Nucleic Acids Research, 2004, Vol. 32, Web Server issue ***W569–W571***
DOI: 10.1093/nar/gkh481

CHOP: parsing proteins into structural domains

Jinfeng Liu[1,2,4] and Burkhard Rost[1,2,3,*]

[1]CUBIC and [2]North East Structural Genomics Consortium (NESG), Department of Biochemistry and Molecular Biophysics, Columbia University, 650 West 168th Street BB217, New York, NY 10032, USA, [3]Columbia University Center for Computational Biology and Bioinformatics (C2B2), Russ Berrie Pavilion, 1150 Saint Nicholas Avenue, New York, NY 10032, USA and [4]Department of Pharmacology, Columbia University, 630 West 168th Street, New York, NY 10032, USA

Received February 14, 2004; Revised April 16, 2004; Accepted May 4, 2004

ABSTRACT

Sequence-based domain assignment is one of the most important and challenging problems in structural biology. We have developed a method, CHOP, that chops proteins into domain-like fragments. The basic idea is to cut proteins from entirely sequenced organisms beginning from very reliable experimental information (Protein Data Bank), proceeding to expert annotations of domain-like regions (Pfam-A) and completing through cuts based on termini of native protein ends. The CHOP server takes protein sequences as input and returns the dissections supported by homology transfer. CHOP results are precompiled for many entirely sequenced proteomes. The service is available at http://www.rostlab.org/services/CHOP/.

INTRODUCTION

Domains are the structural units of proteins. Many large proteins can be viewed as combinatorial arrangements of protein domains. Domains can be defined as semi-independent three-dimensional (3D) units in proteins. The assumption is that such units fold independently. Structural domains often have particular functions, and are observed to be genetically mobile. Knowing the domain organization of a protein sequence is often a crucial starting point for advancing the understanding of its structure and function by experimental and computational means. For example, for a swift and accurate experimental determination of 3D structures of multidomain proteins, it is often necessary to split the proteins into domains and then determine the structures for these independently. Knowing the domain arrangement of a protein also improves the reliability of comparative sequence analysis. Other potential uses of domain information include yeast two-hybrid systems: basing constructs on domains rather than on entire proteins is likely to improve the sensitivity and accuracy of such techniques.

Methods that identify domain-like regions. Numerous methods succeed largely in identifying structural domains from the detailed co-ordinates of 3D structures. In contrast, sequence-based domain assignment remains one of the most challenging problems in structural biology. One of the problems with many of these methods is that the resulting fragments are in fact much shorter than typical structural domains (1). We have developed CHOP, a homology-based method that chops proteins into domain-like fragments (2,3). CHOP is designed to be conservative in the sense that it does not dissect proteins without strong reasons. While CHOP dissects most proteins, ~20–40% of the fragments that CHOP generates are likely to contain more than one domain.

METHODS AND RESULTS

Algorithm

CHOP implements three hierarchical steps that are applied in order of decreasing confidence in the accuracy of the information, namely, the structural domain extracted by PrISM (4) from PDB (Protein Data Bank) (5), the sequence domain as defined in Pfam-A (6) and the termini of SWISS-PROT (7) proteins. The detailed rationale and procedure have been published elsewhere (2). Here, we can only briefly sketch the basic concept. First, the query protein is BLASTed against 3D structure domains identified by PrISM (4). Fragments overlapping at least 80% of the PrISM domain with significant sequence similarity are spliced out of the protein and two fragments N- and C-terminal to this chop are considered in the subsequent steps. Second, all remaining fragments are used independently to search against Pfam-A (6) entries longer than 30 residues using HMMer (8). Again, fragments that match over at least 80% of the Pfam-A entry are spliced out. Third, only the fragments that did not match in the previous two steps are BLASTed against full-length proteins in SWISS-PROT (7).

