Nucleic Acids Research, 2006, Vol. 34, Web Server issue ***W225–W230***
doi:10.1093/nar/gkl121

sgTarget: a target selection resource for structural genomics

Ana P. C. Rodrigues, Barry J. Grant and Roderick E. Hubbard*

Structural Biology Laboratory, University of York, York YO10 5YW, UK

Received February 14, 2006; Revised February 22, 2006; Accepted March 13, 2006

ABSTRACT

sgTarget (http://www.ysbl.york.ac.uk/sgTarget) is a web-based resource to aid the selection and prioritization of candidate proteins for structure determination. The system annotates user submitted gene or protein sequences, identifying sequence families with no homologues of known structure, and characterizing each protein according to a range of physicochemical properties that may affect its expression, solubility and likelihood to crystallize. Summaries of these analyses are available for individual sequences, as well as whole datasets. This type of analysis enables structural biologists to iteratively select targets from their genomic sequences of interest and according to their research needs. All sequence datasets submitted to sgTarget are available for users to select and rank using their choice of criteria. sgTarget was developed to support individual laboratories collaborating in structural and functional genomics projects and should be valuable to structural biologists wishing to employ the wealth of available genome sequences in their structural quests.

INTRODUCTION

The first step in any structure determination project is to select the appropriate molecule for study. Selection strategies vary according to the scientific context and aims of the project. In structural genomics, which aims to determine the structure of all important bio-molecules, the large number of potential candidates complicates the selection process. It is therefore important to identify the molecules for which a structure (normally of a protein) will provide the highest new information content and, where possible, quantify measures of how tractable each molecule is for structure determination (1,2). Evolutionary constraints can be used to identify proteins that may adopt similar conformations to known protein structures. For these proteins, modeling approaches may provide sufficient information to understand structure and mechanism. Certain sets of protein characteristics can be inferred from its sequence and employed in the identification of proteins that may pose problems during the various stages of structure determination. For example, fibrous domains can frustrate single crystal formation protocols and may frequently be identified by examining the protein's amino acid sequence (e.g. certain coiled coils).

Structural biology groups wishing to select and prioritize targets from raw sequence data may currently use genomic annotation servers, such as PEDANT (3) or 3D-Genomics (4). These automated services contain gene and protein annotations for a number of completed genomes. Although they detail annotations of relevance to the selection procedure no user accessible mechanism exists for generating target lists.

sgTarget was specifically designed to enable structural biologists to submit their sequence of interest and to select and rank targets according to their choice of criteria. A simple web interface can be used to generate and download target lists that may be iteratively refined by users. The resource was developed to assist individual laboratories participating in structural and functional genomics consortiums, as necessitated by our laboratory's involvement in the Structural Proteomics IN Europe (SPINE) consortium (http://www.spineurope.org/) and the *Plasmodium* Functional Genomics Initiative (http://www.sanger.ac.uk/PostGenomics/plasmodium/).

THE sgTarget ANNOTATION PIPELINE

A sequence annotation pipeline forms the core of the resource. This carries out the determination and prediction of properties and relationships that can be used in the selection of suitable

*To whom correspondence should be addressed. Tel: +44 1904 328267; Fax: +44 1904 328266; Email: rod@ysbl.york.ac.uk
Present addresses:
Ana P. C. Rodrigues, Burnham Institute for Medical Research, La Jolla CA 92037, USA
Barry J. Grant, Department of Chemistry & Biochemistry, University of California San Diego, La Jolla CA 92037, USA

targets. The pipeline consists of a set of bioinformatics methods that were selected and incorporated into the resource's framework, as follows:

- Methods to predict protein fold, function and prevalence. These help to identify targets, such as proteins for which fold predictions cannot yet be established, those with unknown functions, or ORFan proteins.
- Assessment of known protein expression and crystallization issues. Nucleotide sequence based calculations determine the encoding gene's GC content, codon usage and its compatibility with that of the host expression system (the Codon Adaptation Index). These metrics can highlight potential problems for protein expression. Similarly, sequence based prediction of protein instability, solubility and half-life can identify issues for high throughput structure determination.
- Assessment of known protein structure issues. Protein sequence based calculations predict the locations of intrinsically disordered, fibrous or transmembrane regions. The presence of these features can pose challenges for structure determination.

The majority of protocols employed by the annotation pipeline use established bioinformatics methods and databases (listed in Table 1). A novel procedure for the identification of intrinsically disordered regions was developed (5) and is described briefly below. In addition, tailored thresholds were established for GC content (between 26.9 to 66.8% for the expression host *Escherichia coli*), Codon Adaptation Index (above 0.084 for expression in *E.coli*, and above 0.357 for high levels of expression) and *E*-value cutoffs to assess the structural significance of BLAST alignments (two cutoffs are employed by the resource: 2.07×10^{-11}, a conservative threshold and 2.15×10^{-4}, a 'natural' threshold with a false positive rate of 0.2%).

