Making decisions for structural genomics

Ana Rodrigues and Roderick E. Hubbard

Date received: 24th February 2003

Abstract

A large number of structural genomics programmes have been established worldwide with the common aim of large-scale, high-throughput protein structure determination. Due to the considerable challenges posed by the experimental methods of structural determination (primarily X-ray crystallography and nuclear magnetic resonance spectroscopy) it is important to select and prioritise candidate molecules that will maximise the information gained from each new structure. This paper describes the scientific principles that underlie target selection and the various bioinformatics tools that may be employed in such selection procedures. Then follows a discussion of the availability of resources incorporating these methods and a description of the design and application of a purpose-built target selection resource for structural genomics.

INTRODUCTION

The advances in gene mapping and sequencing of the 1990s are delivering the complete genome sequence for an increasing number of organisms.1 This avalanche of genomic data provides the starting point to develop methods for exploring the functions, interactions and interrelationships between genes and their protein products. This combination of functional genomics and proteomics will lay the foundation for an integrated and extensive view of biology at the functional level.2,3

In many ways, understanding the structure of proteins provides the most detailed view of this integrated biology, where the mechanism of protein action can be explored and related to the interactions and chemistry that underpin biological function. Structural studies can provide detailed descriptions of many features, such as the nature of the specific molecular surfaces for protein, nucleic acid or small molecule recognition, the nature and mechanistic consequences of conformational change in a protein or the details of the structural interactions that catalyse specific chemical reactions. The pace of technical developments in genomics and proteomics has been dramatic and the past 10 years have seen extraordinary advances in the speed and quality of measurements of gene sequence, level of protein expression and the functional consequence of individual proteins. The first protein structures were determined over 40 years ago and although there have been important advances in the past 10 years, the rate at which protein structures can be determined is dramatically slower than the speed at which important and interesting genes and functions are identified.

The past five years has seen the initiation worldwide of a number of programmes in structural genomics (see Table 1). The common feature of all these projects is the development and application of high-throughput methods for determining a large number of protein structures. However, the scientific rationale for these projects varies:

