ProteinSplit: splitting of multi-domain proteins using prediction of ordered and disordered regions in protein sequences for virtual structural genomics

J. Phys.: Condens. Matter **19** (2007) 285222 (8pp) doi:10.1088/0953-8984/19/28/285222

ProteinSplit: splitting of multi-domain proteins using prediction of ordered and disordered regions in protein sequences for virtual structural genomics

Lucjan S Wyrwicz[1,2,5], **Grzegorz Koczyk**[1,3], **Leszek Rychlewski**[1] and **Dariusz Plewczynski**[1,4]

[1] Bioinformatics Laboratory, BioInfoBank Institute, ulica Limanowskiego 24A, 60-744 Poznan, Poland
[2] Department of Gastroenterology, Medical Center for Postgraduate Education, Maria Sklodowska Curie Memorial Cancer Center, and the Institute of Oncology, ulica Roentgena 5, 02-781 Warsaw, Poland
[3] Institute of Plant Genetics, Polish Academy of Sciences, ulica Strzeszynska 34, 60-479 Poznan, Poland
[4] Interdisciplinary Centre for Mathematical and Computational Modeling, University of Warsaw, ulica Zwirki i Wigury 93, 02-089 Warsaw, Poland

E-mail: lucjan@bioinfo.pl

Received 28 September 2006, in final form 2 January 2007
Published 25 June 2007
Online at stacks.iop.org/JPhysCM/19/285222

Abstract

The annotation of protein folds within newly sequenced genomes is the main target for semi-automated protein structure prediction (virtual structural genomics). A large number of automated methods have been developed recently with very good results in the case of single-domain proteins. Unfortunately, most of these automated methods often fail to properly predict the distant homology between a given multi-domain protein query and structural templates. Therefore a multi-domain protein should be split into domains in order to overcome this limitation. ProteinSplit is designed to identify protein domain boundaries using a novel algorithm that predicts disordered regions in protein sequences. The software utilizes various sequence characteristics to assess the local propensity of a protein to be disordered or ordered in terms of local structure stability. These disordered parts of a protein are likely to create interdomain spacers. Because of its speed and portability, the method was successfully applied to several genome-wide fold annotation experiments. The user can run an automated analysis of sets of proteins or perform semi-automated multiple user projects (saving the results on the server). Additionally the sequences of predicted domains can be sent to the Bioinfo.PL Protein Structure Prediction Meta-Server for further protein three-dimensional structure and function prediction. The program is freely accessible as a web service at http://lucjan.bioinfo.pl/proteinsplit together with detailed benchmark results on

[5] Author to whom any correspondence should be addressed.

0953-8984/07/285222+08$30.00 © 2007 IOP Publishing Ltd Printed in the UK

the critical assessment of a fully automated structure prediction (CAFASP) set of sequences. The source code of the local version of protein domain boundary prediction is available upon request from the authors.

(Some figures in this article are in colour only in the electronic version)

1. Introduction

Due to the enormous complexity of cell metabolism and signalling, eukaryotic proteins usually consist of several protein domains responsible for different molecular functions [1]. During a typical protein modelling study, the most accurate template is aligned to a given query. In most cases a short local similarity of protein chains is used as a seed for global alignment and is expanded to the regions around the primary seeds. In the case of distantly related proteins such an approach results in only partially correct gapped alignment (such as a compilation of locally optimal short alignments). Several years of fold recognition experiments have clearly shown that automated structure prediction algorithms are not prepared to deal with multi-domain targets [2]. Users are strongly advised to split a query protein sequence into domains and iteratively submit all single-domain subsequences to automated prediction servers [3, 4].

Most existing protein splitting algorithms focus on the prediction of regions in protein sequences that are not structuralized. Such a local propensity to become a disordered region was found to be a surprisingly good approximation of global structural analysis of the domain arrangement in the three-dimensional structure of multi-domain proteins. Other experimental observations suggest that large ordered regions, when they are divided by shorter parts of disordered regions in a protein chain, are likely to be separate domains. In most cases the complete three-dimensional structure of the whole protein is not available. Therefore there is a strong need for algorithms that use only sequences for finding the domain splitting. Based on experimental findings, we conclude that disordered local sequence segments are likely to be linkers between protein domains. The protein domain splitting can be performed at those residues located in the middle of a disordered region if some global conditions are fulfilled, such as clear separation of two large ordered regions.

