N- and C-terminal Truncations to Enhance Protein Solubility and Crystallization: Predicting Protein Domain Boundaries with Bioinformatics Tools

Christopher D. O. Cooper
Department of Biological Sciences
School of Applied Sciences
University of Huddersfield
Queensgate
Huddersfield
West Yorkshire
United Kingdom
HD1 3DH
c.d.cooper@hud.ac.uk

Brian D. Marsden
Structural Genomics Consortium
Nuffield Department of Medicine
University of Oxford
Old Road Campus Research Building
Roosevelt Drive,
Oxford
Oxfordshire
United Kingdom
OX3 7DQ

and

Kennedy Institute of Rheumatology
Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences
University of Oxford
Old Road Campus
Roosevelt Drive
Oxford
Oxfordshire
United Kingdom
OX3 7FY

brian.marsden@sgc.ox.ac.uk

Running head: Bioinformatics Tools for Soluble Protein Expression.

Abstract

Soluble protein expression is a key requirement for biochemical and structural biology approaches to study biological systems *in vitro*. Production of sufficient quantities may not always be achievable if proteins are poorly soluble which is frequently determined by physico-chemical parameters such as intrinsic disorder. It is well known that discrete protein domains often have a greater likelihood of high-level soluble expression and crystallizability. Determination of such protein domain boundaries can be challenging for novel proteins. Here we outline the application of bioinformatics tools to facilitate the prediction of potential protein domain boundaries, which can then be used in designing expression construct boundaries for parallelized screening in a range of heterologous expression systems.

Key words: Bioinformatics, Protein expression, Protein solubility, Protein structure, Domain, BLAST, PSIPRED, Hidden Markov Model (HMM), Alignment, Secondary structure.

1 Introduction

In order to study proteins by structural, biochemical or biophysical approaches, a key requirement is the ability to produce sufficient levels of purified protein, ranging from the microgram to milligram levels depending on the technique in question [1]. It is costly, inefficient and often impossible to obtain sufficiently pure and adequate quantities from native sources [2]. Modern approaches frequently utilize heterologous protein expression systems such as *Escherichia coli*, optimized to produce large quantities of protein from plasmid expression vectors containing a cloned and defined sequence [3, 4]. It is well known, however, that sequence of the protein is one of the most important determinants of successful protein expression, solubility or crystallization potential [1, 5]. Results vary greatly between the expression constructs used (encoding fragments of defined protein sequence length and context) [6] due to differing protein physicochemical properties and biological factors such as protein folding, export or toxicity in the host cell. Indeed, studies on heterologous expression

in *E. coli* show that less than half of proteins from prokaryotes and one fifth from eukaryotes can be expressed in a soluble form as full-length proteins [7].

In such circumstances researchers often turn to alternative expression hosts, often closer to the original organism of the protein of interest [8], such as other bacterial systems (e.g., *Bacillus* [9] and *Lactococcus* [10]), or eukaryotic systems (e.g., baculovirus/insect cells [11] and protozoa [12]). Furthermore, a wide range of solubility-enhancing and affinity fusion tags have also been successfully applied to heterologous expression systems, such as GST, MBP and thioredoxin [13]. Different levels of expression between fusion tags and target proteins in comparative screens however suggest the necessity of screening multiple tags [14].

Eukaryotic proteins are often comprised of modular structures of defined, folded domains, linked by flexible or unstructured stretches of sequence. Protein domains are thought to fold independently, exhibit globularity (e.g., contain a hydrophobic core and hydrophilic exterior) and perform a specific function (e.g., binding), such that the combination and juxtapositioning of domains determines overall protein function [15]. There is a long-held premise that well-ordered or compact domains or fragments will yield better-behaving proteins than full-length proteins for protein expression and structural studies, in relation to solubility and crystallization potential [7]. For instance, rigid proteins have a greater propensity to crystallize than flexible or highly disordered proteins [5], resulting from increased flexibility either between domains in multi-domain proteins, or from within domains (e.g.. unstructured N- or C-termini or internal loops) entropically hampering crystallization [16]. Furthermore, many proteins exist in complexes with other partners, exhibiting poor expression or solubility when expressed alone and/or in alternative hosts due to, for example, the exposing of hydrophobic patches that the interacting partner normally protects [17]. This may occur even if such regions are localized to a single domain.

Therefore, delineation of independent, folded and compact protein domains for expression as individual units is a key tool in protein and structural biochemistry. Significant attempts have been undertaken to predict optimal protein constructs for expression, many of which involve multiple truncations of full-length proteins from either, or both, the N- and C-termini to express individual domains [7]. Parallel analysis of multiple domains and domain fragments has been simplified with the advent of high-throughput cloning and expression/purification methods [18]. Iterative but random trial and error approaches towards construct N- or C-terminal truncation however can be costly and time-consuming.

A more informed approach, which we call 'domain boundary analysis' or DBA, involves the interrogation of multiple bioinformatics methods to predict protein structural features. This targeted approach to delimit protein domain boundaries and their subsequent combinatorial arrangement is more likely to result in ordered, defined and globular protein fragments [6, 19]. DBA has been very successful in our hands, with nearly half of human proteins attempted being successfully expressed and purified, and around 20 % of those attempted resulting in a solved high-resolution X-ray structure [1]. Here we take the reader through practical usage of a range of common bioinformatics approaches used in DBA, towards defining well-behaving protein domains for biochemical and structural analysis.

2 Materials

All analyses described here can be performed on any standard PC, Mac OS X or Linux-based operating system on a standard desktop or laptop computer with an internet connection. Most common web browsers (Explorer, Safari, Chrome etc.) work with the bioinformatics servers described. Many of the platforms described can be downloaded and installed locally on Linux-based systems or incorporated into bespoke web services, but we are restricting our descriptions to individual web-based analyses for ease of use. The sole requirement from the user is the protein sequence of interest, with residues represented in the IUPAC single letter code format [20]. In a minority of cases, it may be necessary to provide the sequence in

FASTA format [21] which can be facilitated by the simple addition of an identifier (name) preceded with the character ">", required as the first and separate line in the sequence:

```
>sequence_name
MTGHYTHHAYGRETYIPSDFGNMKILPSSWQ
```

Protein three-dimensional structure visualization can be performed also using web-based software or via software that is either provided specifically for an operating system (e.g., Windows, OS/X, Linux) or in an independent form using a platform such as Java.

3 Methods

Our approach to defining construct boundaries by DBA utilizes a range of common bioinformatics approaches, all freely available online. A hierarchical approach is taken to define boundaries (**Fig. 1**), initially identifying domains using a combination of homology-based and Hidden Markov Model (HMM) approaches, supplemented by disorder prediction to suggest protein globularity, a reliable indicator of folded domains. Once potential domains are identified, multiple finer-grained boundaries are defined using predicted secondary structural elements as termini, again supplemented with disorder propensity information. Sequence and structural homology information can further supplement to help guide the determination of likely soluble or crystallizable protein boundaries.

[Fig 1 near here]

Parallel testing of multiple constructs with different domain boundaries can increase experimental success (**Fig. 2**). [1]. Our DBA approach is designed to be used in conjunction with Ligation-Independent Cloning (LIC) or other high-throughput cloning methods to construct N- and C-terminal tagged fusions, combined with small-scale parallel expression in multiple systems (*E. coli*, baculovirus-infected insect cells) [1, 7, 18, 22]. The number of

domain boundaries attempted is determined by the researcher in relation to resources and time available but, from our experience, 12-40 constructs per domain is typical, normally matched to multiple domain-defining secondary structural elements [1]. If multiple tandem domains are present, the respective N- and C-terminal boundaries can also be combined for multiple-domain constructs (**Fig. 2**). In addition, it is also worth attempting the full-length protein itself in expression trials, perhaps with multiple small N- and C-terminal DBA-defined truncations.

[Fig 2 near here]

3.1 Prediction of Protein Secondary Structure and Domains using Sequence and Structural Homology

Since the concept of the "domain hypothesis", a number of experimental and *de novo* computational/statistical methods have been used to attempt to predict protein domain boundaries [15]. The simplest approach to assign boundaries however, is often by similarity to previously defined domains. Hence, the approach we take for DBA uses a number of complementary approaches, either based on direct sequence-based homology (BLAST [23], Conserved Domain Database (CDD) [24]), or profile HMM-based approaches (SMART [25], PFAM [26]). The CDD is a database of annotated multiple sequence alignments, allowing alignment of query sequences to previously detected or characterized domains. The HMM-based SMART and PFAM databases provide a complementary, but often more sensitive, detection of domains including many not found in the CDD, alongside a number of predicted but uncharacterized "Domains of Unknown Function" (DUFs). These approaches are particularly useful to identify "core" domain regions, the precise boundaries of which can be subsequently explored with disorder/secondary element prediction tools described later.

