Protein Expression and Purification 72 (2010) 175–178

Contents lists available at ScienceDirect

Protein Expression and Purification

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

A software tool to accelerate design of protein constructs for recombinant expression

Johanna Sagemark [1], Per Kraulis, Johan Weigelt *

Structural Genomics Consortium, Karolinska Institutet, Department of Medical Biochemistry and Biophysics, 171 77 Stockholm, Sweden

ARTICLE INFO

Article history:
Received 3 December 2009
and in revised form 19 March 2010
Available online 30 March 2010

Keywords:
Sequence analysis
Structure prediction
Construct design
Protein expression
Protein crystallization

ABSTRACT

Structural and biochemical analysis of proteins requires access to purified protein material. Modern molecular biology technologies facilitate straightforward molecular cloning and expression analysis of multiple protein constructs in parallel, and such approaches have proven very efficient to identify samples suitable for further analysis.

A variety of information can be used to support rational design of protein constructs. This includes, e.g. prediction of secondary structure elements, protein domain predictions, and structure prediction methods such as threading. To fully access the available information, collation of data extracted from several different sources is required. This can be cumbersome and sometimes also confusing due to for example different implementation of amino acid residue numbering schemes. The SGC Domain Boundary Analyser tool provides a graphical interface that simplifies and accelerates rational design of protein expression constructs.

Introduction

Structural, biochemical and biophysical characterization of proteins require access to protein material. Over the last three decades recombinant technologies for overexpression of proteins in non-natural host cells have replaced the traditional methods of obtaining protein samples from their natural sources. In order to successfully purify a protein to enable further analyses and characterization it is imperative to produce the protein in sufficient amounts and in soluble form. It is well known that recombinant expression levels, protein solubility as well as other properties, such as crystallization propensity, varies vastly between proteins. Moreover even for the same protein expression levels, solubility, biochemical activity, stability, crystallization propensity etc. will vary between different expression constructs, e.g. [1–4]. For example, the ability to successfully express a full length version of a protein may differ from the ability to express one functional domain of a multi-domain protein. In particular, variation of the N- and C-terminal tailing residues of a protein construct by truncation of the amino acid sequence at different locations may critically affect the protein properties. Researchers interested in a particular protein therefore often attempt to produce several different versions of the protein in order to arrive at a sample suitable for their needs. Identification of protein constructs amenable for crystallization has traditionally been performed by trial and error through testing of different protein constructs sequentially in an iterative manner. Modern molecular biology technologies facilitate straightforward molecular cloning and expression screening of multiple protein constructs in parallel, and recently several reports have described the benefits of evaluating several different constructs in parallel [5,6].

A straightforward approach to optimize the production of soluble protein by multiple construct design is to alter the start and stop positions of the expression construct by truncating the full length coding sequence. It is not possible to confidently predict which constructs will yield soluble or crystallizable protein but a variety of information can be used to support rational construct design. This includes prediction of secondary structure elements [7,8], comparative protein modeling such as homology modeling and threading [9–11], and identification of protein domains annotated in Pfam [12,13]. This information is typically collected from a variety of databases and prediction algorithms. In order to fully access it collation of data extracted from several different sources is thus required. This task may not only be tedious but sometimes also confusing. The various prediction tools often lack a standardized way of presenting the output. For example the conventions for numbering of amino acid residues differ between tools. This complicates the process of combining the results to obtain a full picture of the available information.

Here we introduce a software tool that visually presents the results from an assortment of prediction algorithms in one single

* Corresponding author. Fax: +46 8 524 86868.
E-mail address: johan.weigelt@ki.se (J. Weigelt).
[1] Present address: AstraZeneca R&D Mölndal, Discovery Information, 431 83 Mölndal, Sweden.

1046-5928/$ - see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.pep.2010.03.020

graphical view. This provides an overview of predicted structural features to simplify and enable design of protein expression constructs.

Materials and methods

The SGC Domain Boundary Analyser (DBA[2]) program was written in C# using Microsoft Visual Studio 2005 (Microsoft Corporation) and is dependent on local installations of Blast [14] and HMMER [15], and access to formatted Pfam [12,13] and Conserved Domains (CDD[2]) [16] databases. The current implementation of the DBA tool is based on the following prediction methods:

BioInfoBank Structure Prediction Meta Server

The DBA tool uses the output from the BioInfoBank Structure Prediction Meta Server [17]. This server accepts the protein amino acid residue sequence as input and runs various fold recognition, function prediction and local structure prediction methods. It generates alignments and calculates a similarity score (3D-Jury score or J-score) for each prediction. The latter is based on a comparison between different models generated by different methods. The J-score has been shown to correlate significantly with the number of correctly predicted residues [18]. The output results from the meta server are downloaded in html format for subsequent import and parsing by the DBA tool.