In this work, we outline a rapid method designed to predict ordered and disordered regions in a protein sequence in order to locate domain boundaries. Our algorithms can be used in a typical high-throughput modelling environment in large-scale genome annotation projects. A consensus algorithm utilized by ProteinSplit allows for fast scanning of a given protein sequence dataset. Our implementation of the method also provides a single protein mode for detailed, manual analysis and a semi-automated mode with multiple-user access for mid-scale projects.

2. ProteinSplit web server

The input of the ProteinSplit server is a collection of protein sequences. The server identifies all residues in the query protein sequence that are likely to be ordered or disordered in terms of local structural conformation, and presents an interactive visual representation of those regions along the query protein sequence. This algorithm is capable of splitting a long protein sequence into single, structural domains (regions of ordered structure). In figure 1 we present a diagram of the structure of the ProteinSplit web service. Our approach constitutes a compilation of several methods used to identify interdomain spacers. The tool evaluates the residues using

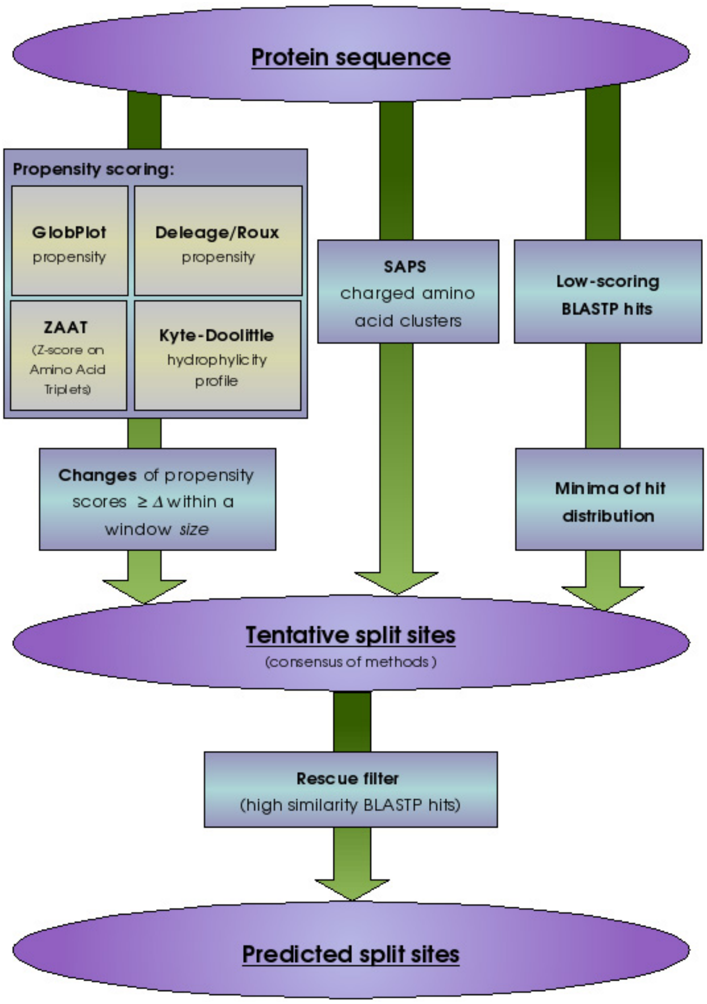
Protein sequence
Propensity scoring:
GlobPlot propensity
Deleage/Roux propensity
ZAAT (Z-score on Amino Acid Triplets)
Kyte-Doolittle hydrophylicity profile
SAPS charged amino acid clusters
Low-scoring BLASTP hits
Changes of propensity scores ≥ Δ within a window size
Minima of hit distribution
Tentative split sites (consensus of methods)
Rescue filter (high similarity BLASTP hits)
Predicted split sites

Figure 1. Schematic block diagram of the information flow in the ProteinSplit service.

cumulative scores presenting various protein characteristics, including non-globularity amino acid scores [5, 6], highly charged patches [7], and oligopeptide scores for the identification of domain linkers. In addition, the web service provides an interface to select the regions of a protein sequence that are incorporated into globular domains, using various sequence characteristics and the presence of non-globular regions. This is accomplished by conducting a comparison of the query protein with the representative set of domain sequences from the Pfam database [8] as well as proteins deposited in PDB (the Protein Data Bank database) [9].

