Due to rapid progress of the genome projects, whole genome sequences of many organisms and a draft of human genome sequence have been already determined. But, the determination of the whole genome sequences does not mean the end of analysis of genetic code. In order to understand the meaning behind the genetic code, we have been developing algorithms for analyzing proteomics data and genomics data. Recently, we focus on the following topics: data mining from chemical reaction data, protein structure prediction, motif extraction, inference of metabolic pathways and genetic networks, and analysis of two-dimensional electrophoresis gel images.

Grants

Remote homology detection for proteins based on support vector machines

Recently, several methods were developed for remote homology detection for protein sequences using SVMs (support vector machines). We propose a new SVM based method (SVM-SW), which uses the SW algorithm as a kernel function (Fig. 1). Though we do not yet succeed to prove that the SW score is always a valid kernel, SVM-SW worked successfully in all cases we tested and was better than several existing methods.

1. H. Saigo et al., Comparison of SVM-based methods for remote homology detection, Genome Informatics, 13, 396-397 (2002).

Point matching under non-uniform distortions and protein side chain packing based on an efficient maximum clique algorithm

We developed maximum clique-based algorithms for spot matching for two-dimensional gel electrophoresis images, protein structure alignment and protein side chain packing. Algorithms based on direct reductions to the maximum clique can find optimal solutions for instances of size (the number of points or residues) up to 50-150 using a standard PC. We also developed pre-processing techniques to reduce the sizes of graphs. Combined with some heuristics, many realistic instances can be solved approximately.


On the complexity of deriving position specific score matrices from examples

PSSMs (Position-Specific Score Matrices) have been applied to various problems in Bioinformatics. We study the following problem: given positive examples (sequences) and negative examples (sequences), find a PSSM which correctly discriminates between positive and negative examples (Fig. 2). We proved that this problem is solved in polynomial time if the size of a PSSM is bounded by a constant. On the other hand, we proved that this problem is NP-hard if the size is not bounded. We obtained similar results on deriving mixture of PSSMs.