- Determining all the structures of genes identified in a particular genome;4
- Attaining a complete structure description of a specific biochemical pathway;5–7
<table>
<thead>
<tr>
<th>Consortium</th>
<th>Focus organisms</th>
<th>Selection criteria</th>
<th>No. of targets</th>
<th>Resources</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>USA</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NIH Protein Structure Initiative</td>
<td>Mycoplasma genitalium, Mycoplasma pneumoniae</td>
<td>Structural novelty, Prevalence</td>
<td>345</td>
<td>PRESAGE</td>
</tr>
<tr>
<td>Berkeley Structural Genomics Center</td>
<td>Arabidopsis thaliana</td>
<td>Structural novelty, Functional novelty, Experimental tractability</td>
<td>1,782</td>
<td>Sesame</td>
</tr>
<tr>
<td>Center for Eukaryotic Structural Genomics</td>
<td>Theriotoga maritima, Caenorhabditis elegans</td>
<td>Genome coverage, Structural novelty, Technology development, Experimental tractability</td>
<td>5,380</td>
<td>PSCA, DAPS, TPM, FSS</td>
</tr>
<tr>
<td>Joint Center for Structural Genomics</td>
<td>Theriotoga maritima</td>
<td>Structural novelty, Functional novelty, Technology development, Experimental tractability</td>
<td>2,834</td>
<td>–</td>
</tr>
<tr>
<td>Midwest Center for Structural Genomics</td>
<td>Bacillus subtilis, Haemophilus influenza, Escherichia coli</td>
<td>Structural novelty, Medical importance</td>
<td>2,834</td>
<td>–</td>
</tr>
<tr>
<td>New York Structural Genomics Research Consortium</td>
<td>Homo sapiens, Model organisms</td>
<td>Structural novelty, Experimental tractability, Information on protein's pathway, expression, and interactions</td>
<td>839</td>
<td>MAGPIE, SANDPIPER, IceDB</td>
</tr>
<tr>
<td>Northeast Structural Genomics Consortium</td>
<td>Saccharomyces cerevisiae, Drosofila melanogaster</td>
<td>Structural novelty, Prevalence</td>
<td>5,391</td>
<td>ZebsView, TAP, SPINE, ReportDB</td>
</tr>
<tr>
<td>Southeast Collaboratory for Structural Genomics</td>
<td>Homo sapiens, Caenorhabditis elegans, Pterococcus furinosus, Plasmodium falciparum</td>
<td>Structural novelty, Multidomain proteins, Membrane proteins, Structural novelty, Medical importance</td>
<td>3,769</td>
<td>–</td>
</tr>
<tr>
<td>Structural Genomics of Pathogenic Protozoa</td>
<td>Trypanosoma brucei, Trypanosoma cruzi, Leishmania sp.</td>
<td>Medical importance</td>
<td>274</td>
<td>–</td>
</tr>
<tr>
<td>TB Structural Genomics Consortium</td>
<td>Mycobacterium tuberculosis</td>
<td>Essential proteins, Mutants affecting host/parasite interaction, Structural novelty</td>
<td>1,389</td>
<td>Online progress report</td>
</tr>
<tr>
<td><strong>Others</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Structure 2 Function Project</td>
<td>Haemophilus influenza</td>
<td>Functional novelty, Experimental tractability</td>
<td>331</td>
<td>Online progress report</td>
</tr>
<tr>
<td>Structural GenomiX</td>
<td>Bacterial pathogens</td>
<td>Functional novelty, Apo- and co-complexes, Protein kinases, Nuclear hormone receptors, Membrane proteins</td>
<td>Not divulged</td>
<td>Proprietary</td>
</tr>
<tr>
<td>Syrrx</td>
<td>None</td>
<td>Disease causing proteins</td>
<td>Not divulged</td>
<td>Proprietary</td>
</tr>
<tr>
<td><strong>Canada</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bacterial Structural Genomics Initiative</td>
<td>Escherichia coli</td>
<td>Small molecule pathways, Structural novelty</td>
<td>475</td>
<td>On-line progress report</td>
</tr>
<tr>
<td>Ontario Center for Structural Proteomics</td>
<td>Methanosarciom trichomacrophilum, Caenorhabditis elegans, Escherichia coli</td>
<td>Structural novelty, Sarcosarcasmicin, Functional novelty</td>
<td>2,700</td>
<td>HSQC Catalogue</td>
</tr>
<tr>
<td>Montreal Network for Pharmacology-Proteomics &amp; Structural Genomics</td>
<td>Mammalian cell</td>
<td>Endoplasmic reticulum protein (known and novel identified through mass spectrometry)</td>
<td>50–100</td>
<td>On-line progress report soon to be available</td>
</tr>
<tr>
<td><strong>EU</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Marseilles Structural Genomics Program</td>
<td>Escherichia coli, Mycobacterium tuberculosis</td>
<td>Functional novelty, Structural novelty, Glycobiology (cutinases, lipases, glycanases and glycosyltransferases, G-protein coupled receptors</td>
<td>313</td>
<td>Online progress report</td>
</tr>
</tbody>
</table>

© HENRY STEWART PUBLICATIONS 1467-5463. BRIEFINES IN BIOINFORMATICS. VOL. 4. NO. 2. 150-167. JUNE 2003
### Table 1: (continued)

