GoodMotif: The System for Finding High Quality Motifs That Can Discriminate a Unique Protein Family

Makiko Suwa suwa@hri.co.jp
Asaf Saramov asaf@hri.co.jp
Tetsuo Nishikawa nisikawa@hri.co.jp
Mark Swindells swintech@hri.co.jp
Helix Research Institute, Inc.
1532-3, Yana, Kisarazu, Chiba, Japan

Abstract

A new computational system was developed, which find high quality motif that can discriminates the unique protein family without false entries. In this system, motif patterns were plotted to the dispersion diagram between observed and expected motif numbers in SWISSPROT database. It was revealed that the motifs with different quality could be clearly assigned on different areas by discrimination line due to the experimental assignment in PROSITE database, together with the alignment analysis of each members of family, relating motifs. The result suggested that the evaluation of motif is feasible using the relation between the observed and the expected of motif match and the similarity scores of family proteins.

1 Introduction

Recent large progress in genome project leads us to the new biological way which compares the protein functions being contained in several genome sets. In that situation, a method for predicting protein functions from amino acid sequences is strongly required. One of the powerful tools for this purpose is the protein motif analysis. Protein motifs are partial sequences that are strongly conserved within relating protein families, with specific function. Therefore, once some motif is discovered, it may be possible to predict the function of the amino acid sequence, which has quite low similarity with other sequences, or is not full length one. In order to predict the function for such sequences, we need to know the sensitivity of each motif that can tell whether the families, linking to motifs, are unique ones or not. Sternberg et al. showed a good evaluation of motif sensitivity [1], using the expected and observed number of patterns in SWISSPROT, scanned by each motif patterns in PROSITE database [2]. However their work can not catch up with the growth of the sequence database, become the evaluation critically, depends with the accuracy of description in PROSITE of old version, assigning true and false positive to the matches. Therefore, absolute evaluation of motifs is required, which calculates and predicts the evaluation value from only amino acid sequences. We developed the computational system to find automatically, the high quality motifs which promise the unique protein families with no false entries from the databases, with locational update.

2 System and Method

2.1 Method

Procedures for motif evaluation are integrated by the software system "Good Motif", which finds out the motif pattern from amino acid sequence, with gathering related sequences and judges whether protein groups is the unique family or not. Detailed procedures are following. First, it scans sequence database by each patterns in motif database, counting the match number (Ni) of the motif

i . Furthermore, expected match of motif i (Nei) was calculated as follows. When the pattern i is "DNA polymerase" pattern, for example, described by (YA)-x-D-T-D-S-(LIVM) as the regular expression, the probability that the motif has is, Pi= (PY+PA)*1*PD*PT*PD*PS*(PL+PI+PV+PM) (1) where, PA, PY, PD, PT, PS, PL, PI, and PM are the probability of each residues: A, Y, D, S, L, I, V and M respectively. x represents the symbol in which any kind of residues are available. The expected match of motif in a scan of Ntot residues in amino acid sequence database is, Nei=Pi.*Ntot (2) At the next step, protein families linked by the motif i were collected. Then each families are judged to be the unique families or not, according to the Ni- Nei dispersion diagrams, together with alignment analysis for the members in each families ([3, 4]).

2.2 Database analysis for determining judging criteria.

SWISSPROT (release 34.) was scanned by motif patterns in PROSITE (release 13. [5]) For the patterns whose positive and false entries are experimentally revealed, observed number of each motif were plotted in Ni-Nei dispersion diagram. First, we excluded the patterns in area where Ni was almost the same to Nei, which was expected by chance, since the patterns are biologically meaningless. Most of protein families with false entries were concentrated in the area where Nei is relatively large (N<1000 and 1.0<Ne<1000), while, other families in the area where (N<100, Ne<<1.0) are almost the unique families with no false entries. For each families, the pair wise alignment scores: the lower limit of related sequences and the upper limit of non-related sequences, were determined.

3 Summary and Conclusions

We developed the system for finding the high quality motif, which can suggest the unique protein family without false entries. In this system, we defined three parameters; The observed and the expected number of motif matches in amino acid sequence, and the scores of alignment analysis. It was suggested that these parameters can be the useful criteria to evaluate the quality of motifs which will be newly determined from sequences, which are rapidly increasing.

References

[1] Sternberg, M.J.E. "Library of common protein motif." *Nature*, 349:111, 1991.

[2] Baioch, A. "Prosite: a Dictionary of Protein sites and Patterns." (Department de Biochimie medicale. univrsite de Genova. 1990)

[3] Smith, T. F. and Waterman, M. S., "Identification of common molecular subsequences." *J. Mol. Biol.*, 147:195–197, 1981.

[4] Taylor, W.R. *J. Mol. Evol.*, 28:161, 1988

[5] Bairoch, A. and Bucher, P., "PROSITE: recent developments." *Nucleic Acids Res.*, 22:3538–3589, 1994.