Classification of Possible Proteins from Genome Sequencing

Yukihiro Eguchi   Yuzo Ueda
ueda@kojia.mitsui-knowledge.co.jp

Research Institute, Mitsui Knowledge Industry Co.,Ltd.
Kojimachi 3-7-4, Chiyoda-ku, Tokyo 102

The development of software tools classifying possible protein sequences from genome projects is our recent attempt. Especially, methods based on their amino-acid or oligopeptide compositions are focused on and considered to be complemental methods for those of sequnece homology and motif matching.

METHOD

We already reported a classification method using similarity indices calculated from amino-acid or oligopeptide compositions. In order to develop more accurate methods, multivariate statistical analyses were applied to this classification problem. In this attempt, first, variables useful for classifying among protein classes were selected, and second, such variables were used in discriminant analysis to classify sequences.

Superfamilies of the PIR database were used as protein classes. Superfamilies consisting of more than 10 sequence entries were picked up (208 superfamilies) and used as training data (about 5500 sequences). Variables were compositions of amino-acids, dipeptides, and a pair of amino acids intervened by any amino acids.

The SAS system (SAS Institute Inc.) was used to apply multivariate statistical analyses.

The STEPDISC procedure was used to select variables making good contribution to the classification power. According to a given variable-chosing criteria (A lower bound of the squared partial correlation was used in this case), and by a given selection method (Stepwise selection was used in this case), this procedure selects useful variables, using Wilks' lambda as indises of discriminatory power. It should be noted that in the computation we could not run the STEPDISC with all variables at once because the number of variables were beyond the CPU capacity of the workstation used. Then we used a following variable reduction procedure to overcome this problem. All variables were partitioned into smaller groups. And variables

江口 至洋，上田 裕三：三井情報開発株式会社 総合研究所，〒102 東京都千代田区麴町 3-7-4

selected by the STEPDISC from a couple of smaller groups were combined into a group to execute another STEPDISC. This procedure was continued until the STEPDISC could run with all variables selected by previous computations.

The DISCRIM procedure was used to classify sequences into superfamilies, using variables selected by the STEPDISC. A variety of discriminant-analysis methods are optional. In this case, we specified group-specfic densities were calculated by Mahalanobis distances (specified to be calculated by the pooled covariance matrix) between a sequence and a superfamily mean. The accuracy of this method was evaluated by crossvalidation-classification error rates.

RESULT

The maximum number of intervening amino acids between a pair of amino acids was set to 3, then total number of variables was 1620.

When a lower bound of the squared partial correlation was set to 0.3 as the variable-chosing criteria, the STEPDISC procedure selected 112 variables. The following table shows top 10 variables (compositions of oligopeptides) in order of the squared partial correlation or Wilks' lambda (where X signifies any intervening amino acid).

	Partial	Wilks'
Variable	R**2	Lambda
RR	0.9678	0.03215344
C	0.9222	0.00250198
CXC	0.8372	0.00040723
GXXG	0.7523	0.00010085
CXXC	0.7005	0.00003021
K	0.6984	0.00000911
R	0.6750	0.00000296
H	0.6534	0.00000103
A	0.6277	0.00000038
QQ	0.6185	0.00000015

Using the selected variables, crossvalidation-classification error rate was about 9%, as a result of the discriminant-analysis.

We will also report the effect of reductions of the number of variables, the effect of of increments of the number of intervening amino acids, and the effect of changes of the discriminant-analysis method (espacially use of other density functions), and discuss about the application of this method to superfamilies consisting of the small number of sequence entries.