THE UNIVERSITY  OF HONG KONG

Department of Mathematics

Numerical Mathematics and Applied Analysis Group Seminar (NMAA)

Biochemically-Weighted Kernel in Glycan Motif Extractions

Miss Jiang Hao
Department of Mathematics, HKU

on Wednesday, September 22, 2010 at 2:00pm
in Room 309, Run Run Shaw Building, HKU

Abstract

Carbohydrates, one of the most abundant and structurally diverse biopolymers, constitute the third major class of biomolecules. However, the study of carbohydrate sugar chains has lagged behind compared to that of DNA and proteins, mainly due to their inherent structural complexity. In order to glean some light into glycan function based on carbohydrate structure, kernel methods have been developed. The recently developed weighted q-gram methodexhibits good performance on glycan structure classification while having some limitations in feature selection. In order to avoid the issue of non-positive-semi-definiteness in the kernel matrix, the "Weighted q-gram method" deals with the similarity matrix. Biologically speaking the kernel matrix used in training a Support Vector Machine (SVM) in principle should involve the similarity matrix. Moreover, the method neglects biochemical information when considering the similarity between two q-grams. Therefore, we propose a biochemically-weighted tree kernel which is based on the original similarity matrix and incorporates the biochemical information of individual q-grams in constructing the kernel matrix. We further applied our new method for the classification and recognition of motifs on publicly available glycan data.

Our novel tree kernel (BioLK-method) using a Support Vector Machine (SVM) improves the classification performance over the previous Linkage Kernel (LK-method), a representative of the weighted q-gram method. Furthermore, our newly developed method is capable of detecting biologically important motifs accurately. It was tested on three glycan data sets from the Consortium for Functional Glycanomics (CFG) and Kyoto Encyclopedia of Genes and Genomes (KEGG) GLYCAN and showed that the results are consistent with the literature. Our results obtained here indicate that direct incorporation of the original similarity matrix contributes to the improvement of classification performance which is consistent with intuition as well as the existing literature. The incorporation of biochemical information of q-grams further shows the flexibility and capability of the novel kernel in feature extraction, which may aid in the prediction of glycan biomarkers.

All are welcome