*To whom correspondence should be addressed at Department of Biochemistry and Molecular Biophysics, Columbia University, 650 W 168th Street, BB-217, New York, NY 10032, USA. Tel: +1 212 305 4018; Fax: +1 212 305 7932; Email: rost@columbia.edu

The online version of this article has been published under an open access model. Users are entitled to use, reproduce, disseminate, or display the open access version of this article provided that: the original authorship is properly and fully attributed; the Journal and Oxford University Press are attributed as the original place of publication with the correct citation details given; if an article is subsequently reproduced or disseminated not in its entirety but only in part or as a derivative work this must be clearly indicated.

© 2004, the authors

Again, significant similarities to full-length SWISS-PROT proteins lead to a chop. At each step fragments that are shorter than 30 residues are removed from the stack of 'remaining fragments'. The final set of fragments is the combination of all fragments identified in the three steps and all remaining fragments that are longer than 30 residues.

Results

We applied the three steps of CHOP to all proteins/open reading frames (ORFs) from 62 entirely sequenced organisms (2): 150 308 (63%) of the 238 492 proteins were dissected by CHOP (Figure 1A). CHOP fragments on average resembled much more generic structural domains than the fragments generated by other domain-dissection methods (1–3). In particular, the subset of proteins for which we find no reason to chop are obviously enriched in single-domain proteins. Similarly, the fragments remaining after the three steps of CHOP are also enriched in single domains. Thus, CHOP succeeds in a first pass to break proteins into their structural components. The major problem that remains is the limited coverage: ~20–40% of all CHOP fragments could be dissected further if we had more information. While this number will shrink with the growth of PDB and Pfam-A, we have recently also developed a method (CHOPnet) that identifies domain boundaries in the absence of annotations by predicting domains from sequence. Although CHOP remains incomplete, the vast majority of all proteins in the 62 entirely sequenced proteomes that we have analysed contain more than one domain-like fragment. In fact, <30% of the subset of all proteins chopped contained only one fragment (Figure 1B). Two other results from our previous work are remarkable. First, most structural domains appear to be only about 100 residues long. Second, there appears to be a significant difference between the length of the longest and all other domains in a protein. Our observation may suggest that there is some minimal length for proteins with one domain and that proteins with more than one domain on average are built of one long and many short domains.

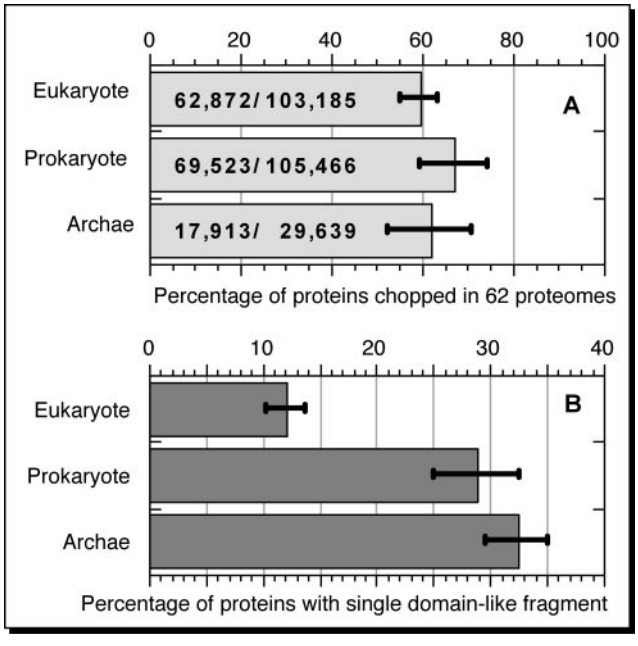

Figure 1. Percentage of proteins chopped in 62 entire proteomes. (**A**) About two-thirds of the proteins from 62 proteomes can be chopped (absolute number of proteins in bars). (**B**) For the subset of proteins that can be chopped, about 30% of the prokaryotic and archaean proteins contain only one fragment that is homologous to either PrISM domains or Pfam-A fragments; this single domain fraction is considerably lower for eukaryotic proteins. Bars indicate the standard deviation of the distribution for these numbers over all three kingdoms.