Identification of intrinsically disordered regions

Intrinsically disordered domains can cause a multitude of adverse effects in structural determination studies, including purification difficulties due to hypersensitivity to protease digestion, missing electron density due to incoherent X-ray scattering, hindered crystallization, extreme broadening of side chain NMR peaks and lack of chemical shift dispersion of NMR backbone data. Some of these segments may become ordered upon interaction with binding partners to perform specific functions (6). Their structural characterization would, however, be difficult even if prior knowledge of the required cofactors was available.

The annotation pipeline employs the charge-hydrophobicity phase-space boundary of Uversky *et al.* (7), complemented by the putative lower bound complexity threshold of Romero and colleagues (8), to predict regions of intrinsic disorder. The low-complexity detection software SEG isolates subsequences with high or low-complexity on the basis of information content (9). In sgTarget, SEG is employed to detect any subsequences of at least 45 residues and a complexity value lower than 2.90. Such regions are annotated as probable non-globular protein stretches. For the remaining subsequences the mean hydrophobicity [the sum of the normalized hydrophobicities from (10) divided by the number of residues] and the mean net charge at pH 7.0 are calculated, and used in Equation 1, to predict if a subsequence is likely to be intrinsically disordered. Uversky and colleagues found that disordered proteins have low overall hydrophobicity and high net charge, always falling below the boundary:

$$\langle H \rangle = \frac{\langle R \rangle + 1.151}{2.785} \qquad \mathbf{1}$$

where $\langle H \rangle$ is the mean hydrophobicity and $\langle R \rangle$ is the mean net charge (7,11).

The performance of sgTarget's disorder prediction method on the CASP5 disorder benchmark was evaluated (12). sgTarget's disorder predictions for those targets that are least related to a protein with known structure, achieved an accuracy of 0.77 (where accuracy is the arithmetic mean of sensitivity and specificity measured on a per residue basis), which compares favorably to previously reported methods.

Table 1. Software, databases and selected protocols employed in sgTarget's annotation pipeline

Software	Application
CodonW[a]	Calculate the relative conformance of a gene to an organism's genome (the Codon Adaptation Index)
BLAST (18)	Perform local protein sequence similarity searches against PDB and NRDB sequences
InterProScan (19)	Run sequence comparison methods required to search the InterPro database (as well as NCOILS (20) to identify coiled-coil domains)
SEG (9)	Detect and isolate subsequences with high or low-complexity
TMHMM (21)	Predict the location and topology of protein transmembrane regions
Database	Description
PDB SEQRES (22)	Protein sequences derived from the SEQRES card of PDB files
InterPro (23)	Integrated collection of the protein domain family databases (Pfam, PRINTS, ProDom, PROSITE, SMART, TIGRFAMs and PANTHER)
GO (24)	Function ontology database with mappings to InterPro
NRDB	Collection of protein sequence databases (PIR, SWISS-PROT, TrEMBL and PDB SEQRES)
Taxonomy[b]	Taxonomical classification of organisms cross-referenced by NRDB
Protocol	Description & Application
Instability index (25)	The instability index is a length-scaled measure of the occurrence of all dipeptides in a protein sequence. Guruprasad and colleagues found a correlation between this measure and protein stability: in general, stable proteins have instability indices smaller than 40.
Estimate half-life using the N-end rule (26)	Estimates of *in vivo* half-life for proteolysis of proteins in prokaryotes can be made by the N-end rule. This considers the presence of a destabilizing N-terminal residue that provides an N-degron degradation signal.
Wilkinson–Harrison solubility index (27,28)	The revised Wilkinson–Harrison statistical solubility model depends on two parameters: the fraction of residues with a high index for forming turns and the approximate average charge of the protein *in vivo*. This model has been shown to be useful in the selection of proteins with high solubility.

[a] CodonW (http://codonw.sourceforge.net/).
[b] Taxonomy (http://www.ncbi.nlm.nih.gov/Taxonomy/taxonomyhome.html).

Hence, the method is suitable to analyze datasets where there may be many new folds, such as the complete genomes that serve as input to the resource.

In summary, the annotation methods employed by sgTarget allow the identification and prediction of a wide range of properties for each putative target. These enable users to filter and prioritize proteins and genes, generating lists of targets to suit diverse requirements.

