Making decisions for structural genomics

Ana Rodrigues and Roderick E. Hubbard

Date received: 24th February 2003

Ana Rodrigues
is a member of the York Structural Biology Laboratory. She is a graduate student with training in bioinformatics, and research interests in computational structural biology.

Roderick E. Hubbard
is a member of the York Structural Biology Laboratory. His research interests are in molecular graphics, modelling and analysis of protein–ligand interactions and structure-based drug discovery.

Keywords: *structural genomics, target selection, informatics resource*

Ana Rodrigues,
Structural Biology Laboratory,
Department of Chemistry,
University of York,
Heslington,
York YO10 5YW, UK

Tel: +44 (0)1904 328 279
Fax: +44 (0)1904 328 266
E-mail: rodrigues@ysbl.york.ac.uk

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 1: World-wide projects in structural genomics and their target selection strategies.

Consortium	Focus organisms	Selection criteria	No. of targets	Resources
USA				
NIH Protein Structure Initiative				
Berkeley Structural Genomics Center[12]	*Mycoplasma genitalium* *Mycoplasma pneumoniae*	Structural novelty Functional novelty Prevalence	345	PRESAGE[13]
Center for Eukaryotic Structural Genomics[14]	*Arabidopsis thaliana*	Structural novelty Functional novelty Experimental tractability	1,782	Sesame[15]
Joint Center for Structural Genomics[16]	*Thermotoga maritima* *Caenorhabditis elegans*	Genome coverage Structural novelty Technology development Experimental tractability Signalling proteins	5,380	PSCA[17] DAPS[18] TPM[19] FSS[20]
Midwest Center for Structural Genomics[21]	*Bacillus subtilis* *Thermotoga maritima* *Haemophilus influenza* *Escherichia coli*	Structural novelty Medical importance	2,834	–
New York Structural Genomics Research Consortium[22]	*Homo sapiens* Model organisms	Structural novelty Experimental tractability Functional information on protein's pathway, expression, family and interactions	839	MAGPIE[23] SANDPIPER[23] ModBase[24] IceDB[25]
Northeast Structural Genomics Consortium[26]	*Saccharomyces cerevisiae* *Drosophila melanogaster* *Caenorhabditis elegans*	Structural novelty Prevalence Use prokaryotic homologues	5,391	ZebaView[27] TAP[28] SPINE[29]
Southeast Collaboratory for Structural Genomics[30]	*Homo sapiens* *Caenorhabditis elegans* *Pyrococcus furiosus*	Experimental tractability Multidomain proteins Membrane proteins	3,769	ReportDB[31]
Structural Genomics of Pathogenic Protozoa[32]	*Plasmodium falciparum* *Trypanosoma brucei* *Trypanosoma cruzi* *Leshmania* sp.	Structural novelty Medical importance	274	–
TB Structural Genomics Consortium[33]*	*Mycobacterium tuberculosis*	Essential proteins Mutants affecting host/parasite interaction Structural novelty	1,389	Online progress report[34]
Others				
Structure 2 Function Project[35]	*Haemophilus influenza*	Functional novelty Experimental tractability	331	Online progress report[36]
Structural GenomiX[37]	Bacterial pathogens	Apo- and co-complexes Protein kinases Nuclear hormone receptors Membrane proteins	Not divulged	Proprietary
Syrrx[38]	None	Disease causing proteins	Not divulged	Proprietary
Canada				
Bacterial Structural Genomics Initiative[39]	*Escherichia coli*	Small molecule pathways Functional novelty	475	On-line progress report[40]
Ontario Center for Structural Proteomics[41]	*Methanobacterium thermoautotrophicum* *Thermotoga maritima* *Arabidopsis thaliana* *Escherichia coli* *Saccharomyces cerevisiae*	Structural novelty	2,700	HSQC Catalogue[42]
Montreal Network for Pharmaco-Proteomics & Structural Genomics[43]	Mammalian cell	Endoplasmic reticulum proteins (known and novel identified through mass spectrometry)	50–100	On-line progress report soon to be available[40]
EU				
France				
Marseilles Structural Genomics Program[44]†	*Escherichia coli* *Mycobacterium tuberculosis*	Functional novelty Structural novelty Glycobiology (cutinases, lipases, glycanases and glycosyltransferases) G-protein coupled receptors	313	Online progress report[45]

Table 1: (*continued*)

Consortium	Focus organisms	Selection criteria	No. of targets	Resources
Yeast Structural Genomics[46]†	*Saccharomyces cerevisiae*	Structural novelty Functional novelty Experimental tractability	250	–
UK				
Oxford Protein Production Facility[47]†	*Homo sapiens* *Herpesviridae*	Structural novelty Experimental tractability cDNA availability Growth factors Immunological molecules	128	–
North West Structural Genomics Center[48]	*Mycobacterium tuberculosis*	Surface proteins	40	–
Germany				
Protein Structure Factory[49]	cDNAs available at the Berlin Resource Center	Structural novelty Functional novelty Experimental tractability	1,280	Online progress report[50]
Structural Proteomics IN Europe[51]	*Bacillus anthracis* *Campylobacter jejuni* *Mycobacterium tuberculosis, leprae* and *bovis* *Herpesviridae* *Homo sapiens*	Virulence genes Host/pathogen interaction Disease-related protein families (kinases, proteases, kiesins, nuclear receptors, cell surface molecules)	600 planned	–
Japan				
RIKEN Structural Genomics and Proteomics Initiative[52]	*Arabidopsis thaliana* *Thermus thermophilus* *Pyrococcus horikoshii*	Structural novelty Eukaryotic specific DNA/RNA binding Cell signalling SNP-bearing Disease related proteins	705	Online progress reports[53, 54]

Consortium	Selection strategy
Canada	
Structure/Function Team of Project CyberCell[55]	Focus on *Escherichia coli* aiming at full genome coverage. Their CC3D[56] resource is available on the world-wide web.
UK	
Ernest Laue Group at the University of Cambridge[57]	Focus on chromatin mediated transcriptional repression, cyclin-dependent kinases and small G-proteins involved in cellular control.
Switzerland	
Structural Biology National Center of Competence in Research Program[58]	Focus on membrane proteins and intermolecular interactions in supramolecular assemblies.
Japan	
Biological Information Research Center Structural Genomics Group[59]	Focus on membrane proteins and ligand–protein interactions. Also developing an integrated database system.
Structural Genomics/Proteomics of Rice	Focus on the *Oryza sativa* genome.
Korea	
Structural Proteomics Research Organisation Program	Focus on the *Mycobacterium tuberculosis* and *Helicobacter pylori* genomes to enable drug discovery.
China	
Structural Genomics Effort	Focus on *Homo sapiens* proteins with emphasis on disease associated ones, as well as novel bacterial proteins.

Totals	Consortia	Focus organisms	No. of targets
	29 projects	>28 genomes	>28,875 proteins

*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;[8]

- Obtaining novel structures to increase coverage of protein fold space.[9–11]

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.