Where strong sequence homology to existing characterized domains may not exist, predicted secondary structure (PSIPRED [27]) and homologies both to close (BLAST/Protein Data

Bank (PDB) [28]) and remote structural templates (pGenTHREADER [29]) can potentially be identified, to guide construct termini design.

3.1.1 Domain Prediction using Homology Searching: BLAST and the CDD

1. Navigate to the NCBI BLAST server web interface (http://blast.ncbi.nlm.nih.gov/Blast.cgi) [23].
2. Select the "protein blast" program, in the Basic BLAST section to open the standard BLAST interface to the blastp algorithm.
3. Copy and paste the full-length query sequence in FASTA or simple text sequence format (or the NCBI protein accession code) into the query box, or select "Choose File" and navigate to the respective file, if the sequence is saved as a text file (*see* **Note 1**).
4. Select the database to be searched from the dropdown menu of the Database option of the Choose Search Set section. Choose "Protein Data Bank proteins (pdb)" to search within potential homologous structures (*see* **Note 2**).
5. The BLAST search can be optionally taxonomically limited should the user require, by starting to type either the common or Latin species/taxon name into the Organism field (e.g., *Homo sapiens*). On typing, taxon options pop up, and select the most relevant one (*see* **Note 3**).
6. Leave the algorithm and general parameters as default for blastp (protein-protein BLAST), with BLOSUM62 matrix and gap parameters as 11/1 (*see* **Note 4**).
7. Press the blue "BLAST" button to run the search.
8. Once the search is complete, the results are graphically displayed as an overview distribution of BLAST hits mapped onto the query sequence (**Fig. 3a**). The color represents the homology between query sequence and identified sequence, with red matches as closest and the longest significant match at the top of the matched sequences (color key is above at the top of the distribution image). Multiple matched regions represent the presence of multiple domains in the query sequence.

9. Select a match on the distribution image to automatically scroll down the page to respective alignment HSP report (**Fig. 3b**), representing a homologous sequence for which a protein structure is present in the PDB database (*see* **Note 5**). The corresponding aligned residue positions of the query and match ("Sbjct") are displayed flanking the alignment.
10. Click on the link beginning "pdb" next to "Sequence ID" in the HSP report to access the corresponding protein structure information, linking to the PDB structure file.
11. The query sequence is also searched against the CDD [24] with the graphical output arranged above the distribution report (top frame, **Fig. 3c**). This displays CDD matches and also strong matches from the SMART and PFAM databases (**see section 3.1.2**). Click on the CDD output image to open a new browser window with the same graphical display and an additional detailed list of matched domains (lower panel, **Fig. 3c**), detailing the boundary regions of the query that matches the domain ("interval") and E-value match significance (*see* **Note 6**).
12. Position the mouse pointer over the domain image in the CDD graphical output, whereby a popup window appears with available biological information (right side window in top frame, **Fig. 3c**). Alternatively, click on the "+" of a domain in the list to expand the list to provide biological descriptions, with an alignment of the query sequence against the consensus for this domain (lower panel, **Fig. 3c**), with the boundaries shown flanking the alignment. Minimize the expansion by clicking "-".

The results from CDD analyses help identify and define domain boundaries (contributing to **step A** of DBA, **Fig. 1**), with BLASTP searches identifying close structural homologues (A, **Fig. 1**). CDD and HSP local sequence alignments help to identify consensus residue positions that might indicate domain boundaries (**steps A and C**, **Fig. 1**).

[Fig 3 near here]

3.1.2 Domain Prediction with HMM Databases: SMART and PFAM

1. Navigate to the SMART webserver (http://smart.embl-heidelberg.de) [25].
2. At the top of the web interface, ensure the SMART mode is set to "NORMAL" and the webpage displays a query box. If not, click on the "NORMAL" link in the "SMART mode" box. Paste the full-length protein sequence into the query box, ensuring all search options are selected in the Sequence Analysis pane (*see* **Note 7**).
3. Run the analysis by selecting the "Sequence SMART" button.
4. SMART output displays a graphical representation of recognized domains from the SMART database, with an approximate residue scale bar (**Fig. 4a**). Mouse over the domain representation to pop-up the residue positions and significance of the match (**Fig. 4a**).
5. If search options were selected (**this section**, **point 2**) domains not present in SMART may be recognized e.g., PFAM and transmembrane (TM) regions (**Fig. 4b**, *see* **Note 8**).
6. Click the domain in the graphical output to link to detailed domain information (**Fig. 4c**).
7. Click on the "Align your sequence against the SMART alignment" button, to generate a similar alignment to the consensus sequence as performed with the CDD software (**section 3.1.1**, **point 12**).

Results from SMART/PFAM searches may identify both characterized and predicted (DUF) domains, with consensus alignments helping delineate domain boundaries (**steps A** and **C**, **Fig. 1**), similar but often more sensitive than CDD (**section 3.1.1**). In addition SMART/PFAM also predict low-complexity sequences (often disordered, **section 3.2**), used in **step B** (**Fig. 1**), (*see* **Note 9**).

[Fig 4 near here]

3.1.3 The PSIPRED Workbench for Protein Domain and Secondary Structure Prediction

PSIPRED [27] and pGenTHREADER [29, 30] are part of the UCL PSIPRED suite of tools

[31], for protein fold and secondary structure prediction (http://bioinf.cs.ucl.ac.uk/psipred/) (*see* **Note 10**). The advantage of this server is that multiple algorithms may be run simultaneously from a single query sequence submission. PSIPRED is amongst the most accurate predictors of protein secondary structural elements, critical for the DBA procedure described here, and in more detail in **section 3.3**. Like BLAST searches of the PDB database (**section 3.1.1**.), pGenTHREADER is particularly useful to find PDB templates for structural considerations in DBA (**section 3.3**), but has the advantage of using PSI-BLAST and threading methods to help determine remote structural homologies (*see* **Note 11**) [32], increasing sensitivity compared with BLAST in our hands.

1. In the web interface, select PSIPRED and pGenTHREADER and paste the protein sequence into the "Input Sequence" window as FASTA or raw sequence format (*see* **Note 12**). Multiple sequences may also be posted.
2. Enter a valid email address in "Submission Details" pane (recommended, *see* **Note 13**) and click "Predict" to run the analysis.
3. Once the submission is complete, the results page (**Fig. 5a**) displays results from different algorithms in different tabs, with the option to download the results (see respective tab) as text or printable PostScipt/PDF files.
4. For pGenTHREADER, click on the respective tab, bringing up a hierarchical display of homologous sequence hits relating to the query sequence (*see* **Note 14**). Click the links under SCOP/CATH codes, CATH entry or on the structure image itself to link to structural information from the SCOP [33], CATH [34] or PDBsum [35] databases.
5. Select the link under "View Alignment" to open a window displaying a structural alignment of the query sequence to the respective match (**Fig. 5b** and *see* **Note 15**).
6. The pGenTHREADER uses a PSIPRED secondary structure prediction in its operation, and full results can be seen or downloaded from the respective results tab (**Fig. 5a**).
7. Raw PSIPRED results (**Fig. 5c**) give a useful graphical superimposition of secondary structural elements on the protein sequence, with a degree of confidence (blue bars).

These secondary elements will determine the exact construct boundaries in the DBA process, described in **section 3.3**.

8. As there is a threshold for query sequence length in PSIPRED, multiple overlapping analyses should be performed where appropriate (*see* **Note 12**).

pGenTHREADER matches thus help identify homologous domains (**step A**, **Fig.1**) and along with resulting PSIPRED predictions, help identify secondary structural elements and fine domain boundaries (**steps C** and **D** respectively, **Fig. 1**).

[Fig 5 near here]

3.2 Protein Domain Identification using Globularity and Disorder Prediction

The methods described for domain identification have so far been based on prior experimental data, often as a consequence of advances in genome sequencing and structural genomics. That is, identifying protein domains using previously identified related or homologous domains using HMM or alignments, or from structural homology to previously solved structures of proteins. However, in order to delineate domains which lack well-defined annotation in the literature, unbiased techniques are required. It is well known that protein domains are usually made up of globular well-ordered cores of secondary structure, with inter-domain linkers often disordered [36]. Here we describe the use of the FoldIndex [37] and GlobPlot 2 [38] webservers which provide complementary approaches to predict order (globularity) to define domain boundaries and regions of proteins that may negatively influence protein crystallization.