THE sgTarget SERVER

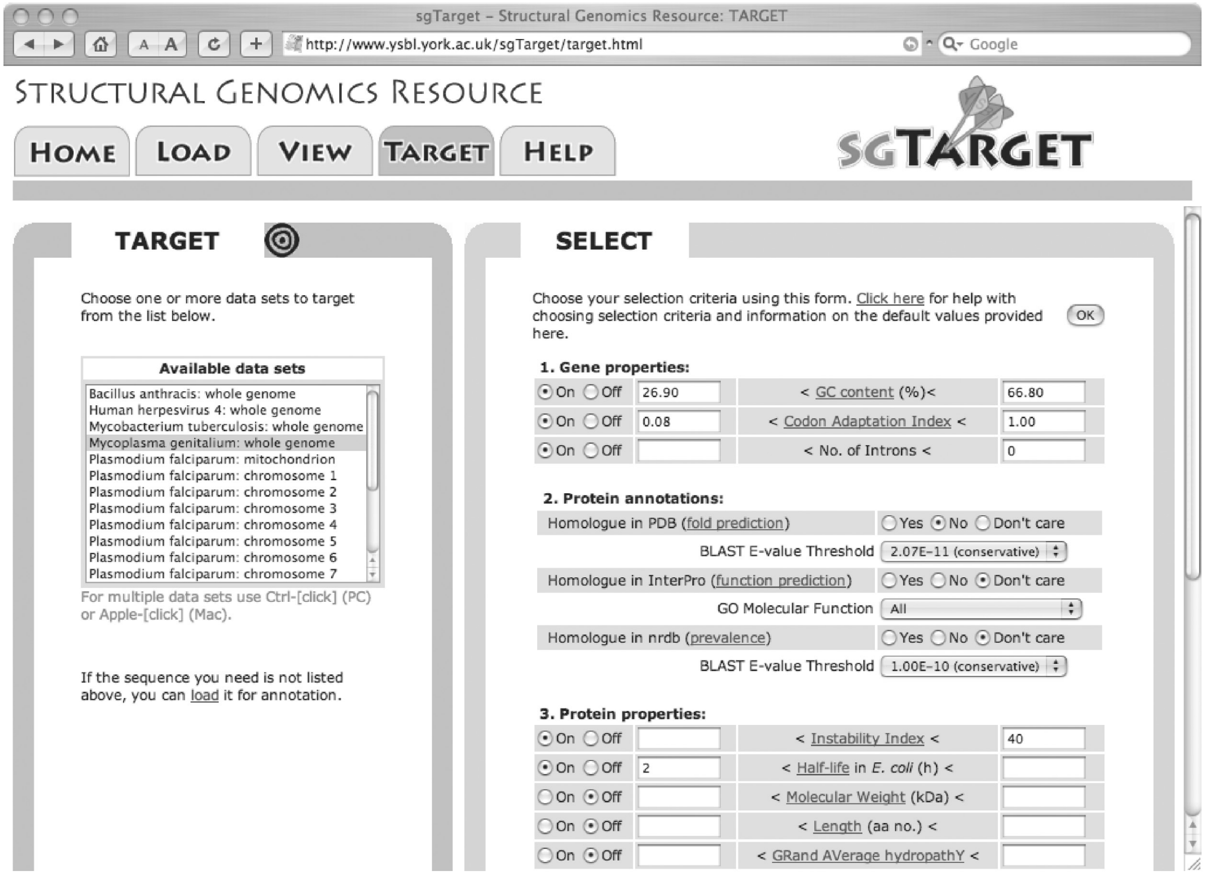
A web-based interface has been developed to interact with the sequence annotation pipeline. This allows users to analyze genomic sequences of interest by submitting them to the server, interact with the resulting data by browsing or searching and to select and prioritize targets for structural determination according to their choice of criteria. The interface is available at http://www.ysbl.york.ac.uk/sgTarget/ and its functionality is divided into three main pages: Load, View and Target.

Load

The Load page allows users to submit their sequences of interest through an anonymous interface. Requests are submitted to the annotation pipeline and processed sequentially. Annotations for an average bacterial chromosome (∼5 Mb or ∼4000 protein coding genes) take ∼24 h to complete. Users can choose to be notified of progress by e-mail on initiation and on completion of annotations. Depending on the level and nature of user requests, there may need to be some prioritization and arbitration on the order and choice of which organisms or datasets are annotated.

View

The View page allows users to analyze the sequence annotations performed by the resource. Users can browse through the annotations for a dataset using the Browse function. Here detailed annotations are available for individual proteins, and global synopses are available for the dataset's characteristics. Browsing the data by protein enables users to investigate the results of all the calculations obtained through the annotation pipeline for a particular gene/protein sequence. This includes gene information, such as GC content and codon usage, protein information, such as function, structure and prevalence predictions, and information on the suitability of the target for structural studies, such as the number of transmembrane, disordered and coiled-coil regions, and the protein's physicochemical properties. Browsing the data


sgTarget – Structural Genomics Resource: TARGET
http://www.ysbl.york.ac.uk/sgTarget/target.html
Google
STRUCTURAL GENOMICS RESOURCE
HOME
LOAD
VIEW
TARGET
HELP
SGTARGET
OK

TARGET

Choose one or more data sets to target from the list below.