The rate of protein structure determination is hindered by a variety of technical challenges

There are many technical challenges for large-scale structure determination (see Heinemann *et al*.[60] for an overall discussion, or the following reviews: Pokala and Handel[61] or Gilbert and Albala[62] on protein production, Hendrickson[63] on X-ray crystallography, Prestegard *et al*.[64] or Al-Hashimi and Patel[65] on nuclear magnetic resonance (NMR) spectroscopy and Baumeister and Steven[66] 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.

It is important to select targets which are tractable and provide the most information return

In this paper, we review target selection, discussing the scientific basis on which it can be performed and suggesting various sequence analysis protocols that may aid its implementation. We also describe a target selection resource developed at York, which employs many of these methods, and provide some preliminary results from our analysis of the malaria genome. Finally, we consider the development of target selection in the context of the emerging structural genomics projects.

CURRENT APPROACHES TO TARGET SELECTION

Our discussion of target selection is broken down into three main sections. First, we review the current understanding of how evolutionary constraints can be used to identify proteins that may adopt similar conformations to known protein structures. For these proteins, modelling approaches may provide sufficient information to understand structure and mechanism. Secondly, we consider how selection strategies depend on the scientific context and aims of the structural project. Finally, we discuss sets of protein characteristics that can be inferred from the sequence and employed in the identification of proteins, which may pose problems during the various stages of 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.[67,68] Where proteins are related by divergent evolution, they share a common ancestor and are said to be homologous, ie 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.[68]

Sequence similarity can be a reliable indicator of protein homology and, hence, structure similarity.[69,70] 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[71]). 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[72] and FastA[73] can be used to rapidly identify homologues with known protein structures and a homology model can be constructed using programs such as MODELLER.[74–76] 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.[77] 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.[78,79] Most structural genomics projects will therefore lower the priority on the experimental structure determination of homologous proteins, unless a detailed study is required.

Low priority is assigned to proteins for which a comparative structure model can be computed

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, ie 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.[80–84] There has been considerable effort over the past 10 years not only to analyse and categorise the fold space[85–89] 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.[90,91] 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[92] 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.

Strategies fall into two broad categories: structural genomics by structure and structural genomics by function

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.[11]

In 'structural genomics by structure', the main aim is the determination of novel protein folds

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,[93] such as PSI-BLAST,[72] hidden Markov model (HMM)-based[94] and profile-based protocols (see for example: Schaffer *et al.*[95] or Yona and Levitt[96]). Implementations of such algorithms, purposefully tuned for fold prediction include SUPERFAMILY[97] and the PDB-Intermediate Sequence Library (PDB-ISL).[98] 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.

Favourite targets for such projects include ORFan proteins and those that will increase fold space coverage

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, ie 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,[99] ProtoMap,[100,101] GeneRAGE[102] and SUPFAM.[103] Such resources can also be employed to, for example, identify those targets whose structure determination will provide structural information for the most proteins (ie 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[104,105] and Pred2ary[106]) to query against a database of known topologies (such as those provided by the TOPS server[107]).

In 'structural genomics by function', the main aim is the determination of structures that elucidate particular functions and/or metabolic pathways

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.

Favourite targets for such projects include prevalent proteins, and those that are involved in pathogenicity or are of biomedical importance

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[108]), using sequence search similarity programs such as BLAST and FastA. More sensitive software tools, such as PSI-BLAST, HMMER[109] and IMPALA[95] (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[110] 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[111,112] and the On-line Mendelian Inheritance in Man (OMIM) database.[113] 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.

The computational identification of problematic proteins or segments allows the filtering of experimentally 'difficult' proteins

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).[114–117] *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.*[118] 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[119] and TMHMM[120,121] 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 hydropathic and physicochemical constraints imposed by the lipid layer, though the available methods are generally more successful in recognising helical membrane segments than strand elements.

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

Regions of a protein with little residue variation are traditionally associated with unstructured regions.[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).[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[118] or CAST.[125] The program COILS2[126,127] 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.[128]

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

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),[129] for example. Others can be estimated using implementations of statistical models derived from empirical data (eg the revised Wilkinson–Harrison statistical solubility model[130]).

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

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 hydropathy, 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,[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[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),[133] ZebaView,[27] the Structural Proteomics In the North East (SPINE) system[134] and ReportDB.[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

Annotation of Genomic Entities (PRESAGE),[135] the Protein Sequence Comparative Analysis (PSCA) system[17] and the Target Analysis and Prioritisation (TAP) database[28] 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 (eg 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.

The tools employed in the selection targets are largely kept within each project's realm

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,[136] the Data Acquisition Prioritisation System (DAPS),[18] the Functional and Structural Space (FSS) tool[20] and the Target PDB Monitor (TMP).[19] 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 FFS can be used to monitor the putative functional and structural coverage that will be conferred by the selected targets.

General purpose genome annotation resources can be employed in the selection of targets for structural genomics

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),[137,138] the Genomes TO Protein structures and functions (GTOP) system,[139] GQServe[140,141] and GeneCensus.[142] 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.[143]

The authors are developing a system that enables structural biologists to select targets according to their research needs

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

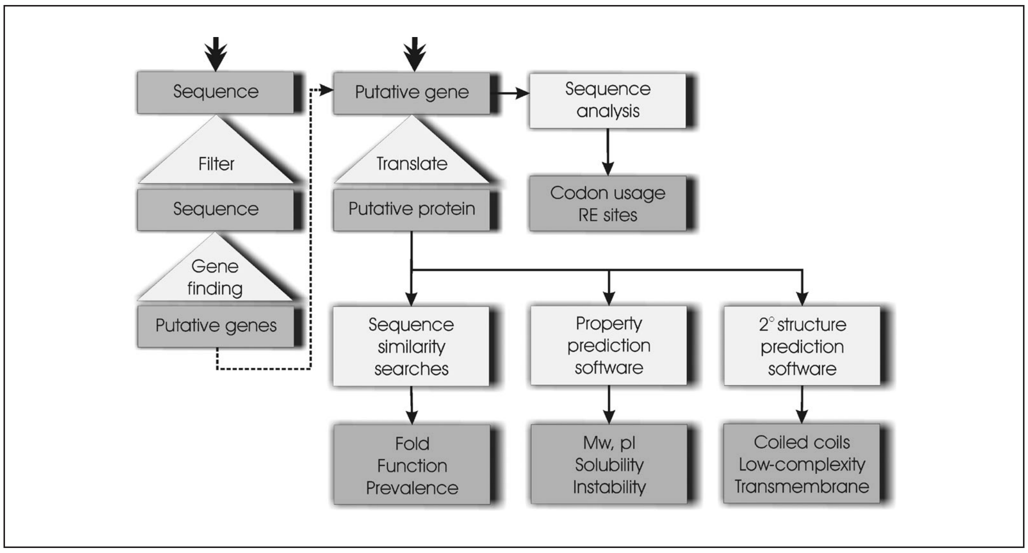


Figure 1: Schematic depiction of the bioinformatics methodologies that can be applied to a nucleotide sequence of interest during the target selection procedure

