BIOINFORMATICS APPLICATIONS NOTE Vol. 19 no. 5 2003, pages 673–674
DOI: 10.1093/bioinformatics/btg031

DomCut: prediction of inter-domain linker regions in amino acid sequences

Mikita Suyama [1,*,†] and Osamu Ohara [1, 2]

[1]Laboratory of DNA Technology, Kazusa DNA Research Institute, 1532-3 Yana, Kisarazu, Chiba 292-0812, Japan and [2]Immunogenomics, RIKEN Research Center for Allergy and Immunology, RIKEN Yokohama Institute, 1-7-22, Suehiro-cho, Tsurumi, Yokohama, Kanagawa, Japan

Received on May 27, 2002; revised on September 18, 2002; accepted on October 25, 2002

ABSTRACT

Summary: DomCut is a program to predict inter-domain linker regions solely by amino acid sequence information. The prediction is made by using linker index deduced from a data set of domain/linker segments. The linker preference profile, which is the averaged linker index along a sequence, can be visualized in the graphical interface.

Availability: The web server, together with supplementary information, is available at http://www.kazusa.or.jp/tech/suyama/domcut. The distribution version of DomCut is also available upon request from the authors.

Contact: suyama@embl-heidelberg.de

INTRODUCTION

The prediction of linker regions can play an important role in structural analysis of large proteins by NMR and X-ray crystallographic studies, where excision of a single functional domain without alteration of its characteristics is strongly desired. Moreover, the prediction of linker regions can also be used in the design of truncated proteins, which are widely used in biochemical studies of proteins to map functional domains on the sequences. Downsizing of proteins without loss of their function is one of the major targets of protein engineering and the linker prediction method presented here offers a powerful tool also for this purpose. Here we developed a simple method, DomCut, which predicts linker regions among functional domains based on the difference in amino acid composition between domain and linker regions.

CALCULATION OF LINKER INDEX

First we analyzed amino acid composition in linker and domain regions. To collect domain and linker sequence elements, we used the term 'DOMAIN' in the feature table (FT) of the entries in the SWISS-PROT database (Bairoch and Apweiler, 2000). Only the domains with sequence length ranging from 50 to 500 residues were taken as domains, and some poorly defined domains such as 'CYTOPLASMIC' and 'GLN-RICH' in FT were excluded. We considered a sequence segment to be a linker if the segment satisfies the following conditions: (1) connecting two adjacent domains defined above; (2) in the range from 10 to 100 residues; and (3) not containing membrane spanning regions. Applying the above conditions we obtained 811 sequences with at least one linker region. Excluding homologous sequences (>30% sequence identity) we finally got the non-redundant sequence set that is comprised of 273 sequences (486 linker and 794 domain segments). The average numbers of amino acid residues in the linker and the domain segments were 35.8 and 122.1, respectively.

To represent the preference for amino acid residues in linker regions, we defined the linker index. The linker index S_i for amino acid residue i is calculated as follows:

$$S_i = -\ln\left(\frac{f_i^{\text{linker}}}{f_i^{\text{domain}}}\right),$$

where $f_i^{\text{linker(domain)}}$ is the frequency of amino acid residue i in the linker or domain region. The negative value of S_i means that the amino acid preferably exists in a linker region (Figure 1a). Proline residues are especially abundant in linker regions ($S_{\text{Pro}} = -0.478$). On the other hand, glycine residues are strongly preferred in domain regions ($S_{\text{Gly}} = 0.331$). This is an unexpected observation because glycine residues, which confer flexibility to a polypeptide chain, are widely used for linkers in artificial proteins.

LINKER PREFERENCE PROFILE

A linker preference profile is generated by plotting the averaged linker index values along an amino acid sequence using a siding window. This is the same procedure

*To whom correspondence should be addressed
† Present address: Biocomputing, EMBL, Meyerhofstrasse 1, 69012 Heidelberg, Germany.