<table>
<thead>
<tr>
<th>Consortium</th>
<th>Focus organisms</th>
<th>Selection criteria</th>
<th>No. of targets</th>
<th>Resources</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>UK</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Oxford Protein Production Facility 47†</td>
<td><em>Homo sapiens</em></td>
<td>Structural novelty</td>
<td>128</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td><em>Herpesviridae</em></td>
<td>Structural novelty</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Experimental tractability</td>
<td></td>
<td></td>
</tr>
<tr>
<td>North West Structural Genomics Center 48</td>
<td><em>Mycobacterium tuberculosis</em></td>
<td>Surface proteins</td>
<td>40</td>
<td>–</td>
</tr>
<tr>
<td><strong>Germany</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Protein Structure Factory 49</td>
<td>cDNAs available at the Berlin Resource Center</td>
<td>Structural novelty</td>
<td>1,280</td>
<td>Online progress report 60</td>
</tr>
<tr>
<td>Structural Proteomics IN Europe 51</td>
<td><em>Bacillus anthracis</em></td>
<td>Virulence genes</td>
<td>600 planned</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td><em>Camptothecin jujuni</em></td>
<td>Host/pathogen interaction</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td><em>Mycobacterium tuberculosis</em>, <em>leprae</em> and <em>bovis</em></td>
<td>Disease-related protein families</td>
<td>(kinases, proteases, kiesins, nuclear receptors, cell surface molecules)</td>
<td></td>
</tr>
<tr>
<td></td>
<td><em>Herpesviridae</em></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td><em>Homo sapiens</em></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Japan</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RIKEN Structural Genomics and Proteomics Initiative 52</td>
<td><em>Arabidopsis thaliana</em></td>
<td>Structural novelty</td>
<td>705</td>
<td>Online progress reports 53, 54</td>
</tr>
<tr>
<td></td>
<td><em>Thermus thermophilus</em></td>
<td>Eukaryotic specific</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td><em>Pyrococcus horikoshii</em></td>
<td>DNA/RNA binding</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Cell signalling</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>SNP-bearing</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Disease related proteins</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Consortium</th>
<th>Selection strategy</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Canada</strong></td>
<td>Focus on <em>Escherichia coli</em> aiming at full genome coverage. Their CC3D 56 resource is available on the world-wide web.</td>
</tr>
<tr>
<td><strong>UK</strong></td>
<td>Focus on chromatin mediated transcriptional repression, cyclin-dependent kinases and small G-proteins involved in cellular control.</td>
</tr>
<tr>
<td><strong>Switzerland</strong></td>
<td>Focus on membrane proteins and intermolecular interactions in supramolecular assemblies.</td>
</tr>
<tr>
<td><strong>Japan</strong></td>
<td>Focus on membrane proteins and ligand–protein interactions. Also developing an integrated database system.</td>
</tr>
<tr>
<td>Biological Information Research Center Structural Genomics Group 59</td>
<td>Focus on the <em>Oryza sativa</em> genome.</td>
</tr>
<tr>
<td><strong>Korea</strong></td>
<td>Focus on the <em>Mycobacterium tuberculosis</em> and <em>Helicobacter pylori</em> genomes to enable drug discovery.</td>
</tr>
<tr>
<td>Structural Proteomics Research Organisation</td>
<td>Focus on <em>Homo sapiens</em> proteins with emphasis on disease associated ones, as well as novel bacterial proteins.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Totals</th>
<th>Consortia</th>
<th>Focus organisms</th>
<th>No. of targets</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>29 projects</td>
<td>&gt;28 genomes</td>
<td>&gt;28,875 proteins</td>
</tr>
</tbody>
</table>

---

*International effort involving labs from North America, Europe, Russia, India, Asia and New Zealand.
† These efforts are also to some extent included in the Structural Proteomics IN Europe (SPINE) project.
• Studying proteins associated with certain disease states;  
• Obtaining novel structures to increase coverage of protein fold space.  

In addition, many structural biology laboratories worldwide are embarking on large-scale structure determinations as part of major programmes in functional genomics. For example, at York, we are a partner in a major Wellcome Trust-funded project to understand aspects of malaria biology.

There are many technical challenges for large-scale structure determination (see Heinemann et al. for an overall discussion, or the following reviews: Pokala and Handel or Gilbert and Albal on protein production, Hendrickson on X-ray crystallography, Prestegard et al. or Al-Hashimi and Patel on nuclear magnetic resonance (NMR) spectroscopy and Baumeister and Steven on electron microscopy, EM). Despite the ambitious goals of many structural genomics projects, the rate at which protein structures can be determined is still quite low, with the major bottleneck being the reliable production of large homogeneous quantities of functional protein. It is therefore important to identify the genes for which a protein structure will provide the highest new information content and, where possible, quantify measures of how tractable each protein system is for structure determination.

Learning from evolution
Proteins, and protein domains, will often assume similar structural scaffolds. These fold similarities can be the result of both convergent and divergent evolution. Where proteins are related by divergent evolution, they share a common ancestor and are said to be homologous, i.e., they belong to the same protein family. Such proteins can be the product of either post-speciation divergence (known as orthologues, or proteins that perform the same function in different species), or gene duplication events (known as paralogues, or proteins that perform different but related functions within one organism). In both cases, the proteins will sustain some degree of similarity depending on how early in evolution the divergence took place.