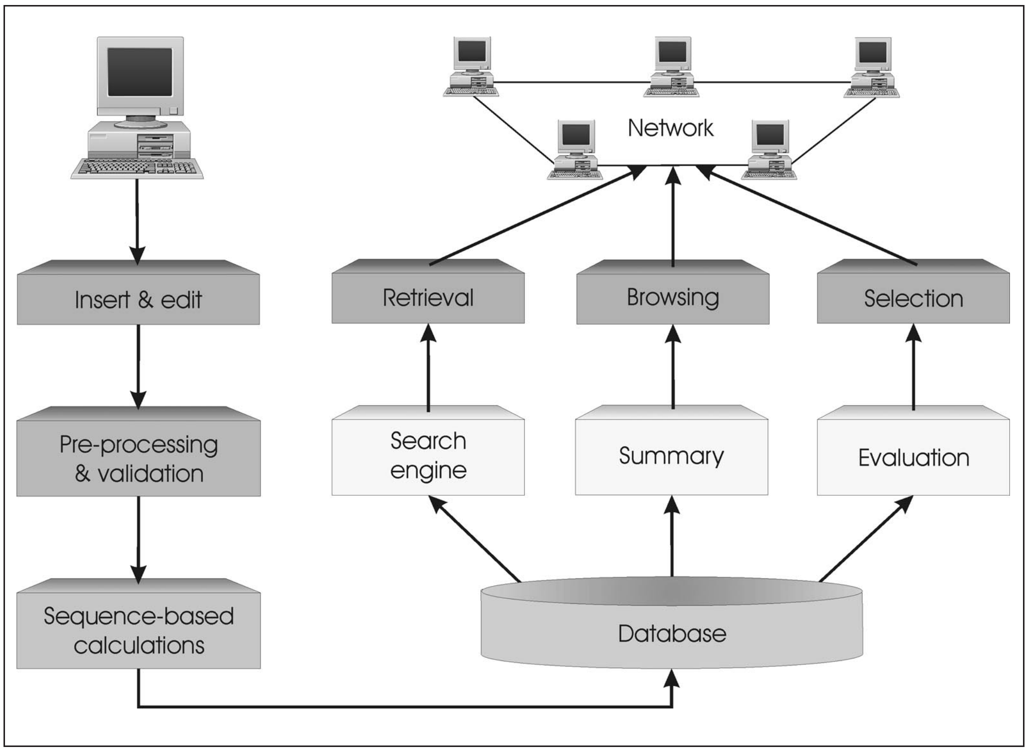


Figure 2: The resource's flow and structure. The first layer of components constitutes the resource's interface. These interact with the core of the resource through a series of pre-processing, validation and post-processing tools. The latter are shown in white

The web-interface allows users to input sequence data for processing

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

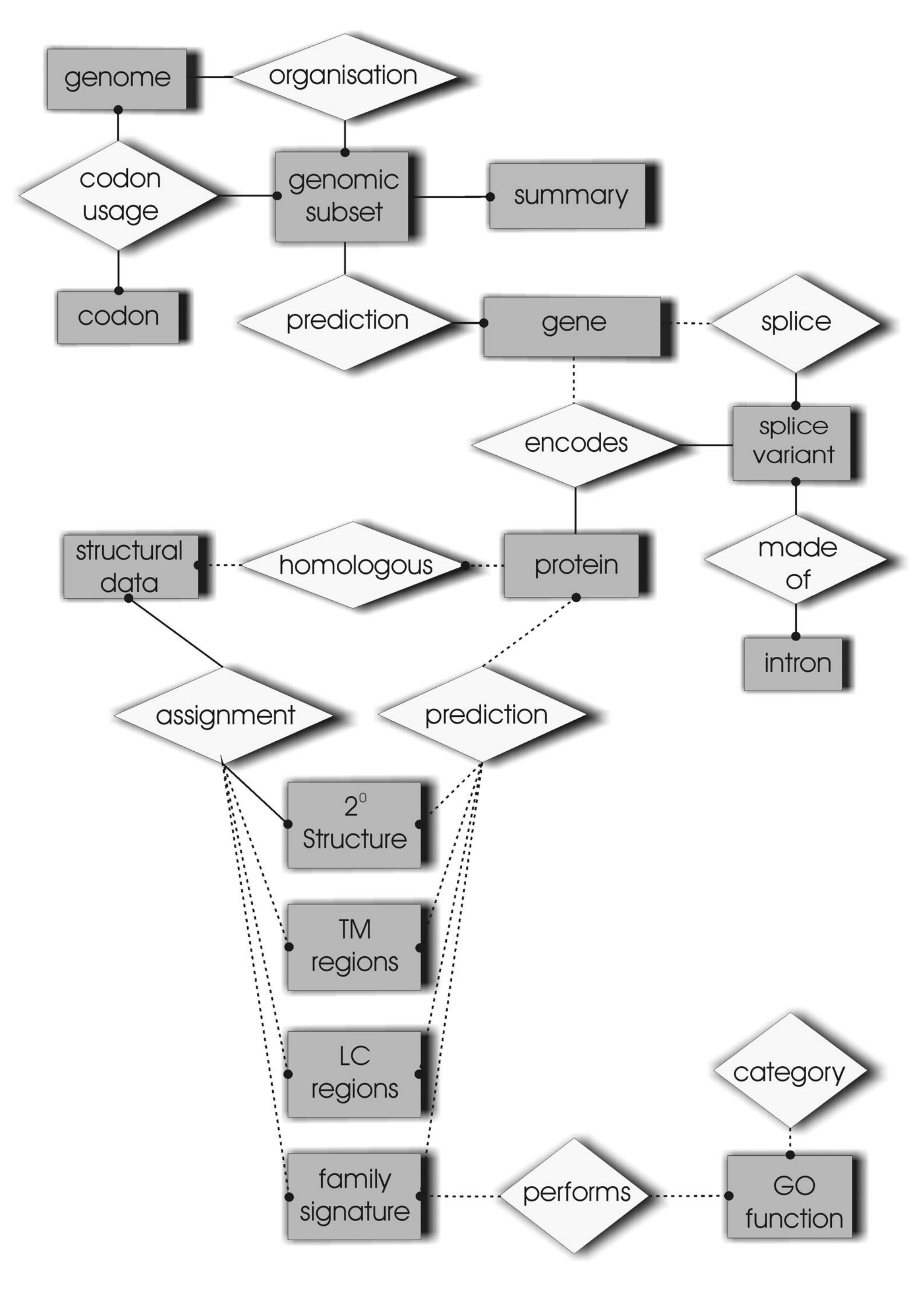


Figure 3: Data model. Entities are shown as rectangles, relationships as diamonds. The dashed lines represent optional relationships whereas solid ones correspond to obligatory ones. In quantitative terms, three different types of relationships can be established: the one-to-one represented by a straight 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

schema depicts the real world entities and relationships that are captured by the database. It was translated into a relational schema using the MySQL Database Management System where relational tables capture the data for every entity and relationship. Several layers of data-warehousing were introduced into the previously normalised schema in order to improve the performance of the user interface.