CHOP insensitive to parameter changes. CHOP is a new method that was developed using all existing data; this makes it very difficult to assess its accuracy. In the context of our involvement in structural genomics, we are currently testing the validity of CHOP assignments experimentally. One way of assessing the reliability of CHOP is to explore its consistency with respect to parameter and sequence changes. We found CHOP assignments to be surprisingly robust with respect to changing the two free CHOP parameters, i.e. the minimal coverage of the known domain by the query alignment (currently 80%) and the minimal level of sequence similarity to consider an alignment (currently BLAST *E*-values of 10^{-3} for PrISM). For fragments created in 62 entirely sequenced proteomes (2), the domain boundaries from similarity to SWISS-PROT proteins were rarely in conflict with those from similarity to PrISM domains (0.2%) and Pfam domains (1%). For proteins that could be chopped according to both PrISM and Pfam-A, the number of domain-like fragments resulting from applying the two methods independently was largely consistent: 40% of the proteins showed no difference, and 34% differed by one domain (2).

Over 90% of linker regions consistent. Finally, we carried out another experiment. First, we found the largest sequence-unique subset of PrISM domains (~5000) and removed all other PrISM domains from the knowledge base used by CHOP. Then we applied CHOP to all remaining (~28 000) PDB chains and assessed the difference between the CHOP results and the original PrISM assignments for the same proteins. For 94% of all multidomain proteins for which the number of domains agreed, the CHOP predictions of the regions between domains were within 10 residues of the linker regions assigned by PrISM. To put this number into perspective, it corresponds to the agreement between SCOP (9) and CATH (10) linker regions. However, the number of domains agreed only for 57% of the assignments (incidentally a similar agreement to that between PrISM and Pfam). Studying these results in more detail, we uncovered that not all disagreements between PrISM and CHOP indicated errors. For example, the T7 RNA polymerase (PDB code: 1aro, chain P) was the protein with the largest difference in our cross-validation test. It was chopped into five domains according to sequence similarity to five different PrISM domains. Different structure-based domain assignment methods gave very different assignments for this protein: PrISM assigned a single protein domain of 878 residues; SCOP (9) indicated this to be a multidomain protein but did not assign domain boundaries, nor specify the number of domains; CATH (10) assigned five domains (one of which was discontinuous in sequence), and most domain boundaries roughly matched those predicted by CHOP. Although this example may not explain all the differences found in our cross-validation experiment, it sheds light on the problems of assigning domains even given the full details of 3D structures.

INPUT, OUTPUT AND JOB OPTIONS

Input and output. The CHOP server accepts sequence information in FASTA sequence format or protein identifiers from SWISS-PROT (11). Sequences can also be uploaded from the user's local machine. Proteins shorter than 30 residues are returned unprocessed. The CHOP output contains information about the putative domains, such as the position of the CHOP fragments in the protein, source of the homology (i.e. the reason why CHOP identified a fragment) and the statistical significance (*E*-value) for the domain assignment. Users have the option of receiving plain text (ASCII) output or HTML formatted results that can be displayed in any web browser. In HTML results, hyperlinks to the corresponding PDB (5) and Pfam (6) entries are provided. Users can choose to receive the results by email or directly online through the same browser window used for submitting the query.

Advanced options. CHOP was not sensitive to our choice of parameters, such as BLAST/PSI-BLAST (12) *E*-values, HMMER (8) *E*-values and the minimum coverage of known domains (2). Advanced users who venture to experiment with these parameters have five options: they can define the BLAST *E*-value threshold (default = 0.01), the HMMER *E*-value threshold (default = 0.01), the minimum coverage of the domain used to chop (default = 80%, i.e. the user's protein has to cover at least 80% of the residues in the known domain), the minimum length of proteins or protein fragments that are processed (default = 30) and the minimum length of fragments that are reported in the results (default = 30).