3.2.1 Disorder Analysis with FoldIndex

1. Paste the protein sequence directly into the “Sequence area” window of the FoldIndex webserver (http://bioportal.weizmann.ac.il/fldbin/findex) [37].

2. Default parameters are advised for the sequence window and step, but enable the "graph Phobic values" and "graph charge values" options (*see* **Note 16**).
3. Select the "Process" button to run the analysis.
4. Predicted folded (ordered, green) and unfolded (disordered, red) regions are graphically displayed, mapped to residue position (**Fig. 6a**), alongside hydrophobic or charged regions if previously selected. This image may be saved as a PNG file.
5. Alongside prediction statistics, (dis)order predictions are mapped onto the primary sequence in the output window (**Fig. 6b**), allowing (dis)order to be mapped onto the sequence (*see* **Note 16**).

3.2.2 Disorder Analysis with GlobPlot

1. Paste the protein sequence directly into the "Sequence" window of the GlobPlot 2 webserver (http://globplot.embl.de) [38].
2. Default parameters are advised, but otherwise enable the "Russell/Linding" disorder propensity option and the "Perform SMART/Pfam domain prediction" options (*see* **Note 17**).
3. Select the "GlobPlot NOW!" button to run the analysis.
4. As with FoldIndex (**section 3.2.1**), ordered/disordered regions are mapped onto the protein primary sequence (**Fig. 6c**), in this case green/black respectively (*see* **Note 18**). In addition, predicted ordered sequences ('GlobDoms") are listed above the sequence.
5. Graphical results (which can also be downloaded in PostScript format) display predicted globularity/disorder as green/blue blocks respectively, alongside residue number (**Fig. 6d**). Disorder propensity is plotted as a white line, with downhill and uphill regions corresponding to predicted globular regions or disorder, respectively.
6. Predicted SMART/PFAM domains are superimposed onto this plot according to the included key, allowing simple combination of *de novo* globularity and HMM approaches.

FoldIndex and GlobPlot approaches thus help identify globular regions, towards identification of (sub)-domains (**step A**, **Fig. 1**) and disordered termini (**step B**, **Fig. 1**), in the domain boundary analysis hierarchy.

[Fig 6 near here]

3.3 Combining Bioinformatics Approaches for Domain Boundary Prediction

Once bioinformatics analyses have been completed, results should be combined cohesively as part of the DBA process. **Figure 1** demonstrates the overall DBA workflow, and the contribution of each bioinformatics tool to the process. Most aspects of the procedure have been duplicated with multiple algorithms, increasing the accuracy of domain boundary prediction. Important considerations are illustrated using human POLQ (DNA polymerase θ, UniProt ID: O75417) as an example (**Fig. 7**) [39].

1. Alignment and HMM-based approaches identify predicted domains by homology (**A**, **Fig. 1**), with improved confidence conferred if multiple servers predict domains in the same sequence neighborhood (e.g., PFAM:DEAD and SMART:DEXDc domains, **Fig. 7a**). Additional non-HMM domains (e.g., "BLAST", **Fig. 7a**) should also be taken into account, even if only found by a single algorithm. Low-complexity sequences are found at the extreme ends of the 1-900aa region and are recommended not to be included in designed constructs (**B**, **Fig. 1**). In this example, the analysis suggests 2-3 domains in POLQ from ~80 to 550 residues.
2. Disorder prediction with both GlobPlot2 and FoldIndex suggests the protein is predominantly globular up to 900aa (**step A**, **Fig. 1** and **Fig. 6**). Biologically inferred data from the most homologous structure (*Archaeoglobus fulgidus* HEL308, found from both BLAST searches to the PDB database and pGenTHREADER), suggests that the entire region from ~70-850aa is globular from its expression and structural determination, hence

the HMM-derived domains such as SMART:DEXDc are likely to be sub-domains (**A**, **Fig. 1**) (*see* **Note 19**).

3. Domain boundaries can in principle focus on the sub-domains, but examination of homologous structures (**Fig. 7b**) suggest that if this was the case, significant biological information would be lost (*see* **Note 20**). Here, the expected substrate (an ATP analogue) is bound between the RecA sub-domains (green/yellow) corresponding to the two predicted PFAM/SMART sub-domains in **Fig. 7a**. Hence, the more biologically relevant domain boundaries should span these two sub-domains. Furthermore, a cryptic domain not detected in HMM-based searches can only be noted by comparison to the homologous HEL308 structure, seen here in the final POLQ structure (helix-hairpin-helix, red in **Fig. 7b**). Hence, analysis of sequence similarity in homologous protein structures can yield important information in addition to sequence-based HMM searches (**step A**, **Fig. 1**).
4. Co-localization of domains to the same region of sequence may have different local boundaries (e.g., PFAM:DEAD 93-274aa and SMART:DEXDc 88-299aa). In such cases we recommend using the longer of the two regions if within 10-20 residues as the boundary (*see* **Note 9**).
5. Once approximate domain boundaries are predicted, use PSIPRED secondary structure predictions to delineate secondary elements as the next level of construct boundary, serially expanding the boundaries in both directions one element at a time (**step C**, **Fig. 1**). It is important to compare PSIPRED predictions to the actual elements in homologous determined structures e.g., with the structural alignment output of pGenTHREADER (*see* **Note 21**), to avoid bisecting secondary structural elements.

[Fig 7 near here]

6. If homologous structures are found from BLAST or pGenTHREADER searches, PSIPRED secondary element predictions should be compared to those in the known structure in case removing a specific element destabilizes the protein (*see* **Notes 22, 23**).

7. The final stage of DBA is to choose the residue positions to determine the precise construct boundaries (**step D**, **Fig. 1**). It is critical that full secondary elements are considered when determining the termini of boundaries, e.g., in this example the first α-helix as a boundary should begin at GRCLK (**Fig. 5c**). If resources allow, a further boundary should be designed by addition of a small amount of coil/non-element structure, e.g., GLGRCLK (**Fig. 5c**). Close additional boundaries may be useful, as such regions are often not structured in crystals and the true secondary element may in fact comprise this additional sequence, amongst other factors (*see* **Note 24**).

[Fig 8 near here]

3.4 Further Methods for Domain Boundary Analysis: Beyond Bioinformatics

The DBA approach we have outlined here to delineate protein domains, is designed to be used in conjunction with high-throughput parallel cloning and expression methods, as described earlier [1]. *E. coli* systems are predominantly used in initial expression screening, moving to baculovirus-mediated insect cell expression if not successful. Although such approaches frequently lead to respectable success rates in small scale tests (**Fig. 8**) [1], re-iteration of the DBA procedure may be required for protein expression optimization for difficult targets. Analogous approaches have been attempted by others, often bringing together similar bioinformatics approaches but in automated pipelines, such as ProteinCCD [19], or by our colleagues at the Structural Genomics Consortium [6]. However for small-scale domain prediction, the use of individual bioinformatics tools allows the user a great deal of analytical flexibility, depending on the protein in question.

A range of experimental data may also be applied to protein domain delineation. If full length protein is available, limited proteolysis combined with mass spectrometric (MS) approaches can determine core folded domains, as connecting unfolded sequence or disordered termini may be trimmed away by proteases, with core domains identified by MS [40]. In addition, the

advent of powerful high-throughput screening of random or combinatorial protein truncation or mutation libraries allows an unbiased approach with no prior knowledge required [41]. Rather than replacing bioinformatics approaches to domain boundary analysis, these experimental techniques may facilitate the accuracy of domain prediction for difficult proteins, especially if used in combination with *in silico* approaches described here.

4 Notes

1. Single or lists of multiple sequences can also be entered in this manner. Sub-sequences may be selected in the "Query subrange" box.
2. The full NCBI protein sequence database can be searched instead if homologous structures are not required or found, by selecting the "Non-redundant protein sequences(nr)" dropdown option.
3. We normally leave the "Organism" option blank, to give the greatest chance of finding a close homologue.
4. blastp algorithm parameters can be changed if using protein sequences with few close homologues, but we find default parameters are adequate for most sequences, especially for mammalian proteins.
5. HSP (High-scoring Segment Pair) is the alignment of the query to database sequence, generally representing a single domain. However multiple HSPs may be present within a domain if variable intervening sequences are present (e.g., loop regions or low-complexity sequences). Significance of matches ("Expect" or "E-value") is greater the smaller the number, with zero being most significant. The length of the match (both for identity and similarity ("positives") is also displayed.
6. Expect (E)-values are an estimate of the significance of a BLAST match i.e. the number of hits expected by chance in a particular database. Hence, the lower the number and closer to zero the E-value, the more significant the match e.g., $1e^{-6}$ is a good starting point for a significant hit.

