TarO: a target optimisation system for structural biology

Ian M. Overton1, C. A. Johannes van Niekerk1, Lester G. Carter2, Alice Dawson1, David M. A. Martin1, Scott Cameron1, Stephen A. McMahon2, Malcolm F. White2, William N. Hunter1, James H. Naismith2 and Geoffrey J. Barton1,*

1School of Life Sciences Research, University of Dundee, Dow Street, Dundee, DD1 5EH and 2Centre for Biomolecular Sciences, School of Biomedical Science, North Haugh, The University, St Andrews, KY16 9ST, UK

Received January 31, 2008; Revised March 13, 2008; Accepted March 16, 2008

ABSTRACT

TarO (http://www.compbio.dundee.ac.uk/taro) offers a single point of reference for key bioinformatics analyses relevant to selecting proteins or domains for study by structural biology techniques. The protein sequence is analysed by 17 algorithms and compared to 8 databases. TarO gathers putative homologues, including orthologues, and then obtains predictions of properties for these sequences including crystallisation propensity, protein disorder and post-translational modifications. Analyses are run on a high-performance computing cluster, the results integrated, stored in a database and accessed through a web-based user interface. Output is in tabulated format and in the form of an annotated multiple sequence alignment (MSA) that may be edited interactively in the program Jalview. TarO also simplifies the gathering of additional annotations via the Distributed Annotation System, both from the MSA in Jalview and through links to Dasy2. Routes to other information gateways are included, for example to relevant pages from UniProt, COG and the Conserved Domains Database. Open access to TarO is available from a guest account with private accounts for academic use available on request. Future development of TarO will include further analysis steps and integration with the Protein Information Management System (PIMS), a sister project in the BBSRC ‘Structural Proteomics of Rational Targets’ initiative

INTRODUCTION

Target selection for structural biology encompasses a variety of analyses, and may include optimisation of the protein target for successful progress in the structure determination pipeline. The evaluation of putative homologues and/or alternative constructs is a key aspect of the optimisation process (1,2). One useful metric that may be applied to this end is estimated crystallisation propensity (3,4). This approach aims to increase the odds of success in the face of attrition rates that typically exceed 90% in structural genomics consortia (5–7). However, target optimisation is also commonplace as a salvage strategy following difficulties with the originally selected protein.

Numerous bioinformatics analyses can be applied during target optimisation, including searching various databases (8–11) and sequence-based prediction of protein properties, such as protein disorder (1). However, the generation, integration and management of results from these analyses are not trivial (1,12). There are many publicly available servers that run individual bioinformatics analysis steps. Websites are also available to provide a single point of access to individual analysis tools, for example Expasy (13), Entrez (14) and OPAL (12). However, target optimisation using these sites is laborious and there is little facility to integrate the results of numerous analyses across many sequences. A greater level of integration over a user-supplied multiple sequence alignment (MSA) is provided by MACSIMS (15), which also propagates annotations by homology inference. However, MACSIMS is not focused on target optimisation and does not generate any ranking of sequences. Also, MACSIMS returns a limited set of annotation types and only annotation that is amenable to display on a MSA is given in a user-friendly format. Servers that focus on target selection are available, such as SGTarget (16), and the more recent XtalPred (17). These provide some integration of data for the user, but are limited in terms of the number of annotation types and the server features. Neither SGTarget nor XtalPred provide an annotated MSA.

We have developed a system (TarO) that offers a single point of reference for key target optimisation analyses. TarO features include gathering and annotation of
putative orthologues and homologues, searching the protein input against the Protein DataBank (18) with PSIBLAST (19), generation of annotated MSA, and presentation of integrated results to the user. TarO was originally developed for the Scottish Structural Proteomics Facility (SSPF) (www.sspf.ac.uk), and plays a key role in the SSPF bioinformatics platform. To date, TarO has processed more than 720 queries and is used by several different research groups outside the SSPF.

METHODS

Overview of TarO

TarO takes a protein sequence as input, which is used to search for putative orthologues and homologues. The input and associated sequences are analysed in a number of annotation steps, and the results stored in a database. The TarO website (www.compbio.dundee.ac.uk/taro) provides access to results, and integrates the Jalview (20,21) program to visualise complex annotation over the MSA. All analyses are run on a local computer cluster. The TarO website, which incorporates the Jalview (20,21) program, facilitates human interpretation of the data. The final ranking is therefore semi-automated, combining electronic and human interpretation of the data.

