A normalised scale for structural genomics target ranking: the OB-Score

Citation for published version:

Digital Object Identifier (DOI):
10.1016/j.febslet.2006.06.015

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Publisher's PDF, also known as Version of record

Published In:
FEBS Letters

Publisher Rights Statement:
Copyright 2006 Published by Elsevier B.V

General rights
Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.
Abstract Target selection and ranking is fundamental to structural genomics. We present a Z-score scale, the “OB-Score”, to rank potential targets by their predicted propensity to produce diffraction-quality crystals. The OB-Score is derived from a matrix of predicted isoelectric point and hydrophobicity values for nonredundant PDB entries solved to ≤3.0 Å against a background of UniRef50. A highly significant difference was found between the OB-Scores for TargetDB test datasets. A wide range of OB-Scores was observed across 241 proteomes and within 7868 PfamA families; 73.4% of PfamA families contain ≥1 member with a high OB-Score, presenting favourable candidates for structural studies.

© 2006 Published by Elsevier B.V. on behalf of the Federation of European Biochemical Societies.

Keywords: HTPX; Structural genomics; Target selection; Pfam; Bioinformatics

1. Introduction

Structural genomics has become a major enterprise in the ‘post-genomic era’ [1–3]. The field currently receives around $200 million in public funding per annum worldwide and this has nurtured significant technological developments in high-throughput systems [4–6]. However, it is common for only 5% of selected protein targets in high-throughput labs to achieve a high-resolution protein model [1,7–9]. Various strategies have been proposed [10–13] to reduce this attrition rate and so increase the return from structural genomics efforts. One approach has been to work with multiple orthologues of a target protein and so increase the probability of finding a protein sequence that will yield its molecular structure [14]. This approach requires strategies to rank protein sequences within such functionally defined groups according to their likely success in the structural genomics pipeline.

Analyses of structural genomics efforts on the proteome of Thermotoga maritima [15] have suggested that predicted isoelectric point (pI) and the grand average of hydrophobicity as calculated by the Kyte–Doolittle scale [16] (GRAVY) are strong indicators of a protein’s propensity to crystallise. GRAVY correlated well with proteins likely to be transmembrane and so more difficult to crystallise [15]. Studies based on decision trees and random forest algorithms have also proposed these measures, among others, as useful indicators for a protein’s successful progress through structural genomics protocols [17].

In this paper, the available structures in the PDB [18] and the UniRef50 database [19] are used to develop a normalised scale, the OB-Score, based on pI and GRAVY. Validation is performed on sequence sets from TargetDB [20], and the OB-Score is applied to 7868 Pfam families [21], 16 Ensembl proteomes [22] as well as 225 proteomes from the comprehensive microbial resource (CMR) database [23]. The OB-Score is found to be descriptive of a target’s progress through structural genomics pipelines to the production of diffraction-quality crystals, which is a key stage on the way to obtain a high-resolution structural model [24–26].

2. Materials and methods

Table 1 summarises the length, pI and GRAVY statistics of the datasets considered. The method was developed on 5454 Protein Data Bank [18] (PDB) chains solved to <3.0 Å resolution and clustered such that no pair of proteins shared more than 40% sequence identity [27,28] (‘Dbrack_PDB’). The threshold of 3.0 Å resolution aimed to filter out structures where modelling side-chain orientations would be challenging. Control data representing the known protein universe were provided by the 794,085 chains in the UniRef50 database [19] (‘UniRef50’). UniProt covers all publicly available protein sequences, whilst UniRef50 is produced by clustering UniProt so that no pair of sequences share more than 50% identity [19]. Test data for proteins expected to produce diffraction-quality crystals were taken from 1212 TargetDB [20] entries for which diffraction-quality crystals had been indicated (TDB_DIFF). Test data for proteins thought not to crystallise were taken from 11,745 TargetDB proteins which were those entries annotated as “Work Stopped” or “Work stopped” and where work had been stopped before the target was crystallised (TDB_WS). Test data to investigate the relationship of OB-Score to soluble expression were taken from 5235 TargetDB proteins where soluble expression had been indicated (TDB_SOL). Not all proteins in the TDB_WS dataset are necessarily difficult to crystallise, since work may have been stopped for a variety of reasons. However, the TargetDB datasets were searched with PSIBLAST [29] against a database of 72,314 PDB sequences embedded in the UniRef50 database, in order to remove any sequences similar to those deposited in the PDB. The UniRef50-embedded PDB database was filtered by the SEG program [30] to generate the PDB_U50 database. Sequences were screened from the TDB_WS, TDB_SOL and TDB_DIFF datasets based on matches to PDB sequences from PDB_U50 (PSIBLAST expectation value ≤10−5, 90% query coverage, 95% identity). This filtering step eliminated 440 proteins from the TDB_WS dataset, 55 proteins from the TDB_SOL dataset and 125 proteins from the TDB_DIFF dataset. The UniRef50, TargetDB, Dbrack_PDB and PDB data were downloaded on 2/2/2005, 23/2/2006, 21/6/2005 and 5/1/2006, respectively.