7. Optional tick boxes engage additional database searching, including PFAM [26], membrane protein signal sequences [42], repeats and outlier homologues.
8. Identification of TM regions is beneficial, as following their high hydrophobicity, their removal increases the likelihood of soluble protein domain expression.
9. IMPORTANT: CDD/SMART/PFAM methods and domain definitions are very conservative, often defining domains as core regions and hence removing surrounding regions that may in fact be true domain boundaries. Hence, if multiple methods coincide with approximate boundaries, the longest prediction should be used. Furthermore, predicted secondary structural elements (**section 3.1.3**) around these predicted domain boundaries should extend away from, rather than into these regions, in order to prevent shortened and therefore erroneous domain boundary predictions.
10. Additional software, useful for construct design and run simultaneously, are available in the PSIPRED workbench package [31], particularly for transmembrane helix and topology prediction (e.g., MEMSAT3/MEMSATSVM) and additional orthogonal disorder prediction (DISOPRED3), but out of the scope of these protocols.
11. Although pGenTHREADER is useful for detecting remote structural homologies in the case of low sequence similarity, care should be taken in the interpretation of, or using such remote homologies, as false positive hits may be prevalent with some hits bearing no real functional similarity.
12. An upper sequence length limit of 1500 residues exists for PSIPRED workbench servers. Hence longer proteins should be broken down into shorter fragments for submission, ideally not comprising multiple domains. These should be arranged to tile of fragment predictions with 200-500 residue overlaps, to ensure that positioning at fragment ends does not influence prediction accuracy.
13. The PSIPRED workbench algorithms are computationally intensive and may take up to two hours to run, hence it is recommended to supply an email address for delivery of a weblink to results.

14. The color code on the left panel for pGenTHREADER results (**Fig. 5a**) gives a rapid idea of match confidence, with green being firm hits, followed by orange then yellow (weak). Orange/weak hits should only be used if green and confident matches are not found, suggesting only remote structural homology has been found.
15. pGenTHREADER structural alignments are especially useful when only remote homologies are matched to query sequences, guiding alignment on the basis of (predicted) structure, rather than potentially biased or misguided poor sequence similarity. In such circumstances, the use of multiple weak/average matches should be used to reduce bias in PDB template choice.
16. Graphing the hydrophobic and charged regions in FoldIndex gives further information to solubility propensity i.e. hydrophobic/charged regions are likely to negatively/positively influence protein solubility respectively.
17. The SMART/PFAM search is useful in GlobPlot, superimposing HMM-based domain searches (**section 3.1.2**) onto globularity/disorder predictions and the query sequence.
18. Copying the colored alignment from FoldIndex and GlobPlot and pasting into word processing or text editing software with the "Courier" font preserves text formatting and spacing for useful documentation.
19. It should be noted that although a stretch of protein may be predicted to be (globally) globular, it could in fact comprise a string of local globular domains with very small linkers that do not show up in disorder prediction.
20. Many protein structure visualization platforms may be freely downloaded, and although this is out of the scope of this chapter, the authors recommend Chimera (cgl.ucsf.edu/chimera/) [43] or PyMOL (pymol.org).
21. If only remote homologues exist, such structural alignments in pGenTHREADER will considerably increase the accuracy of secondary element prediction.
22. Removing specific secondary structural elements could expose significant regions of hydrophobicity (or remove favorable charged regions), both of which could diminish protein solubility.

23. In parallel β-sheets in particular, the strand arrangement from one side to another does not necessarily follow the N- to C-terminal order. Hence, removal of the most N-terminal strand could destabilize a whole β-sheet if juxtaposed centrally in the β-sheet, with increased likelihood of protein insolubility (e.g., removal of N-terminal β1 or β2 in a POLQ would split the β-sheet, **Fig. 7c**).
24. Terminal residue composition may influence protein expression [44], hence a range of alternative but close boundaries may be beneficial. Even if soluble protein is produced, some terminal residues may negatively influence crystal packing, e.g., PPPGLGRCLK (**Fig. 5c**) may cause a sharp N-terminal kink increasing disorder or decrease potential packing, due to the high proline content.

Acknowledgements

The SGC is a registered charity (number 1097737) that receives funds from AbbVie, Bayer Pharma AG, Boehringer Ingelheim, the Canada Foundation for Innovation, Genome Canada, GlaxoSmithKline, Janssen, Lilly Canada, Merck & Co., the Novartis Research Foundation, the Ontario Ministry of Economic Development and Innovation, Pfizer, São Paulo Research Foundation-FAPESP, Takeda, and the Wellcome Trust [092809/Z/10/Z]. C.D.O.C. thanks the University of Huddersfield for support.

References

1. Savitsky P, Bray J, Cooper CD et al (2010) High-throughput production of human proteins for crystallization: the SGC experience. J Struct Biol 172: 3-13

2. Mesa P, Deniaud A, Montoya G et al (2013) Directly from the source: endogenous preparations of molecular machines. Curr Opin Struct Biol 23: 319-325

3. Makrides SC (1996) Strategies for achieving high-level expression of genes in Escherichia coli. Microbiol Rev 60: 512-538

4. Terpe K (2006) Overview of bacterial expression systems for heterologous protein production: from molecular and biochemical fundamentals to commercial systems. Appl Microbiol Biotechnol 72: 211-222

5. Dale GE, Oefner C, D'Arcy A (2003) The protein as a variable in protein crystallization. J Struct Biol 142: 88-97

6. Sagemark J, Kraulis P, Weigelt J (2010) A software tool to accelerate design of protein constructs for recombinant expression. Protein Expr Purif 72: 175-178

7. Graslund S, Sagemark J, Berglund H et al (2008) The use of systematic N- and C-terminal deletions to promote production and structural studies of recombinant proteins. Protein Expr Purif 58: 210-221

8. Fernandez FJ, Vega MC (2013) Technologies to keep an eye on: alternative hosts for protein production in structural biology. Curr Opin Struct Biol 23: 365-373

9. Zweers JC, Barak I, Becher D et al (2008) Towards the development of Bacillus subtilis as a cell factory for membrane proteins and protein complexes. Microb Cell Fact 7: 10

10. Morello E, Bermudez-Humaran LG, Llull D et al (2008) Lactococcus lactis, an efficient cell factory for recombinant protein production and secretion. J Mol Microbiol Biotechnol 14: 48-58

11. Mahajan P, Strain-Damerell C, Gileadi O et al (2014) Medium-throughput production of recombinant human proteins: protein production in insect cells. Methods Mol Biol 1091: 95-121

12. Fernandez-Robledo JA, Vasta GR (2010) Production of recombinant proteins from protozoan parasites. Trends Parasitol 26: 244-254

13. Esposito D, Chatterjee DK (2006) Enhancement of soluble protein expression through the use of fusion tags. Curr Opin Biotechnol 17: 353-358

14. Hammarstrom M, Hellgren N, van Den Berg S et al (2002) Rapid screening for improved solubility of small human proteins produced as fusion proteins in Escherichia coli. Protein Sci 11: 313-321

15. Ingolfsson H, Yona G (2008) Protein domain prediction. Methods Mol Biol 426:117-143

16. Derewenda ZS (2010) Application of protein engineering to enhance crystallizability and improve crystal properties. Acta Crystallogr D Biol Crystallogr 66: 604-615

17. Gopal GJ, Kumar A (2013) Strategies for the production of recombinant protein in Escherichia coli. Protein J 32: 419-425

18. Gileadi O, Burgess-Brown NA, Colebrook SM et al (2008) High throughput production of recombinant human proteins for crystallography. Methods Mol Biol 426: 221-246

19. Mooij WT, Mitsiki E, Perrakis A (2009) ProteinCCD: enabling the design of protein truncation constructs for expression and crystallization experiments. Nucleic Acids Res 37: W402-405

20. IUPAC-IUB Commission on Biochemical Nomenclature. A one-letter notation for amino acid sequences. Tentative rules (1969) Biochem J 113: 1-4