Sequence similarity can be a reliable indicator of protein homology and, hence, structure similarity. This relationship allows the structure of a protein to be predicted if the three-dimensional coordinates of one of its homologues has been determined (see, for example, Swindells and Thornton). In more general terms, when the structure of a protein family member is determined,
the overall fold of all other members of the family can be inferred. Sequence similarity search tools, such as BLAST and FastA, can be used to rapidly identify homologues with known protein structures and a homology model can be constructed using programs such as MODELLER. The empirical cut-off for obtaining a reasonable homology model for a protein with a known structural homologue is widely accepted to be 40 per cent sequence identity over a considerable alignment span. At this level of homology, the model of the structure of a protein will reliably predict its overall fold. In addition, depending on the extent and nature of sequence conservation, the model may be sufficient to make predictions about the function and properties of the new protein.

Most structural genomics projects will therefore lower the priority on the experimental structure determination of homologous proteins, unless a detailed study is required.

The increasing number of known protein structures has also helped identifying cases where nature develops similar structural or mechanistic solutions from intrinsically different starting points, i.e. the convergent evolution of proteins that have no common ancestor and thus possess distinct sequences. It is now recognised that many proteins with very different sequences adopt the same fold, presumably because there is a limited number of stable folds. There has been considerable effort over the past 10 years not only to analyse and categorise the fold space but also to develop fold recognition methods. There is a wide variety of approaches, though most involve assessing how well a novel protein sequence will fit into each of a representative set of folds. These threading methods rely heavily on alignment methods and in particular on scoring functions that assess how stable a fold is. Such types of calculations are challenging and are not sufficiently robust for target selection. However, one of the outcomes of current structural genomics efforts will be knowledge of an increased number of structures and folds that will improve these prediction methods.

Deciding on a strategy

Two distinct trends can be identified in the goals of current structural genomics projects namely: structural genomics by structure and structural genomics by function.

For most projects in ‘structural genomics by structure’, the main task is to identify proteins likely to have a novel fold. For example, particularly appealing targets are proteins that have no recognisable homologues, so-called ORFan proteins, that may assume novel folds and perform previously unperceived functions.

Sequence similarities to proteins with a known three-dimensional structure often do not comply with the comparative homology model threshold described above. These can range from a high sequence identity with a small alignment length to a low sequence identity with extensive alignment length. Such matches can be false positives, but can also correspond to conserved structural and/or functional motifs or distant homologues, respectively. The true positives can be differentiated, to some extent, through the use of more sophisticated sequence comparison algorithms, such as PSI-BLAST, hidden Markov model (HMM)-based and profile-based protocols (see for example: Schaffer et al. or Yona and Levitt). Implementations of such algorithms, purposefully tuned for fold prediction include SUPERFAMILY and the PDB-Intermediate Sequence Library (PDB-ISL). Though alignments identified through these methods are indicative of fold similarities, and can thus help predict the likelihood of a protein sequence to assume a novel fold, the proteins are usually not related enough to allow a homology model to be computed.

For structural genomics projects to uncover the richness of structural space,
the three-dimensional structure of a representative protein from each family (where a family contains proteins with 40 per cent and more similarity over a large span of their sequences, i.e. family members are within ‘homology-modelling distance’) will have to be experimentally determined. A series of databases and tools have been devised to cluster all known protein sequences into such families, namely ProTarget, ProtoMap, GeneRAGE and SUPFAM. Such resources can also be employed to, for example, identify those targets whose structure determination will provide structural information for the most proteins (i.e. the largest family). If obtaining structures for proteins which assume novel folds is the main drive of the project, one can also resort to the use of secondary structure predictions (using programs such as PHDsec and Pred2ary) to query against a database of known topologies (such as those provided by the TOPS server).

In ‘structural genomics by function’, priority is given to specific protein families, those that participate in particular metabolic pathways, or all proteins that perform a generic function of interest. Those protein families, for which a representative has been identified in all thus-far sequenced organisms, are especially attractive targets. The family’s prevalence suggests that these proteins may be essential to life. Among pathogenic organisms’ genomes, proteins associated with virulence or host interactions are another class of highly desirable candidates.