A

Amino Acid	S_i	Amino Acid	S_i
Pro*	-0.478	Asn	0.027
Ser*	-0.177	Ile	0.060
Thr*	-0.163	Leu*	0.138
Glu*	-0.128	His*	0.195
Lys*	-0.112	Phe*	0.240
Gln	-0.073	Met*	0.275
Ala	-0.058	Tyr*	0.322
Val	-0.052	Gly*	0.331
Arg	0.000	Cys*	0.447
Asp	0.016	Trp*	0.564

B

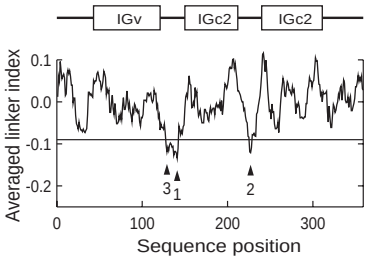

Fig. 1. Linker index and an example of linker preference profile. (a) Linker index. Amino acids are sorted by their linker index values. The numbers of individual amino acids were compared between the linker and the domain regions by a χ^2 test. Amino acids with significant difference ($P < 10^{-3}$) are indicated by an asterisk (*). (b) An example of linker preference profile generated by DomCut. The protein (Accession number: Q24372) is comprised of three domains: one immunoglobulin-like V-type domain (IGv) and two immunoglobulin-like C2-type domains (IGc2). The domain organization is drawn to scale at the top of the profile. A horizontal line is drawn at the averaged linker index value -0.09. Troughs lower than this value are indicated by triangles with their ranks.

used in the hydropathy plot (Kyte and Doolittle, 1982). In DomCut, the window size $w = 15$ is used since it gives the best performance. An example of the output is shown in Figure 1b. This protein, Lachesin (Accession number: Q24372), does not have a highly similar sequence (>40% sequence identity) in the data set, from which the linker index was calculated. The linker regions clearly correspond to the troughs of the profile.

EVALUATION OF LINKER PREDICTION ACCURACY

To evaluate the accuracy of the prediction of linker positions on the basis of the linker preference profiles, a jack-knife test was applied to the 273 representative sequences, i.e. all but one of the sequences were used to calculate the linker index and the remaining one sequence was subjected to the prediction. This procedure was repeated until all the sequences were predicted. A linker was taken to be correctly predicted if there was a trough in the linker region and the averaged linker index value at the minimum of the trough was lower than the threshold value. The regions without domain or linker assignment were taken to be uncertain regions, and a prediction which fell within that region was not counted neither as a correct prediction nor as a false one. The prediction accuracy varied with the threshold value of averaged linker index for the trough detection. At the threshold value -0.09, sensitivity (proportion of the total number of successfully predicted linkers against the total number of linkers) is 53.5% and selectivity (proportion of correct predictions in all predictions) is 50.1%. If a trough has a very low averaged linker index, the prediction is highly confident: for example, for troughs with the averaged linker index less than -0.15, the selectivity of prediction is 73.4%.

There are several methods for protein domain boundary or linker predictions, such as the SEG program (Wootton, 1994), the DGS algorithm (Wheelan *et al.*, 2000) and the SnapDRAGON program (George and Heringa, 2002). It is difficult to directly compare the accuracy of the predictions because all of these programs use different criteria for assessing the predictive power. Moreover, these programs use completely different characteristics in the prediction: DomCut uses the difference of amino acid composition between domain and linker regions, while SEG, DGS and SnapDRAGON are based on complexity of sequence, length distribution of known three-dimensional (3D) domain structures and *ab initio* 3D model construction, respectively. Since these prediction methods predict domain boundaries or linker regions from different aspects, combined use of these methods would improve accuracy and reliability of the prediction as a whole.

IMPLEMENTATION

The web server takes an amino acid sequence as input. There are three output formats that can be specified by the user. The first format is GIF. This is the default setting, and the results are directly visualized in the web browser. The second format is PostScript, which is suitable for printing and can also be edited by drawing software. The third format is raw data in text. The data in this format can be imported to a spreadsheet to draw user's own graph by graph drawing software. There is also a distribution version (Perl scripts) to run in the command line.

ACKNOWLEDGEMENTS

We thank Hidekazu Hiroaki for valuable suggestions and discussion.

REFERENCES

Bairoch,A. and Apweiler,R. (2000) The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.*, **28**, 45–48.

George,R.A. and Heringa,J. (2002) SnapDRAGON: a method to delineate protein structural domains from sequence data. *J. Mol. Biol.*, **316**, 839–851.

Kyte,J. and Doolittle,R.F. (1982) A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.*, **157**, 105–132.

Wheelan,S.J., Marchler-Bauer,A. and Bryant,S.H. (2000) Domain size distributions can predict domain boundaries. *Bioinformatics*, **16**, 613–618.

Wootton,J.C. (1994) Non-globular domains in protein sequences: automated segmentation using complexity measures. *Comput. Chem.*, **18**, 269–285.