21. Lipman DJ, Pearson WR (1985) Rapid and sensitive protein similarity searches. Science 227: 1435-1441

22. Keates T, Cooper CD, Savitsky P et al (2012) Expressing the human proteome for affinity proteomics: optimising expression of soluble protein domains and in vivo biotinylation. N Biotechnol 29: 515-525

23. Altschul SF, Gish W, Miller W et al (1990) Basic local alignment search tool. J Mol Biol 215: 403-410

24. Marchler-Bauer A, Derbyshire MK, Gonzales NR et al (2015) CDD: NCBI's conserved domain database. Nucleic Acids Res 43: D222-226

25. Schultz J, Milpetz F, Bork P et al (1998) SMART, a simple modular architecture research tool: identification of signaling domains. Proc Natl Acad Sci U S A 95: 5857-5864

26. Finn RD, Coggill P, Eberhardt RY et al (2016) The Pfam protein families database: towards a more sustainable future. Nucleic Acids Res 44: D279-285

27. Jones DT (1999) Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol 292: 195-202

28. Rose PW, Prlic A, Bi C et al (2015) The RCSB Protein Data Bank: views of structural biology for basic and applied research and education. Nucleic Acids Res 43: D345-356

29. Lobley A, Sadowski MI, Jones DT (2009) pGenTHREADER and pDomTHREADER: new methods for improved protein fold recognition and superfamily discrimination. Bioinformatics 25: 1761-1767

30. Jones DT (1999) GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences. J Mol Biol 287: 797-815

31. Buchan DW, Minneci F, Nugent TC et al (2013) Scalable web services for the PSIPRED Protein Analysis Workbench. Nucleic Acids Res 41: W349-357

32. McGuffin LJ, Jones DT (2003) Improvement of the GenTHREADER method for genomic fold recognition. Bioinformatics 19: 874-881

33. Murzin AG, Brenner SE, Hubbard T et al (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 247: 536-540

34. Sillitoe I, Lewis TE, Cuff A et al (2015) CATH: comprehensive structural and functional annotations for genome sequences. Nucleic Acids Res 43: D376-381

35. Laskowski RA, Hutchinson EG, Michie AD et al (1997) PDBsum: a Web-based database of summaries and analyses of all PDB structures. Trends Biochem Sci 22: 488-490

36. Dosztanyi Z, Tompa P (2008) Prediction of protein disorder. Methods Mol Biol 426: 103-115

37. Prilusky J, Felder CE, Zeev-Ben-Mordehai T et al (2005) FoldIndex: a simple tool to predict whether a given protein sequence is intrinsically unfolded. Bioinformatics 21: 3435-3438

38. Linding R, Russell RB, Neduva V et al (2003) GlobPlot: Exploring protein sequences for globularity and disorder. Nucleic Acids Res 31: 3701-3708

39. Newman JA, Cooper CD, Aitkenhead H et al (2015) Structure of the Helicase Domain of DNA Polymerase Theta Reveals a Possible Role in the Microhomology-Mediated End-Joining Pathway. Structure 23: 2319-2330

40. Gao X, Bain K, Bonanno JB et al (2005) High-throughput limited proteolysis/mass spectrometry for protein domain elucidation. J Struct Funct Genomics 6: 129-134

41. Hart DJ, Tarendeau F (2006) Combinatorial library approaches for improving soluble protein expression in Escherichia coli. Acta Crystallogr D Biol Crystallogr 62: 19-26

42. Petersen TN, Brunak S, von Heijne G et al (2011) SignalP 4.0: discriminating signal peptides from transmembrane regions. Nat Methods 8: 785-786

43. Pettersen EF, Goddard TD, Huang CC et al (2004) UCSF Chimera--a visualization system for exploratory research and analysis. J Comput Chem 25: 1605-1612

44. Bivona L, Zou Z, Stutzman N et al (2010) Influence of the second amino acid on recombinant protein expression. Protein Expr Purif 74: 248-256

Figure Captions

Fig. 1. Representation of the hierarchical approach to domain boundary analysis. The workflow is shown by boxed rectangles (A to D) connected by solid black arrows. The involvement of bioinformatics tools at various pipeline stages (dark gray boxes, grouped by type of method (rounded light gray boxes)), is represented by grey arrows. Dashed gray arrows represent iteration of secondary element/fine boundary redesign following cloning and protein test expression, where necessary. Abbreviations: *p-HMM*, profile-Hidden Markov Model; *MSA*, Multiple Sequence Alignment; PDB, Protein Data Bank.

Fig. 2. Representation of domain boundary analysis. Individual domains in a full-length protein sequence are identified (blue/orange), then combinatorial sets of N- and C-terminal truncations are made. Constructs containing tandem domains (red) may also be used.

Fig. 3. Screenshot from NCBI BLAST output using the human POLQ protein as input to search against the PDB database. (a) Distribution of BLAST hits mapped onto the input sequence, color coded for strength of alignment. (b) Detailed BLAST HSP alignment. (c) CDD output (top frames, domain annotations with example pop up window for cd06140 CDD

entry; lower frames, domain lists with example expansion showing input sequence alignment against CDD consensus).

Fig. 4. Screenshot from SMART output, using human POLQ protein as input. (a) Graphical output showing recognized SMART domain, with popup window on mouse over. (b) Graphical output showing recognized transmembrane region (blue) and PFAM domain, with popup window on mouse-over. (c) Expansion on clicking SMART domain from Fig. 4(a).

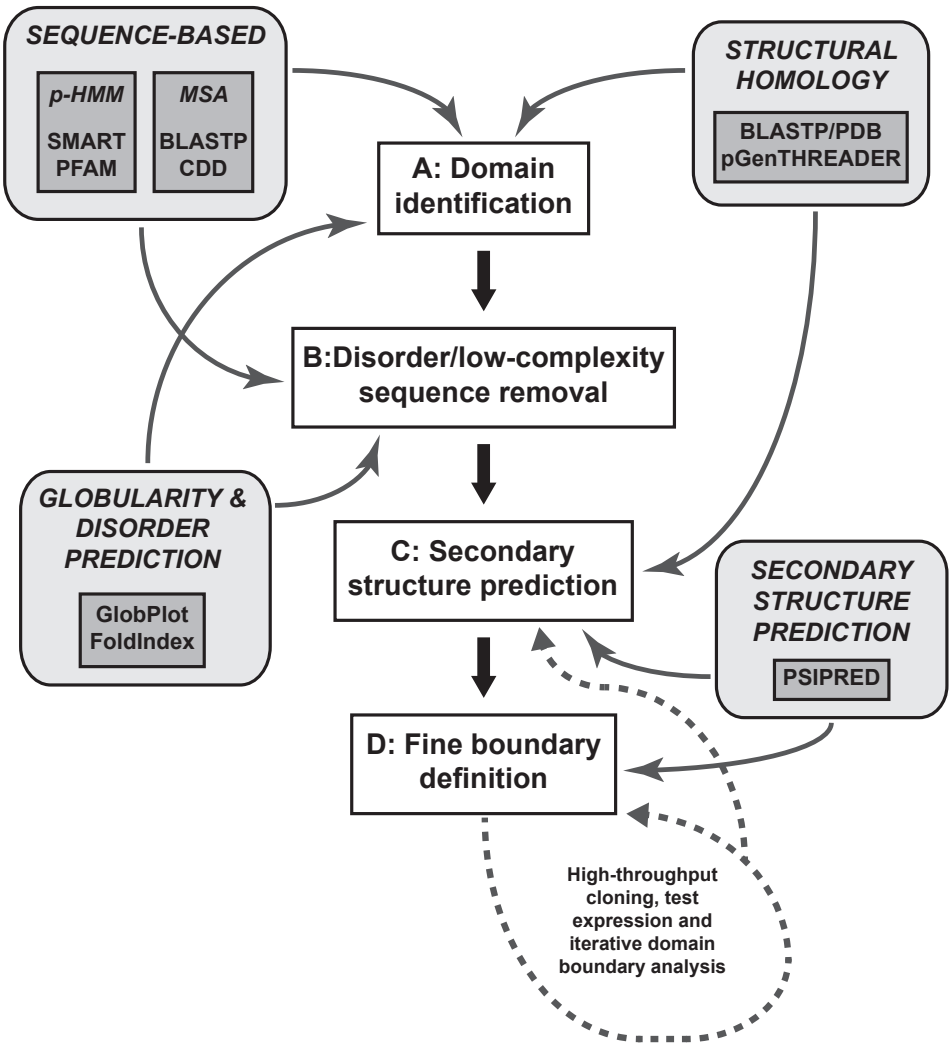
Fig. 5. Screenshots of graphical outputs from the PSIPRED suite of programs. (a) pGenTHREADER table output, with most identical/homologous sequence ranked highest (lowest p-value is most significant), with high confidence hits in green (medium in orange and weak in red, not shown). (b) Structural alignment output following selection of "View Alignment" in (a). Predicted or structurally determined α-helices (purple) and β-strands (yellow) are mapped onto query and matched sequences, respectively. (c) Detailed PSIPRED output for query sequence with same color scheme as for (b), with secondary elements definitions: C, coil, H, α-helix, E, β-strands, and "Conf" representing prediction confidence.

