PURE: a web server for querying the relationship between pre-existing domains and unassigned regions in proteins

CURRENT STATUS: POSTED

Chilamakuri C Sekhar Reddy
National Centre for Biological Sciences, Tata Institute of Funda ental Research, 2 Laboratoire de Biochimie et Genetique Moleculaire, Universite de La Reunion

Khader Shameer
National Centre for Biological Sciences, Tata Institute of Fundamental Research, GKVK Campus, Bellary Road, Bangalore 560 065, India

Offmann Bernard
Laboratoire de Biochimie et Genetique Moleculaire, Universite de La Reunion, 15 avenue Rene Cassin, BP 7151, 97715 Saint Denis Messag, Cedex 09 La Reunion, France

Ramanathan Sowdhamini
National Centre for Biological Sciences, Tata Institute of Fundamental Research, GKVK Campus, Bellary Road, Bangalore 560 065, India

DOI:
10.1038/nprot.2007.486

SUBJECT AREAS
Biochemistry *Computational biology and bioinformatics*

KEYWORDS
protein structure

Introduction

Protein domains are the structural and functional units of proteins. The ability to parse proteins into different domains is important for effective classification, understanding of protein structure, function and evolution and is hence biologically relevant. Several computational methods are available to identify domains in the sequence. Domain finding algorithms often employ stringent thresholds to recognize sequence domains. Identification of additional domains can be tedious involving intense computation and manual intervention but can lead to better understanding of overall biological function. In this context, the problem of identifying new domains in the unassigned regions of a protein sequence assumes a crucial importance. We report the availability of a convenient server for the domain prediction in unassigned regions in proteins (PURE) which can be accessed at "http://caps.ncbs.res.in/PURE/":http://caps.ncbs.res.in/PURE/

Introduction

Protein domains are the structural and functional units of proteins and represent one of the most useful levels to understand protein function. Analysis of proteins at the level of domain families has had a profound impact on the study of individual proteins. The availability of information on additional co-existing domains can provide knowledge about the overall function of a protein. The word ‘domain’ was coined by Wetlaufer [1] to suggest the presence of compact substructures within protein folds: such domains are called ‘structural domains’. Soon, biochemical experiments proved that some domains are capable to be an independent folding unit: such domains are called ‘folding domains’. Amongst several proteins involved in signal transduction, individual domains might be responsible for carrying out a specific function such that the whole gene is capable of multi-tasking in response to the requirements and be recruited in particular biochemical pathways: such domains are called ‘functional domains’. In many instances, these domains and their boundaries might coincide. They can be defined using multiple criteria, or combinations of criteria, including evolutionary conservation, discrete functionality and the ability to fold independently [2]. Several reports have reviewed the domain architectures of members of protein families to suggest overall function of whole proteins (for example [3-6], for phosphatases).

Protein domain discovery using various computational approaches has been progressing steadily over the past 35 years [7]. The identification of protein domains within a polypeptide chain can be achieved in several ways. Methods applied by classification databases such as the Dali Domain Dictionary [8], CATH [9], SCOP [10], DIAL [11], HOMSTRAD [12] employ structural data to locate and assign domains. Such structural data could be queried using objective algorithms (for example, SCOP information is organized in SUPERFAMILY database as Hidden Markov Models (HMM) [13]). Identification of domains at the sequence level most often relies on the detection of global and local sequence alignments between a given target sequence and domain sequences found in databases such as Pfam [14]. These organized databases of sequence domain families can also be queried using objective HMM algorithms. Continuing efforts to improve domain identification have produced wealth of different algorithms like the very recently developed DOMAC [15], DOMPRED [16] or DomainDiscovery [17] (see Bioinformatics Link Directory for a comprehensive list of available servers in this field[18]). However, difficulties in elucidating the domain content of a given sequence still arise when searching the target sequence against sequence or structural databases resulting in a lack of significant matches. For example, *Mycoplasma genitalium* is a small genome with 483 proteins but only 386 protein sequences have known Pfam hits with 56% residue coverage [19] emphasizing the need to further explore other methods for domain assignment from sequence. These are largely due to reasons such as high evolutionary divergence, distant similarities and incomplete or unequal representation of protein families in sequence space [20, 21]. Though, similar approaches of integrating multiple, sensitive database search to detect distant homologues has been reported as a successful method to establish remote homology[22, 23], we have recently shown that it is possible to enhance prediction of domains by 25% through indirect connections, namely consulting the domain architecture of sequence homologues [24].

In this paper, we report the availability of a bioinformatics protocol as a web server called **PURE** (domain Prediction in unassigned regions in proteins), which will enhance the domain predictions.

PURE protocol utilizes the concept of Intermediate Sequence Search (ISS) [22] to assign functional domain to a given unassigned region with the help of connecting sequences. Whereas the concept of ISS is classically applied for sequence similarity at the whole protein level[22], in this protocol, we make use of the property that homologous sequences could adopt similar domain architectures and where sequence homology exists at different domains, such an extrapolation of domain architectures is plausible.

Unassigned regions in proteins are examined in the non-redundant sequence (NR) database, using PSI-BLAST, for homologues at stringent thresholds. Representative homologues, identified at the unassigned region, are traced back to their full-length sequences and subject to time-intensive hmmpfam search that enables the delineation of domain architectures of all homologues. Where a direct hmmpfam search could not yield any relationship to a pre-existing protein domain family, we have earlier shown that as much as 25% of connections could be obtained by following the indirect assignment of domains. These are termed as ‘indirect connections’. Indirect connections between the query and distantly related domain is established through a powerful procedure using PSI-BLAST hits which are individually routed through a rigorous hmmpfam search against Pfam database [25]. In addition, PURE server automatically provides other structurally relevant findings such as the prediction of coiled coils and transmembrane helices.

Equipment

A computer with access to the internet and a web browser.

EQUIPMENT SETUP

Data

Your input protein sequences should be written in the standard one-letter code. Thus, the allowed characters are ACDEFGHIKLMNPQRSTVWY. You can choose to upload your sequences either through copy-and-paste in an input window or by submitting an entire sequence file (which must be in plain text format). The file must be in FASTA format: sequence should be preceded by a line beginning with the ‘‘>’’ sign followed by the sequence name.