3. The algorithm for identifying domain boundaries

Here we present a short description of the algorithm used in the ProteinSplit web server. The main components of our tool are as follows (see figure 1).

(i) Three scoring methods based on the previously characterized amino acid characteristics are applied independently: GlobPlot [5] and Deleage/Roux [6] disorder propensity scales and

Kyte–Doolittle hydrophilicity profile [10]. The first two single amino acid scores correlate with observed disordered regions of proteins solved crystallographically and deposited in the PDB database [9]. The last scale describes the relative fraction of hydrophilic amino acids. This is used since the high accumulation of hydrophilicities is common in interdomain spacers [11].

(ii) A novel domain propensity scoring scheme designed for ProteinSplit (ZAAT score) is applied. This score is calculated as a Z-score of the distribution of amino acid triplet frequency ratios with the numerator corresponding to the frequency in a database of defined domains (PfamA; Sanger) [8] and the denominator corresponding to the frequency in a database of full-length protein sequences (NR; NCBI) [12]. Redundancy in the databases was filtered prior to the calculations with CD-HI [13] at an identity threshold set to 90% (creating the so-called NR90 database). ZAAT scores are available at http://lucjan.bioinfo.pl/proteinsplit/ZAAT.

(iii) SAPS [7] is used to identify non-random patterns of charged amino acids.

(iv) General amino acid composition is tested with low-threshold BLAST searches [14] against a representative subset of the PfamA [15] database consisting of one sequence from each protein family as well as a non-redundant subset of the PDB [9] database (clustered prior to the calculations with CD-HI [13] at an identity threshold set to 90%). For this purpose the maximum E-value of the reported BLAST hits was set to 1000. Application of other databases is optional (PfamA, PfamB, clustered databases: PDB90; NR70; NR90).

(v) For the input sequence, adequate scores calculated with scales described in points (i) and (ii) are used to pinpoint tentative domain boundaries (split sites). As the scores correlate with potential disorder in the folded protein, the increase of score for the delta parameter (set to 1.5, by default) within a sliding window of a given number of residues (set to 20, by default) creates a tentative split site.

(vi) Additional tentative split sites are set according to SAPS results (point (iii)) and the results of low-score BLAST searches (regions of minimum hit density).

(vii) A ‘rescue filter’ for highly similar proteins is created based on the results of BLASTP [14] searches against the database described in point (iv). The E-value threshold here is default set to 1×10^{-3}, by default. Minima of the distribution of hits are excluded from the filter.

(viii) All the results of method assigning per residue scores (as described in points (i), (ii), (iv) and (vii)) are smoothed by averaging in a given size of window (set to ±5 residues, by default).

(ix) From the full list of potential split sites the ProteinSplit server selects those sites predicted by the highest number of methods (i)–(iv), if the minimal size of the resulting potential domains is preserved by such a choice of split site (this size is set to 100 amino acids, in default ProteinSplit configuration). The domain boundaries cannot be located within regions excluded from the prediction by the ‘rescue filter’.

The ProteinSplit method is, to a large degree, customizable—a user is able to easily change several parameters: window and delta, minimum and maximum domain size (set to 800, by default, as this value is a common maximum query size for protein structure prediction tools), BLAST E-value threshold, scanned reference database and data-smoothing window.

