatively independent of each other, while the last part, the Safety Management Subsystem, is incorporated into the whole system. There are mainly three kinds of users, administrator, laboratory users and normal users, with different rights for browsing and management of the data.

For a typical laboratory user, although they can read and access most of the experimental data produced and maintained by the LIMS, they can only input or modify the information for their own projects.

3.2. Output of HTP cloning from the PKU SG platform

The upper part of Table 1 summarizes about 22 months of work using the TOPO/Gateway HTP cloning systems and the lower part presents about 15 months of work (including about three months of test runs) with the conventional HTP cloning methods on the Beckman Coulter Biomek FX system; however, the solved and refined structures were mostly finished during 2004–2005. The efficiency gain can be estimated as roughly fivefold. The results shown in the table do not match the cloning capacity of the Biomek FX, which could be estimated as at least 1000 genes per month in our experience. This capacity is much higher than we need for cloning work, thus we are exploring new applications such as protein-expression checking and crystallization screening using this robot. We could of course extend its capability to applications such as HTP drug screening when suitable drug targets and assays have been identified.

By careful selection of the restriction sites according to the costs of restriction enzymes and the distribution of the restriction sites among the open reading frames (ORFs) in the genomes, we could make conventional restriction-enzyme-based methods work perfectly well on the Biomek FX robot,

as shown in Fig. 2, with a total material costs in the range of US $15–20 per gene, which is about a quarter to one-fifth of the material costs of the TOPO/Gateway cloning systems.

Since the downstream process of the conventional cloning batches is still under way (lower part of Table 1), the overall output of the PKU SG platform is compared with the TOPO/Gateway system (upper part of Table 1). The overall success rate of our SG platform, from cloning to structures, is similar to that achieved by the international SG consortia (Service, 2005), with an overall number of solved structures equivalent to more than 10% of the targeted genes for the bacterium *B. subtilis* and much lower (about 6%) for human proteins.

In the first half of 2006, about 800 more genes have been selected to be cloned into pET28 and have increased the total number of genes to about 2300 for our SG platform.

Table 2
A list of structures solved within the PKU SG platform.

The structures listed were determined between early 2003 and December 2005, with the majority of structures solved during 2004–2005. PDB codes for deposited structures are also listed.

		Solution methods†				
No. of structures	Sequence identity < 30%	SAD	MAD	SIRAS	MR	PDB codes
40	11	7	3	3	27	1rkb, 1r3u, 1yfz, 1rn7, 1roa, 2bb0, 2g3f, 2f07, 2d4g, 2baz, 2ayd, 2b79, 2b78, 2fkn

† Structures determined using in-house data include one by sulfur-SAD, one by SIRAS using a mercury derivative and 16 by MR methods.

3.3. Protein purification and crystallization screens

The majority (above 90%) of soluble proteins were purified *via* nickel-column chromatography using the His_6 tag at the N-terminus of the recombinant proteins. A small fraction of proteins could not be tagged or could not be affinity-purified on a nickel column. For these proteins, other means such as GST tags or ion-exchange columns were then tried. Proteins in this category are normally collaborative projects and have functions that are known to some extent.

In the system shown in Fig. 4, a four-sample two-step purification protocol is running with nickel-column affinity chromatography followed by gel filtration, which is a default setting. The automatic runs normally take 12–18 h each and can be conveniently set to run overnight. Apart from the company default protocols, we have also tried to modify and develop other methods and protocols on this system, *e.g.* for non-His_6-tagged proteins or His_6-tagged proteins that failed to bind to the nickel column. Sometimes ion-exchange columns work well and we have been routinely using a two-step purification protocol with anion-exchange chromatography (Hitrap Q columns) followed by gel filtration and have obtained quite good results.

Fig. 6 shows some recorded pictures of crystals produced in the PKU SG platform. The images were recorded manually before our automatic imaging system was constructed. Most of the crystals shown here have produced good diffraction data and several reports on these crystals have already been published (Duan *et al.*, 2005; Nan *et al.*, 2006; Ren *et al.*, 2004; You *et al.*, 2003; Yu *et al.*, 2006; Zhou *et al.*, 2006).

3.4. The determination of the crystal structures

The general strategy adapted in the PKU SG platform for structure determination is as follows.

Figure 6
A panel of protein crystals obtained within the PKU SG platform. The structures of most crystals shown here have been solved.

(i) When a diffracting crystal is identified, a few commonly used heavy-atom compounds, including metal and halide soaks, are tried first before producing SeMet-substituted crystals.

(ii) High-redundancy sulfur-SAD data sets are collected from well diffracting crystals.

