STRUCTURAL GENOMICS drives demand for high throughput protein purification

The rapid growth in proteomic and structural and functional genomic research is driving demand for purified proteins that far exceeds the industry's ability to scale-up conventional protein production technologies. This demand has sparked innovation in protein separation and purification technologies. We discuss how the advent of improved systems for protein expression and high throughput parallel protein purification can deliver the opportunity to ignite a revolution in structural and functional genomics, proteomics and high-value drug target discovery.

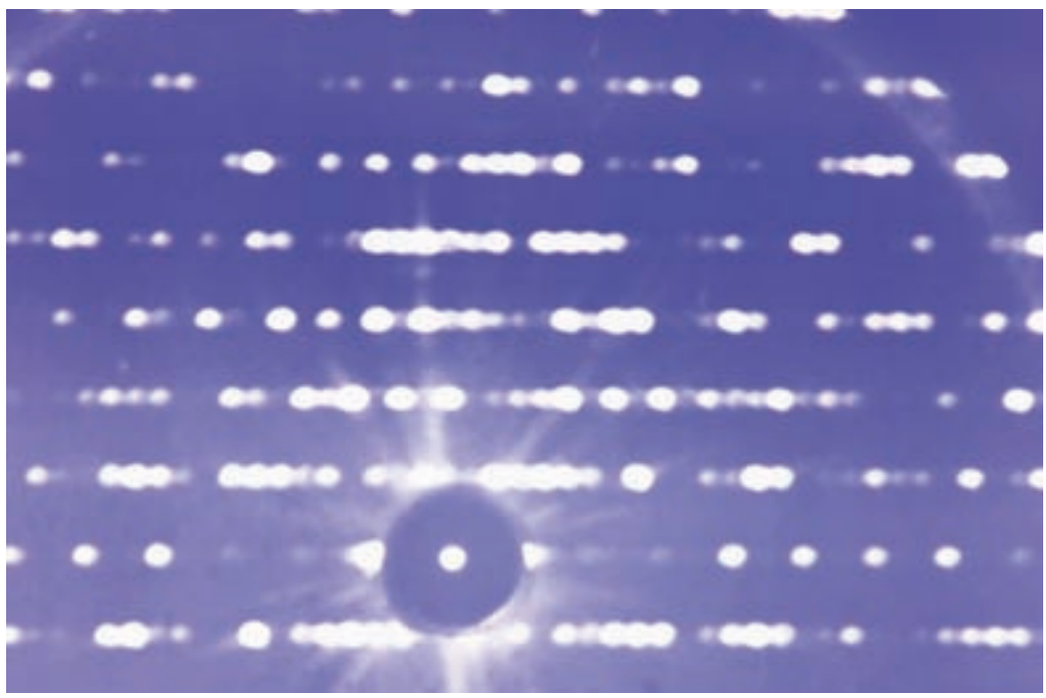
By Karsten Fjarstedt

Particular phenotypes and disease states reflect differences in the types, levels and activities of proteins within specific cells. Drug discovery relies on finding compounds that can suppress, amplify or modulate the activities of specific proteins. The Human Genome Project revealed the nucleotide sequence of all human genes and thus the amino acid sequence of all the body's proteins. But the genomics revolution has failed to deliver the hoped-for flood of novel medical entities entering the pharma pipeline. To realise the massive potential of sequencing the human genome, we also need to determine the three-dimensional structure and function of the proteins the genome encodes. The 1990s was the decade of DNA; we are now at the beginning of the protein epoch.

Structural genomics refers to the elucidation of the three-dimensional structure of the proteins expressed by the genome with the objective of linking how genetic variation predisposes an individual to particular diseases, the recognition of the proteins involved in that disease and the identification of those which are candidates for drug targets. The ultimate goal is to target those proteins specific to a particular individual, or a group of individuals with a certain disease.

Protein function is largely determined by the molecule's three-dimensional structure and its chemical and conformational modifications. Prediction models can be used to approximate a protein's folding pattern from its amino acid sequence, but the most sophisticated means for visualising a protein's structure, whether in isolation or bound to its target, is x-ray crystallography (**Figure 1**).