The interface allows users to browse and search results, as well as generate lists of targets

Data post-processing is a general term encompassing all the methodologies required for the analysis of stored data. These comprise the facilities that support the components and functions that the end-user accesses through the interface, including: a data summarising feature (to enable data browsing), a search engine (to enable data retrieval) and a module capable of reasoning the results of sequence annotations (to enable data selection). All post-processing features are also implemented using the Perl programming language. The DBI (Database Interface) and DBD-Mysql (Database Driver for the MySQL DBMS) modules provide interaction with the database. Some post-processing tools generate graphical depictions and summaries of the results (structural assignments and their functional distribution, for example, as shown in Figure 4c). These pictures are automatically generated through the use of the graphics package gd interfaced by the GD and GD::Graph Perl modules, which were modified for that purpose. Image maps for each of the pictures are also generated, via the GD::Graph::Map module. Pictures and corresponding image maps are stored in the database and are retrieved by the interface when the user activates a particular function (browsing by structural and functional assignments, in this case).

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

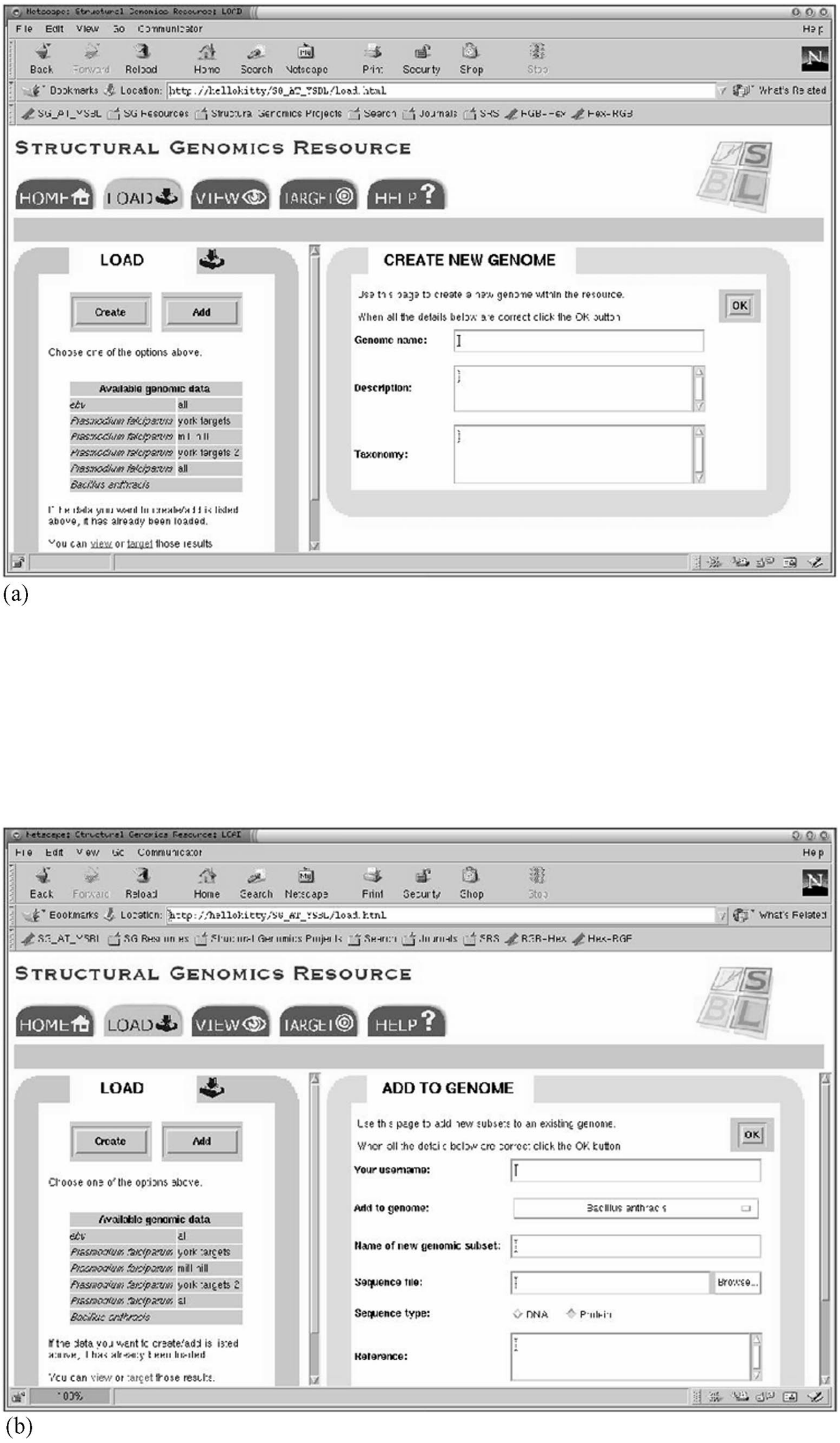
Application to the malaria genome

An example of what can be achieved with the Target Selection Resource is provided by work in our laboratory on the *Plasmodium falciparum*'s proteome (the principal malaria causing organism). The entire genome[144–149] 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.[144–149] 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*.[150] or Pizzi and Frontali[151] 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*,[144] 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.


Netscape: Structural Genomics Resource: LOAD
File Edit View Go Communicator Help
Back Forward Reload Home Search Netscape Print Security Shop Stop
Bookmarks Location: http://hellokitty/SG_AT_YSBL/load.html What's Related
SG_AT_YSBL SG Resources Structural Genomics Projects Search Journals SRS RGB-Hex Hex-RGB
STRUCTURAL GENOMICS RESOURCE
HOME LOAD VIEW TARGET HELP
YSBL
LOAD
Create Add
CREATE NEW GENOME
OK

Choose one of the options above.

Available genomic data	
ebv	all
Plasmodium falciparum	york targets
Plasmodium falciparum	mill hill
Plasmodium falciparum	york targets 2
Plasmodium falciparum	all
Bacillus anthracis	

If the data you want to create/add is listed above, it has already been loaded.

You can view or target those results

Use this page to create a new genome within the resource.

When all the details below are correct click the OK button

Genome name:

Description:

Taxonomy:

(a)

Netscape: Structural Genomics Resource: LOAD
File Edit View Go Communicator Help
Back Forward Reload Home Search Netscape Print Security Shop Stop
Bookmarks Location: http://hellokitty/SG_AT_YSBL/load.html What's Related
SG_AT_YSBL SG Resources Structural Genomics Projects Search Journals SRS RGB-Hex Hex-RGB
STRUCTURAL GENOMICS RESOURCE
HOME LOAD VIEW TARGET HELP
YSBL
LOAD
Create Add
ADD TO GENOME
OK
Browse...
100%

Choose one of the options above.