Programs

In practice, PURE server processes query sequences using a computationally intensive bioinformatics protocol. A detailed flow-chart of the protocol is provided in Figure 1. The web interface is developed in HTML and Java Script. The wrapper scripts for external programs like PEPCOILS26, TMAP26, PSIPRED27, ScanProsite28, PSI-BLAST29, CD-HIT30 and hmmpfam31, while core programs that integrate results, CGI programs and automated e-mail programs are coded in Perl. Multiple alignment visualization of PSI-BLAST output is enabled using MView32. The graphics that integrate the results are generated using Bio::Graphics module from Bioperl33.

Procedure

1. If you need to perform domain identification in unassigned regions in your protein sequence, go to PURE server page - "http://caps.ncbs.res.in/PURE/PURE.html":ttp://caps.ncbs.res.in/PURE/PURE.html
2. You have two input options (Figure 2): in the first option, you can upload a file and in the second option, you can paste your sequence. Click on the option which you want. In both these options, a single sequence must be provided in the FASTA format: sequence information (see Equipment Setup section above) should be preceded by a line beginning with the ‘‘>’’ sign followed by the sequence name. Multiple sequences cannot be processed by PURE because it is a very computer intensive method. If multiple sequences are provided, this will generate an error message. a. First option, Upload sequence : you can upload a file containing a single sequence; sequence file must be a text file.! CAUTION Uploaded file must be a plain text file (generally using ASCII or Unicode schemes) while other rich text formats like those produced by most text editing tools e.g Microsoft Word , cannot be processed by PURE server.

b. Second option, Paste sequence : type in or paste your sequence information in the text box area. Provide a query name to your job. This is mandatory. If you do not fill up this information, your request will not be processed and an error message will be displayed.

3. Enter your valid Email id; PURE results, once available, will be automatically

forwarded to this email account. Make sure that you have provided a valid Email id. This is mandatory whether you provide the sequence information by uploading a file or you paste in sequence in the text box. If you do not fill up this information, your request will not be processed and an error message will be displayed.

4. PURE Server is using TMAP and Pepcoils to predict transmembrane regions and coiled coils region in your query sequence. Users have an option to Switch-off this filter, so that the query sequence will not be filtered using TMAP and COILS programs.
5. Pick up an appropriate database for PSI-BLAST search. You can choose different databases for PSI-BLAST search from database menu ; click on the database button to see the different databases available (like NR database (non-redundant database) [34] or SWISSPROT database [35]) and click on the database you want to search for. If you do not select any database by default search is against NR database. ! CAUTION if you choose NR database for PSI-BLAST search, it may take long time for the database search and subsequently it takes long time to get PURE results.
6. Choose E-Value for PSI-BLAST search. You can choose range of E-values for PSI-BLAST search from E-value menu, click on the value you want to choose and the default is 0.001. E-value is a statistical significance threshold for reporting matches against database sequences. The smaller this value the more stringent will be criteria for selecting the hits and the less hits you will get. The higher this value, the less stringent will be your search criteria and the larger the number of hits that will be picked. However, hits with large E-values, typically above 0.1, can be detected solely by chance. Default E-value proposed by PURE (0.001) is a relatively rigorous criterion.
7. Select a CD-HIT threshold value for clustering the output from PSI-BLAST; CD-HIT will cluster sequences into groups based on sequence identity. A representative

sequence from each group will then be used for the subsequent Hmmpfam search. This helps not to provide redundant sequences or very close homologues for the Hmmpfam search. It hence allows a reduction of the number of searches to be performed and helps to optimize computational time. You can choose from a range of CD-HIT threshold value for clustering (between 0.4 and 1). If you provide a small value, there will be less number of clusters subsequently less number of representative sequences for Hmmpfam search and vice versa.

8. Choose an appropriate E-Value for Hmmpfam search. You can choose range of E-values for Hmmpfam search from E-value. The default is set to 0.01. Check comments about E-value in step 6 above.
9. Indicate the number of cluster representatives from step 7 above that you had like to subject for Hmmpfam search. The default is set to 50. ! CAUTION Hmmpfam search is computationally intensive and rate limiting step, therefore chooses lower number of sequences for Hmmpfam search, if you do not get satisfactory results resubmit the query with higher number of hits for Hmmpfam search but you may have to wait long time to get PURE results.
10. Click on the “PURE” button to run the analysis once you fill up the required information. If you need to modify filled up information you can reset the page by clicking on “Clear Form” button.
11. Check for any messages. Depending on the success or failure of sequence information submission different messages will be displayed. If you have not given required information, an error message will be displayed, you need to recheck and fill the required information for successful submission. If your sequence gets uploaded successfully to the PURE server, you can see the message “Query Sequence: “your query name” has been successfully uploaded”. You can also see the

parameters which you have selected for PURE run.

12. Once your submission is successful, the sequence will be processed by server and the results will be sent to the Email address which you have provided. Time required for processing of your query highly depends on the length of your query and on the parameters you have chosen (please see the section on Time Taken to obtain an indication).

13. Check your email regularly for results from PURE server. Object of messages from PURE server begins with the following “PURE Server Results for Query : your query name“. In the PURE results emailed, you can see two URL links: first link leads to concise result (Figure 3); you can see table and graphical form of the results. In the table form of the result, First column of the table indicates different domains in your sequence. If your query sequence is not picking up any domain “No Domains Predicted by PURE” message appear in the 2nd row and 1st column of the table. If your sequence is picking up domains, depending on the number of different domains in query, the number of rows in the table will change. Second column of the table indicates the frequency of the occurrence of each domain (a snapshot of an example result File is shown in Figure 3). In this example, there are two different domains predicted in the query; first domain is TPR_1 which is predicted 3 times out of 12 times. Third column gives Pfam link to the domain. In the other section of the results, graphical representation of PURE results is shown; you can see the result in the graphics form which is the integrated representation of the outputs. You can see the lengths of your query and any fragments after the prediction of coiled coil and transmembrane regions. You can also see the different predicted domains, connecting sequences or hits responsible for the indirect connections and E values. Details of the graphics are shown in Figure 3.