To facilitate easy scaling of ProteinSplit usage in annotation projects the user can select one of the two main layouts of work: either the manual prediction (designed for a low number of queries run by submitting the sequences in multiple FASTA format) or by running the tool in the batch mode (the user can upload the whole data set of proteins and will be informed via e-mail when the prediction results are ready). Additionally all users of the service can store the results of their predictions on the server, where it can be accessed by providing the name of the

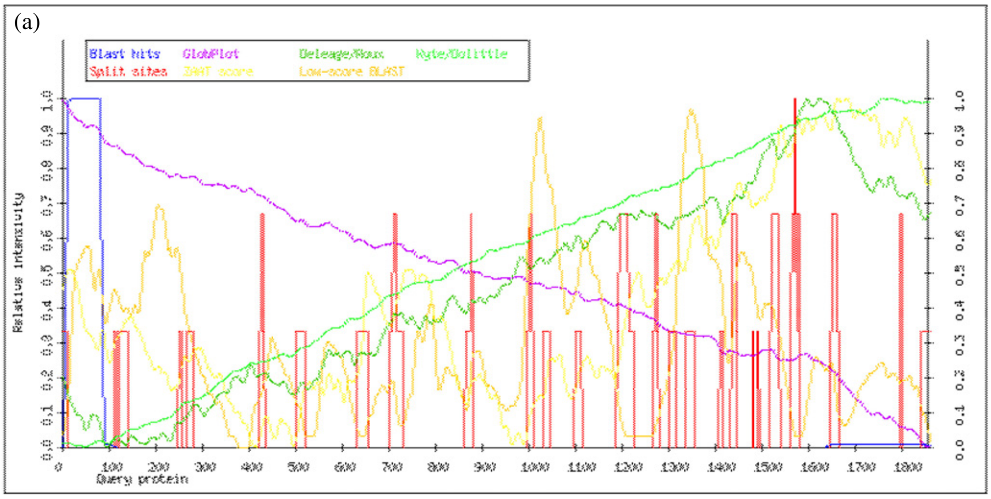

Figure 2. (a) The output of a ProteinSplit prediction of ordered/disordered regions in a query protein involved in breast cancer (gi: 6552299) with sequence. ((b), (c)) The list of ordered segments is provided below.

project and a password. The example output web page for a typical query protein sequence is presented in figure 2.

4. Conclusions

The high-throughput genome and transcriptome sequencing projects provide large sets of proteomic data. There is a need for robust technologies of annotation of novel functions within these sets and automatic selection of proteins of yet unknown folding for structural genomics projects [16]. The recent developments in the field of fold prediction [17] allowed the researchers to introduce methods of sequence-based annotation of protein function, using the distant sequence similarity [18]. Still, in order to recognize a given protein fold, there is a need for high-throughput protein modelling [19]. Such ‘virtual structural genomics’ is a critical approach to the annotation of a function in divergent genomes like eukarotic *Plasmodium spp.* [20], the divergent and fast evolving *Mollicutes* bacteria [21, 22] and large viral genomes like *Herpesviridae* [23].

In this paper, we have presented a fast algorithm to address the needs of defining domain boundaries, and we provide its implementation as a web service that can be used to scan for globular domains within multi-domain proteins. We have introduced a novel method of scoring of interdomain spacers (ZAAT) based on a Z-score of the distribution of amino acid triple frequency ratios in a defined domain to full length protein. Our approach was successfully applied in the annotation of complex genomes (diatomes *Thalassiosira pseudonana* and *Phaeodactylum tricornutum*, circa 22 000 proteins) and divergent species like human-pathogenic *Herpesviridae* (eight species; total about 600 proteins). The ProteinSplit web server has considerably speeded up the protein structure prediction and allowed for rapid selection of potential domain boundaries needed by large-scale virtual genomics studies.