Fig. 6. Output from FoldIndex and GlotPlot servers, using residues 1-1500 or full length human POLQ as a query sequence, respectively. (a) FoldIndex PNG file graphical output, with green and red regions as folded/unfolded respectively. Hydrophobic and charge propensity are plotted as blue and pink traces respectively. (b) FoldIndex output screenshot with predicted ordered/disordered regions plotted onto the query sequence as green/red text respectively. (c) GlobPlot output screenshot with predicted globular/disordered regions plotted onto the query sequence as green capitalized/black small case text respectively. (d) GlobPlot graphical output for full-length POLQ as query sequence. Globular domains are green blocks, disordered regions as blue blocks and recognized SMART domains according to the key. Disorder propensity is plotted as the white line, described in the main text.

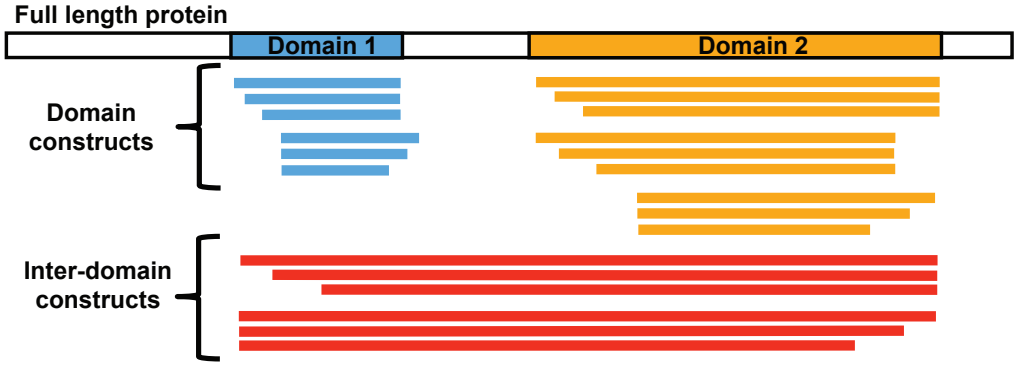
Fig. 7. Considerations in domain boundary analysis. (a) Representation of PFAM and SMART detected domains mapped to the first 1000 residues of human POLQ (base image generated by SMART server [25]). Numbers in parentheses denotes predicted domain boundaries from respective analyses, with low-complexity regions in purple. The closest structure homologue is PDB:2P6R *A. fulgidus* HEL308. (b) (Sub) domain crystallized structure of human POLQ (~residues 70-900, PDB:5AGA [39]), showing RecA and helix-hairpin-helix subdomains rendered in green/yellow and red, respectively. (c) Parallel β-sheet from human POLQ structure showing non-contiguous β-strand arrangement, with strands numbered from N- to C-terminus (β1 to β7). Images in (b) and (c) were rendered with Chimera [43].

Fig. 8. Typical small-scale protein expression screening. SDS-PAGE analysis of 3ml test expression from Sf9 insect cell of various N- and C-terminal construct truncations of human POLQ, following no soluble expression in *E. coli*. Red arrows denote successful and correctly-sized proteins.

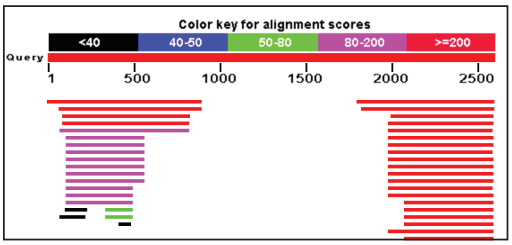
1


SEQUENCE-BASED
p-HMM
SMART
PFAM
MSA
BLASTP
CDD
STRUCTURAL HOMOLOGY
BLASTP/PDB
pGenTHREADER
A: Domain identification
B:Disorder/low-complexity sequence removal
GLOBULARITY & DISORDER PREDICTION
GlobPlot
FoldIndex
C: Secondary structure prediction
SECONDARY STRUCTURE PREDICTION
PSIPRED
D: Fine boundary definition
High-throughput cloning, test expression and iterative domain boundary analysis

2


Full length protein
Domain 1
Domain 2
Domain constructs
Inter-domain constructs

3a



3b

Chain A, Crystal Structure Of The Helicase Domain Of Human Dna Polymerase Theta, Apo-form

Sequence ID: pdb|5A9J|A Length: 896 Number of Matches: 1

See 3 more title(s)

Range 1: 3 to 896 GenPept Graphics

Next Match Previous Match

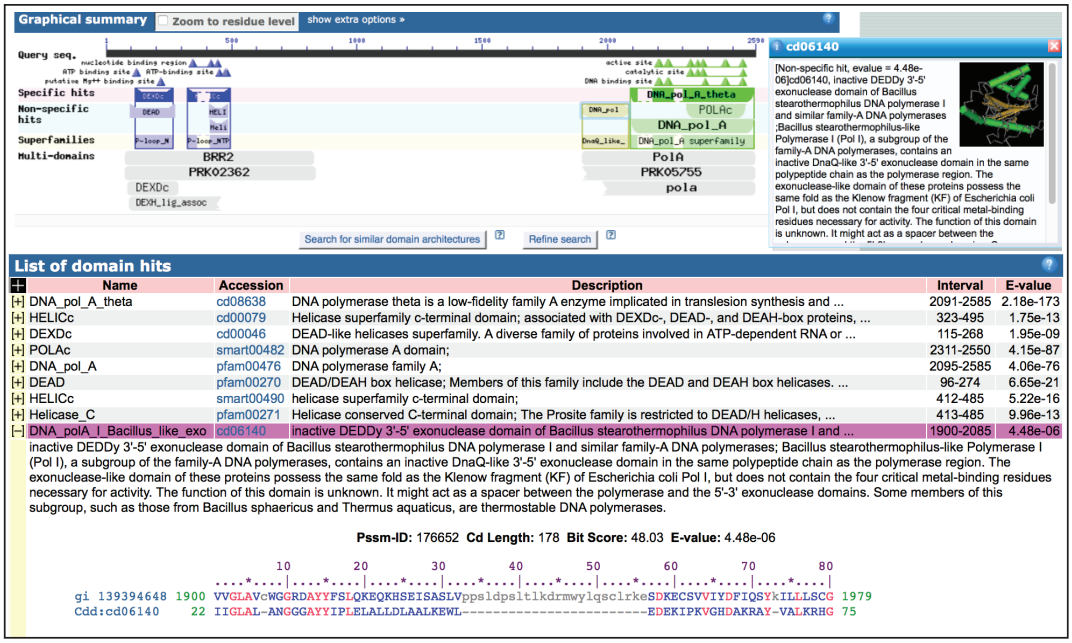
Score	Expect	Method	Identities	Positives	Gaps
1851 bits(4794)	0.0	Compositional matrix adjust.	894/894(100%)	894/894(100%)	0/894(0%)

```
Query  1    MNLLRRSGKRRRSESGSDSFSGSGGDSSASPQFLSGSVLSPPPGLGRCLKAAAAGECKPT  60
            MNLLRRSGKRRRSESGSDSFSGSGGDSSASPQFLSGSVLSPPPGLGRCLKAAAAGECKPT
Sbjct  3    MNLLRRSGKRRRSESGSDSFSGSGGDSSASPQFLSGSVLSPPPGLGRCLKAAAAGECKPT  62

Query  61   VPDYERDKLLLANWGLPKAVLEKYHSFGVKKMFEWQAECLLLGQVLEGKNLVYSAPTSAG  120
            VPDYERDKLLLANWGLPKAVLEKYHSFGVKKMFEWQAECLLLGQVLEGKNLVYSAPTSAG
Sbjct  63   VPDYERDKLLLANWGLPKAVLEKYHSFGVKKMFEWQAECLLLGQVLEGKNLVYSAPTSAG  122

Query  121  KTLVAELLILKRVLEMRKKALFILPFVSVAKEKKYYLQSLFQEVGIKVDGYMGSTSPSRH  180
            KTLVAELLILKRVLEMRKKALFILPFVSVAKEKKYYLQSLFQEVGIKVDGYMGSTSPSRH
Sbjct  123  KTLVAELLILKRVLEMRKKALFILPFVSVAKEKKYYLQSLFQEVGIKVDGYMGSTSPSRH  182
```