Searching for CHOP fragments in the PEP database. We have applied CHOP to over 100 entirely sequenced organisms; for about 70 of these the data is available through our PEP database of predictions for entire proteomes (13). PEP is accessed through an SRS interface [Sequence Retrieval System (14)]. This allows queries with over 40 different data fields such as name, function, length, number of membrane segments, subcellular localization and sequence. Flat files with CHOP assignments are also available for download.

Future extensions. We hope to extend the CHOP server in the near future by integrating a method that predicts structural domains directly from sequence. We will also update our database of CHOP predictions for entire proteomes [PEP (13)] in order to ease access to these data.

ACKNOWLEDGEMENTS

Thanks to An-Suei Yang (Columbia) for his help in using PrISM and to our experimental colleagues at the Northeast Structural Genomics Consortium (NESG) for their advice and strong support of our project. In particular, thanks to Guy Montelione (Rutgers) for his invaluable optimism in leading the NESG team, to Barry Honig (Columbia) and Diana Murray (Cornell) for the fruitful collaboration on the NESG target strategy and to the teams around Tom Acton (Rutgers), Cheryl Arrowsmith and Aled Edwards (Toronto) for testing our method. Thanks also to all those who deposit their experimental data in public databases, and to those who maintain these databases. This work was supported by the grants 1-P50-GM62413-01, RO1-GM63029-01 and R01-GM64633-01 from the National Institutes of Health (NIH), and R01-LM07329-01 from the National Library of Medicine (NLM).

REFERENCES

1. Liu,J. and Rost,B. (2003) Domains, motifs, and clusters in the protein universe. *Curr. Opin. Chem. Biol.*, **7**, 5–11.
2. Liu,J. and Rost,B. (2004) CHOP proteins into structural domain-like fragments. *Proteins*, **55**, 678–688.
3. Liu,J., Acton,T., Goldsmith,S., Honig,B., Montelione,G.T. and Rost,B. (2004) Automatic target selection for structural genomics on eukaryotes. *Proteins* (in press).
4. Yang,A.S. and Honig,B. (2000) An integrated approach to the analysis and modeling of protein sequences and structures. II. On the relationship between sequence and structural similarity for proteins that are not obviously related in sequence. *J. Mol. Biol.*, **301**, 679–689.
5. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.
6. Bateman,A., Birney,E., Cerruti,L., Durbin,R., Etwiller,L., Eddy,S.R., Griffiths-Jones,S., Howe,K.L., Marshall,M. and Sonnhammer,E.L. (2002) The Pfam protein families database. *Nucleic Acids Res.*, **30**, 276–280.
7. Bairoch,A. and Apweiler,R. (2000) The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.*, **28**, 45–48.
8. Eddy,S.R. (1998) Profile hidden Markov models. *Bioinformatics*, **14**, 755–763.
9. Lo Conte,L., Brenner,S.E., Hubbard,T.J., Chothia,C. and Murzin,A.G. (2002) SCOP database in 2002: refinements accommodate structural genomics. *Nucleic Acids Res.*, **30**, 264–267.
10. Pearl,F.M., Bennett,C.F., Bray,J.E., Harrison,A.P., Martin,N., Shepherd,A., Sillitoe,I., Thornton,J. and Orengo,C.A. (2003) The CATH database: an extended protein family resource for structural and functional genomics. *Nucleic Acids Res.*, **31**, 452–455.
11. Boeckmann,B., Bairoch,A., Apweiler,R., Blatter,M.C., Estreicher,A., Gasteiger,E., Martin,M.J., Michoud,K., O'Donovan,C., Phan,I. *et al.* (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.*, **31**, 365–370.
12. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
13. Carter,P., Liu,J. and Rost,B. (2003) PEP: predictions for entire proteomes. *Nucleic Acids Res.*, **31**, 410–413.
14. Etzold,T., Ulyanov,A. and Argos,P. (1996) SRS: information retrieval system for molecular biology data banks. *Meth. Enzymol.*, **266**, 114–128.