For all these applications, a detailed annotation of the protein’s function assumes prime importance. Functional annotation can be achieved through sequence comparison with proteins of known function (found in curated databases such as SWISS-PROT), using sequence search similarity programs such as BLAST and FastA. More sensitive software tools, such as PSI-BLAST, HMMER and IMPALA (profile-profile based method), allow the detection of remote homologies within the ever-increasing sequence data sets. Methods such as those combined in the InterPro database increase the reliability of the predictions by utilising curated protein domain family information, developed to enable sequence comparisons at the domain level (thus avoiding misannotations due to the modular nature of proteins). Further information on the protein’s function, such as its metabolic role, and its part, if any, in human disease, can also be obtained through sequence similarity searches, using web-based resources such as the Kyoto Encyclopaedia of Genes and Genomes (KEGG) metabolic pathways and the On-line Mendelian Inheritance in Man (OMIM) database. An indication of the prevalence of the protein can also be obtained through the use of such algorithms, by scanning the distribution of the gene product in sequenced genomes from all kingdoms of life.

Coping with limitations
Unfortunately, not every protein is tractable to structure determination and the experimental process has many potential bottlenecks. These range from difficulties experienced while cloning, expressing and purifying a protein, to issues related to the structural determination technique per se, such as crystal growth (in X-ray crystallography) or size limitations (in solution state NMR spectroscopy). A priori identification of problematic proteins or protein segments can remove the more obvious experimentally difficult proteins. Integral membrane proteins have proved to be particularly troublesome (see Creuzet et al. for a success story). The main difficulty is the production of large quantities of homogeneous, functional protein, and purification and crystallisation are hampered by solubility issues. Programs such as TMAP and TMHMM can be used to predict the location of transmembrane regions in a protein sequence. The identification of...
such segments is relatively straightforward due to the hydrophobic and physico-chemical constraints imposed by the lipid layer, though the available methods are generally more successful in recognising helical membrane segments than strand elements. Regions of a protein with little residue variation are traditionally associated with unstructured regions.\textsuperscript{104,123} These so-called low-complexity regions are, therefore, less amenable to structural studies. Low-complexity regions of a highly repetitive nature are, in fact, underrepresented in the Protein Data Bank (PDB).\textsuperscript{124} Non-globular segments, such as low-complexity regions and coiled-coils, can also be identified using primary sequence information. Low-complexity sequences can be distinguished using low-complexity segment identification algorithms, such as SEG\textsuperscript{118} or CAST\textsuperscript{125}. The program COILS\textsuperscript{2126,127} can be used to predict the likelihood of sequence segments to form left-handed two-stranded coiled-coils, though the more generic SEG algorithm can also be employed in the detection of such regions.\textsuperscript{128}

Within families of interest there will be proteins possessing physical and chemical characteristics more or less desirable according to the experimental procedure to be employed. Taking attributes such as size, predicted stability and solubility into account may help to reduce the failure rate of the structure determination process. Several of these properties can be predicted or derived based on the protein’s amino acid sequence alone. Some can be calculated using a sequence analysis software package like the European Molecular Biology Open Software Suite (EMBOSS),\textsuperscript{129} for example. Others can be estimated using implementations of statistical models derived from empirical data (eg the revised Wilkinson–Harrison statistical solubility model\textsuperscript{130}). Certain protein characteristics may not be necessary for selection, but might provide useful information to guide experimental procedures such as the protein’s extinction coefficient, molecular weight, grand average hydrophathy, isoelectric point and chemical composition. Software to compute each of these characteristics is also available in the EMBOSS software package. Nucleotide sequence properties, such as codon usage or the GC content of a gene, which can be calculated with little effort, can also be valuable for identifying potential issues in protein production.

**TARGET SELECTION RESOURCES**

Information about the targets selected by structural genomics projects worldwide is centrally stored at TargetDB,\textsuperscript{131} a target registration database developed and maintained by the PDB. The data, currently over 24,000 protein targets, are organised according to the International Task Force in Target Tracking recommendations\textsuperscript{132} and can be searched in a variety of ways (including through sequence similarity), as well as downloaded in XML format.

Most structural genomics consortia have also established on-line progress reports which contain details on, and reflect the current experimental status of, each of their targets. Examples of such resources are the Integrated Consortium Experimental Database (IceDB),\textsuperscript{133} ZebaView,\textsuperscript{27} the Structural Proteomics In the North East (SPINE) system\textsuperscript{134} and ReportDB.\textsuperscript{31} These web-based resources can be accessed, to a greater or lesser extent, by the general public, and contain varying degrees of information on the targets. Data regarding determined structures and homology models derived from the newly solved structures are generally retrievable, whereas information on the calculations performed for each target to enable its selection is mostly kept within each consortium’s domain.