14. Click on the second URL which is in the Email for detailed results of PURE, or click on the link in the concise result which leads to detailed results. In the detailed results section, we present the results of different programs to the user for clarity (Figure 4). Because it is a multi-step process, we have divided the detailed results into different subsections:

i. Check the Pepcoils program results. This program is for the prediction of coiled coil regions that works around COILS36 algorithm. Click on pepcoils output I to see the original pepcoils output file. In the pepcoils output II, if your sequence has predicted coiled coil region, the coiled coil region will be substituted with “=” symbol. If your sequence does not have predicted coiled coil regions, the original sequence will be displayed without any modification.

ii. Analyze Tmap program results. Tmap is program for the prediction of transmembrane regions in the sequence. Click on the Tmap output I for the original Tmap output file. In the Tmap output II, if your sequence has predicted transmembrane region, the transmembrane region will be substituted with “x” symbol. If your sequence does not have predicted transmembrane regions, the original sequence will be displayed without any modification.

iii. Check the integrated Pepcoils and Tmap output. If you wish to see both coiled coil and transmembrane regions in your sequence, click on the integrated output file, you can see both the coiled coil and transmembrane regions, where “=” symbol indicated coiled coil region and symbol “x” indicates transmembrane regions.

iv. Examine the sequence segments extracted from query sequence after filtering using Pepcoils and Tmap. Sequence is split into different fragments based on the presence of coiled coil and transmembrane regions, each fragment should have at least 30 residues to be considered for further analysis. Click on the sequence segment to see the segment sequence. Different sequence segments will be considered independently for the further analysis. ! CAUTION If you select “Switch-off Tmap and Coils filter” option, results i-iv will not be displayed.

v. Check the PSIPRED results; PSIPRED is a program for secondary structure prediction. If you wish to see the output of PSIPRED program for different sequence segments, click on the corresponding segments.

vi. Assess the PSI-BLAST results. You can see the PSI-BLAST output for different sequence segments by clicking on the different sequence segments you wish to see. You can also see the information regarding number of hits in the PSI-BLAST search. If your query sequence is picking up less than 500 hits, you can also visualize the PSI-BLAST output in the MView format by clicking on the MView output link for different sequence segments.

vii. Check the PROSITE Results. In this section, you can see the functional motifs predicted by PROSITE scan 28.

viii. Look at CD-HIT Results. In this section, you can obtain information regarding the number of representative homologues after clustering.

ix. Analyze HMMPFAM Results. Here you can see the Hmmpfam search for the representative homologues. ! CAUTION number of Hmmpfam search results shown depends on the maximum number of hits you have selected for Hmmpfam search while submitting sequence, this is the maximum number of sequences that will be displayed in this section.

x. Check for the domains assigned by PURE. In this section of detailed output, you can see the domains assigned to different segments of your query, if any. Click on the link to see the details of the domains assignment. (Figure 5) As shown in figure 5, you can see your query name above the table for which domains have been assigned. In the table, first column shows number of different connecting sequences, second column shows the probable domain in your query , third and forth column gives start and end residue number where probable domain can be associated. Fifth column gives the intermediate sequence id through which domain is predicted in your query, click on the sequence id to know more about it. Last column provides the E value of the Hmmpfam search at which domain is associated to the intermediate sequence. For example in the (Figure 5), there are 12 connecting sequences through which the domains are picked up. In the table shows (row 2) there is a probable TPR_2 domain from 1 to 34 residues of your sequence; ZP_00107619 is the connecting sequence through which TPR_2 is associated to your sequence.

xi. Examine the graphical representation of PURE results; you can see the result in the graphics form which is the integrated representation of the all the programs output. You can see the lengths of your

query and its fragments. You can also see the coiled coil and transmembrane regions along with the different predicted domains connecting sequences and E values. Details of this graphic shown in the (Figure 3).

xii. Check the concluding remarks; if you want to skip the detailed output of the different programs and see only the final concise output, you can click the link on this section, where you can see the tabular form of the results. We have explained this table in the earlier section.

Timing

Time taken highly depends on the length of the protein sequence length, number of homologues obtained and parameters chosen by the user. For example, sequence with 182 residues, which has picked only five hits in PSI-BLAST search, has taken only four minuets, in another example, sequence with 226 residues, with 50 Hmm searches initiated, took 80 minutes for PURE run. On an average, sequence with 100 residues long with 50 hmm searches starting from homologues take around one hour for PURE run.

Troubleshooting

1. The default minimum length of an unassigned region to be considered for PURE runs is 30 residues. CAUTION! If sequence is shorter, this can give rise to several sub-optimal similarities and false positives.
2. If the unassigned region is split into small segments (< 70 residues) due to the prediction of transmembrane (TM) or coiled coil (CC) regions, the server will not pass through the rest of PURE run. If you still wanted to search this region against domain families, you can switch-off the option of predicting TM and CC regions. This will straightaway take your unassigned region for searching in sequence database for homologues. CAUTION! This can give rise to false positives since many TM-containing and CC-containing protein sequences could be obtained as homologues potentially leading to drifts or wrong connections.
3. If there are too few homologues obtained, you can relax the thresholds, like E-values,

in PSI-BLAST.

4. If too few homologues are entering in domain assignments in the HMM searches, you can increase the CD-hit thresholds to include more representatives.
5. If there are too many homologous sequences obtained, the server will accept the first 50 hits alone by default since the subsequent HMM runs are time-intensive. If you did not obtain new relationships between this unassigned region and pre-existing domain families by indirect connections, you can include more homologues in HMMsearch. This is possible by increasing the number of homologues from 50 to higher numbers. You can also increase the number of homologues and simultaneously decrease the CD-hit threshold so that fewer representatives are alone employed for domain finding. This is a better option since the sequence space will be sampled better. CAUTION: increase in the number of homologues can delay the output of results. Please refer to the section on ‘Time taken’ and be very discreet at this stage since it can also freeze the machine!
6. Sometimes, the connections to pre-existing domains could be significant according to E-values but the alignment lengths suggest a partial similarity. Such possible relationships could be due to sub-optimal similarities or due to discontinuous domains or novel domain arrangements and have to be viewed with caution.

Anticipated Results

The analysis of unassigned regions in proteins by PURE protocol allows for quick and accurate prediction of protein domains (for output details see Table 1). Where possible, within the stringent thresholds applied, it is possible to associate domains to unassigned regions (see Case 6 or 7 as in Table 1). It is possible that structure prediction terminates, either due to insufficient length after prediction of coiled-coils and TM helices (Case 1) or due to low predicted secondary structural content (Case 2) or few homologous sequences (Case 3) or inherent sequence alignment length requirements and discontinuous domains (Case 5). It is also possible that fewer homologues were alone considered

for time-intensive HMMSEARCH (Case 4). We are providing relaxations to each of these parameters so that the users can interfere at different stages, if required, and provide relaxation to cutoffs for these features. However, one has to be cautious that it is a trade-off between obtaining enhanced connections and encountering false positives or even drifts.