Benchmark results show clearly that our tool is able to properly identify ordered and disordered regions of a protein sequence. The testing set of protein sequences was acquired from critical assessment of the fully automated structure prediction (CAFASP) web site at

(b)

```
>gi|6552299|ref|NP_009225.1| breast cancer 1, early onset; breast-ovarian cancer, included
[Homo sapiens]
MDLSALRVEEVQNVINAMQKILECPICLELIKEPVSTKCDHIFCKFCMLKLLNQKKGPSQCPLCKNDITK
RSLQESTRFSQLVEELLKIICAFQLDTGLEYANSYNFAKKENNSPEHLKDEVSIIQSMGYRNRAKRLLQS
EPENPSLQETSLSVQLSNLGTVRTLRTKQRIQPQKTSVYIELGSDSSEDTVNKATYCSVGDQELLQITPQ
GTRDEISLDSAKKAACEFSETDVTNTEHHQPSNNDLNTTEKRAAERHPEKYQGSSVSNLHVEPCGTNTHA
SSLQHENSSLLLTKDRMNVEKAEFCNKSKQPGLARSQHNRWAGSKETCNDRRTPSTEKKVDLNADPLCER
KEWNKQKLPCSENPRDTEDVPWITLNSSIQKVNEWFSRSDELLGSDDSHDGESESNAKVADVLDVLNEVD
EYSGSSEKIDLLASDPHEALICKSERVHSKSVESNIEDKIFGKTYRKKASLPNLSHVTENLIIGAFVTEP
QIIQERPLTNKLKRKRRPTSGLHPEDFIKKADLAVQKTPEMINQGTNQTEQNGQVMNITNSGHENKTKGD
SIQNEKNPNPIESLEKESAFKTKAEPISSSISNMELELNIHNSKAPKKNRLRRKSSTRHIHALELVVSRN
LSPPNCTELQIDSCSSSEEIKKKKYNQMPVRHSRNLQLMEGKEPATGAKKSNKPNEQTSKRHDSDTFPEL
KLTNAPGSFTKCSNTSELKEFVNPSLPREEKEEKLETVKVSNNAEDPKDLMLSGERVLQTERSVESSSIS
LVPGTDYGTQESISLLEVSTLGKAKTEPNKCVSQCAAFENPKGLIHGCSKDNRNDTEGFKYPLGHEVNHS
RETSIEMEESELDAQYLQNTFKVSKRQSFAPFSNPGNAEEECATFSAHSGSLKKQSPKVTFECEQKEENQ
GKNESNIKPVQTVNITAGFPVVGQKDKPVDNAKCSIKGGSRFCLSSQFRGNETGLITPNKHGLLQNPYRI
PPLFPIKSFVKTKCKKNLLEENFEEHSMSPEREMGNENIPSTVSTISRNNIRENVFKEASSSNINEVGSS
TNEVGSSINEIGSSDENIQAELGRNRGPKLNAMLRLGVLQPEVYKQSLPGSNCKHPEIKKQEYEEVVQTV
NTDFSPYLISDNLEQPMGSSHASQVCSETPDDLLDDGEIKEDTSFAENDIKESSAVFSKSVQKGELSRSP
SPFTHTHLAQGYRRGAKKLESSEENLSSEDEELPCFQHLLFGKVNNIPSQSTRHSTVATECLSKNTEENL
LSLKNSLNDCSNQVILAKASQEHHLSEETKCSASLFSSQCSELEDLTANTNTQDPFLIGSSKQMRHQSES
QGVGLSDKELVSDDEERGTGLEENNQEEQSMDSNLGEAASGCESETSVSEDCSGLSSQSDILTTQQRDTM
QHNLIKLQQEMAELEAVLEQHGSQPSNSYPSIISDSSALEDLRNPEQSTSEKAVLTSQKSSEYPISQNPE
GLSADKFEVSADSSTSKNKEPGVERSSPSKCPSLDDRWYMHSCSGSLQNRNYPSQEELIKVVDVEEQQLE
ESGPHDLTETSYLPRQDLEGTPYLESGISLFSDDPESDPSEDRAPESARVGNIPSSTSALKVPQLKVAES
AQSPAAAHTTDTAGYNAMEESVSREKPELTASTERVNKRMSMVVSGLTPEEFMLVYKFARKHHITLTNLI
TEETTHVVMKTDAEFVCERTLKYFLGIAGGKWVVSYFWVTQSIKERKMLNEHDFEVRGDVVNGRNHQGPK
RARESQDRKIFRGLEICCYGPFTNMPTDQLEWMVQLCGASVVKELSSFTLGTGVHPIVVVQPDAWTEDNG
FHAIGQMCEAPVVTREWVLDSVALYQCQELDTYLIPQIPHSHY
```