Some consortia do divulge such annotations, through information repositories that can be searched and queried by any user. Resources such as the Protein Resource Entailing Structural

Prediction of the physical and chemical properties of proteins enables the prioritisation of targets according to the experimental pipeline to be employed

The target registration database, TargetDB, holds information about targets selected by structural genomics projects worldwide

Most consortia have established on-line progress reports containing information on the current experimental status of each of their targets

\hspace{1cm}

\hspace{1cm}© HENRY STEWART PUBLICATIONS (1467-5463). BRIEFINGS IN BIOINFORMATICS. VOL 4. NO 2. 150–167. JUNE 2003
Annotation of Genomic Entities (PRESAGE), the Protein Sequence Comparative Analysis (PSCA) system and the Target Analysis and Prioritisation (TAP) database allow the scientific community to select proteins within their target list according to specific characteristics, such as functional and structural annotations, experimental status or sequence properties (e.g., length or theoretical isoelectric point). The PRESAGE database also allows external registered users to add annotations to the targets, while a number of TAP suite tools can be rerun against up-to-date data sets by any user.

A few consortia have developed resources that enable not only the consultation of the annotations for each of their targets, but also the reprioritisation of this target list based on the annotations, namely: Sesame, the Data Acquisition Prioritisation System (DAPS), the Functional and Structural Space (FSS) tool and the Target PDB Monitor (TMP). The ability to generate new lists, with new ranking orders for the selected targets can be used by researchers within the consortium to help define their own working targets. DAPS, for example, enables the prioritisation of crystallised proteins according to a variety of factors ranging from the protein’s structural novelty to its length, whereas FSS can be used to monitor the putative functional and structural coverage that will be conferred by the selected targets.

The resources described above were developed to support specific structural genomics consortia. Although some allow a certain amount of reprioritisation, the lists of targets are essentially preselected by the consortia. Such resources do not allow external users to generate their own list of targets from raw genomic or proteomic data. Structural biology groups wishing to do so can use a number of genomic annotation resources, which were not specifically built to support structural genomics projects, but do provide appropriate information to aid in the selection and prioritisation of targets, namely: the Protein Extraction, Description and Analysis Tool (PEDANT), the Genomes TO Protein structures and functions (GTOP) system, GQServe and GeneCensus. Each of these automatic resources contains exhaustive annotations of gene and protein sequences for a large number of genomes, including some of the structural, functional and property information for each protein that is required during the selection procedure. Indeed, in a recent study conducted by Frishman, a large-scale target selection experiment using a novel clustering methodology (STRUcture DEtermination Logic or STRUDEL) was achieved using the genome analysis data available within PEDANT for 32 prokaryotic organisms.

## DEVELOPING A TARGET SELECTION RESOURCE

The authors are developing an informatics resource capable of performing target selection through the implementation of the methodologies and protocols outlined in previous sections and represented schematically in Figure 1. Our objective is to establish a system that enables structural biologists to select targets from their genomic sequence of interest according to their own research needs.

The resource is a fully automated system, the structure of which is depicted in Figure 2. It involves the coordination of five distinct areas:

- user interface;
- pre-processing and evaluation of data;
- sequence-based calculations;
- post-processing of data; and
- data storage.

The web-based interface allows end-users to interact with the resource by inserting and editing genomic data, as well as iteratively analysing the resulting...
calculations by browsing, searching or selecting particular proteins or protein characteristics. The interface is implemented through the Perl programming language, utilising the Common Gateway Interface (CGI) specification to generate dynamic Web content. The number of server runs is minimised through the use of JavaScript error checking functions wherever viable.

Users can insert sequence data corresponding to the coding regions of a whole genome, an entire proteome or even every protein sequence encoded by a particular genomic subset (such as a chromosome) (the interface for these features is shown in Figures 4a and 4b). The resource uses a variety of simple scripts, implemented in the Perl programming language, to ensure the
Sequences and their annotations are stored in the resource’s underlying database.

Correct pre-processing and validation of the input data.

The sequence annotation tools are then used to derive new information about the input. The calculations are incorporated into the resource’s procedures via a wrapper script. The wrapper’s functions are: to coordinate the use of the selected external programs (as well as the parsing scripts required to format their input and output) and to populate the resource’s underlying database.