Available data sets

Bacillus anthracis: whole genome
Human herpesvirus 4: whole genome
Mycobacterium tuberculosis: whole genome
Mycoplasma genitalium: whole genome
Plasmodium falciparum: mitochondrion
Plasmodium falciparum: chromosome 1
Plasmodium falciparum: chromosome 2
Plasmodium falciparum: chromosome 3
Plasmodium falciparum: chromosome 4
Plasmodium falciparum: chromosome 5
Plasmodium falciparum: chromosome 6
Plasmodium falciparum: chromosome 7

For multiple data sets use Ctrl-[click] (PC) or Apple-[click] (Mac).

If the sequence you need is not listed above, you can load it for annotation.

SELECT

Choose your selection criteria using this form. Click here for help with choosing selection criteria and information on the default values provided here.

1. Gene properties:

● On ○ Off	26.90	< GC content (%)<	66.80
● On ○ Off	0.08	< Codon Adaptation Index <	1.00
● On ○ Off		< No. of Introns <	0

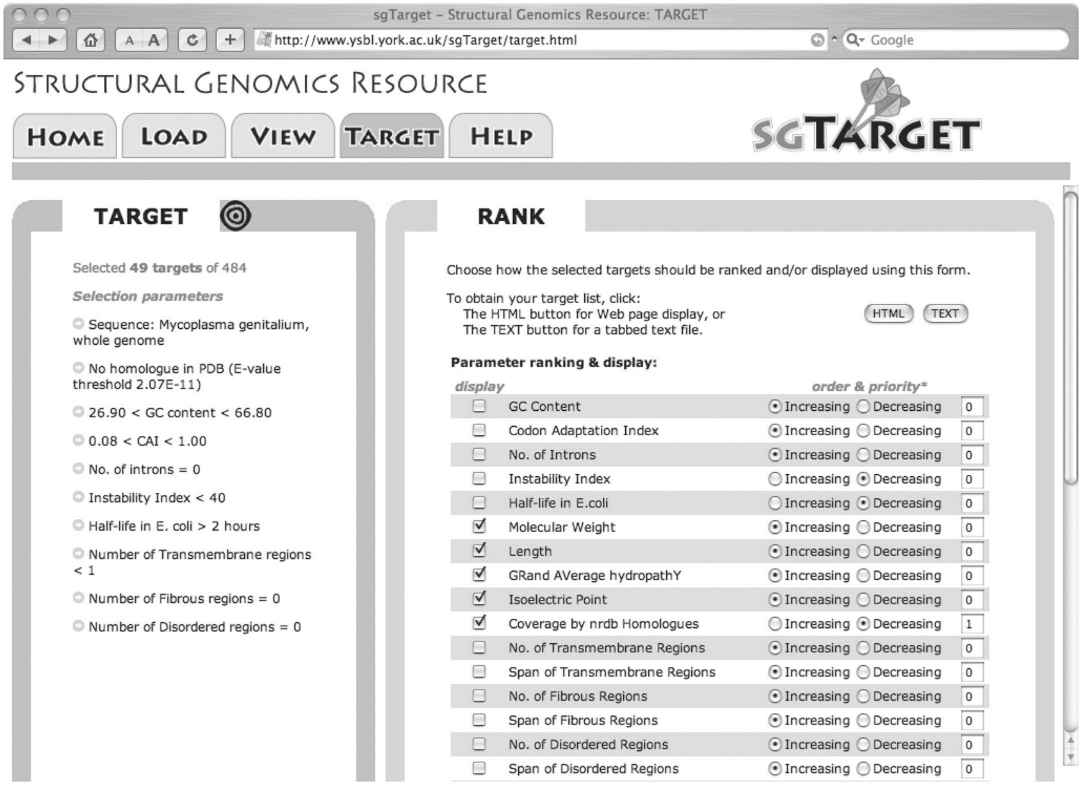
2. Protein annotations:

Homologue in PDB (fold prediction)	○ Yes ● No ○ Don't care
BLAST E-value Threshold	2.07E–11 (conservative)
Homologue in InterPro (function prediction)	○ Yes ○ No ● Don't care
GO Molecular Function	All
Homologue in nrdb (prevalence)	○ Yes ○ No ● Don't care
BLAST E-value Threshold	1.00E–10 (conservative)

3. Protein properties:

● On ○ Off		< Instability Index <	40
● On ○ Off	2	< Half-life in *E. coli* (h) <	
○ On ● Off		< Molecular Weight (kDa) <	
○ On ● Off		< Length (aa no.) <	
○ On ● Off		< GRand AVerage hydropathY <	

Figure 1. Target page with Select function activated. The menu area (on the left) allows users to choose one or more sequence datasets to target. The work area (on the right) allows users to specify selection criteria. In this example, the *Mycoplasma genitalium* genome has been chosen for targeting. The selection criteria specify that genes must have a GC content and CAI that is optimal for *E.coli*, and proteins have no homologues with known structure, are likely to be stable, viable in *E.coli* for at least 2 h, have at most one transmembrane region, and no fibrous or disordered regions (sgTarget's default selection criteria). When users click the OK button they are presented with the Rank function, and asked to choose how the target list should be prioritized and displayed (shown in Figure 2).