Figure 2. (Continued.)

http://cafasp4.cse.buffalo.edu/dp/update.html. The domain prediction section included the evaluation of 58 protein sequences, including 41 single-domain targets and 17 two-domain targets. The secondary structure was identified for those protein sequences by the PSIPRED tool and was known from the experimental crystallized structure. The list of ordered and disordered regions and list of domain prediction of nearly 12 different domain prediction algorithms (for example, DomSSEA and DPS + DomSSEA by Jones *et al* [24], mateo by Lexa and Valle [25], dopro by von Ohsen, SSEP by Gewehr and Zimmer [26], ADDA by Heger and Holm [27], or armadillo by Dumontier *et al* [28]) was then compared with prediction of our algorithm. The initial results show that our tool is not able to perform better than most of those algorithms, yet it provides richer information about ordered and disordered regions of a protein sequence, and it is much faster and therefore can be used in virtual high-throughput experiments. The detailed results of the comparison are available on the ProteinSplit web server pages. We conclude that the ProteinSplit algorithm is able to predict domain boundaries by splitting a protein sequence into local segments with low and high structural stability.

Acknowledgments

This work has been supported by MEiN and the European Commission with grants LSHG-CT-2004-512035 and LSHG-CT-2003-503265 and MEiN with grant 2P05A00130. LSW is a fellow of the Young Scientist Award from The Foundation for Polish Science.