The implementation of a relational database to support such data avoids problems of organisation, efficiency, concurrency and reliability. The database is set up according to the data model shown in Figure 3. This conceptual

Figure 3: Data model. Entities are shown as rectangles, relationships as diamonds. The dashed lines represent optional relationships whereas solid ones correspond to obligatory ones. In quantitative terms, three different types of relationships can be established: the one-to-one represented by a straight line, the many-to-many depicted with a circle at both ends, and the one-to-many shown as a combination of the other two notations.
The interface allows users to browse and search results, as well as generate lists of targets.

For 39.7 per cent of *Plasmodium falciparum* proteins no structural or functional assignments could be established.

Application to the malaria genome

An example of what can be achieved with the Target Selection Resource is provided by work in our laboratory on the *Plasmodium falciparum* proteome (the principal malaria causing organism). The entire genome was loaded into the resource and initial calculations revealed that for 39.7 per cent of proteins no structural or functional assignments could be established (see Figure 4c). Targeting such proteins could be a rewarding strategy both from a ‘structural genomics by structure’ and a ‘structural genomics by function’ point of view.

*Plasmodium falciparum* supports a peculiar genome, with an extremely high A + T content (~80 per cent overall, with ~75 per cent in coding regions and ~90 per cent in introns and intergenic regions) and an uncommonly biased composition of dinucleotides. Its proteome appears to be somewhat unusual too. A characteristic of a large number of predicted malaria proteins is the presence of long stretches of biased amino acid composition or low-complexity regions (see, for example, White *et al.* or Pizzi and Frontali for a comprehensive study). These large tracts (>30 amino acids) are often inserted directly into globular domains, which are otherwise conserved among a variety of organisms. Although experimental studies have shown that it is likely that most, if not all, of these regions are expressed *in vivo*, their function and mechanism of evolution are not known.

The resource has allowed researchers in our laboratory to generate a list of targets by refining the selection choices to consider the GC content of the encoding gene (a GC content that is very divergent from the one used by the expression system will lead to expression problems) and whether it contains any such insert regions (non-globular regions are unstructured and thus not amenable for structural studies). Each *Plasmodium* transcript (of a total of 5,334 gene products) was filtered and prioritised according to the following characteristics:

- At the gene level: single exon gene and 30–70 per cent GC content.
Figure 4: Example views of the resource’s interface. (a) The ‘LOAD’ web page with the ‘Create’ function activated. The palette area (left-hand side) within this component summarises the genomes and data sets for which data have been calculated. The work area (right-hand site) shows an input form that allows the user to create new genome entries within the resource, so that sequence data can be added and the calculations initiated. (b) The ‘LOAD’ web page with the ‘Add’ function activated. The work area shows an input form that allows the user to add new genomic subsets to a genome entry already created in the resource. (c) The ‘By characteristic’ view of the ‘Structural and functional assignments’ characteristic of the ‘Browse’ function within the ‘VIEW’ component of the resource. This page shows the distribution of structural assignments for all the proteins in a genomic subset (Plasmodium falciparum’s genome in this case), as well as the breakdown of those proteins making up each of the structural annotation classes into their functional categories.
At the protein level: no transmembrane regions, no long non-globular hydrophilic regions and novel fold. This selection procedure, based on choosing those *Plasmodium* proteins that are most suited to experimental studies (namely: expression and crystallisation) and most likely to assume a novel fold, generated a list of 62 protein targets for structural studies (see the web site\textsuperscript{152} for further information).

PROSPECTS FOR TARGET SELECTION

The experimental structure determination pipeline has numerous bottlenecks, which account for the patterns of discovery reported by the ongoing structural genomics projects. Target selection reports a large number of candidate proteins, which are then dramatically reduced during the cloning (50–60 per cent of the selected targets), expression (<80 per cent of the cloned targets), purification (50–60 per cent of the expressed targets), diffraction and structure solving processes (<10 per cent of the purified targets) (see, for example, Chance \textit{et al.}\textsuperscript{133}).