sgTarget – Structural Genomics Resource: TARGET
http://www.ysbl.york.ac.uk/sgTarget/target.html
Google
STRUCTURAL GENOMICS RESOURCE
HOME
LOAD
VIEW
TARGET
HELP
sgTARGET

TARGET

Selected **49 targets** of 484

Selection parameters

- Sequence: Mycoplasma genitalium, whole genome
- No homologue in PDB (E-value threshold 2.07E-11)
- 26.90 < GC content < 66.80
- 0.08 < CAI < 1.00
- No. of introns = 0
- Instability Index < 40
- Half-life in E. coli > 2 hours
- Number of Transmembrane regions < 1
- Number of Fibrous regions = 0
- Number of Disordered regions = 0

RANK

Choose how the selected targets should be ranked and/or displayed using this form.

To obtain your target list, click:
The HTML button for Web page display, or
The TEXT button for a tabbed text file.

HTML TEXT

Parameter ranking & display:

display		*order & priority**	
□	GC Content	● Increasing ○ Decreasing	0
□	Codon Adaptation Index	● Increasing ○ Decreasing	0
□	No. of Introns	● Increasing ○ Decreasing	0
□	Instability Index	○ Increasing ● Decreasing	0
□	Half-life in E.coli	○ Increasing ● Decreasing	0
☑	Molecular Weight	● Increasing ○ Decreasing	0
☑	Length	● Increasing ○ Decreasing	0
☑	GRand AVerage hydropathY	● Increasing ○ Decreasing	0
☑	Isoelectric Point	● Increasing ○ Decreasing	0
☑	Coverage by nrdb Homologues	○ Increasing ● Decreasing	1
□	No. of Transmembrane Regions	● Increasing ○ Decreasing	0
□	Span of Transmembrane Regions	● Increasing ○ Decreasing	0
□	No. of Fibrous Regions	● Increasing ○ Decreasing	0
□	Span of Fibrous Regions	● Increasing ○ Decreasing	0
□	No. of Disordered Regions	● Increasing ○ Decreasing	0
□	Span of Disordered Regions	● Increasing ○ Decreasing	0

Figure 2. Target page with Rank function activated. The menu area (on the left) shows a summary of the results returned by the Select function. The work area (on the right) allows users to specify which data to display for the selected targets, and how to rank those targets by specifying the priority of each annotation. Users can choose to view the prioritized target list as a Web page (by clicking the HTML button) or, alternatively, as a tabbed text file (by clicking the TEXT button). In this case, 49 targets were selected with the criteria specified in Figure 1. The target list is to be ranked with decreasing coverage by NRDB database (i.e. proteins with more of their length annotated as similar to a protein in the NRDB database have higher priority) and a number of protein physicochemical properties are to be displayed along with the default attributes (off the screen in this screenshot) (see Figure 3 for resulting page).

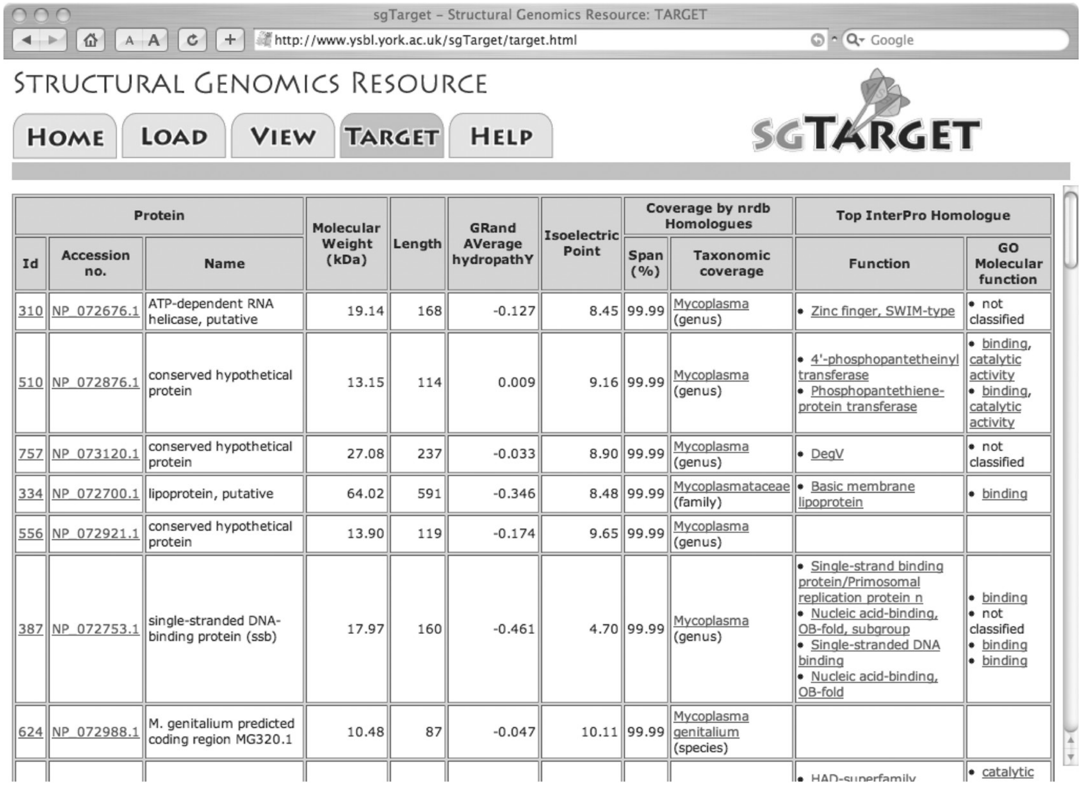
by characteristic enables users to investigate the results of a particular set of calculations for that dataset. This includes global statistics for gene expression predictions, structural and functional annotations, prevalence assignments, transmembrane and non-globular regions predictions, as well as physicochemical properties. Within the View page, users can also search each subset using the Search function. It allows users to find proteins using the resource's own identifier, as well as other identifiers (GenBank accession no.) and names (sequencing center naming), as provided by the sequence input files.