(iii) If both (i) and (ii) fail, new crystal forms will be sought and SeMet-substituted variants will be prepared.

(iv) If all the above fail, different protein constructs will be made, hoping to obtain a more suitable protein and crystal form to work on.

The above strategy has worked well on our platform, with about 80% of well diffracting crystals solved in a few months. This way, the in-house machine has been used not only for screening crystals for SR beamlines, but also as an important addition to the structure-determination process. More than 40% of the structures in Table 2 were solved using data collected on the in-house machine, employing MR, SIRAS and sulfur-SAD phasing methods (Ren *et al.*, 2004, 2005).

With limited access to SR beamlines, sulfur-SAD and different soaking strategies can be performed to make full use of the in-house resources. Generally, such an approach follows the strategy described above. When the in-house source is not sufficiently intense, it is still helpful to learn the crystal properties in preparation for data collection at an SR source.

Although our wet laboratory and dry laboratory computer systems are built with two different strategies, as described above, the software packages installed on each system are very similar. In each case, software was installed and updated only once on the respective server. After the update, users can directly use or copy the software to their own computers. This procedure saves time and helps new users to get started easily. Since some of the installed crystallography software already provides powerful packages with a pipeline-like structure-solution capability, we have not developed an HTP pipeline for structure determination on our platform. Furthermore, students need to be trained to solve and refine some structures manually in order to understand protein crystallography.

For the 40 unique structures solved so far on our SG platform, the estimated total cost per structure is about US $20 000. We are confident that this cost can be reduced to US $10 000 in one or two years.

4. Conclusions and perspectives

The main task for structural genomics projects is to catalogue the structural folds of unique proteins or their domains. The number of unique folds is believed to be in the range 2000–3000, whereas the number of protein families under 30% sequence identity is in the range 30 000–40 000 according to the available genome-sequence data. A complete protein domain/fold set will increase our understanding not only of the architecture and function of proteins, but also of their origin and evolution and will particularly shed light on the protein-folding problem.

While pursuing the SG approach, we realised that the SG pipeline can not only serve as a fast lane for solving protein structures, but more importantly is also a protein-production fast lane that can accumulate hundreds and thousands of soluble proteins. Furthermore, the HTP automatic multi-gene system used in present SG laboratories can be readily extended to HTP automatic multi-constructs by rational or random design of mutations for one particular gene with important function. In summary, it is hoped that the SG methods should greatly enhance our ability to solve protein structures at large, firstly the easier obtained 'low-hanging fruits' and subsequently more and more difficult protein structures.

Genomic approaches can be practiced not only by large consortia, but may also be adapted by normally funded university laboratories. We wished to provide an example that an average-funded structural laboratory can also work in the HTP mode and achieve the cost of about US $10 000 per solved protein crystal structure.

We wish to express our sincere gratitude to all laboratory members participating in the construction of our SG platform, particularly Drs Shanyun Lü, Xiaofeng Zheng, Haitao Ding, Hui Ren, Jianli An and Xiaoyan Zhang, graduate students Dan Li, Yamei Yu, Qiang Chen, Wei Mi, Linglong Ma, Jian Lei, Yanfeng Zhou, Xiangyu Liu, Kaituo Wang, Xiaoyan Liu, Juan Wang, Rui Li and others. The data-collection facilities at EMBL Outstation in Hamburg, Germany and MAX-Lab in Lund, Sweden have continuously provided great help to us. This work is supported by grants from the National Natural Science Foundation of China (NSFC; Nos. 30325012 and 30530190), Chinese High Tech Research and Development (863) program (Nos. 2001AA233011 and 2002BA711A13) and Peking University Grants 985 and 211. We also wish to thank Professor Xiaocheng Gu at Peking University and Professor Ming Luo at the University of Alabama at Birmingham, USA for invaluable help during the establishment of the structural genomics laboratory at Peking University and to thank Dr Na Yang for help in critical reading of the manuscript.

References

Adams, P. D., Grosse-Kunstleve, R. W., Hung, L. W., Ioerger, T. R., McCoy, A. J., Moriarty, N. W., Read, R. J., Sacchettini, J. C., Sauter, N. K. & Terwilliger, T. C. (2002). *Acta Cryst.* D**58**, 1948–1954.

Ajdic', D., McShan, W. M., McLaughlin, R. E., Savic', G., Chang, J., Carson, M. B., Primeaux, C., Tian, R., Kenton, S., Jia, H., Lin, S., Qian, Y., Li, S., Zhu, H., Najar, F., Lai, H., White, J., Roe, B. A. & Ferretti, J. J. (2002). *Proc. Natl Acad. Sci. USA*, **99**, 14434–14439.