Earlier in our study on class III Adenylyl cyclase proteins, we considered 254 unassigned regions that were at least 30 residues and retain structural content more than 15%. We found that 67 of the 254 unassigned regions can be associated with pre-existing Pfam families (Table 2). These connections could be established only after considering homologous sequences and by consulting their domain architectures. Most of these connections pertain to distant relationships with pre-existing domain families, that are not often recorded in SWISSPFAM 19 with less than 30% sequence identity with known members but are reliable predictions24. Such relationships to additional domains could enhance our understanding of the overall biological function of the gene product. For example, Q8F152 protein is a 770 residue long guanylyl cyclase protein, with a single guanylyl cyclase domain from 567 to 753, there is a long unassigned regions at the amino terminus region of the protein from 1 to 566. When we applied the PURE protocol to this unassigned region it is picking up probable CHASE2 domain in the region of 110 to 480 residues by indirect connections through homologous sequences. A direct search would not have reliably associated this domain due to poor E-values (0.026). CHASE domain always occurs N-terminally in extracellular or periplasmic locations, followed by an intracellular tail housing diverse enzymatic signaling domains such as histidine kinases and adenylyl cyclases 37. The presence of a putative CHASE2 domain at the N-terminus suggesting that this domain might be involved in extracellular ligand and may regulate the activity of cytoplasmic guanylyl cyclase domain activity.

In another example, Q8F690 which is a 492 residues long guanylyl cyclase protein, there is a single guanylyl cyclase domain at the carboxyl terminal of the protein, there is a 307 residues long unassigned segment at the amino terminus region of the protein. When we applied the PURE protocol to this unassigned region of the protein, it is picking up GAF domain. All the 169 connecting sequences are picking up GAF domain, however, a direct search would not have proposed this

relationship with significance (E-value for direct search is 0.072). In several species, GAF domains, which are widely expressed small-molecule-binding domains that regulate enzyme activity, are known to bind cyclic nucleotides in few proteins it may also participate in the dimerization 38. The newly assigned amino terminal GAF domain to this gene might suggest that this domain participates either in the cyclic nucleotide binding or in the protein dimerization function.

In the Q8EXW3 protein there is an assigned region in the carboxyl terminal part of the protein from 903 t0 979, by applying the PURE protocol to this unassigned region, it is picking up either TPR_1 domain (3 out of 12 connecting sequences points to TPR_1) or TPR_2 domain (9 out of 12 connecting sequences points to TPR_2 domain). The tetratrico peptide repeats (TPR) is a degenerate 34-amino acid repeated motif 39 that is widespread in the evolution and known to mediate protein - protein interactions.

In the protein O83498, initially, there are two domains 'HAMP' and guanylate cyclase domain, leaving behind a fairly large 288 residues long amino terminal unassigned region. We identified one 'Cache' domain on the N-terminal unassigned region by applying PURE protocol. In most cases, Cache is a periplasmic or an extracellular domain. Extracellular cache domains are known to bind to small molecules and transmit the signals to the cytoplasmic catalytic domain 40. Therefore, the putative N-terminal cache domain may act as signal sensing domain and may regulate the guanylyl cyclase domain activity along with HAMP domain.

FUTURE WORK

The domain profile library in PURE server will be periodically updated to be abreast with PFAM database19. We also plan to upgrade the server speed so that more intense job processes can be handled.

References

1. Wetlaufer DB. Nucleation, rapid folding and globular intrachain regions in proteins. *Proc Nat1 Acad Sci USA* **70**:697-701 (1973).
2. Holm, L. & Sander, C. Parser for protein folding units. *Proteins* **19**:256-268 (1994).
3. Denu, J. M. and J. E. Dixon. Protein tyrosine phosphatases: mechanisms of catalysis

and regulation. *Curr Opin Chem Biol* **2**:633-41 (1998).

4. Andersen, J. N., O. H. Mortensen, et al. Structural and evolutionary relationships among protein tyrosine phosphatase domains. *Mol Cell Biol* **21**:7117-36 (2001).
5. Bhaduri, A. and R. Sowdhamini . A genome-wide survey of human tyrosine phosphatases. *Protein Eng* **16**:881-8 (2003).
6. Bhaduri, A. and R. Sowdhamini. Genome-wide survey of prokaryotic O-protein phosphatases. *J Mol Biol* **352**:736-52 (2005).
7. Copley, R.R., Doerks, T., Letunic, I., & Bork, P. Protein domain analysis in the era of complete genomes. *FEBS Lett*. **513**:129–134 (2002).
8. Dietmann, S. & Holm, L. Identification of homology in protein structure classification. *Nature Struc. Biol*. **8**:953–957 (2001) .
9. Orengo, C., *et al*. CATH—a hierarchic classification of protein domain structures. *Structure* **5**:1093–1108 (1997).
10. Murzin, A.G., Brenner, S.E., Hubbard, T., & Chothia, C. SCOP: A structural classification of protein database for the investigation of sequences and structures. *J. Mol. Biol*. **247**:536–540 (1995).
11. Sowdhamini, R. & Blundell, T.L. Automatic identification and analysis of domains in proteins of known crystal structure. *Prot. Sci*. **4**:506-520 (1995).
12. de Bakker PI I et al. HOMSTRAD: adding sequence information to structure-based alignments of homologous protein families. *Bioinformatics*, **8**:748-749, (2001).
13. Gough, J. and C. Chothia. SUPERFAMILY: HMMs representing all proteins of known structure. SCOP sequence searches, alignments and genome assignments. *Nucleic Acids Res* **30**:268-72 (2002).
14. Bateman, A., *et al*. The Pfam protein families database. *Nucleic Acids Res*. **28**:263–266 (2000).

15. Cheng, J. DOMAC: an accurate, hybrid protein domain prediction server. *Nucleic Acids Res.*, **35** (Web Server issue):W354-6, (2007).

16. Bryson, K., Cozzetto, D., Jones DT. Computer-assisted protein domain boundary prediction using the DomPred server. *Curr Protein Pept Sci.* **2**:181-8, (2007).

17. Sikder, A.R., Zomaya, A.Y. Improving the performance of DomainDiscovery of protein domain boundary assignment using inter-domain linker index. *BMC Bioinformatics*, **7**:S6, (2006).

18. Fox, J.A., Butland, S.L., McMillan, S., Campbell, G., Ouellette, B.F. The Bioinformatics Links Directory: a compilation of molecular biology web servers. *Nucleic Acids Res.*, **33**(Web server issue):W3-24, (2005).