Target

The Target page enables users to select and prioritize targets. The Select function is used to specify the datasets to target, which gene and protein properties the targets should possess, and what parameters and thresholds should be employed in the selection (Figure 1). All annotations established through the annotation pipeline can be employed as selection parameters. Upon selecting targets, users are presented with the Rank function, which enables them to perform target prioritization (Figure 2). This function also allows users to choose the format and layout of the target list, which is finally presented to them (Figure 3).

APPLICATION

sgTarget has underpinned the selection of targets for our laboratory's collaboration in the *Plasmodium* Functional Genomics Initiative. The resource was employed to annotate the genome of *Plasmodium falciparum*, the organism that causes the most fatal form of human malaria (Figure 4). This enabled the generation of a target list by refining the selection choices to consider parameters selected by researchers in the group. The initial list of 73 targets consists of malaria proteins encoded by single exon genes with GC contents higher than 30%, no transmembrane regions and no long non-globular hydrophilic regions. GC content and intron number are the most selective of the parameters, together reducing the number of possible targets by 98%. These selection criteria were chosen to identify proteins likely to express in *E.coli*, and initial results obtained by the group indicate that the target list has been successful on those terms (13). Thus far, the group have initiated work on 10 of these targets, successfully cloned and expressed 8, purified 6, of which 1 is in crystallization trials [and has also been shown to be crucial for the parasite's invasion of human red blood cells, (14)] and 3 have already yielded high-resolution structures (15,16) and Boucher, I., Brzozowski, A.M., Brannigan, J.A., Schnick, C., Smith, D., Kyes, S. and Wilkinson, A.J., manuscript in preparation.


sgTarget – Structural Genomics Resource: TARGET
http://www.ysbl.york.ac.uk/sgTarget/target.html
Google
STRUCTURAL GENOMICS RESOURCE
HOME
LOAD
VIEW
TARGET
HELP
SGTARGET

Protein			Molecular Weight (kDa)	Length	GRand AVerage hydropathY	Isoelectric Point	Coverage by nrdb Homologues		Top InterPro Homologue	
Id	Accession no.	Name					Span (%)	Taxonomic coverage	Function	GO Molecular function
310	NP_072676.1	ATP-dependent RNA helicase, putative	19.14	168	-0.127	8.45	99.99	Mycoplasma (genus)	• Zinc finger, SWIM-type	• not classified
510	NP_072876.1	conserved hypothetical protein	13.15	114	0.009	9.16	99.99	Mycoplasma (genus)	• 4'-phosphopantetheinyl transferase • Phosphopantethiene-protein transferase	• binding, catalytic activity • binding, catalytic activity
757	NP_073120.1	conserved hypothetical protein	27.08	237	-0.033	8.90	99.99	Mycoplasma (genus)	• DegV	• not classified
334	NP_072700.1	lipoprotein, putative	64.02	591	-0.346	8.48	99.99	Mycoplasmataceae (family)	• Basic membrane lipoprotein	• binding
556	NP_072921.1	conserved hypothetical protein	13.90	119	-0.174	9.65	99.99	Mycoplasma (genus)		
387	NP_072753.1	single-stranded DNA-binding protein (ssb)	17.97	160	-0.461	4.70	99.99	Mycoplasma (genus)	• Single-strand binding protein/Primosomal replication protein n • Nucleic acid-binding, OB-fold, subgroup • Single-stranded DNA binding • Nucleic acid-binding, OB-fold	• binding • not classified • binding • binding
624	NP_072988.1	M. genitalium predicted coding region MG320.1	10.48	87	-0.047	10.11	99.99	Mycoplasma genitalium (species)		
									• HAD-superfamily	• catalytic

Figure 3. Target page showing a target list. The selected targets are ranked according to the order and priority specified for the different annotations, and a table of prioritized targets is built using the annotations that were chosen for display. In this case, a list of 49 targets (selected from *M.genitalium*'s genome with the criteria specified in Figure 1) was ranked by decreasing coverage by NRDB database proteins, and a table constructed showing the target's identifier (in sgTarget), accession number, name, molecular weight, length, GRAVY score, isoelectric point, coverage by NRDB database proteins (including the span of the alignments on the target and the top taxonomic group which encompasses all reported alignments) and function annotation (the top InterPro hit and its GO high-level molecular function) (as specified in Figure 2).