X-ray crystallography requires crystals that are singular and of perfect quality, and thus demands an extremely pure sample of the protein. At present, a researcher wishing to determine a protein's



Alfred Pasieka/Science Photo Library

Figure 1
Protein structure determination x-ray crystallography studies reveal the structure of a protein. Spots (white) show the diffraction pattern that forms as x-rays pass through the protein. The diffraction occurs because the protein has been prepared in a crystalline form with its atoms forming a regular lattice arrangement. The structure of the protein determines the pattern seen, which allows complex mathematics to be used to determine the structure of the protein. Milligramme quantities of purified protein are required for x-ray crystallography studies

structure has no alternative than to obtain a pure sample of that protein. Herein lies the challenge in the drug discovery workflow. A single gene may express hundreds of proteins, differing in amino acid sequence due to splice variants, in three-dimensional conformation, or in post-translational modifications. Different cells express different proteins, and even the proteins expressed in a given cell will vary throughout its life depending on the signals it receives. Evaluating a target gene of interest necessitates expressing, purifying and evaluating every one of the hundreds of proteins it may encode to determine which has the greatest potential as a drug target. In a given therapeutic area, researchers will need to screen and study the function of several hundreds, if not thousands, of proteins.

The need for high throughput

The traditional approach to structural genomics focused on identifying and studying proteins of interest one at a time. But lagging productivity in pharmaceutical discovery and development has squeezed bottom lines and put pressure on discovery groups to generate more high-value information with rapid turnaround. The explosive growth in proteomic and structural and functional genomic research is driving demand for purified proteins that far exceeds the industry's ability to scale-up conventional protein production technologies.

Similar to the historical bottleneck caused by a paucity of chemical entities to screen against drug targets, which led to the advent of combinatorial chemistry and high throughput compound synthesis, and to the logjam in compound screening that stimulated the development of high throughput screening technology, growing demand for pure samples of proteins has sparked innovation in protein separation and purification technologies.

An alternative to the traditional, methodical approach to structural genomics is a shotgun strategy aimed at identifying proteins and determining their three-dimensional structure in as short a time as possible. The ultimate goal is to compile a library of structures that serves as a database for a predictive model of protein function. By comparing novel proteins to a library of protein molecules known to have a role in disease, it will be possible to predict which proteins are most likely to be relevant in biochemical pathways linked to pathology. This shotgun approach to structural genomics – and the need to build a structural database of sufficient magnitude – is a primary driving force behind the development of high throughput protein production.

High throughput protein production is one element in the workflow of cloning, protein expression, screening, protein purification and x-ray crystallography/nuclear magnetic resonance, to provide the structure of the target protein. At any of these stages the process may fail. Typically for researchers to obtain one structure, they must

begin working with 18 gene targets at the protein expression stage, even for relatively easy to handle proteins. However, more than 60% of drug targets are membrane proteins, hydrophobic and insoluble. These remain notoriously difficult to crystallise, although portions of such proteins (lacking the hydrophobic region) have been successfully cloned, expressed and crystallised.

Issues in protein expression

Protein expression systems must be optimised for high yield of a specific soluble protein. Whether carried out in bacterial, insect or mammalian cells, recombinant protein expression must take into account factors such as yield, proper folding, post-translational modification, and ease of collecting and purifying the protein product.

Selection of an expression system and optimisation of conditions for maximum yield often requires a trial-and-error approach, with several alternative expression systems and sets of conditions being screened to identify the optimum approach. Expression screening typically relies on liquid handling and other robotic systems to maximise throughput. These studies are performed at low volumes with only small amounts of protein. Subsequent crystallography, nuclear magnetic resonance, and functional proteomic experiments aimed at defining protein structure and function require scale-up of the expression system to allow for the production of much larger quantities (10 to 50mg) of protein. Such purification systems need to be fully integrated, automated and robust to handle the demands put upon them.

At present, the bacterium *Escherichia coli* (*E. coli*) is the most commonly used organism for protein expression. *E. coli* is readily amenable to high throughput protein expression due to its ease and speed of use and low cost. Its limitations include the inability to perform eukaryotic post-translational modifications and poor solubility of some proteins. Cell-free expression systems (usually lysates of bacterial cells) have advantages for high throughput applications as the absence of a cell membrane simplifies transfection and product acquisition. Protein yields from cell-free expression systems have improved substantially, and yields of up to 6mg/ml of individual protein are now achievable.

For expressing biologically active human proteins, mammalian cells such as CHO or HEK293 cells offer advantages in terms of proper folding and post-translational modification. However, their use in high throughput protein expression systems must also take into consideration their negative aspects such as slow growth, lower productivity, transfection inefficiency, reliance on costly reagents and fragility in high-throughput processes. Researchers have also demonstrated successful protein expression in insect cells and in the yeast *Saccharomyces cerevisiae*, although both have limitations, for example differences in post-translational modifications and the greater complexity of transfection mechanisms.

On-demand expression of large numbers of functional proteins in a high throughput format will become increasingly achievable in the future as rules for predicting the success of an expression system become more powerful, analytical tools for assessing the physical and functional integrity of purified proteins become more sophisticated, and methods improve for pinpointing the simplest and least costly strategy for producing large amounts of quality proteins. Concurrently, advances in bioinformatics will provide the database and data mining tools needed to optimise the design of expression systems and to catalogue and compare protein parameters such as structural features, solubility and biochemical characteristics.