19. Pfam Genome Distribution website : "http://www.sanger.ac.uk/cgi-bin/Pfam/genome_dist.pl":http://www.sanger.ac.uk/cgi-bin/Pfam/genome_dist.pl

20. Holm, L. and C. Sander. Mapping the protein universe. *Science* **273**:595-603 (1996).

21. Sowdhamini, R., D. F. Burke, et al., CAMPASS: a database of structurally aligned protein superfamilies. *Structure* **6**:1087-94 (1998).

22. Park, J., Teichmann, S.A., Hubbard, T.& Chothia, C. Intermediate sequences increase the detection of homology between sequences. *J Mol Biol*, **273**(1):349-354 (1997).

23. Salamov, A.A, Suwa, M, Orengo, C.A, & Swindells M.B. Combining sensitive database searches with multiple intermediates to detect distant homologues. *Protein Eng* **12**:95-100 (1999).

24. Reddy, C.S, Manonmani, A., Babu, M. & Sowdhamini, R. Enhanced structure prediction of gene products containing class III adenylyl cyclase domains. *In Silico Biology* **6**(5):351-62 (2006).

25. Bateman, A., *et al*. The Pfam protein families database. *Nucleic Acids Res.* **32**:D138-D141 (2004)

26. Rice, P., Longden, I. & Bleasby, A. EMBOSS: The European Molecular Biology Open Software Suite *Trends in Genetics* **16**:276—277 (2000).

27. McGuffin, L. J., Bryson, K. & Jones, D. T. The PSIPRED protein structure prediction server. *Bioinformatics* **16**:404-405 (2000).

28. Edouard de Castro et al. ScanProsite: detection of PROSITE signature matches and ProRule-associated functional and structural residues in proteins. *Nucleic Acids Res*. **34**:W362-W365 (2006).

29. Altschul, S. F. *et. Al* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*. **25**:3389-3402 (1997).

30. Weizhong Li & Adam Godzik Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, **22**:1658-9 (2006).

31. Eddy, S. Profile hidden markov models *Bioinformatics* **14**:755–763 (1998).

32. Brown, N.P., Leroy, C. & Sander, C. MView: A Web compatible database search or multiple alignment viewer. *Bioinformatics* **14**:380-381 (1998).

33. Stajich, J.E. *et al*. The Bioperl toolkit: Perl modules for the life sciences. *Genome Res* . **12**:1611-8 2002.

34. NCBI-NR Database : "ftp://ftp.ncbi.nih.gov/blast/db/FASTA/nr.gz":ftp://ftp.ncbi.nih.gov/blast/db/FASTA/nr.gz

35. Bairoch, A., Boeckmann, B., Ferro S., & Gasteiger E. Swiss-Prot: Juggling between evolution and stability Brief. *Bioinform*. **5**:39-55 (2004).

36. Lupas, A., Van Dyke, M., & Stock, J. Predicting coiled coils from protein sequences. *Science* **252**:1162-1164 (1991).

37. Anantharaman, V. and Aravind, L. The CHASE domain: a predicted ligand-binding module in plant cytokinin receptors and other eukaryotic and bacterial receptors. *Trends Biochem. Sci*. **26**:579-582 (2001).

38. Martinez, S. E. et al. Crystal structure of the tandem GAF domains from a cyanobacterial adenylyl cyclase: Modes of ligand binding and dimerization. *Proc. Natl. Acad. Sci. USA* **102**:3082-3087 (2005).

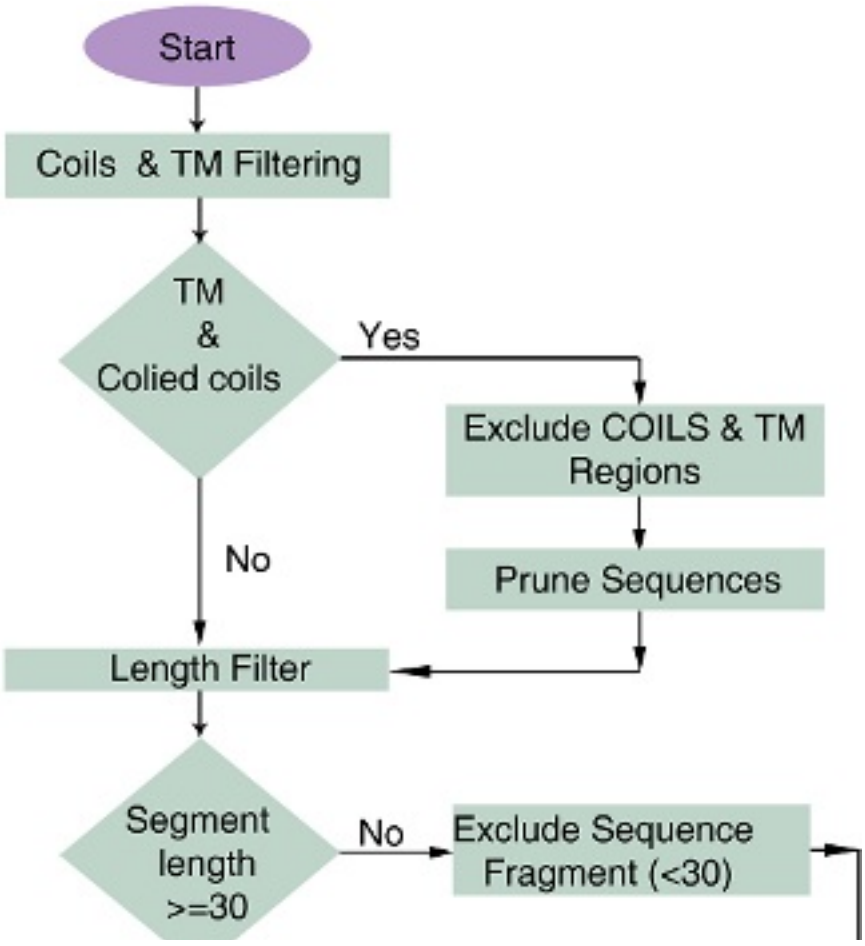
39. Goebl, M. and Yanagida, M. The TPR snap helix: a novel protein repeat motif from mitosis to transcription. *Trends Biochem. Sci.* **16**:173-177 (1991).