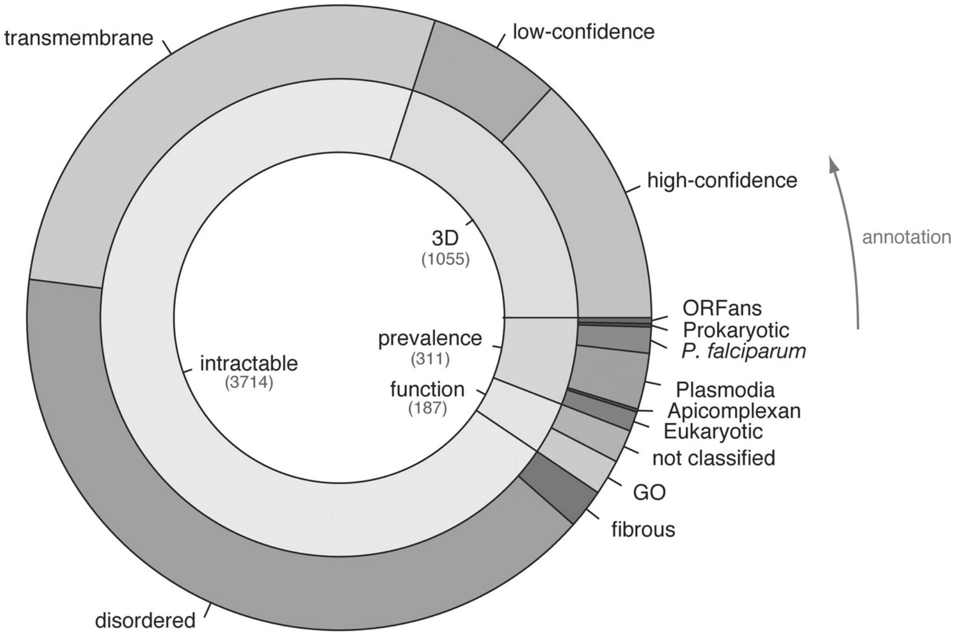


Figure 4. *P.falciparum* annotation wheel, with an emphasis on structural annotation. Annotations are displayed anti-clockwise as follows: A total of 1055 proteins have structural annotations, 691 high-confidence and 364 low-confidence (PDB SEQRES, release 05/2002); Of the remaining proteins, 3714 are likely to be intractable: 1475 have transmembrane regions, a further 2131 have disordered regions and the other 108 have fibrous regions; For the remainder of the proteome, 187 proteins have function annotations, although only 97 of these are classified by GO; Most other proteins are found in other organisms (295), except for 16 ORFan proteins.

In addition, sgTarget has been employed to select a set of *Bacillus anthracis* target proteins for the SPINE consortium. Here, the resource was used in tandem with the bioinformatics tools available at the Oxford Protein Production Facility (http://www.oppf.ox.ac.uk/bioinformatics.php) to select a set of proteins of desirable molecular weight (20 to 55 kDa), which are likely to be soluble (insolubility probability smaller than 0.7) (17).

We encourage structural biologists to submit sequence datasets to sgTarget and contact us regarding suggestions on software and databases for the annotation pipeline, the annotation views provided by sgTarget and the functionality of the Target page.

ACKNOWLEDGEMENTS

We would like to thank the authors and curators of the software and databases used in this work. A.R. was supported by a grant to the York Structural Biology Centre by Accelrys Inc, San Diego. Funding to pay the Open Access publication charges for this article was provided by Accelrys.

Conflict of interest statement. None declared.

