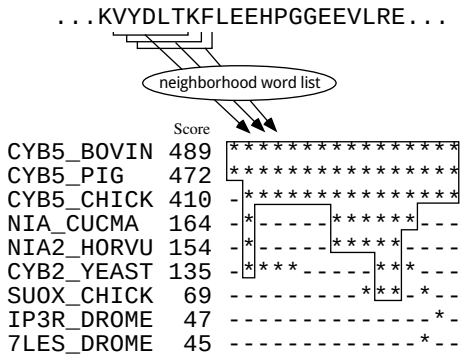

algorithm, every neighborhood word set w_i at position i of the query sequence is a candidate of motif. We set the word length $k = 5$. After the homology search by BLAST algorithm is done, sequences belonging to group g_i are collected at every position i by adding sequences one by one in order of the homology score until occurrence of the word set w_i is greater than a given threshold. Each group g_i is refined by succeeding procedures which eliminate sequences with low similarity by evaluating multiple sequence similarity. Finally, too small groups are discarded because of statistical insignificance.

To test this algorithm, we selected one sequence from the DR field of each PROSITE entry and searched it against SWISS-PROT database by above algorithm to collect related segments. When we used 736 non-overlapping PROSITE groups containing 18094 SWISS-PROT entries, our algorithm found 13455 correct entries (74% coverage) and additional 6948 entries. In many of the latter cases, our algorithm found relationships other than the original PROSITE groups, although some false positives were also found. Since our method can be applied to any query sequence, it is able to collect motifs independently of PROSITE catalog or any other groups.

Acknowledgment

This work was supported in part by a Grant-in-Aid for Scientific Research on Priority Areas 'Genome Informatics' from The Ministry of Education, Science, Sports and Culture in Japan. The Computation time was provided by the Supercomputer Laboratory, the Institute for Chemical Research, Kyoto University.

References

[1] A. Bairoch, P. Bucher, "PROSITE: recent developments" *Nucleic Acids Res.*, Vol. 22, pp. 3583-3589, 1994.

[2] S. Henikoff, J. G. Henikoff, "Automated assembly of protein blocks for database searching" *Nucleic Acids Res.*, Vol. 19, pp. 19-23, 1991.

[3] I. Uchiyama, A. Ogiwara, Z. Ohkubo, M. Kanehisa, "Automatic procedure to extract signature pentapeptides from the protein sequence database" *Proc. Genome Informatics Workshop IV* pp. 255-263, 1993.

[4] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, D. J. Lipman "Basic local alignment search tool" *J. Mol. Biol.*, Vol. 215, pp. 403-410, 1990

Protein Sequence Grouping by Peptide Word Motifs

I. Uchiyama [1]
T. Takagi [1]
uchiyama@ims.u-tokyo.ac.jp
takagi@ims.u-tokyo.ac.jp

A. Ogiwara [1]
M. Kanehisa [2]
ogi@ims.u-tokyo.ac.jp
kanehisa@kuicr.kyoto-u.ac.jp

[1] Human Genome Center,
Institute of Medical Science, the University of Tokyo
4-6-1 Shirokanedai, Minato-ku, Tokyo 108 Japan

[2] Institute for Chemical Research, Kyoto University
Uji, Kyoto, 611 Japan

Abstract

Methods for collecting related segments from the protein sequence database using strongly conserved peptide words as well as sequence homology was applied to the problem of reconstruction of PROSITE catalog [1] from the sequence database. In many case our results were well consistent with PROSITE although some additional relationships were also found.

1 Introduction

Protein motif search becomes one of the most useful approaches to give functional interpretations to newly determined sequences. Because contents of motif dictionaries such as PROSITE are limited by the current knowledge of protein functional sites, automatic discovery of motifs has been a challenging theme in order to expand motif libraries. Two steps are required to construct motif library: 1) grouping sequences and 2) detecting common patterns or conserved regions in these sequence groups. There are several methods automating the second step. BLOCKS [2] is a well known example where conserved segments were automatically extracted from unaligned sequence sets. However, because these blocks were extracted only from sequence groups cataloged in PROSITE, it could not expand PROSITE catalog. It is important to develop a procedure to group sequences.

We have developed methods for compiling related sequences from the entire protein sequence database by very local sequence features [3]. Here we apply our method to the problem of redefining PROSITE catalog to test the ability of automatic grouping.

2 Methods & Results

Our method is based on BLAST algorithm [4]. BLAST initially constructs a word list for initial search which contains all k-length words in the query sequence and their neighborhood words. In our

[1]内山郁夫、荻原淳、高木利久：東京大学医科学研究所ヒトゲノム解析センター，〒 108 東京都港区白金台 4-6-1
[2]金久實：京都大学化学研究所，〒 611 京都府宇治市五ヶ庄