Sequence redundancy filtering was applied independently to each of the PDB-filtered TargetDB datasets by an RPSBLAST [29,31] search of the Pfam profiles from the conserved domain database (CDD) [32]. Each TargetDB query sequence was assigned to the CDD Pfam profile that gave the lowest expectation value (E-value) below a threshold of 10−6. The redundancy-filtered datasets were independently constructed by taking one TargetDB sequence representative per matched Pfam profile, as well as all sequences in that TargetDB dataset without

---

*Corresponding author. Fax: +44 1382 345764. E-mail address: geoff@compbio.dundee.ac.uk (G.J. Barton).
Table 1
Statistics for the datasets studied

<table>
<thead>
<tr>
<th>Dataset</th>
<th>N</th>
<th>Length</th>
<th>Min.</th>
<th>Median</th>
<th>Mean</th>
<th>Max.</th>
<th>pI</th>
<th>Median</th>
<th>Range</th>
<th>GRAVY</th>
<th>Median</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>TDB_DIF</td>
<td>728</td>
<td>12</td>
<td>252.5</td>
<td>282.3</td>
<td>1727</td>
<td>5.9</td>
<td>8.4</td>
<td>−0.23</td>
<td>2.4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(1087)</td>
<td>(12)</td>
<td>(262.0)</td>
<td>(289.1)</td>
<td>(1727)</td>
<td>(5.9)</td>
<td>(8.4)</td>
<td>(−0.21)</td>
<td>(2.4)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TDB_WS</td>
<td>6025</td>
<td>24</td>
<td>194.0</td>
<td>242.2</td>
<td>2514</td>
<td>7.2</td>
<td>10.7</td>
<td>−0.36</td>
<td>5.0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(11305)</td>
<td>(24)</td>
<td>(238.0)</td>
<td>(277.9)</td>
<td>(6048)</td>
<td>(6.9)</td>
<td>(10.7)</td>
<td>(−0.31)</td>
<td>(5.0)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TDB_SOL</td>
<td>3667</td>
<td>12</td>
<td>224.5</td>
<td>272.8</td>
<td>2695</td>
<td>6.7</td>
<td>9.6</td>
<td>−0.37</td>
<td>2.6</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(5180)</td>
<td>(12)</td>
<td>(242.0)</td>
<td>(284.2)</td>
<td>(2695)</td>
<td>(6.6)</td>
<td>(9.6)</td>
<td>(−0.34)</td>
<td>(2.6)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dbrack_PDB</td>
<td>5454</td>
<td>20</td>
<td>254.4</td>
<td>220.5</td>
<td>1733</td>
<td>6.4</td>
<td>9.9</td>
<td>−0.30</td>
<td>4.9</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PfamA</td>
<td>961,405</td>
<td>5</td>
<td>150.5</td>
<td>117.0</td>
<td>2799</td>
<td>6.8</td>
<td>11.6</td>
<td>−0.16</td>
<td>6.13</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>UniRef50</td>
<td>794,085</td>
<td>11</td>
<td>360.1</td>
<td>265.0</td>
<td>34350</td>
<td>7.5</td>
<td>12.3</td>
<td>−0.25</td>
<td>7.4</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Abbreviated column headings are as follows: N, total number of sequences; pI, isoelectric point; GRAVY, hydrophobicity. Abbreviations used for datasets are as follows: TDB_DIF, TargetDB diffraction-quality crystals; TDB_WS, TargetDB work stopped before crystals were obtained; TDB_SOL, TargetDB soluble; Dbrack_PDB, Dunbrack culled PDB. For the rows describing TargetDB datasets, statistics in brackets refer to the PDB-filtered raw data prior to redundancy filtering. Comparisons of the redundancy filtered and raw TargetDB datasets reveal similar statistics describing length, GRAVY and pI. However, the redundancy filtering does result in a substantial reduction in N. These data were calculated using the R package [41,42].