The various structural genomics endeavours are testing and introducing methods to improve the success rate of each of these steps. The biophysical characterisation of each expressed target, for example, is being used to predict the likelihood of crystallisation.\textsuperscript{133} A preferable approach, however, would be to predict such characteristics during target selection (ie before experimental time is invested on a target) thus helping to reduce the ‘funnelling’ effect of the structure determination process. The inherent large-scale nature of structural genomics projects delivers a wealth of data on the performance of each of the structural determination pipeline experimental procedures. Through mining this abundance of data researchers can increase the accuracy, sensitivity and scope of target selection procedures. Christendat and colleagues, for example, used experimental data obtained through a prototype structural genomics project to
derive solubility and crystallisability decision trees based on protein sequence attributes (such as size, amino acid composition, similarity to other proteins, measures of hydrophobicity and polarity and regions of low sequence complexity). They were able to develop simple sequence-based prediction rules, which can enhance the probability of selecting targets that will be both soluble and amenable to crystallisation. They also report that the reliability of the discrimination achieved through the solubility rules was higher due to the availability of a larger data set, and were able to improve these rules less than a year later by virtue of the growth on the information base.

As structural genomics projects evolve, valuable experimental data will be accumulated, thus presenting researchers with a unique opportunity to establish improved predictive methods for a protein’s chemical and physical behaviour based on its amino acid sequence. It is essential for laboratories producing such data to keep track of both ‘successful’ and ‘unsuccessful’ results, so that these can be fed back into the structural determination pipeline through the improvement of the target selection procedures.

Acknowledgments
We would like to thank Barry Grant for discussions on the resource and for critically reviewing the manuscript, as well as the Malaria team at York for their interest and feedback on the resource. Ana Rodrigues is supported by a grant to the York Structural Biology Centre by Accelrys Inc, San Diego.

References
12. BSGC (URL: http://www.strgen.org/).
13. PRESAGE (URL: http://presage.berkeley.edu/).
14. CESG (URL: http://www.uwstructuralgenomics.org/).
15. Sesame (URL: http://kamba.nmr.fas.wisc.edu/Sesame/).
16. JCSG (URL: http://www.jcsg.org/).
18. DAPS (URL: http://www.jcsg.org/daps/).
19. TPM (URL: http://www1.jcsg.org/tpm/).
20. FSS (URL: http://www1.jcsg.org/fss/).
21. MCSG (URL: http://www.mcsg.anl.gov/).
22. NYSGRC (URL: http://www.nysgrc.org/).
26. NESG (URL: http://www.nesg.org/).
27. ZebaView (URL: http://www-nmr.cabm.rutgers.edu/bioinformatics/ZebaView/).
28. TAP (URL: http://www-nmr.cabm.rutgers.edu/bioinformatics/cogo/).
29. SPINE (URL: http://spine.inbb.yale.edu/spine/sum.php3).
30. SECSG (URL: http://www.secsg.org/).
35. S2F (URL: http://s2f.umbi.umd.edu/).
38. Syrrx (URL: http://www.syrrx.com).
41. OCSP (URL: http://www.uhnres.utoronto.ca/proteomics/).
42. HSQC Catalogue (URL: http://www.uhnres.utoronto.ca/proteomics/nmr/hsqc_catalogue_home.html).
44. Marseilles Structural Genomics Program (URL: http://afibn.cns-nr.fr/stgen/).
46. YSG (URL: http://genomics.eu.org/).
47. OPPF (URL: http://www.OPPF.ox.ac.uk/).
48. NWSGC (URL: http://www.nwsgc.ac.uk/).
49. PSF (URL: http://www.proteinstrukturnfabrik.de/).
50. PSF On-line Progress Report (URL: http://www.proteinstrukturnfabrik.de/public/PSF_PUBLICSTATUS0.html).
51. SPINE (URL: http://www.spine.org/).
52. RIKEN SGI (URL: http://www.rsgl.riken.go.jp/).
56. CC3D (URL: http://redpoll.pharmacy.ualberta.ca/~bahram/CCDB.html).
57. Ernest Laue Group at the University of Cambridge (URL: http://www.bio.cam.ac.uk/~ed11/).
58. Structural Biology NCCR Program (URL: http://www.structuralbiology.unizh.ch/).
59. JBBRC Structural Genomics Group (URL: http://www.aist.go.jp/aist_e/research_units/research_center/birc/birc_main.html).
70. Wood, T. C. and Pearson, W. R. (1999), ‘Evolution of protein sequences and...


92. Critical Assessment of Techniques for Protein Structure Prediction (CASP) (URL: http://predictioncenter.llnl.gov/).


111. KEGG metabolic pathways (URL: http://www.genome.ad.jp/kegg/metabolism.html).