Available genomic data	
ebv	all
Plasmodium falciparum	york targets
Plasmodium falciparum	mill hill
Plasmodium falciparum	york targets 2
Plasmodium falciparum	all
Bacillus anthracis	

If the data you want to create/add is listed above, it has already been loaded

You can view or target those results.

Use this page to add new subsets to an existing genome.

When all the details below are correct click the OK button

Your username:

Add to genome: Bacillus anthracis

Name of new genomic subset:

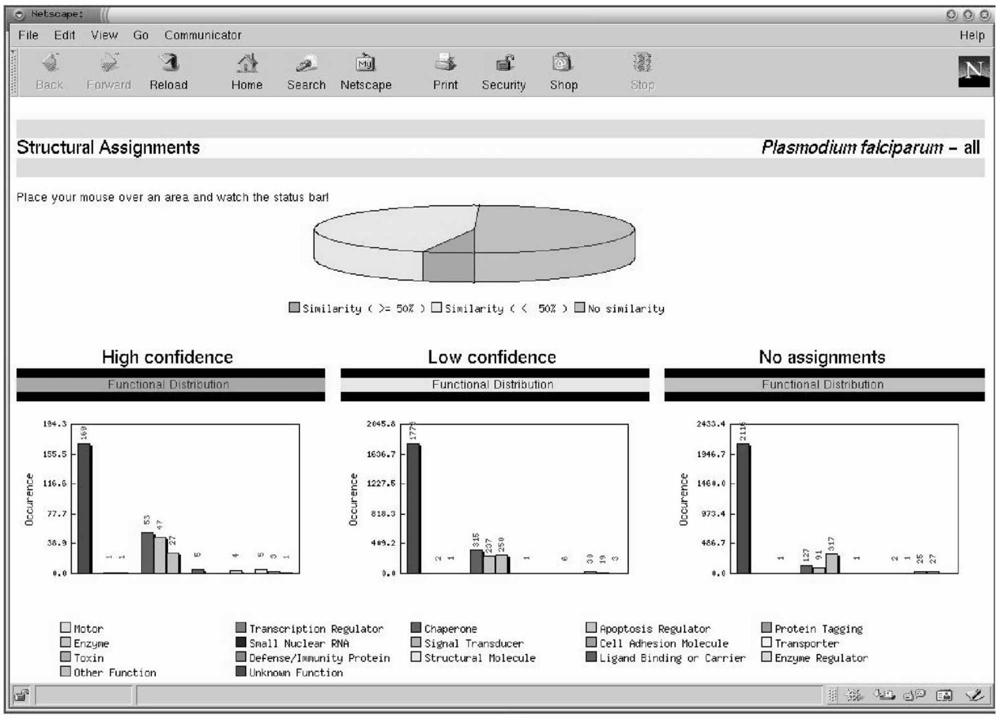
Sequence file:

Sequence type: ◇ DNA ◇ Protein

Reference:

(b)

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



(c)

Figure 4: (*continued*)

- At the protein level: no transmembrane regions, no long non-globular hydrophilic regions and novel fold.

Selection of those *Plasmodium* proteins that are most suited to experimental studies and most likely to have a novel fold delivered a list of 62 targets

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[152] for further information).

PROSPECTS FOR TARGET SELECTION

Target selection procedures will be improved through examination of the results obtained by structural genomics projects at each step of the structure determination pipeline

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 *et al*.[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.[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).[153] 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.[134]

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

1. NCBI Genomes (URL: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Genome).
2. Woychik, R. P., Klebig, M. L., Justice, M. J. *et al.* (1998), 'Functional genomics in the post-genome era', *Mutat. Res.*, Vol. 400(1–2), pp. 3–14.
3. Delneri, D., Brancia, F. L. and Oliver, S. G. (2001), 'Towards a truly integrative biology through the functional genomics of yeast', *Curr. Opin. Biotechnol.*, Vol. 12(1), pp. 87–91.
4. Kim, S. H. (2000), 'Structural genomics of microbes: An objective', *Curr. Opin. Struct. Biol.*, Vol. 10(3), pp. 380–383.
5. Erlandsen, H., Abola, E. E. and Stevens, R. C. (2000), 'Combining structural genomics and enzymology: completing the picture in metabolic pathways and enzyme active sites', *Curr. Opin. Struct. Biol.*, Vol. 10(6), pp. 719–730.
6. Bonanno, J. B., Edo, C., Eswar, N. *et al.* (2001), 'Structural genomics of enzymes involved in sterol/isoprenoid biosynthesis', *Proc. Natl Acad. Sci. USA*, Vol. 98(23), pp. 12896–12901.
7. Burley, S. K. and Bonanno, J. B. (2002), 'Structural genomics of proteins from conserved biochemical pathways and processes', *Curr. Opin. Struct. Biol.*, Vol. 12(3), pp. 383–391.
8. Sunyaev, S., Lathe, W. 3rd and Bork, P. (2001), 'Integration of genome data and protein structures: Prediction of protein folds, protein interactions and "molecular phenotypes" of single nucleotide polymorphisms', *Curr. Opin. Struct. Biol.*, Vol. 11(1), pp. 125–130.
9. Sali, A. (1998), '100,000 protein structures for the biologist', *Nat. Struct. Biol.*, Vol. 5(12), pp. 1029–1032.
10. Cort, J. R., Koonin, E. V., Bash, P. A. and Kennedy, M. A. (1999), 'A phylogenetic approach to target selection for structural genomics: solution structure of YciH', *Nucleic Acids Res.*, Vol. 27(20), pp. 4018–4027.
11. Fischer, D. (1999), 'Rational structural genomics: affirmative action for ORFans and the growth in our structural knowledge', *Protein Eng.*, Vol. 12(12), pp. 1029–1030.
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.nmrfam.wisc.edu/Sesame/).
16. JCSG (URL: http://www.jcsg.org/).
17. PSCA (URL: http://www1.jcsg.org/psat/help/document.html).
18. DAPS (URL: http://www.jcsg.org/dasp/).
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/).
23. MAGPIE/SANDPIPER (URL: http://genomes.rockefeller.edu/research.shtml).
24. ModBase (URL: http://pipe.rockefeller.edu/modbase/index.shtml).

25. IceDB (URL: http://www.nysgrc.org/nysgrc/icedb.html).

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/cogs/).

