Statistics in Molecular Biology
IMS Lecture Notes – Monograph Series (1999) Volume 33

PROTEIN FOLD CLASS PREDICTION IS A NEW FIELD FOR STATISTICAL CLASSIFICARION AND REGRESSION

BY LUTZ EDLER AND JANET GRASSMANN

Biostatistics Unit, Research Program on Genome Research and Bioinformatics, German Cancer Research Center, Heidelberg, Germany and BRAIN², Germany

Protein structure classification and prediction is introduced and elaborated for the application of standard and new statistical classification, discrimination and regression methods. With the sequence to structure to function paradigm in the background, methods of secondary and tertiary structure prediction will be reviewed and super-secondary classes and of fold classes will defined. We apply two branches of statistical classification - methods based on posterior probabilities and methods based on class conditional probabilities - and we will explore the role of artificial neural networks for the protein structure prediction. The procedures will be applied to a set of 268 previously described protein sequences for their statistical performance in the prediction of the four super-secondary classes and also in the prediction of 42 fold structure classes.

1. Introduction. Knowledge of the three-dimensional (3D) structure of a protein is essential for describing and understanding its function and for its use in molecular modeling [Fasman (1989)]. The impact of the structural knowledge for medical interventions and the understanding of diseases and their evolution has been clearly demonstrated [Branden and Tooze (1991), Gierasch and King (1990)]. Knowledge of the 3D structure of hemoglobine [e.g. Perutz (1978) Dickerson and Geis (1983)] enabled researchers to increase its oxygen capacity. This was the first and crucial step of a development which resulted in a synthetic hemoglobin substitute with consequences for blood transfusion [Mickler and Longnecker (1992)]. On the other hand, sickle cell anemia is caused by a single mutation in the amino acid sequence of hemoglobin, a change from *Glu* to *Val* on the surface of the globin fold [see Branden and Tooze (1991) p. 39] which causes movements of the α-helices relative to each other and makes the cell membrane more permeable to potassium ions. The disease is lethal for homozygotes, but increases the resistance to malaria in heterozygotes by killing the parasite through the drop of the potassium ion concentration. Therefore, the determination of structure is useful in different aspects: altering an existing protein's function (protein engineering), creating a protein *de novo* (protein de-

AMS 1991 *subject classifications.* Primary 92A10; secondary 62H30.
Key words and phrases. Protein structure, fold class, classification and prediction, discriminants, regression, neural networks, cross-validation.