3c



List of domain hits

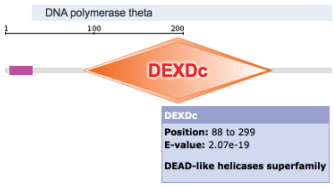
	Name	Accession	Description	Interval	E-value
[+]	DNA_pol_A_theta	cd08638	DNA polymerase theta is a low-fidelity family A enzyme implicated in translesion synthesis and ...	2091-2585	2.18e-173
[+]	HELICc	cd00079	Helicase superfamily c-terminal domain; associated with DEXDc-, DEAD-, and DEAH-box proteins, ...	323-495	1.75e-13
[+]	DEXDc	cd00046	DEAD-like helicases superfamily. A diverse family of proteins involved in ATP-dependent RNA or ...	115-268	1.95e-09
[+]	POLAc	smart00482	DNA polymerase A domain;	2311-2550	4.15e-87
[+]	DNA_pol_A	pfam00476	DNA polymerase family A;	2095-2585	4.06e-76
[+]	DEAD	pfam00270	DEAD/DEAH box helicase; Members of this family include the DEAD and DEAH box helicases. ...	96-274	6.65e-21
[+]	HELICc	smart00490	helicase superfamily c-terminal domain;	412-485	5.22e-16
[+]	Helicase_C	pfam00271	Helicase conserved C-terminal domain; The Prosite family is restricted to DEAD/H helicases, ...	413-485	9.96e-13
[–]	DNA_polA_I_Bacillus_like_exo	cd06140	inactive DEDDy 3'-5' exonuclease domain of Bacillus stearothermophilus DNA polymerase I and ...	1900-2085	4.48e-06

inactive DEDDy 3'-5' exonuclease domain of Bacillus stearothermophilus DNA polymerase I and similar family-A DNA polymerases; Bacillus stearothermophilus-like Polymerase I (Pol I), a subgroup of the family-A DNA polymerases, contains an inactive DnaQ-like 3'-5' exonuclease domain in the same polypeptide chain as the polymerase region. The exonuclease-like domain of these proteins possess the same fold as the Klenow fragment (KF) of Escherichia coli Pol I, but does not contain the four critical metal-binding residues necessary for activity. The function of this domain is unknown. It might act as a spacer between the polymerase and the 5'-3' exonuclease domains. Some members of this subgroup, such as those from Bacillus sphaericus and Thermus aquaticus, are thermostable DNA polymerases.

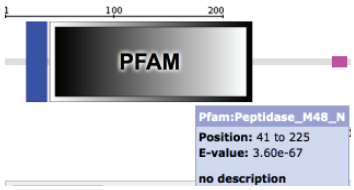
Pssm-ID: 176652 **Cd Length:** 178 **Bit Score:** 48.03 **E-value:** 4.48e-06

```
                    10        20        30        40        50        60        70        80
                    ....*....|....*....|....*....|....*....|....*....|....*....|....*....|....*....|
gi 139394648 1900 VVGLAVcWGGRDAYYFSLQKEQKHSEISASLVppsldpsltlkdrmwylqsclrkeSDKECSVVIYDFIQSYkILLLSCG 1979
Cdd:cd06140    22 IIGLAL-ANGGAYYIPLELALLDLAALKEWL----------------------EDEKIPKVGHDAKRAY-VALKRHG 75
```

4a



4b



4c

The domain within your query sequence starts at position 88 and ends at position 299; the E-value for the DEXDc domain shown below is 2.07e-19.

```
GVKKMFEWQAECLLLGQVLEGKNLVYSAPTSAGKTLVAELLILKRVLEMRKKALFILPFV
SVAKEKKYYLQSLFQEVGIKVDGYMGSTSPSRHFSSLDIAVCTIERANGLINRLIEENKM
DLLGMVVVDELHMLGDSHRGYLLELLLTKICYITRKSASCQADLASSLSNAVQIVGMSAT
LPNLELVASWLNAELYHTDFRPVPLLESVKVG
```

BLAST with Domain | Align your sequence against the SMART alignment

DEXDc

DEAD-like helicases superfamily

SMART accession number:	SM00487
Description:	-
Interpro abstract (IPR014001):	Helicases have been classified in 5 superfamilies (SF1-SF5). All of the proteins bind ATP and, consequently, all of them carry the classical Walker A (phosphate-binding loop or P-loop) and Walker B (Mg2+-binding aspartic acid) motifs. For the two largest groups, commonly referred to as SF1 and SF2, a total of seven characteristic motifs has been identified [(PUBMED:2546125)]. These two superfamilies encompass a large number of DNA and RNA helicases from archaea, eubacteria, eukaryotes and viruses that seem to be active as monomers or dimers. RNA and DNA helicases are considered to be enzymes that catalyse the separation of double-stranded nucleic acids in an energy-dependent manner [(PUBMED:11839499)]. The various structures of SF1 and SF2 helicases present a common core with two alpha-beta RecA-like domains [(PUBMED:11839499), (PUBMED:11087862)]. The structural homology with the RecA recombination protein covers the five contiguous parallel beta strands and the tandem alpha helices. ATP binds to the amino proximal alpha-beta domain, where the Walker A (motif I) and Walker B (motif II) are found. The N-terminal domain also contains motif III (S-A-T) which was proposed to participate in linking ATPase and helicase activities. The carboxy-terminal alpha-beta domain is structurally very similar to the proximal one even though it is bereft of an ATP-binding site, suggesting that it may have originally arisen through gene duplication of the first one. This entry represents the DNA-binding domain of classical SF1 and SF2 helicases.

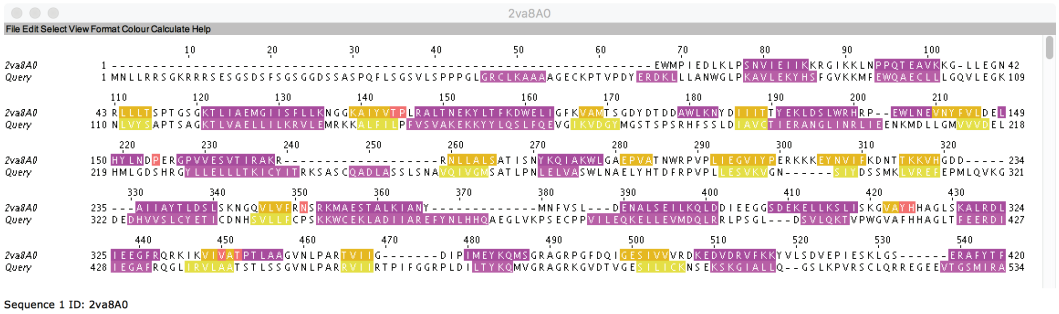
5a

Summary | pGenTHREADER | PSIPRED | Downloads

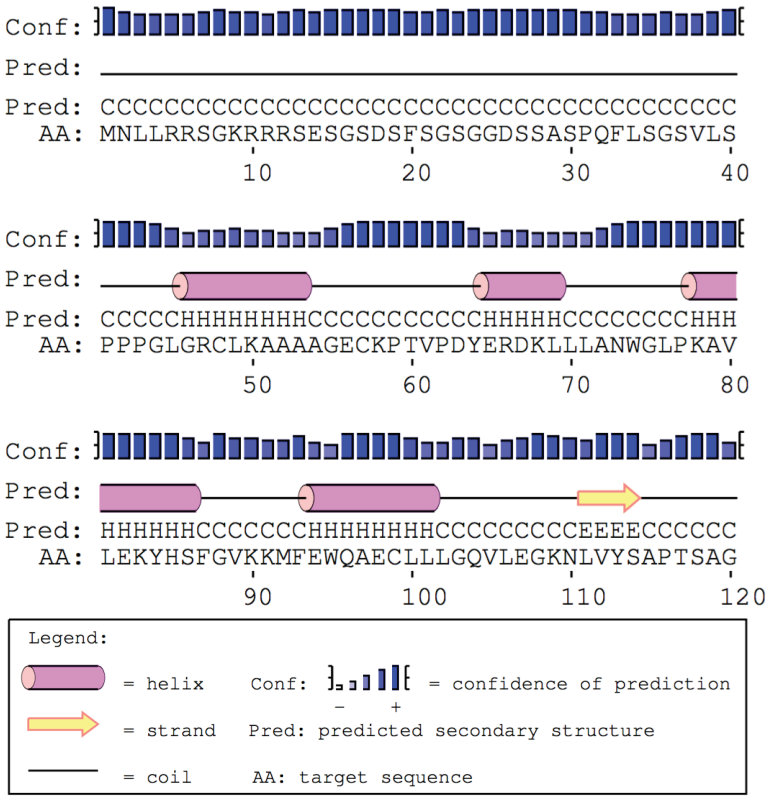
pGenTHREADER Scores (alignments displayed against the first of the input sequences)