29. SPINE (URL: http://spine.mbb.yale.edu/spine/sum.php3).

30. SECSG (URL: http://www.secsg.org/).

31. ReportDB (URL: http://www.secsg.org/cgi-bin/report.pl).

32. SGPP (URL: http://depts.washington.edu/sgpp/main.html).

33. TBX (URL: http://www.doe-mbi.ucla.edu/TB/index.php).

34. TB On-line Progress Report (URL: http://www.doe-mbi.ucla.edu/TB/notebook_status.php).

35. S2F (URL: http://s2f.umbi.umd.edu/).

36. S2F On-line Progress Report (URL: http://s2f.umbi.umd.edu/cgi-bin/status.cgi).

37. Structural GenomicX (URL: http://www.stromix.com).

38. Syrrx (URL: http://www.syrrx.com).

39. BSGI (URL: http://euler.bri.nrc.ca/brimsg/bsgi.html).

40. BSGI On-line Progress Report (URL: http://euler.bri.nrc.ca/brimsg/target_new.html?opt=list/).

41. OCSP (URL: http://www.uhnres.utoronto.ca/proteomics/).

42. HSQC Catalogue (URL: http://www.uhnres.utoronto.ca/proteomics/nmr/hsqc_catalogue_home.html).

43. Montreal Network for Pharmaco-Genomics and Structural Genomics (URL: http://www.genomequebec.com/eng/recherches/projets.htm).

44. Marseilles Structural Genomics Program (URL: http://afmb.cnrs-mrs.fr/stgen/).

45. Marseilles On-line Progress Report (URL: http://afmb.cnrs-mrs.fr/stgen/tglist.html).

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.proteinstrukturfabrik.de/).

50. PSF On-line Progress Report (URL: http://www.proteinstrukturfabrik.de/public/PSF_PUBLICSTATUS0.html).

51. SPINE (URL: http://www.spine.org/).

52. RIKEN SGI (URL: http://www.rsgi.riken.go.jp/).

53. Riken SGPI On-line Progress Report (URL: http://www.htpf.harima.riken.go.jp).

54. Riken SGPI On-line Progress Report (URL: http://protein.gsc.riken.go.jp).

55. Structure/Function Team, Project CC (URL: http://www.projectcybercell.com).

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/~edl1/).

58. Structural Biology NCCR Program (URL: http://www.structuralbiology.unizh.ch/).

59. JBIRC Structural Genomics Group (URL: http://www.aist.go.jp/aist_e/ressearch_units/research_center/birc/birc_main.html).

60. Heinemann, U., Illing, G. and Oschkinat, H. (2001), 'High-throughput three-dimensional protein structure determination', *Curr. Opin. Biotechnol.*, Vol. 12(4), pp. 348–354.

61. Pokala, N. and Handel, T. M. (2001), 'Review: protein design – where we were, where we are, where we're going', *J. Struct. Biol.*, Vol. 134(2–3), pp. 269–281.

62. Gilbert, M. and Albala, J. S. (2002), 'Accelerating code to function: Sizing up the protein production line', *Curr. Opin. Chem. Biol.*, Vol. 6(1), pp. 102–105.

63. Hendrickson, W. A. (2000), 'Synchrotron crystallography', *Trends Biochem. Sci.*, Vol. 25(12), pp. 637–643.

64. Prestegard, J. H., Valafar, H., Glushka, J. and Tian, F. (2001), 'Nuclear magnetic resonance in the era of structural genomics', *Biochemistry*, Vol. 40(30), pp. 8677–8685.

65. Al-Hashimi, H. M. and Patel, D. J. (2002), 'Residual dipolar couplings: Synergy between NMR and structural genomics', *J. Biomol. NMR*, Vol. 22(1), pp. 1–8.

66. Baumeister, W. and Steven, A. C. (2000), 'Macromolecular electron microscopy in the era of structural genomics', *Trends Biochem. Sci.*, Vol. 25(12), pp. 624–631.

67. Jacob, F. (1977), 'Evolution and tinkering', *Science*, Vol. 196(4295), pp. 1161–1166.

68. Ridley, M. (1996), 'Evolution', 2nd edn, Blackwell Science, USA.

69. Chothia, C. and Lesk, A. M. (1986), 'The relation between the divergence of sequence and structure in proteins', *Embo J.*, Vol. 5(4), pp. 823–826.

70. Wood, T. C. and Pearson, W. R. (1999), 'Evolution of protein sequences and

structures', *J. Mol. Biol.*, Vol. 291(4), pp. 977–995.

71. Swindells, M. B. and Thornton, J. M. (1991), 'Modelling by homology', *Curr. Opin. Struct. Biol.*, Vol. 1, pp. 219–223.

72. Altschul, S. F., Madden, T. L., Schaffer, A. A. *et al.* (1997), 'Gapped BLAST and PSI-BLAST: A new generation of protein database search programs', *Nucleic Acids Res.*, Vol. 25(17), pp. 3389–3402.

73. Pearson, W. R. and Lipman, D. J. (1988), 'Improved tools for biological sequence comparison', *Proc. Natl Acad. Sci. USA*, Vol. 85(8), pp. 2444–2448.

74. Sali, A. and Blundell, T. L. (1993), 'Comparative protein modelling by satisfaction of spatial restraints', *J. Mol. Biol.*, Vol. 234(3), pp. 779–815.

75. Fiser, A., Do, R. K. and Sali, A. (2000), 'Modeling of loops in protein structures', *Protein Sci.*, Vol. 9(9), pp. 1753–1773.

76. Marti-Renom, M. A., Stuart, A. C., Fiser, A. *et al.* (2000), 'Comparative protein structure modeling of genes and genomes', *Annu. Rev. Biophys. Biomol. Struct.*, Vol. 29, pp. 291–325.

77. Xu, D. and Xu, Y. (2000), 'Protein tertiary structure prediction', *Curr. Protoc. Prot. Sci.*, unit 2.7, pp. 2.7.1–2.7.17.

78. Oldfield, T. J., Murray-Rust, P. and Hubbard, R. E. (1993), 'Model structures and action of interleukin 1 and its antagonist', *Protein Eng.*, Vol. 6(8), pp. 865–871.

79. Matsumoto, R., Sali, A., Ghildyal, N. *et al.* (1995), 'Packaging of proteases and proteoglycans in the granules of mast cells and other hematopoietic cells. A cluster of histidines on mouse mast cell protease 7 regulates its binding to heparin serglycin proteoglycans', *J. Biol. Chem.*, Vol. 270(33), pp. 19524–19531.

80. Chothia, C. (1992), 'Proteins. One thousand families for the molecular biologist', *Nature*, Vol. 357(6379), pp. 543–544.

81. Wang, Z. X. (1998), 'A re-estimation for the total numbers of protein folds and superfamilies', *Protein Eng.*, Vol. 11(8), pp. 621–626.

82. Zhang, C. and DeLisi, C. (1998), 'Estimating the number of protein folds', *J. Mol. Biol.*, Vol. 284(5), pp. 1301–1305.

83. Govindarajan, S., Recabarren, R. and Goldstein, R. A. (1999), 'Estimating the total number of protein folds', *Proteins*, Vol. 35(4), pp. 408–414.

84. Wolf, Y. I., Grishin, N. V. and Koonin, E. V. (2000), 'Estimating the number of protein folds and families from complete genome data', *J. Mol. Biol.*, Vol. 299(4), pp. 897–905.