40. Anantharaman, V. and Aravind, L. Cache - a signaling domain common to animal Ca2+ - channel subunits and a class of prokaryotic chemotaxis receptors. *Trends Biochem. Sci.* **25**:535-537 (2000).

Acknowledgements

RS would like to thank Wellcome Trust (U.K.) and NCBS (TIFR) for financial support. CSR is supported by a PhD grant from Conseil Regional de La Reunion. The authors would like to thank Prof. N. Srinivasan and Dr. Yamuna Krishnan for helpful inputs and discussions.

Figures



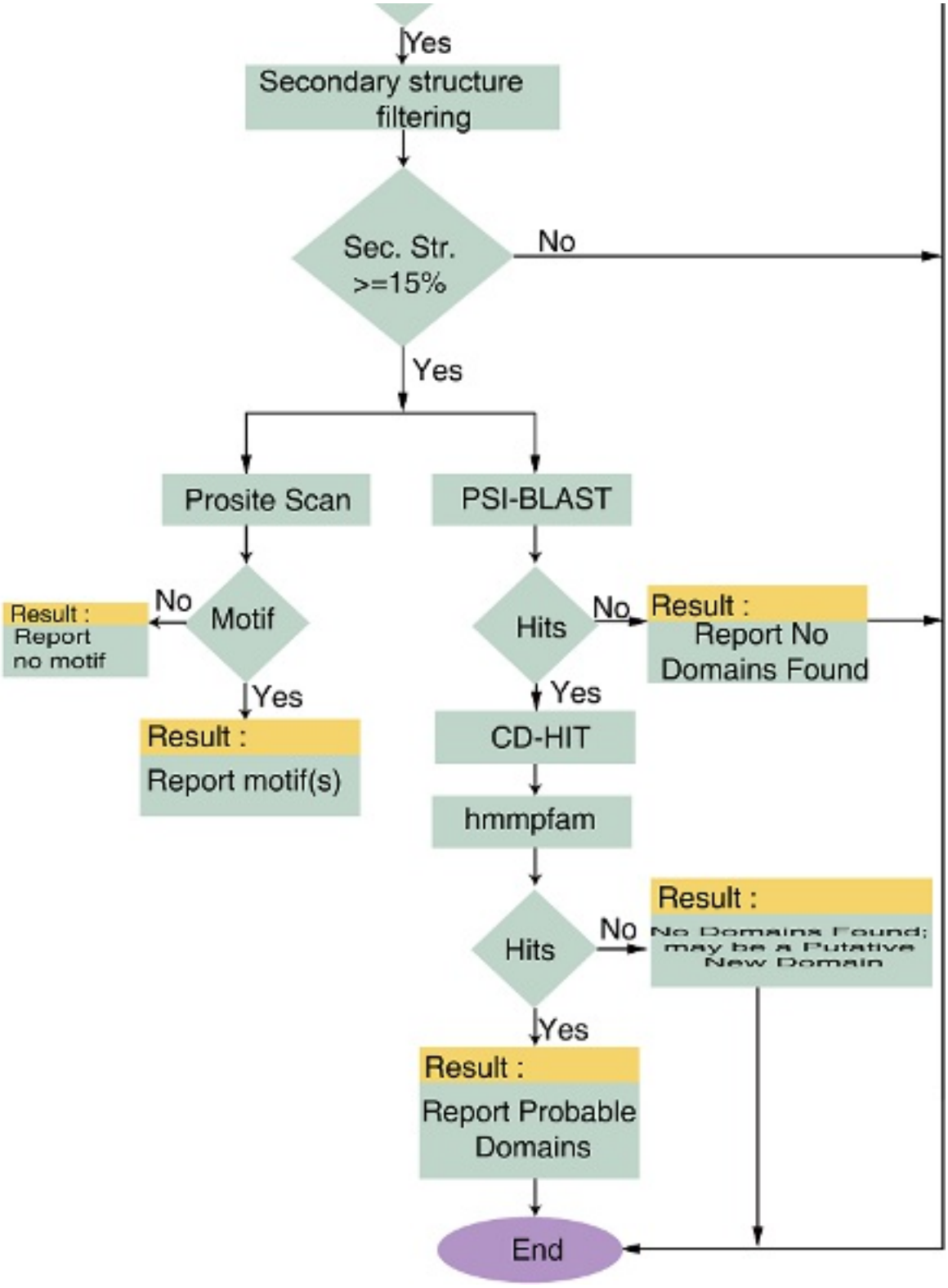

Yes
Secondary structure filtering
Sec. Str. >=15%
No
Yes
Prosite Scan
PSI-BLAST
Result : Report no motif
No
Motif
Yes
Result : Report motif(s)
Hits
No
Result : Report No Domains Found
Yes
CD-HIT
hmmpfam
Hits
No
Result : No Domains Found; may be a Putative New Domain
Yes
Result : Report Probable Domains
End