Conf.	Net Score	p-value	PairE	SolvE	Aln Score	Aln Len	Str Len	Seq Len	View Alignment	SCOP Codes	CATH Codes	Structure	CATH Entry
CERT	240.870	2e-23	-1520.7	-26.3	1047.0	681	693	1000	Display 2va8A0 Alignment	Search SCOP for 2va8A	Search CATH for 2va8		CATH Summary
CERT	240.003	3e-23	-1356.5	-34.0	1070.0	683	700	1000	Display 2zj8A0 Alignment	Search SCOP for 2zj8A	Search CATH for 2zj8		CATH Summary
CERT	213.950	1e-20	-1088.8	-14.8	994.0	670	683	1000	Display 2p6rA0 Alignment	Search SCOP for 2p6rA	Search CATH for 2p6r		CATH Summary
CERT	179.838	4e-17	-1132.2	-11.5	735.0	773	1723	1000	Display 4f92B0 Alignment	Search SCOP for 4f92B	Search CATH for 4f92		CATH Summary

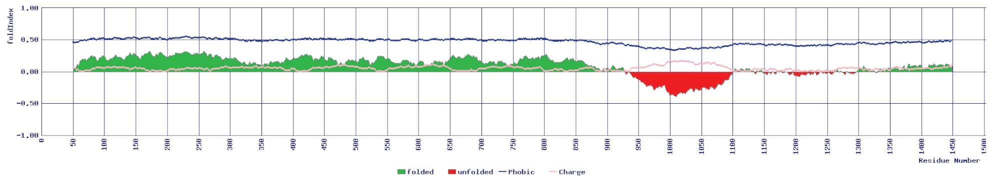
5b



5c



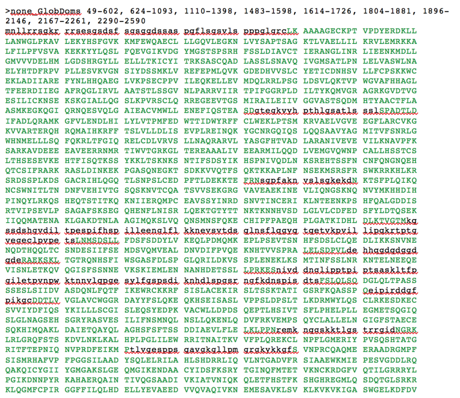
6a


foldIndex
Residue Number
folded
unfolded
Phobic
Charge

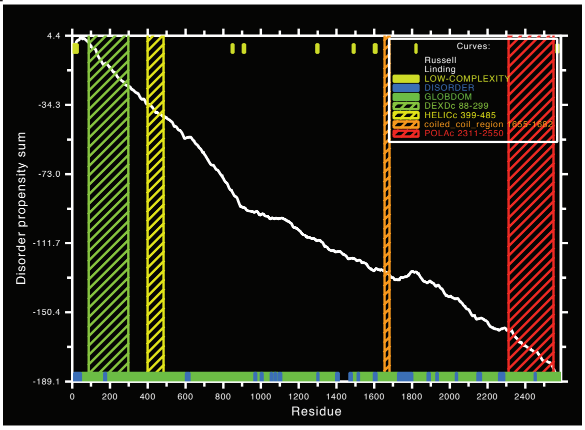
6b


(Predicted disordered segment)

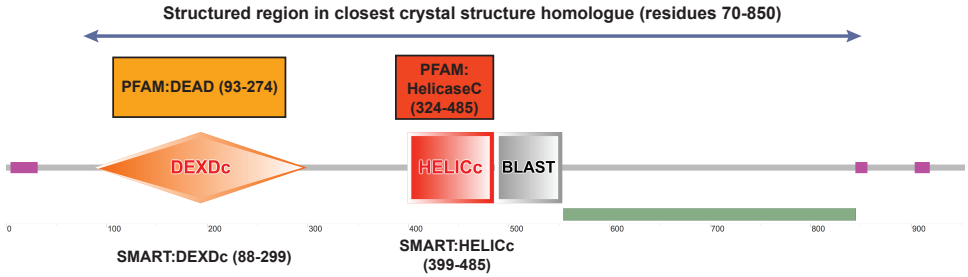
6c



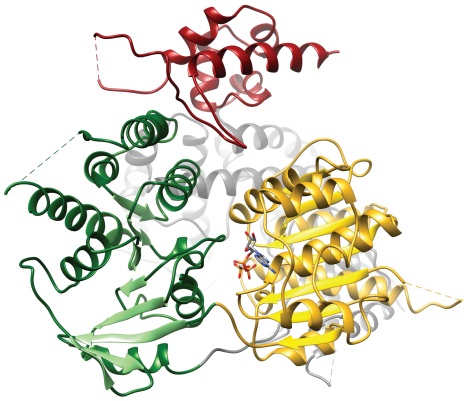
6d


Curves:
Russell
Linding
LOW-COMPLEXITY
DISORDER
GLOBDOM
DEXDc 88-299
HELICc 399-485
coiled_coil_region 1655-1682
POLAc 2311-2550
Disorder propensity sum
Residue

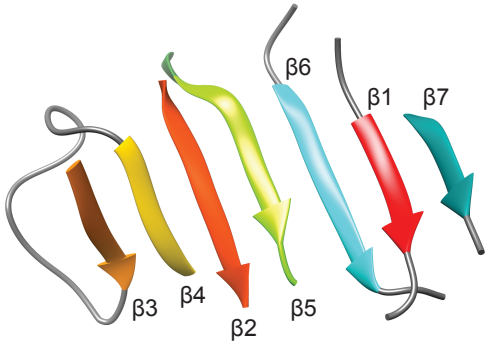
7a


Structured region in closest crystal structure homologue (residues 70-850)
PFAM:DEAD (93-274)
PFAM: HelicaseC (324-485)
DEXDc
HELICc
BLAST
0
100
200
300
400
500
600
700
800
900
SMART:DEXDc (88-299)
SMART:HELICc (399-485)

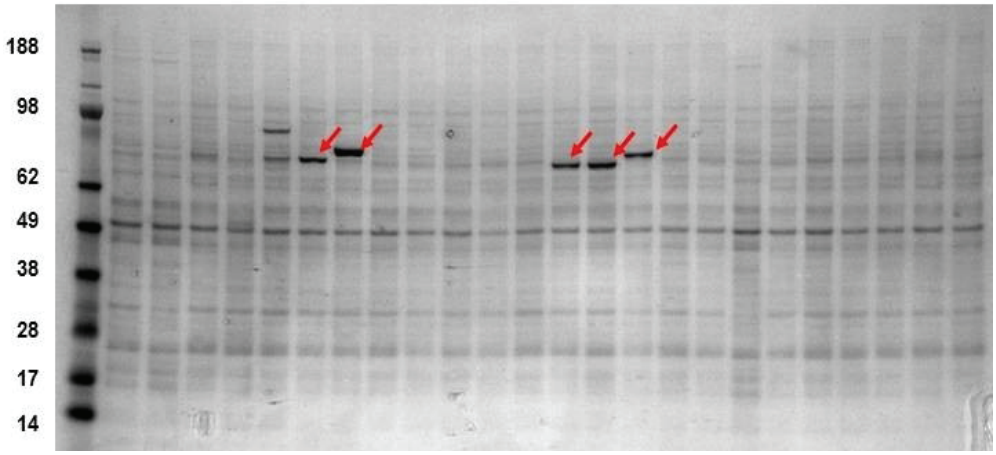
7b



7c


β6
β1
β7
β3
β4
β2
β5

8


188
98
62
49
38
28
17
14