85. Murzin, A. G., Brenner, S. E., Hubbard, T. and Chothia, C. (1995), 'SCOP: A structural classification of proteins database for the investigation of sequences and structures', *J. Mol. Biol.*, Vol. 247(4), pp. 536–540.

86. Holm, L. and Sander, C. (1996), 'Mapping the protein universe', *Science*, Vol. 273(5275), pp. 595–603.

87. Orengo, C. A., Michie, A. D., Jones, S. *et al.* (1997), 'CATH – a hierarchic classification of protein domain structures', *Structure*, Vol. 5(8), pp. 1093–1108.

88. Zhang, C. and DeLisi, C. (2001), 'Protein folds: Molecular systematics in three dimensions', *Cell. Mol. Life Sci.*, Vol. 58(1), pp. 72–79.

89. Rost, B. (2002), 'Did evolution leap to create the protein universe?', *Curr. Opin. Struct. Biol.*, Vol. 12(3), pp. 409–416.

90. Jones, D. T. and Thornton, J. M. (1996), 'Potential energy functions for threading', *Curr. Opin. Struct. Biol.*, Vol. 6(2), pp. 210–216.

91. Smith, T. F., Lo Conte, L., Bienkowska, J. *et al.* (1997), 'Current limitations to protein threading approaches', *J. Comput. Biol.*, Vol. 4(3), pp. 217–225.

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

93. Rost, B. (1999), 'Twilight zone of protein sequence alignments', *Protein Eng.*, Vol. 12(2), pp. 85–94.

94. Eddy, S. R. (1998), 'Profile hidden Markov models', *Bioinformatics*, Vol. 14(9), pp. 755–763.

95. Schaffer, A. A., Wolf, Y. I., Ponting, C. P. *et al.* (1999), 'IMPALA: Matching a protein sequence against a collection of PSI-BLAST-constructed position-specific score matrices', *Bioinformatics*, Vol. 15(12), pp. 1000–1011.

96. Yona, G. and Levitt, M. (2002), 'Within the twilight zone: A sensitive profile-profile comparison tool based on information theory', *J. Mol. Biol.*, Vol. 315(5), pp. 1257–1275.

97. Gough, J., Karplus, K., Hughey, R. and Chothia, C. (2001), 'Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure', *J. Mol. Biol.*, Vol. 313(4), pp. 903–919.

98. Teichmann, S. A., Chothia, C., Church, G. M. and Park, J. (2000), 'Fast assignment of protein structures to sequences using the intermediate sequence library PDB-ISL', *Bioinformatics*, Vol. 16(2), pp. 117–124.

99. Portugaly, E., Kifer, I. and Linial, M. (2002), 'Selecting targets for structural determination

by navigating in a graph of protein families', *Bioinformatics*, Vol. 18(7), pp. 899–907.

100. Yona, G., Linial, N. and Linial, M. (2000), 'ProtoMap: Automatic classification of protein sequences and hierarchy of protein families', *Nucleic Acids Res.*, Vol. 28(1), pp. 49–55.

101. Yona, G., Linial, N. and Linial, M. (1999), 'ProtoMap: Automatic classification of protein sequences, a hierarchy of protein families, and local maps of the protein space', *Proteins*, Vol. 37(3), pp. 360–378.

102. Enright, A. J. and Ouzounis, C. A. (2000), 'GeneRAGE: A robust algorithm for sequence clustering and domain detection', *Bioinformatics*, Vol. 16(5), pp. 451–457.

103. Pandit, S. B., Gosar, D., Abhiman, S. *et al.* (2002), 'SUPFAM – a database of potential protein superfamily relationships derived by comparing sequence-based and structure-based families: Implications for structural genomics and function annotation in genomes', *Nucleic Acids Res.*, Vol. 30(1), pp. 289–293.

104. Rost, B. and Sander, C. (1993), 'Prediction of protein secondary structure at better than 70% accuracy', *J. Mol. Biol.*, Vol. 232(2), pp. 584–599.

105. Rost, B. and Sander, C. (1994), 'Combining evolutionary information and neural networks to predict protein secondary structure', *Proteins*, Vol. 19(1), pp. 55–72.

106. Chandonia, J. M. and Karplus, M. (1999), 'New methods for accurate prediction of protein secondary structure', *Proteins*, Vol. 35(3), pp. 293–306.

107. Gilbert, D., Westhead, D., Viksna, J. and Thornton, J. (2001), 'A computer system to perform structure comparison using TOPS representations of protein structure', *Comput. Chem.*, Vol. 26(1), pp. 23–30.

108. Boeckmann, B., Bairoch, A., Apweiler, R. *et al.* (2003), 'The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003', *Nucleic Acids Res.*, Vol. 31(1), pp. 365–370.

109. Eddy, S. R. (1996), 'Hidden Markov models', *Curr. Opin. Struct. Biol.*, Vol. 6(3), pp. 361–365.

110. Apweiler, R., Attwood, T. K., Bairoch, A. *et al.* (2001), 'The InterPro database, an integrated documentation resource for protein families, domains and functional sites', *Nucleic Acids Res.*, Vol. 29(1), pp. 37–40.

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

112. Kanehisa, M., Goto, S., Kawashima, S. and Nakaya, A. (2002), 'The KEGG databases at GenomeNet', *Nucleic Acids Res.*, Vol. 30(1), pp. 42–46.

113. Online Mendelian Inheritance in Man, OMIM $^{(TM)}$ (URL: http://www.ncbi.nlm.nih.gov/omim/).

114. Abola, E., Kuhn, P., Earnest, T. and Stevens, R. C. (2000), 'Automation of X-ray crystallography', *Nat. Struct. Biol.*, Vol. 7(Suppl), pp. 973–977.

115. Edwards, A. M., Arrowsmith, C. H., Christendat, D. *et al.* (2000), 'Protein production: Feeding the crystallographers and NMR spectroscopists', *Nat. Struct. Biol.*, Vol. 7(Suppl), pp. 970–972.

116. Lamzin, V. S. and Perrakis, A. (2000), 'Current state of automated crystallographic data analysis', *Nat. Struct. Biol.*, Vol. 7(Suppl), pp. 978–981.

117. Montelione, G. T., Zheng, D., Huang, Y. J. *et al.* (2000), 'Protein NMR spectroscopy in structural genomics', *Nat. Struct. Biol.*, Vol. 7(Suppl), pp. 982–985.

118. Creuzet, F., McDermott, A., Gebhard, R. *et al.* (1991), 'Determination of membrane protein structure by rotational resonance NMR: Bacteriorhodopsin', *Science*, Vol. 251(4995), pp. 783–786.

119. Persson, B. and Argos, P. (1994), 'Prediction of transmembrane segments in proteins utilising multiple sequence alignments', *J. Mol. Biol.*, Vol. 237(2), pp. 182–192.

120. Sonnhammer, E. L., von Heijne, G. and Krogh, A. (1998), 'A hidden Markov model for predicting transmembrane helices in protein sequences', in 'Proceedings of 6th International Conference on Intelligent Systems for Molecular Biology', AAAI Press, Menlo Park, CA, pp. 175–182.