REFERENCES

1. Brenner,S.E. (2000) Target selection for structural genomics. *Nature Struct. Biol.*, **7**, 967–969.
2. Rodrigues,A. and Hubbard,R.E. (2003) Making decisions for structural genomics. *Brief Bioinform*, **4**, 150–167.
3. Frishman,D., Mokrejs,M., Kosykh,D., Kastenmuller,G., Kolesov,G., Zubrzycki,I., Gruber,C., Geier,B., Kaps,A., Albermann,K. *et al.* (2003) The PEDANT genome database. *Nucleic Acids Res.*, **31**, 207–211.
4. Fleming,K., Muller,A., MacCallum,R.M. and Sternberg,M.J. (2004) 3D-GENOMICS: a database to compare structural and functional annotations of proteins between sequenced genomes. Nucleic Acids Research 32:D245-D250. *Nucleic Acids Res.*, **32**, D245–D250.
5. Rodrigues,A.P.C. (2004) Target Selection in Structural Genomics. PhD Thesis. University of York, York, UK.
6. Wright,P.E. and Dyson,H.J. (1999) Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. *J. Mol. Biol.*, **293**, 321–331.
7. Uversky,V.N., Gillespie,J.R. and Fink,A.L. (2000) Why are 'natively unfolded' proteins unstructured under the physiological conditions? *Proteins*, **41**, 415–427.
8. Romero,P., Obradovic,Z., Li,X., Garner,E., Brown,C.J. and Dunker,A.K. (2001) Sequence complexity of disordered proteins. *Proteins*, **42**, 38–48.
9. Wootton,J.C. and Federhen,S. (1993) Statistics of local complexity in amino acid sequences and sequence databases. *Comput. Chem.*, **17**, 149–163.
10. Kyte,J. and Doolittle,R. (1982) A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.*, **157**, 105–132.
11. Uversky,V.N. (2002) Natively unfolded proteins: a point where biology waits for physics. *Protein Sci.*, **11**, 739–756.
12. Melamud,E. and Moult,J. (2003) Evaluation of disorder predictions in CASP5. *Proteins*, **53**, 561–565.
13. Brannigan,J.A., Boucher,I., Dodson,G., Rodrigues,A., Schnick,C. and Wilkinson,A.J. (2003) Structural studies of *Plasmodium* proteins by X-ray crystallography. *Exp. Parasitol.*, **105**, 26.
14. Green,J.L., Martin,S.R., Fielden,J., Ksagoni,A., Grainger,M., Yim Lim,B.Y., Molloy,J.E. and Holder,A.A. (2006) The MTIP-myosin A complex in blood stage malaria parasites. *J. Mol. Biol.*, **355**, 933–941.
15. Whittingham,J.L., Leal,I., Nguyen,C., Kasinathan,G., Bell,E., Jones,A.F., Berry,C., Benito,A., Turkenburg,J.P., Dodson,E.J. *et al.* (2005) dUTPase as a platform for anti-malarial drug design: structural basis for the selectivity of a new class of nucleoside inhibitors. *Structure*, **13**, 329–338.
16. Schnick,C., Robien,M.A., Brzozowski,A.M., Dodson,E.J., Murshudov,G.N., Anderson,L., Luft,J.R., Mehlin,C., Hol,W.G., Brannigan,J.A. *et al.* (2005) Structures of *Plasmodium falciparum* purine nucleoside phosphorylase complexed with sulfate and its natural substrate inosine. *Acta. Crystallogr. D. Biol. Crystallogr.*, **61**, 1245–1254.
17. Au,K., Berrow,N.S., Blagova,E., Boyle,M.P., Brannigan,J.A., Carter,L.J., Grenha,R., Levdikov,V.M., Kalliomaa,A.K., Meier,C. *et al.* (2006) Application of high-throughput technologies to a structural-genomics type analysis of *Bacillus anthracis*. *Acta. Crystallogr. D. Biol. Crystallogr.,* in press.
18. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
19. Zdobnov,E.M. and Apweiler,R. (2001) InterProScan–an integration platform for the signature-recognition methods in InterPro. *Bioinformatics*, **17**, 847–848.
20. Lupas,A., Van Dyke,M. and Stock,J. (1991) Predicting coiled coils from protein sequences. *Science*, **252**, 1162–1164.
21. Sonnhammer,E.L., von Heijne,G. and Krogh,A. (1998) A hidden Markov model for predicting transmembrane helices in protein sequences. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **6**, 175–182.
22. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.
23. Mulder,N.J., Apweiler,R., Attwood,T.K., Bairoch,A., Barrell,D., Bateman,A., Binns,D., Biswas,M., Bradley,P., Bork,P. *et al.* (2003) The InterPro Database, 2003 brings increased coverage and new features. *Nucleic Acids Res.*, **31**, 315–318.
24. The GO Consortium (2000) Gene Ontology: tool for the unification of Biology. *Nature Genet.*, **25**, 25–29.
25. Guruprasad,K., Reddy,B.V.B. and Pandit,M. (1990) Correlation between stability of a protein and its dipeptide composition: a novel approach for predicting *in vivo* stability of a protein from its primary sequence. *Protein Eng.*, **4**, 155–161.
26. Tobias,J.W., Shrader,T.E., Rocap,G. and Varshavsky,A. (1991) The N-end rule in bacteria. *Science*, **254**, 1374–1377.
27. Wilkinson,D.L. and Harrison,R.G. (1991) Predicting the solubility of recombinat proteins in *Escherichia coli*. *Biotechnology*, **9**, 443–449.
28. 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.*, **65**, 382–388.