High throughput protein purification

Meeting the need for pure proteins for structural genomics applications, protein microarrays and biochemical studies requires rapid, robust and reproducible methods for purifying tens of milligrammes of pure protein from a cell preparation or cell-free lysate. Downstream processing and chromatographic separation of large-scale process streams has historically been a bottleneck in protein production and a target for automation and software-driven process management strategies. Traditional separation schemes typically resulted in a loss of 50% or more of the target protein during purification: an enormous waste of time and resources. Control software designed to optimise process parameters, combined with chromatographic media capable of extracting a single protein from a complex mixture, and algorithms that assess the results of the separation steps and identify the fraction to take forward, can provide an integrated and automated solution to liquid chromatography protein purification. By minimising manual intervention, exploiting multiple separation technologies, and relying on a knowledge-based, computer control system, liquid chromatography has come of age as a truly high throughput platform for protein separation and purification.

In the past, researchers have had to choose between products that offer sequential purification of single samples, or one-step purification performed

in high throughput formats using liquid handling robots and 96-well microtitre plates. However, now researchers can buy instruments that combine these attributes, integrating multi-step purification with highly parallel processing, and providing researchers with the capability to design a flexible workflow and select the most appropriate chromatographic steps, as well as meeting the need for large quantities of pure protein.

The most commonly used chromatographic separation techniques for protein purification are affinity chromatography (which separates proteins using an affinity tag), desalting or gel filtration (which separates proteins according to size) and ion exchange chromatography (which separates proteins according to charge). Affinity tags attached to recombinant proteins facilitate purification using standard liquid chromatographic techniques. His_6 tags and glutathione-S-transferase (GST) tags are popular choices and account for about 80% of all tagged proteins. Some affinity tags may improve protein solubility. However, tagging may also have negative consequences, such as improper protein folding, aberrant protein function, or unsuccessful crystallisation, and for some applications researchers may wish to remove the tag from the purified protein.

In order to produce enough pure protein for crystallisation and structural genomics studies, a high throughput system needs to be able to purify tens of milligrammes of protein in a single run. Automated sample loading, process monitoring, peak identification, and transfer of fractions to the next purification step are required to minimise the time required for hands-on operations and to standardise the processes, making them more robust, reliable and reproducible. The tag removal process can be integrated into the purification process through a simple enzymatic cleavage step.

High throughput protein purification also requires sophisticated computer control systems based on well-defined process algorithms, preset default settings and watch commands capable of guiding the system. The control systems should be easy to use, guiding the operator in designing a purification protocol through selecting functions and parameters from drop-down menus and check boxes. The default settings should be readily changeable to accommodate particular purification processes specific to individual proteins. A key advantage of a fully automated system is that the operator need not be a scientist with expertise in protein purification. Instead, a trained technician can oversee the day-to-day operations of the system, freeing the expert scientist to pursue other activities. This move to automation has major implications for research institutes and commercial organisations carrying out large-scale high throughput protein purification, and for those companies producing protein purification equipment. The emphasis has changed from employing skilled chromatographers, experts in protein purification, to technicians and system operators. The fully integrated systems for protein purification now have built-in knowledge such that the operator no longer needs to be an expert in protein purification. This process will accelerate. Even today a single non-expert operator is able to control a bank of protein purification systems, running 24 hours a day. Fifty thousand protein purifications per year can be run with a single system. The throughput requirements of structural genomics research are so vast that this automation is essential if a sound understanding of protein structure is to be achieved at the necessary speed.

Conclusion

A rapid, shotgun approach to structural genomics that enables scientists to define the structure and function of proteins playing a role in disease and to identify those that might prove to be valuable drug targets has become a realistic goal with the advent of improved systems for protein expression and high-throughput, parallel protein purification. Similar to the way automation, computer control systems and parallel processing combined to launch chemical library synthesis and compound screening to a new level of productivity, so high throughput protein purification delivers the opportunity to ignite a revolution in structural and functional genomics, proteomics and high-value drug target discovery. In years to come, understanding the structure and function of proteins will enable researchers to design drugs targeting specific proteins, even to the proteins specific to a particular individual or group of individuals, ultimately delivering the vision of personalised medicine offering greater efficacy, fewer side-effects and fewer costly treatment failures. DDW

Karsten Fjarstedt *has more than 15 years' experience in the field of protein purification. He is currently Director of Product Management, Separations Instruments with GE Healthcare (formerly Amersham Biosciences). Based in Uppsala, Sweden, Karsten holds an MSc in Chemical Technology from the Royal Institute of Technology, Stockholm.*