121. Krogh, A., Larsson, B., von Heijne, G. and Sonnhammer, E. L. (2001), 'Predicting transmembrane protein topology with a hidden Markov model: Application to complete genomes', *J. Mol. Biol.*, Vol. 305(3), pp. 567–580.

122. Wootton, J. C. and Federhen, S. (1993), 'Statistics of local complexity in amino acid sequences and sequence databases', *Comput. Chem.*, Vol. 17, pp. 149–163.

123. Wootton, J. C. (1994), 'Non-globular domains in protein sequences: Automated segmentation using complexity measures', *Comput. Chem.*, Vol. 18(3), pp. 269–285.

124. Berman, H. M., Westbrook, J., Feng, Z. *et al.* (2000), 'The Protein Data Bank', *Nucleic Acids Res.*, Vol. 28(1), pp. 235–242.

125. Promponas, V. J., Enright, A. J., Tsoka, S. *et al.* (2000), 'CAST: An iterative algorithm for the complexity analysis of sequence tracts. Complexity analysis of sequence tracts', *Bioinformatics*, Vol. 16(10), pp. 915–922.

126. Lupas, A., Van Dyke, M. and Stock, J. (1991), 'Predicting coiled coils from protein sequences', *Science*, Vol. 252(5010), pp. 1162–1164.

127. Lupas, A. (1997), 'Predicting coiled-coil regions in proteins', *Curr. Opin. Struct. Biol.*, Vol. 7(3), pp. 388–393.

128. Wootton, J. C. and Federhen, S. (1996), 'Analysis of compositionally biased regions in sequence databases', *Methods Enzymol.*, Vol. 266, pp. 554–571.

129. Rice, P., Longden, I. and Bleasby, A. (2000), 'EMBOSS: The European Molecular Biology Open Software Suite', *Trends Genet.*, Vol. 16(6), pp. 276–277.

130. Davis, G. D., Elisee, C., Newham, D. M. and Harrison, R. G. (1999), 'New fusion protein systems designed to give soluble expression in *Escherichia coli*', *Biotechnol. Bioeng.*, Vol. 65(4), pp. 382–388.

131. Westbrook, J., Feng, Z., Chen, L. *et al.* (2003), 'The Protein Data Bank and structural genomics', *Nucleic Acids Res.*, Vol. 31(1), pp. 489–491.

132. International Structural Genomics Organisation (URL: http://www.isgo.org/).

133. Chance, M. R., Bresnick, A. R., Burley, S. K. *et al.* (2002), 'Structural genomics: A4 pipeline for providing structures for the biologist', *Protein Sci.*, Vol. 11(4), pp. 723–738.

134. Bertone, P., Kluger, Y., Lan, N. *et al.* (2001), 'SPINE: An integrated tracking database and data mining approach for identifying feasible targets in high-throughput structural proteomics', *Nucleic Acids Res.*, Vol. 29(13), pp. 2884–2898.

135. Brenner, S. E., Barken, D. and Levitt, M. (1999), 'The PRESAGE database for structural genomics', *Nucleic Acids Res.*, Vol. 27(1), pp. 251–253.

136. Zolnai, Z., Lee, P. T., Li, J. *et al.* (2003), 'Project management system for structural and functional proteomics: Sesame', *J. Struct. Funct. Genomics*, in press.

137. Frishman, D., Albermann, K., Hani, J. *et al.* (2001), 'Functional and structural genomics using PEDANT', *Bioinformatics*, Vol. 17(1), pp. 44–57.

138. Frishman, D., Mokrejs, M., Kosykh, D. *et al.* (2003), 'The PEDANT genome database', *Nucleic Acids Res.*, Vol. 31(1), pp. 207–211.

139. Kawabata, T., Fukuchi, S., Homma, K. *et al.* (2002), 'GTOP: A database of protein structures predicted from genome sequences', *Nucleic Acids Res.*, Vol. 30(1), pp. 294–298.

140. Andrade, M. A., Brown, N. P., Leroy, C. *et al.* (1999), 'Automated genome sequence analysis and annotation', *Bioinformatics*, Vol. 15(5), pp. 391–412.

141. Hoersch, S., Leroy, C., Brown, N. P. *et al.* (2000), 'The GeneQuiz web server: protein functional analysis through the Web', *Trends Biochem. Sci.*, Vol. 25(1), pp. 33–35.

142. Lin, J., Qian, J., Greenbaum, D. *et al.* (2002), 'GeneCensus: Genome comparisons in terms of metabolic pathway activity and protein family sharing', *Nucleic Acids Res.*, Vol. 30(20), pp. 4574–4582.

143. Frishman, D. (2002), 'Knowledge-based selection of targets for structural genomics', *Protein Eng.*, Vol. 15(3), pp. 169–183.

144. Gardner, M. J., Tettelin, H., Carucci, D. J. *et al.* (1998), 'Chromosome 2 sequence of the human malaria parasite *Plasmodium falciparum*', *Science*, Vol. 282(5391), pp. 1126–1132.

145. Bowman, S., Lawson, D., Basham, D. *et al.* (1999), 'The complete nucleotide sequence of chromosome 3 of *Plasmodium falciparum*', *Nature*, Vol. 400(6744), pp. 532–538.

146. Hyman, R. W., Fung, E., Conway, A. *et al.* (2002), 'Sequence of *Plasmodium falciparum* chromosome 12', *Nature*, Vol. 419(6906), pp. 534–537.

147. Gardner, M. J., Shallom, S. J., Carlton, J. M. *et al.* (2002), 'Sequence of *Plasmodium falciparum* chromosomes 2, 10, 11 and 14', *Nature*, Vol. 419(6906), pp. 531–534.

148. Hall, N., Pain, A., Berriman, M. *et al.* (2002), 'Sequence of *Plasmodium falciparum* chromosomes 1, 3–9 and 13', *Nature*, Vol. 419(6906), pp. 527–531.

149. Gardner, M. J., Hall, N., Fung, E. *et al.* (2002), 'Genome sequence of the human malaria parasite *Plasmodium falciparum*', *Nature*, Vol. 419(6906), pp. 498–511.

150. White, J. H., Kilbey, B. J., de Vries, E. *et al.* (1993), 'The gene encoding DNA polymerase alpha from *Plasmodium falciparum*', *Nucleic Acids Res.*, Vol. 21(16), pp. 3643–3646.

151. Pizzi, E. and Frontali, C. (2001), 'Low-complexity regions in *Plasmodium falciparum* proteins', *Genome Res.*, Vol. 11(2), pp. 218–229.

152. URL: http://www.ysbl.york.ac.uk/~rodrigues/targets.html

153. Christendat, D., Yee, A., Dharamsi, A. *et al.* (2000), 'Structural proteomics of an archaeon', *Nat. Struct. Biol.*, Vol. 7(10), pp. 903–909.