Figure 1

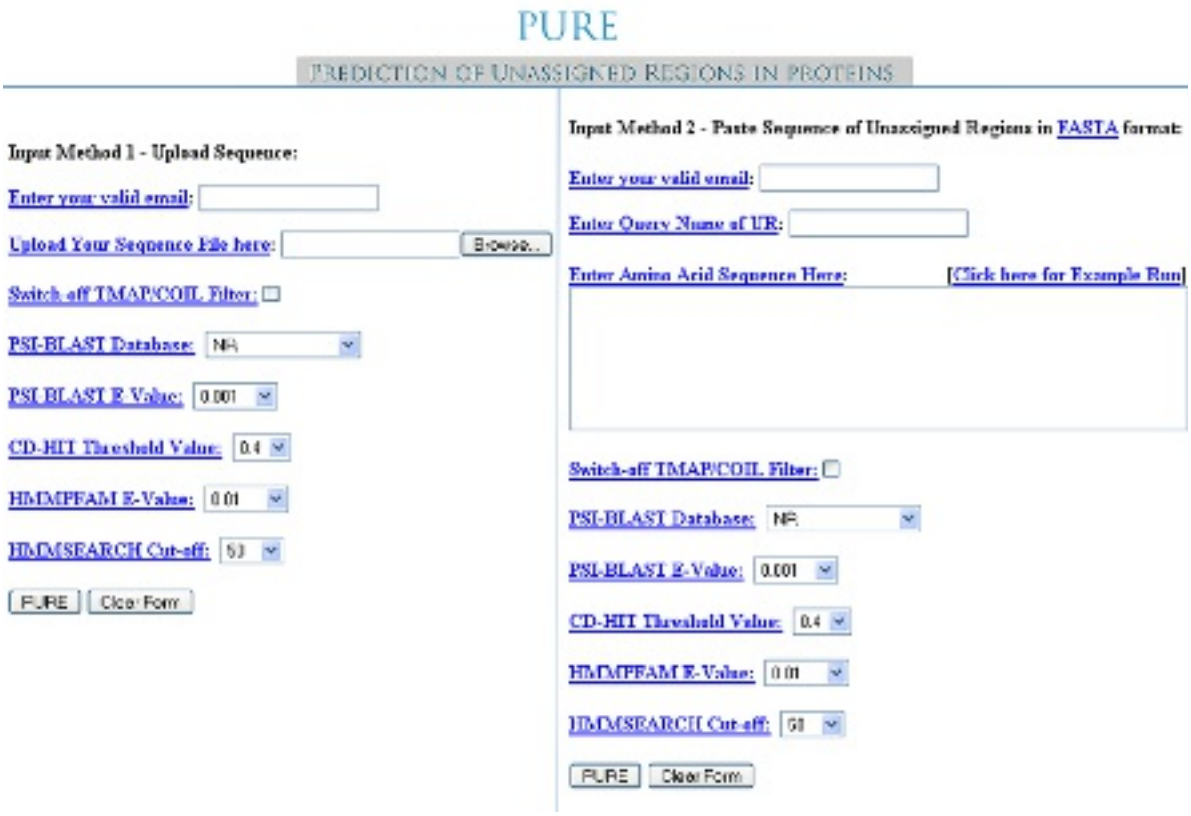


Figure 2

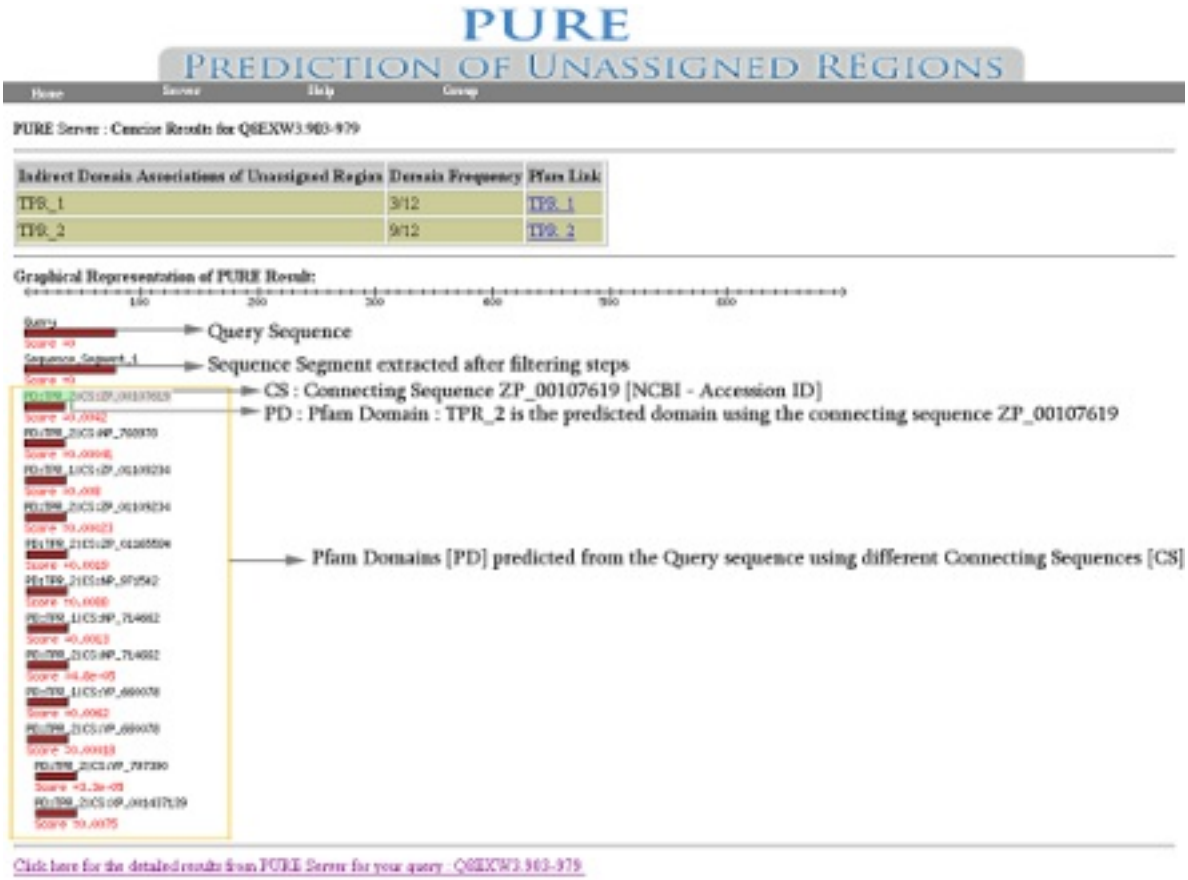


Figure 3

PURE

PREDICTION OF UNASSIGNED REGIONS IN PROTEINS

Home | Server | Help | Group

PURE : Results for QSEVAS 903-979

Click here for consensus output for your query.

PURE - Detailed Output Files:
Query Sequence : QSEVAS 903-979 has been successfully uploaded and processed

Detailed Analysis Results:
1. PSPCOILS - Prediction of Coiled Coil Regions Results:
PSPCOILS output : [Click here for PSPCOILS OUTPUT - I]
PSPCOILS output : Coiled Coil regions substituted with 'x' [Click here for PSCOILS OUTPUT - II]

2. TMAP - Prediction of Transmembrane Regions Results:
TMAP output [Click here for TMAP OUTPUT - I]
TMAP output : Transmembrane regions substituted with 'x' [Click here for TMAP OUTPUT - II]

3. Integrated Output - Coiled Coils and Transmembrane regions substituted with 'x' and 'z':
[Click here for the Integrated Output file]
Note : Based on the presence of TM and Coils region we split the sequence in to different fragments; sequence fragments should have at least 70 residues for further analysis
4. Sequence Segments extracted from query sequence after filtering using PSPCOILS & TMAP :
Sequence segment 1 : [Click here for Sequence Segment 1]

5. PSIPRED Results for sequence segments extracted after PSPCOILS & TMAP Run :
PSIPRED Result for Sequence Segment 1 : [Click here for PSIPRED output for segment : 1]
Note : Only sequences with >= 15% secondary structural content are used for further analysis.