*Statistics for dataset prior to sequence redundancy filtering is given in brackets.

3. Results and discussion

The PDB-filtered, non-redundant TDB_DIF and TDB_WS datasets were used as test sets to investigate the power of the OB-Score to distinguish between proteins that produce diffraction-quality crystals via structural genomics pipelines, and those that are abandoned before the crystallisation stage. The TargetDB data are not perfect test sets as they are biased by structural genomics target selection and deselection procedures [35]. The PDB-filtering step (see Section 2) partially ameliorates this problem by removing deselected sequences. Also, the PDB-filtering aims to eliminate overlap between the Dbrack_PDB and the TargetDB datasets. Fig. 2 shows the OB-Score distributions for the TDB_DIF and TDB_WS datasets. As expected TDB_DIF is significantly enriched in high OB-Score values and depleted in low OB-Score values when compared to TDB_WS. The Wilcoxon rank sum test was used to compare the OB-Score distributions of TDB_DIF and TDB_WS, finding a highly significant difference between these datasets (two-tailed P-value <2.2e−16). The median OB-Scores for TDB_DIF and TDB_WS were 5.03 and 0.42, respectively. Additionally, the PDB-filtered, nonredundant TDB_SOL dataset was found to be slightly enriched in higher OB-Scores compared to TDB_WS. TDB_SOL median OB-Score was 1.95. TDB_SOL was found to be significantly different to TDB_DIF and TDB_WS (respective Wilcoxon two-tailed P-values: 8.882e−16, <2.2e−16).

In order to investigate whether certain organisms may be better candidates for high-throughput crystallography, the OB-Score was applied to the 225 proteomes from version 16.0 of the CMR database [23] and the 16 Ensembl [22,36] proteomes available on 10/6/05. Of the 241 organisms examined, only six had positive OB-Scores for >70% of their annotated proteome. The most extreme was Buchnera aphidicola an endosymbiont of the aphid Schizaphis graminis with only 7% of its proteins having positive OB-Scores. Two further Buchnera sp. have <13% positive OB-Scores, while three Mycoplasma sp. had...
<30%. The remainder of the low positive set of proteomes was made up of *Wigglesworthia glossinidia brevipalpis* an endosymbiont of the tsetse fly (7%), and *Blochmannia floridanus*. The eukaryotic proteomes all showed median OB-Scores close to zero.

The OB-Score is anticipated to contribute to more efficient use of structural genomics resources by aiding the selection of the most crystallisable sequence(s) for a given target function. Additionally, the OB-Score may help efforts to identify representative template structures for all protein families [12,13,38], by allowing the prioritisation of those sequences within a given family that have the highest OB-Score. Fig. 3 illustrates the OB-Score statistics across all 7868 families in Pfam version 17.0. A large number of families are seen to have a wide range of OB-Scores. The median range is 10.4, the median inter-quartile range is 3.0, while 949 (12.1%) families have representatives at both the minimum and maximum OB-Score values. It is encouraging that 73.4% (5777) of Pfam families have at least 1 member with an OB-Score $\geq 5$, and so are predicted to be relatively amenable to structural investigations. There is no apparent correlation between the median OB-Score and the number of organisms in, nor the average sequence length of a Pfam family (data not shown). Visual inspection of the Pfam functional descriptions for those families with the most extreme values of median OB-Score did not reveal any clear trends, except for a higher frequency of membrane proteins within those families that have the lowest median OB-Score values.

In summary, a normalised scale, the OB-Score, has been developed to indicate the similarity of a protein’s $pI$–GRAVY combination to that of previously crystallised proteins. High positive OB-Scores may be used to indicate proteins that should be more likely to succeed in the process that runs from cloning through expression and purification to structure deter-
The OB-Score data for 7868 PfamA families, as well as software to calculate the OB-Score are freely available from (http://www.compbio.dundee.ac.uk/obscore).

Acknowledgements: We thank Profs. J. Naismith and M. White for helpful discussions. We also thank Drs. T. Walsh and J. Monk for computational advice, and the UK BBSRC (Biotechnology and Biological Sciences Research Council) Structural Proteomics of Rational Targets initiative for financial support (Grant BBS/B/14434).

References