Low complexity regions

The functional aspects of low complexity regions in protein sequences are not fully understood but comparing the abundance of these regions in sequence databases to the abundance in the PDB they are clearly under-represented in the latter. It has also been shown that X-ray protein structures are more disordered in regions with low complexity [19]. Moreover, it is commonly assumed that the crystal contacts that are needed to form the crystal lattice may be less favored in low complexity regions. Hence analysis of low complexity regions should be included while making construct design decisions for structural analysis. The SEG algorithm of the BLAST package is used to identify low complexity regions in the protein sequence [14,20].

Pfam domains

The Pfam database [12,13] provides hidden Markov model (HMM) profiles for protein domain families. Pfam is built with manually curated seed alignments and HMM is used to align members to the families. The *hmmpfam* algorithm of the HMMER package [15] is used by the SGC DBA tool to search a profile HMM database to annotate Pfam domains in the protein sequence.

Conserved Domains Database

The Conserved domains database [16] is provided by the NCBI and holds multiple sequence alignments of conserved protein domains. It consists of alignment data from Pfam [12,13], SMART [21,22], COGs [23,24] and Protein Clusters [25]. The most abundant residues in aligned columns are calculated and reported for the consensus sequence of each conserved domain. This is used to calculate position-specific score matrices (PSSM). The query protein sequence is compared to the PSSM using the reverse position-specific BLAST algorithm and the result is added to the DBA tool output.

The protein amino acid residue sequence to be analysed is first submitted to the meta server which distributes the query to a number of different prediction servers and collects the results. Once results from the meta server analyses have been downloaded they can be read into the DBA tool. The query sequence is then automatically analysed with respect to low complexity regions and occurrence of Pfam and CDD domains. This analysis is only executed the first time a protein is loaded and the results are stored for future use. If re-analysis of the protein is required the result files can be deleted from the file structure associated with the program. This will trigger a new analysis of the protein upon loading of the meta server output.

Results and discussion

The DBA tool implementation is based on the output of the BioInfoBank Structure Prediction Meta Server but could be adapted to accept the output of other protein structure prediction engines. The BioInfoBank meta server was chosen since it includes several well established prediction protocols. Its scoring protocol has also been shown to be reliable [18]. The DBA tool parses the results from the meta server and creates a graphical overview of the results. In addition it analyses the amino acid sequence for the occurrence of low complexity regions and conserved domains (Pfam & CDD). All information is presented to the user in one single graphical view. This facilitates manual analysis of all available information in parallel, and the user can interactively select start and stop positions for different constructs. For future reference these selections can be exported as a text file by the program.

A snapshot of the graphical interface of the DBA tool is shown in Fig. 1A. The query amino acid sequence is displayed both at the top and the middle of the graphical view to simplify the interactive analysis. Predicted low complexity regions and occurrence of domains annotated in Pfam or CDD are displayed in the top half of the window. The meta server output contains PSIPRED [26] and PROFSEC [B. Rost unpublished, http://www.rostlab.org] secondary structure predictions based on the amino acid sequence. In the DBA tool these are displayed above the query sequence at the center of the graphical display. Predicted alpha helices and beta sheets are marked with different colors (helices are colored red and sheets blue). All 3D structure predictions, sorted by J-scores, are displayed in the lower part of the interactive window following the same coloring scheme for helices and sheets. Conserved amino acid residues are marked with a small vertical line over the residue identifier in the 3D structure prediction panel (Fig. 1B).

The graphical interface can be used for careful manual analysis of the available predictions and annotations to design expression constructs. While selecting constructs for protein expression, purification and crystallization it is often beneficial only to work on single domains or pairs/groups of domains that form a functional unit. A draggable ruler is available to simplify the analysis. The program indicates the position of the ruler by displaying the amino acid residue identifier (one letter code + residue number) next to the ruler (Fig. 1C). The sequence position of a particular residue is also readily available by clicking on the sequence. Once decisions for start and stop positions of expression constructs have been made, these can be marked in the sequence by right-clicking at the selected positions (Fig. 1D). The analysis can be saved to a text file that includes the marked start and stop positions.

Obviously the different prediction data carry different weights in the construct design process. In order to reliably design a set of protein constructs good structural models must have been identified. The secondary structure predictions and identification of

[2] *Abbreviations used*: DBA, Domain Boundary Analyser; CDD, Conserved Domains; HMM, hidden Markov model; PSSM, position-specific score matrices.