Detection and annotation of functionally related sequences

Detection of functionally and structurally similar proteins helps in the selection of sequences that are more amenable to structural studies. Orthologues frequently share substantial functional similarity, and this assumption may be cautiously extended to all homologues (22,23). Part of the assessment of functional relationships involves examination of the patterns of annotation and conserved residues, or ‘functional signatures’, on the sequences. This process is assisted by an annotated MSA constructed from the input sequence and the putative orthologues/homologues. The annotated MSA is displayed in Jalview (20,21). Scores from BLAST (19) sequence alignments also provide a rough metric for estimating functional similarity in TarO.

TarO detects putative orthologues by searching the input sequence against COG/KOG (11) with BLASTP (19). Matches for both the orthologue and homologue searches are defined from thresholds selected to infer protein structural similarity (24). In addition, all matches must have BLAST expectation values of $10^{-3}$ or better. The top-scoring COG/KOG match forms the basis to infer a COG/KOG orthologue cluster; all sequences in the relevant orthologue cluster are thus assigned as putative orthologues of the input protein. Subsequently, the input sequence as well as any putative orthologues are searched against the UniRef100 (8) database with PSIBLAST (three iterations, default values) (19).

The input sequence and any putative orthologues/homologues found are searched against the Protein DataBank (PDB) (18) with PSIBLAST and BLASTP, respectively. The input and associated sequences are also searched against TargetDB (25) with BLASTP, thereby highlighting any similar targets that have been registered by Structural Genomics consortia. The searches of TargetDB and the PDB both use the thresholds for structural similarity (24) and expectation value as described above. RPSBLAST (19) is also used to search all query-associated sequences against the Conserved Domains Database (CDD) (26,27), which includes profiles from Pfam (9,10), SMART (28,29) and COG/KOG (11). RPSBLAST matches to domain profiles are defined by an expectation value threshold of $10^{-3}$. Elementary chemical properties [e.g. average GES hydrophobicity (30)] are calculated with custom perl code, Bioperl (31) and PEPPSTATS (32). Sequences are assigned to phylogenetic classifications in order to allow for SignalP (33) prediction of signal peptide (default parameters). This classification is based on the data provided by COG/KOG and UniRef100. Where phylogenetic classification is not available, SignalP is run using all of the possible classifications. Only the first 70 amino acids of each sequence are taken as input to SignalP in order to reduce false positives. Additionally, predictions for the input and all associated sequences are obtained for NetOglyc (34), NetPhos (35), RONN (36), Disembl (37), Globplot (38), Jpred (39,40) and NetNglyc (http://www.cbs.dtu.dk/services/NetNGlyc/), with the default settings for each algorithm. It is important to note that NetNglyc and NetOglyc glycosylation predictions should be treated with caution when a signal peptide is not also predicted (34) http://www.cbs.dtu.dk/services/NetNGlyc/. TarO gives a warning when displaying the list of predicted glycosylation sites for a sequence without a predicted signal peptide. The MSA is generated from the input and associated sequences by running MUSCLE (41). Reliably generating a MSA from automatically obtained search results can be difficult, so sequences are only included in the MSA if their BLAST alignment to the input sequence has an expectation value $\leq 10^{-20}$, and if their sequence length...
is no more than 125% of the input sequence length. Also, sequences are chosen for inclusion into the MSA according to the order of priority: input > putative orthologues > putative homologues. This order is followed until the user-specified maximum number of sequences is reached (default 100), or until all of the query-associated sequences have been examined. We plan further development of the strategy for generating the MSA which will be incorporated into later releases of TarO.

TarO also annotates the input and associated sequences with information that is useful through the course of ‘wet-lab’ stages in the structure determination pipeline. The predicted extinction coefficient at 280 nm is calculated by PEPSTATS (32), to assist with protein purification. Counts of the amino acids histidine, cysteine and methionine are given, which may be relevant for protein purification and deriving phases by anomalous scattering approaches. Other information in this category includes molecular weight, sequence length, hydrophobicity and isoelectric point. Table 1 summarises the various algorithms and databases currently employed in TarO.

