Methods 55 (2011) 3–11

Contents lists available at SciVerse ScienceDirect

Methods

journal homepage: www.elsevier.com/locate/ymeth

Computational approaches to selecting and optimising targets for structural biology

Ian M. Overton [a,*], Geoffrey J. Barton [b]

[a] MRC Human Genetics Unit, Institute of Genetics and Molecular Medicine, Western General Hospital, Crewe Road, Edinburgh EH4 2XU, United Kingdom
[b] College of Life Sciences, University of Dundee, Dundee DD1 5EH, United Kingdom

ARTICLE INFO

Article history:
Available online 27 August 2011

Keywords:
Target selection
Crystallisation
Structural genomics
Structural biology
Bioinformatics
Construct design

ABSTRACT

Selection of protein targets for study is central to structural biology and may be influenced by numerous factors. A key aim is to maximise returns for effort invested by identifying proteins with the balance of biophysical properties that are conducive to success at all stages (e.g. solubility, crystallisation) in the route towards a high resolution structural model. Selected targets can be optimised through construct design (e.g. to minimise protein disorder), switching to a homologous protein, and selection of experimental methodology (e.g. choice of expression system) to prime for efficient progress through the structural proteomics pipeline.

Here we discuss computational techniques in target selection and optimisation, with more detailed focus on tools developed within the Scottish Structural Proteomics Facility (SSPF); namely XANNpred, ParCrys, OB-Score (target selection) and TarO (target optimisation). TarO runs a large number of algorithms, searching for homologues and annotating the pool of possible alternative targets. This pool of putative homologues is presented in a ranked, tabulated format and results are also visualised as an automatically generated and annotated multiple sequence alignment. The target selection algorithms each predict the propensity of a selected protein target to progress through the experimental stages leading to diffracting crystals. This single predictor approach has advantages for target selection, when compared with an approach using two or more predictors that each predict for success at a single experimental stage. The tools described here helped SSPF achieve a high (21%) success rate in progressing cloned targets to diffraction-quality crystals.

© 2011 Elsevier Inc. All rights reserved.

1. Introduction

Of all techniques applied in molecular biology, macromolecular crystallography reveals the most exquisite details about the machines of life. Advances in X-ray sources, computational methods and cryo-techniques over the last 20 years have led to a dramatic increase in the rate at which a protein structure may be determined once diffracting crystals have been obtained. Unfortunately, expressing proteins at levels suitable for structural studies and obtaining crystals that diffract remain the major bottlenecks in most structural biology laboratories [1–3]. Accordingly, computational sequence analysis and similar methods are often applied to increase the chances of success. Common strategies are to seek out related proteins that might fare better than the preferred target (e.g. orthologues, pathway members), to "optimise" the target protein in some way, or to adjust the laboratory approach (e.g. choice of expression system) [3–13]. If the native protein fails to crystallise, optimisation typically starts with truncation of the protein into likely domains, or the removal of disordered regions, but may include more sophisticated engineering. These strategies rely on the application of computational tools for sequence analysis and alignment in conjunction with the structural biologist's experience. Although a single investigator might spend days studying options to try on their protein, in a high-throughput structural proteomics environment it is necessary to streamline this process by introducing a higher degree of automation. In this article, we examine computational approaches for selecting and optimising proteins for crystallography with emphasis on those developed at the University of Dundee [12,14–16] as part of the Scottish Structural Proteomics Facility (SSPF) [3]. Although developed with high-throughput crystallography in mind, most of the tools described here are equally applicable to smaller-scale structural studies.

2. Influence of project scope on structural proteomics target selection

The overall approach to selecting targets is dictated by the scope of the project. In the subsections below we outline a few examples of how the research aims may impact on target selection

Abbreviations: MSA, Multiple Sequence Alignment; PTM, Post Translational Modification; SSPF, Scottish Structural Proteomics Facility; MCC, Matthew's correlation coefficient; AROC, Area Under the Receiver Operator Characteristic curve.

* Corresponding author. Fax: +44 1314678456.
E-mail address: ian.overton@hgu.mrc.ac.uk (I.M. Overton).

1046-2023/$ - see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.ymeth.2011.08.014

and optimisation. A common principle in target selection is to identify proteins (e.g. orthologues) that both satisfy the project aims and are relatively amenable to structural characterisation. Target optimisation is applicable to almost every project and is discussed in greater detail in Section 6 of this article. Indeed analyses enabled by the Target Optimisation Utility (TarO) [12], such as prediction of domain boundaries, are useful in any structural biology laboratory – even for work that focuses specifically on a single target. Further information on current structural proteomics projects can be obtained by exploring links from the International Structural Genomics Organisation (ISGO) list of active initiatives [17].