(c) A sequence is then splitted into ordered segments:

```
>gi|6552299|ref|NP_009225.1| (1-127)
MDLSALRVEEVQNVINAMQKILECPICLELIKEPVSTKCDHIFCKFCMLKLLNQKKGPSQCPLCKNDITKRSLQESTRFSQLVEELLKIICAFQ
LDTGLEYANSYNFAKKENNSPEHLKDEVSIIQS
>gi|6552299|ref|NP_009225.1| (128-272)
MGYRNRAKRLLQSEPENPSLQETSLSVQLSNLGTVRTLRTKQRIQPQKTSVYIELGSDSSEDTVNKATYCSVGDQELLQITPQGTRDEISLDSA
KKAACEFSETDVTNTEHHQPSNNDLNTTEKRAAERHPEKYQGSSVSNLHVE
>gi|6552299|ref|NP_009225.1| (273-430)
PCGTNTHASSLQHENSSLLLTKDRMNVEKAEFCNKSKQPGLARSQHNRWAGSKETCNDRRTPSTEKKVDLNADPLCERKEWNKQKLPCSENPRD
TEDVPWITLNSSIQKVNEWFSRSDELLGSDDSHDGESESNAKVADVLDVLNEVDEYSGSSEKID
>gi|6552299|ref|NP_009225.1| (431-560)
LLASDPHEALICKSERVHSKSVESNIEDKIFGKTYRKKASLPNLSHVTENLIIGAFVTEPQIIQERPLTNKLKRKRRPTSGLHPEDFIKKADLA
VQKTPEMINQGTNQTEQNGQVMNITNSGHENKTKGD
>gi|6552299|ref|NP_009225.1| (561-710)
SIQNEKNPNPIESLEKESAFKTKAEPISSSISNMELELNIHNSKAPKKNRLRRKSSTRHIHALELVVSRNLSPPNCTELQIDSCSSSEEIKKKK
YNQMPVRHSRNLQLMEGKEPATGAKKSNKPNEQTSKRHDSDTFPELKLTNAPGSFT
>gi|6552299|ref|NP_009225.1| (711-878)
KCSNTSELKEFVNPSLPREEKEEKLETVKVSNNAEDPKDLMLSGERVLQTERSVESSSISLVPGTDYGTQESISLLEVSTLGKAKTEPNKCVSQ
CAAFENPKGLIHGCSKDNRNDTEGFKYPLGHEVNHSRETSIEMEESELDAQYLQNTFKVSKRQSFAPFSNPGNA
>gi|6552299|ref|NP_009225.1| (879-1003)
EEECATFSAHSGSLKKQSPKVTFECEQKEENQGKNESNIKPVQTVNITAGFPVVGQKDKPVDNAKCSIKGGSRFCLSSQFRGNETGLITPNKHG
LLQNPYRIPPLFPIKSFVKTKCKKNLLEENF
>gi|6552299|ref|NP_009225.1| (1004-1109)
EEHSMSPEREMGNENIPSTVSTISRNNIRENVFKEASSSNINEVGSSTNEVGSSINEIGSSDENIQAELGRNRGPKLNAMLRLGVLQPEVYKQS
LPGSNCKHPEIK
>gi|6552299|ref|NP_009225.1| (1110-1273)
KQEYEEVVQTVNTDFSPYLISDNLEQPMGSSHASQVCSETPDDLLDDGEIKEDTSFAENDIKESSAVFSKSVQKGELSRSPSPFTHTHLAQGYR
RGAKKLESSEENLSSEDEELPCFQHLLFGKVNNIPSQSTRHSTVATECLSKNTEENLLSLKNSLNDCSNQ
>gi|6552299|ref|NP_009225.1| (1274-1438)
VILAKASQEHHLSEETKCSASLFSSQCSELEDLTANTNTQDPFLIGSSKQMRHQSESQGVGLSDKELVSDDEERGTGLEENNQEEQSMDSNLGE
AASGCESETSVSEDCSGLSSQSDILTTQQRDTMQHNLIKLQQEMAELEAVLEQHGSQPSNSYPSIISDSSA
>gi|6552299|ref|NP_009225.1| (1439-1572)
LEDLRNPEQSTSEKAVLTSQKSSEYPISQNPEGLSADKFEVSADSSTSKNKEPGVERSSPSKCPSLDDRWYMHSCSGSLQNRNYPSQEELIKVV
DVEEQQLEESGPHDLTETSYLPRQDLEGTPYLESGISLFS
>gi|6552299|ref|NP_009225.1| (1573-1732)
DDPESDPSEDRAPESARVGNIPSSTSALKVPQLKVAESAQSPAAAHTTDTAGYNAMEESVSREKPELTASTERVNKRMSMVVSGLTPEEFMLVY
KFARKHHITLTNLITEETTHVVMKTDAEFVCERTLKYFLGIAGGKWVVSYFWVTQSIKERKMLNEH
>gi|6552299|ref|NP_009225.1| (1733-1863)
DFEVRGDVVNGRNHQGPKRARESQDRKIFRGLEICCYGPFTNMPTDQLEWMVQLCGASVVKELSSFTLGTGVHPIVVVQPDAWTEDNGFHAIGQ
MCEAPVVTREWVLDSVALYQCQELDTYLIPQIPHSHY
```

Figure 2. (Continued.)

Conflict of interest statement. None declared.

References

[1] Ekman D, Bjorklund A, Frey-Skott J and Elofsson A 2005 Multi-domain proteins in the three kingdoms of life: orphan domains and other unassigned regions *J. Mol. Biol.* **348** 231–43

[2] Ginalski K, Grishin N, Godzik A and Rychlewski L 2005 Practical lessons from protein structure prediction *Nucl. Acids Res.* **33** 1874–91

[3] von Grotthuss M, Pas J, Wyrwicz L, Ginalski K and Rychlewski L 2003 Application of 3D-Jury, GRDB, and Verify3D in fold recognition *Proteins* **53** 418–23

[4] Saini H and Fischer D 2005 Meta-DP: domain prediction meta-server *Bioinformatics* **21** 2917–20

[5] Linding R, Russell R, Neduva V and Gibson T 2003 GlobPlot: exploring protein sequences for globularity and disorder *Nucl. Acids Res.* **31** 3701–8

[6] Deleage G and Roux B 1987 An algorithm for protein secondary structure prediction based on class prediction *Protein Eng.* **1** 289–94