The TarO database and external database update management

The results of the various analyses run by TarO, including searches of external databases, are parsed with custom perl code and stored in a relational database. The TarO web server queries this database when presenting results to the user. External databases (Table 1) are stored as flat files and searched locally on a high-performance compute cluster as part of the process of running a TarO query. These external databases are updated on a weekly basis with custom scripts based around the ‘wget’ Unix command. As a consequence, the information gathered by TarO is no more than one week old at the time of running a given query. Results associated with a TarO query reflect the information available at the time that the search was performed. The TargetDB database ‘target status’ information is a special case in this regard, because it is regularly updated into the TarO database. Therefore, the TargetDB ‘target status’ displayed in TarO is updated every week for any matched TargetDB sequence, regardless of the date and time at which the TarO query was run. However, all matches between TarO and TargetDB sequences are identified from a search of the TargetDB database available at the time that the TarO query is run. Regular searches of completed TarO queries are not run against any database, partly because a TarO query is not necessarily an active target. However, the option of periodically searching certain databases (e.g. TargetDB, PDB) may be incorporated in a future release.

**USAGE**

**Submitting a TarO query**

Open access to TarO is available for any user, via a ‘Guest’ area that can be easily accessed from a link on the TarO home page. The ‘New Query’ link in the ‘Guest’ area navigates to a form that will accept TarO queries in ‘FASTA’ format. Queries can be uploaded to the server as a file or pasted into a textbox. There is an input option to specify the maximum number of sequences to include in the MSA (default value is 100). There is also a ‘functional description’ textbox which allows users to more easily identify their submitted queries. Some algorithms do not accept non-standard amino acid characters, and so these are removed from the query sequence input when appropriate. Queries submitted by the ‘Guest’ user are visible to anyone and deleted from the server after a minimum of 8 days. However, free private accounts are available for academic use; see the TarO website (www.compbio.dundee.ac.uk/taro) for further details. We ask that users wait for the results of a submitted query before making a further submission to the server. We estimate that an ‘average’ query will require approximately 100 cpu hours, though these are spread over...
a compute cluster. Given a typical load on the cluster, throughput is in excess of 70 queries per week and a typical query is completed within 4–12 hr.

Tracking query progress and access to results

Figure 2 shows an example of the query sequence information page, which serves as a hub for each TarO query. The table at the top has 47 columns and so extends to about three times the width of the figure. This table includes basic sequence statistics, as well as details of the top-scoring match from COG/KOG (11), the PDB (18), TargetDB (25) and UniRef100 (8). Several links are available within this table, notably to display ranked annotations for putative orthologues and homologues of the input sequence, respectively displayed as the characters ‘O’ and ‘H’. There are also links to relevant pages of the COG/KOG, Dasty2 (42), CDD (26,27) and UniProt (8) websites, as well as links to results of RPSBLAST (19) searches of domain profiles. Clicking on the grey rectangle below this table displays the annotated MSA in the Jalview (20,21) applet (Figure 3). The ‘Query Status’ table allows tracking of the query progress through the various annotation stages, according to a ‘traffic lights’ system. Stages that have started are shown in Amber, Red is used to indicate a failed step, and completed analyses are shown in Green. Inset shows an example ‘Query Status’ table for a query that is in progress. There is extensive context-sensitive help throughout TarO, and the table headings also provide links to the relevant section of the help document. An example query sequence information page is given at http://www.compbio.dundee.ac.uk/taro/cgi-taro/targpipe_display_query_seqs.pl?query=657&funcdesc=Test_Guest1.

Figure 2. Query sequence information page. This page serves as a hub for each TarO query. The table at the top has 47 columns and so extends to about three times the width of the figure. This table includes basic sequence statistics, as well as details of the top-scoring match from COG/KOG (11), the PDB (18), TargetDB (25) and UniRef100 (8). Several links are available within this table, notably to display ranked annotations for putative orthologues and homologues of the input sequence, respectively displayed as the characters ‘O’ and ‘H’. There are also links to relevant pages of the COG/KOG, Dasty2 (42), CDD (26,27) and UniProt (8) websites, as well as links to results of RPSBLAST (19) searches of domain profiles. Clicking on the grey rectangle below this table displays the annotated MSA in the Jalview (20,21) applet (Figure 3). The ‘Query Status’ table allows tracking of the query progress through the various annotation stages, according to a ‘traffic lights’ system. Stages that have started are shown in Amber, Red is used to indicate a failed step, and completed analyses are shown in Green. Inset shows an example ‘Query Status’ table for a query that is in progress. There is extensive context-sensitive help throughout TarO, and the table headings also provide links to the relevant section of the help document. An example query sequence information page is given at http://www.compbio.dundee.ac.uk/taro/cgi-taro/targpipe_display_query_seqs.pl?query=657&funcdesc=Test_Guest1.