2.1. Structural proteomics on a specific organism

Some projects seek to provide structural coverage across the whole proteome of a particular organism, such as *Thermotoga maritima* [18] or *Saccharomyces cerevisiae* [10,19]. This kind of genome-wide approach rules out searching for more tractable orthologues; however, ranking targets according to their predicted success may inform experimental strategy. Optimisation of the construct sequence may also be productive, for example to minimise protein regions predicted to be disordered or to adjust codon usage [11,12,20]. Such optimisation can be useful for all targets, but is more often adopted as a salvage strategy for targets that flounder with a standard approach.

2.2. Structural biology projects for drug discovery and biological chemistry

Some structural proteomics projects focus on targets that might be suitable in drug discovery against a specific pathogen, for example, the *Mycobacterium tuberculosis* structural genomics consortium [21]; even these consortia may have scope for flexibility across different targets amongst pathways and sub-networks. However, prioritisation of druggable targets with favourable properties, such as control of metabolic flux and therapeutic selectivity, limits the choice of alternative structural targets and constructs [22]. Structural characterisation of a biological pathway or a particular enzyme function enjoys greater flexibility, where exploration of different orthologues and constructs (e.g. the catalytic domain) may be helpful.

2.3. Mapping protein structure space

Efforts to extend protein structure space coverage (e.g. [7,23]) have good scope for selecting the most favourable candidates from groups of structurally similar proteins, at least where structural relationships can be reliably inferred. Similar flexibility is available to efforts that focus on particular classes of proteins (e.g. [24,25]). As noted above, target optimisation is also useful in these contexts.

3. Useful protein features in target selection and optimisation

In order to identify favourable targets and constructs, significant attention has been given to exploring biophysical properties and investigating protein selection strategies that correlate with success in obtaining a structure (e.g. [9,14–16,26–30]). To give a few examples, properties influencing soluble expression include isoelectric point (pI), hydrophobicity, and sequence length; properties influencing production of diffraction-quality crystals from purified protein include surface entropy, disordered sequence, and protein post-translational modifications [11,26–32]. Many of these features, as well as relevant algorithms and databases are summarised in Table 1. Properties that impact on success are often correlated. For example, regions that participate in protein–protein interactions have greater hydrophobicity [33,34], and sites of post-translational modification are enriched for disordered regions [31]. In addition, individual biophysical properties have been shown to significantly influence multiple pipeline stages. For example, hydrophobicity affects soluble expression, purification and crystallisation; glycosylation affects soluble expression and crystallisation; while the sequence length has an impact on cloning, soluble expression and crystallisation [11,27]. Moreover, selection or engineering for success at a given experimental stage can hinder progress at other parts of the structure determination pipeline. For example, surface entropy and charge are related because several high entropy residues have charge (e.g. Lys, Glu, Arg). In general, more surface charge, and consequently higher entropy, favours solubility; on the other hand lower surface entropy, and consequently charge, favours crystallisation [28,35]. Therefore, target selection and optimisation would ideally find protein chains that possess the correct balance of properties required for successful progression through all experimental stages leading to a high-resolution structural model. Indeed, algorithms have been developed with this goal in mind [14–16,36]. Algorithms are also available to predict progression at a particular pipeline stage [28–30,37]; for example PXS aims to predict the crystallisation of 'well-behaved' soluble proteins [28]. Section 5, below, gives further discussion of these and other tools.

An assessment of the existing functional annotation available to inform structure interpretation is also useful for target selection. Indeed, new structures are difficult to interpret without some functional knowledge, and so make a less immediate contribution to

Table 1
Estimation of protein characteristics useful for target selection and optimisation.

Protein characteristics	Exemplar algorithms and/or databases
Homology relationships	Algorithms: BLAST[88], SCANPS [95], MUSCLE [87], Magicmatch [96] Databases: eggNOG [55], InParanoid [56], UniProt [78]
Matches to known structures/declared targets	PDB [69], TargetDB/PepcDB [68]
Domains	Algorithms: HMMER [97], RPSBLAST Databases: Pfam [62], CDD [63], SMART [98], Superfamily [99], Biozon [100]
Protein interactions	PIPS [51], STRING [101]
Disorder/low-complexity sequence	Disembl [59], RONN [58], GlobPlot [60], SEG [102]
Signal peptide and transmembrane regions	SignalP [91], Phobius [45], TMHMM2 [44]
Glycosylation sites	NetOGlyc [65], NetNGlyc [90]
Phosphorylation sites	NetPhos [67], Musite [103]
Secondary structure	JPred [61], PSIPRED [104]
Surface entropy	SERp [9]
Chemical properties: isoelectric point (pI), molecular weight, charge, sequence length, extinction coefficient, #Methionines, #Cysteines, #Histidines, hydrophobicity, protease sites	Bioperl [105], PEPSTATS (EMBOSS) [106]
Annotated function	Gene Ontology [38]
Overall tractability (selected to diffraction-quality crystals)	XANNPred [16], XtalPred [36], OB-Score [15], ParCrys [14]

Download English Version:

https://daneshyari.com/en/article/10826223

Download Persian Version:

https://daneshyari.com/article/10826223

Daneshyari.com