[7] Brendel V, Bucher P, Nourbakhsh I, Blaisdell B and Karlin S 1992 Methods and algorithms for statistical analysis of protein sequences *Proc. Natl Acad. Sci. USA* **89** 2002–6

[8] Finn R *et al* 2006 Pfam: clans, web tools and services *Nucl. Acids Res.* **34** D247–51

[9] Berman H, Westbrook J, Feng Z, Gilliland G, Bhat T, Weissig H, Shindyalov I and Bourne P 2000 The protein data bank *Nucl. Acids Res.* **28** 235–42

[10] Kyte J and Doolittle R 1982 A simple method for displaying the hydropathic character of a protein *J. Mol. Biol.* **157** 105–32

[11] Howard M, Ekborg N, Taylor L, Hutcheson S and Weiner R 2004 Identification and analysis of polyserine linker domains in prokaryotic proteins with emphasis on the marine bacterium Microbulbifer degradans *Protein Sci.* **13** 1422–5

[12] Wheeler D *et al* 2006 Database resources of the national center for biotechnology information *Nucl. Acids Res.* **34** D173–80

[13] Goulet B, Watson P, Poirier M, Leduy L, Berube G, Meterissian S, Jolicoeur P and Nepveu A 2002 Characterization of a tissue-specific CDP/Cux isoform, p75, activated in breast tumor cells *Cancer Res.* **62** 6625–33

[14] Lash A, Tolstoshev C, Wagner L, Schuler G, Strausberg R, Riggins G and Altschul S 2000 SAGEmap: a public gene expression resource *Genome Res.* **10** 1051–60

[15] Sonnhammer E, Eddy S and Durbin R 1997 Pfam: a comprehensive database of protein domain families based on seed alignments *Proteins* **28** 405–20

[16] Todd A, Marsden R, Thornton J and Orengo C 2005 Progress of structural genomics initiatives: an analysis of solved target structures *J. Mol. Biol.* **348** 1235–60

[17] Rychlewski L, Fischer D and Elofsson A 2003 LiveBench-6: large-scale automated evaluation of protein structure prediction servers *Proteins* **53** (Suppl. 6) 542–7

[18] Ginalski K, von Grotthuss M, Grishin N and Rychlewski L 2004 Detecting distant homology with Meta-BASIC *Nucl. Acids Res.* **32** W576–81

[19] Kelley L, MacCallum R and Sternberg M 2000 Enhanced genome annotation using structural profiles in the program 3D-PSSM *J. Mol. Biol.* **299** 499–520

[20] Gerloff D, Creasey A, Maslau S and Carter R 2005 Structural models for the protein family characterized by gamete surface protein Pfs230 of Plasmodium falciparum *Proc. Natl Acad. Sci. USA* **102** 13598–603

[21] Rychlewski L, Zhang B and Godzik A 1998 Fold and function predictions for Mycoplasma genitalium proteins *Fold Des.* **3** 229–38

[22] Fischer D and Eisenberg D 1997 Assigning folds to the proteins encoded by the genome of Mycoplasma genitalium *Proc. Natl Acad. Sci. USA* **94** 11929–34

[23] Holzerlandt R, Orengo C, Kellam P and Alba M 2002 Identification of new herpesvirus gene homologs in the human genome *Genome Res.* **11** 1739–48

[24] Marsden R L, McGuffin L J and Jones D T 2002 Rapid protein domain assignment from amino acid sequence using predicted secondary structure *Protein Sci.* **11** 2814–24

[25] Lexa M and Valle G 2003 PRIMEX: rapid identification of oligonucleotide matches in whole genomes *Bioinformatics* **19** 2486–8

[26] Gewehr J E and Zimmer R 2006 SSEP-Domain: protein domain prediction by alignment of secondary structure elements and profiles *Bioinformatics* **22** 181–7

[27] Heger A and Holm L 2003 Exhaustive enumeration of protein families *J. Mol. Biol.* **328** 749–67

[28] Dumontier M, Yao R, Feldman H J and Hogue C W 2005 Armadillo: domain boundary prediction by amino acid composition *J. Mol. Biol.* **350** 1061–73