DISCUSSION

Structural biology projects are highly variable and so there is not a universally applicable target optimisation strategy. However, certain criteria are generally useful. Target optimisation frequently draws upon overlapping information for the evaluation of both alternative constructs and putative homologues. Although NMR is an important technique for structure determination, as of January 2008 85% of all structures in the PDB (18) had been solved by X-ray crystallography. As a consequence, obtaining crystals is a key stage in most structural biology pipelines. Modifying the construct sequence may influence crystallisation propensity, and alternative homologues may be
examined since protein families commonly have members with a wide range of estimated crystallisation propensity (3). The OB-Score (3), ParCrys (4) and Hydrophobicity/pI clustering (43) are all harnessed by TarO to estimate crystallisation propensity, and so guide the evaluation of homologues. Proteins with transmembrane regions or significant disordered sequence are frequently problematic (1,17). Also, posttranslational modifications (PTMs) are commonly associated with protein disorder (44). TarO assists with identification of sequences that are likely to contain these potentially troublesome, but biologically interesting, features.

TarO also assists with the identification of protein domain boundaries, facilitated by an annotated MSA that is viewed in Jalview (20,21). The MSA annotations include matched domains from Pfam (9,10) and the conserved domains database (CDD) (26,27), combined with predicted protein disorder. Predicted transmembrane regions, signal peptide [SignalP (33)], PTMs and secondary structure [JPred (39,40)] are also annotated on the MSA. Other useful information associated with the MSA is provided by the Jalview program. For example, Jalview automatically provides a display of residue conservation at each position of the alignment. In addition, Jalview provides the facility to query numerous Distributed Annotation System (46) servers, and to display any returned annotation on the MSA. The various annotations associated with the MSA are useful to assist with the design of optimised constructs and identification of functionally important residues. Building upon this, a likely future development.
in TarO is the automated design and ranking of optimised construct sequences. Of course, the design of optimised construct sequences may also benefit from information provided by experimental methods such as limited proteolysis (47).

Retaining the functional features that originally stimulated interest in the target is an important consideration during target optimisation. For example, removing part of an enzyme’s active site might make crystals easier to obtain; although the resultant protein structure would be relatively ineffective for studies of the molecular mechanism of catalysis! The range of functional information provided by TarO aims to assist with identification and comparison of functional regions in protein sequences.

A possible future direction is the automated evaluation of sequence features to provide more sophisticated prediction and analysis of the functional conservation for a given protein pair. These predictions could be useful in the context of target optimisation, for example by enabling more advanced protein ranking systems. Different projects have different sets of functional properties that are required to be retained in the optimised target sequence. However, all putative orthologues and homologues currently identified in TarO pass thresholds that aim to preserve a reasonable level of structural similarity (24).

As a screening mechanism to avoid duplication of effort, the protein input and associated sequences are searched against the PDB (18) and TargetDB (25). The discovery of a similar structure in the PDB or TargetDB may be sufficient grounds to eliminate a potential target. On the other hand, identification of a known and related structure could be important; this may provide a model for molecular replacement calculations, or inform on components of multi-domain or multi-subunit systems.

In summary, TarO enables selection of sequences that are likely to be more amenable to structural studies and share functional similarity with the input sequence. Additionally, TarO provides information relevant for many of the structure determination pipeline stages, including design of optimised constructs. The use of TarO accelerates progress in structural proteomics by efficiently providing bioinformatics data to inform decision-making on the prioritisation and optimisation of potential targets. TarO simplifies the gathering, storage and retrieval of data and so frees up research time to make use of the information and to think creatively. Please cite TarO as well as the underlying algorithms and databases, as appropriate. Active development of TarO is continuing to include further analysis steps, improvements to the user interface, and integration with the Protein Information Management System (PIMS) a sister project in the BBSRC Structural Proteomics of Rational Targets (SPoRT) initiative. We also plan to make available a distribution of the TarO source code. We feel that community interactions with the TarO project can lead to further advancement and dissemination of best practices for target optimisation. Access to TarO is from www.compbio.dundee.ac.uk/taro and we are grateful to receive feedback from users.

ACKNOWLEDGEMENTS

Thanks to Drs T. Walsh and C. Cole for computational advice. This work was funded by the UK Biotechnology and Biological Sciences Research Council (BBSRC) Structural Proteomics of Rational Targets (SPoRT) initiative, (Grant BBS/B/14434). Funding to pay the Open Access publication charges for this article was provided by BBSRC.

Conflict of interest statement. None declared.

REFERENCES

PSI-BLAST: a new generation of protein database search programs.

Nucleic Acids Res. 25, 3389-3402.