Sequence Segment 1 : [Click here for PSIPRED output for segment : 1]

6. PSIBLAST Results:
Results for Sequence Segment 1 : [Click here for PSIBLAST output for segment : 1]
Total Unique hits = 80
Overview PSI-BLAST Results of Query Segment 1 : [Mouse over output of PSIBLAST file segment : 1]

7. PROSITE Results:
Protein patterns for Sequence Segment 1 : [Click here for PROSITE output for segment : 1]

8. CD HIT Results:
Homologues from BLAST Hits after CD-HIT [Sequence Segment 1] = 48

9. HMMPFAM Results:
HMMPFAM Results for query 1 [Click here for HMMPFAM output of 527_214662.1.652-721.hmm]
HMMPFAM Results for query 2 [Click here for HMMPFAM output of 527_91305810.1.903-979.hmm]
HMMPFAM Results for query 3 [Click here for HMMPFAM output of [illegible]]
HMMPFAM Results for query 4 [Click here for HMMPFAM output of YP_661073.1.295-361.hmm]
HMMPFAM Results for query 5 [Click here for HMMPFAM output of YP_413935.1.561-713.hmm]
HMMPFAM Results for query 6 [Click here for HMMPFAM output of YP_807986.1.620-655.hmm]

10. Domains Assigned using PURE:
Domains extracted for segment 1 : [Click here for Domains predicted using PURE from segment 1 :]

11. Graphical Representation of PURE Result:

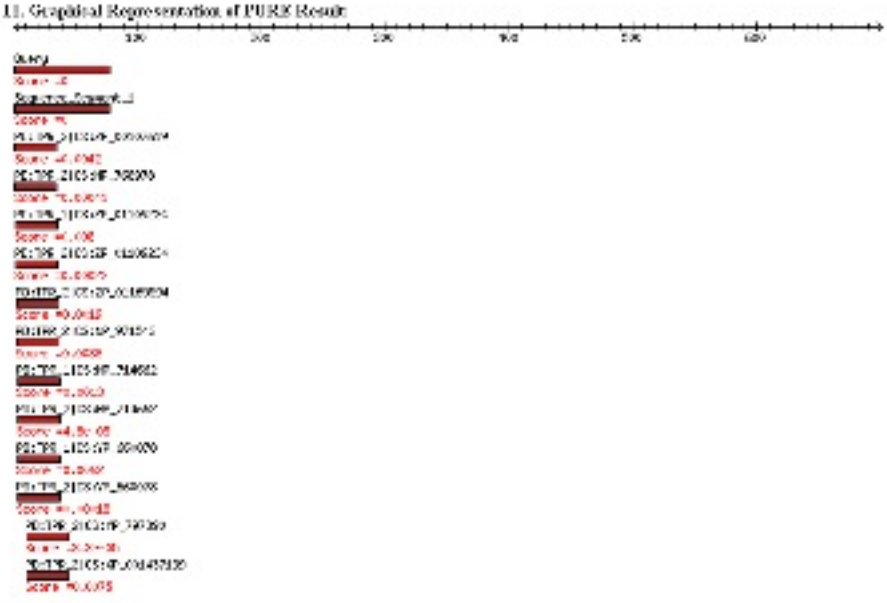
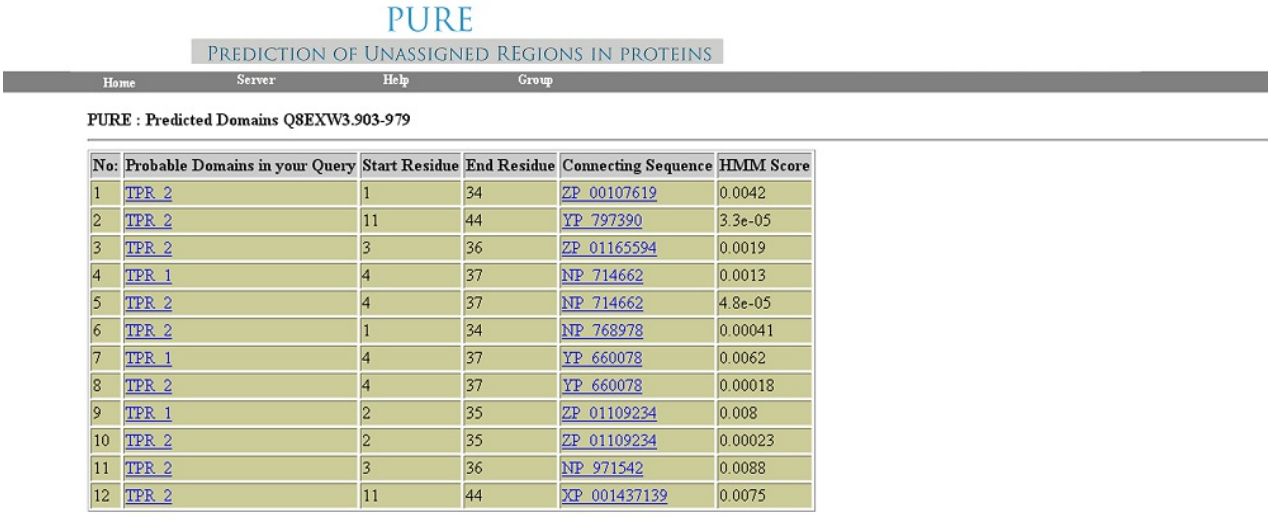


Figure 4



No:	Probable Domains in your Query	Start Residue	End Residue	Connecting Sequence	HMM Score
1	TPR_2	1	34	ZP_00107619	0.0042
2	TPR_2	11	44	YP_797390	3.3e-05
3	TPR_2	3	36	ZP_01165594	0.0019
4	TPR_1	4	37	NP_714662	0.0013
5	TPR_2	4	37	NP_714662	4.8e-05
6	TPR_2	1	34	NP_768978	0.00041
7	TPR_1	4	37	YP_660078	0.0062
8	TPR_2	4	37	YP_660078	0.00018
9	TPR_1	2	35	ZP_01109234	0.008
10	TPR_2	2	35	ZP_01109234	0.00023
11	TPR_2	3	36	NP_971542	0.0088
12	TPR_2	11	44	XP_001437139	0.0075

Figure 5

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

Tables.doc