Brenner, S. E. (2001). *Nature Rev. Genet.* **2**, 801–809.

Bricogne, G., Vonrhein, C., Flensburg, C., Schiltz, M. & Paciorek, W. (2003). *Acta Cryst.* D**59**, 2023–2030.

Brünger, A. T., Adams, P. D., Clore, G. M., DeLano, W. L., Gros, P., Grosse-Kunstleve, R. W., Jiang, J.-S., Kuszewski, J., Nilges, M., Pannu, N. S., Read, R. J., Rice, L. M., Simonson, T. & Warren, G. L. (1998). *Acta Cryst.* D**54**, 905–921.

Burley, S. K. (2000). *Nature Struct. Biol.* **7**, *Suppl.*, 932–934.

Chandonia, J. M. & Brenner, S. E. (2006). *Science*, **311**, 347–351.

Collaborative Computational Project, Number 4 (1994). *Acta Cryst.* D**50**, 760–763.

Dai, X., Chen, Q., Lian, M., Zhou, Y., Zhou, M., Lu, S., Chen, Y., Luo, J., Gu, X., Jiang, Y., Luo, M. & Zheng, X. (2005). *Biochem. Biophys. Res. Commun.* **332**, 593–601.
Ding, H.-T., Ren, H., Chen, Q., Fang, G., Li, L.-F., Li, R., Wang, Z., Jia, X.-Y., Liang, Y.-H., Hu, M.-H., Li, Y., Luo, J.-C., Gu, X.-C., Su, X.-D., Luo, M. & Lu, S.-Y. (2002). *Acta Cryst.* D**58**, 2102–2108.
Duan, M.-R., Ren, H., Mao, P., Wei, C.-H., Liang, Y.-H., Li, Y. & Su, X.-D. (2005). *Biochim. Biophys. Acta*, **1750**, 14–16.
Emsley, P. & Cowtan, K. (2004). *Acta Cryst.* D**60**, 2126–2132.
Gong, W. M. *et al.* (2003). *J. Struct. Funct. Genomics*, **4**, 137–139.
Jones, T. A., Zou, J.-Y., Cowan, S. W. & Kjeldgaard, M. (1991). *Acta Cryst.* A**47**, 110–119.
Nan, B., Zhou, Y., Liang, Y.-H., Wen, J., Ma, Q., Zhang, S., Wang, Y. & Su, X.-D. (2006). *Biochim. Biophys. Acta*, **1764**, 839–841.
Otwinowski, Z. & Minor, W. (1997). *Methods Enzymol.* **276**, 307–326.
Perrakis, A., Morris, R. & Lamzin, V. S. (1999). *Nature Struct. Biol.* **6**, 458–463.
Ren, H., Liang, Y., Li, R., Ding, H., Qiu, S., Lu, S., An, J., Li, L., Luo, M., Zheng, X. & Su, X. D. (2004). *Acta Cryst.* D**60**, 1292–1294.
Ren, H., Wang, L., Bennett, M., Liang, Y., Zheng, X., Lu, F., Li, L., Nan, J., Luo, M., Eriksson, S., Zhang, C. & Su, X.-D. (2005). *Proc. Natl Acad. Sci. USA*, **102**, 303–308.
Service, R. (2005). *Science*, **307**, 1554–1558.
Stevens, R. C., Yokoyama, S. & Wilson, I. A. (2001). *Science*, **294**, 89–92.
Storoni, L. C., McCoy, A. J. & Read, R. J. (2004). *Acta Cryst.* D**60**, 432–438.
Terwilliger, T. C. (2003). *Methods Enzymol.* **374**, 22–37.
Wang, J.-W., Chen, J.-R., Gu, Y.-X., Zheng, C.-D., Jiang, F. & Fan, H.-F. (2004). *Acta Cryst.* D**60**, 1987–1990.
Winn, M. D., Ashton, A. W., Briggs, P. J., Ballard, C. C. & Patel, P. (2002). *Acta Cryst.* D**58**, 1929–1936.
You, D., Chen, Q., Liang, Y., An, J., Li, R., Gu, X., Luo, M. & Su, X.-D. (2003). *Acta Cryst.* D**59**, 1863–1865.
Yu, Y., Li, L., Zheng, X., Liang, Y.-H. & Su, X.-D. (2006). *Biochim. Biophys. Acta*, **1764**, 153–156.
Zhou, Y.-F., Mi, W., Li, L., Zhang, X., Liang, Y.-H., Su, X.-D. & Wei, S. (2006). *Biochim. Biophys. Acta*, **1764**, 324–326.