Bioinformatics Center
- Biological Information Network -

http://www.bic.kyoto-u.ac.jp/takutsu/index.html

Prof
AKUTSU, Tatsuya
(D Eng)

Instr
UEDA, Nobuhisa
(D Eng)

Students

HAYASHIDA, Morihiro (D1)
K. C., Dukka Bahadur (M2)
FUKAGAWA, Daiji (M2)
SAIGO, Hiroto (M2)
SHIBATA, Yuzo (M2)
YAMAMURA, Masaki (M2)
MOESA, Harry Amri (M1)

Visitor

Prof JIANG, Tao University of California, Riverside, 1 July 2002

Scope of Research

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.

Research Activities (Year 2002)

Presentations

On the complexity of deriving position specific score matrices from examples, Akutsu T, Bannai H (U Tokyo), Miyano S, Ott S (U Tokyo), Annual Symp. Combinatorial Pattern Matching, 3 July.

Inferring a union of halfspaces from examples, Akutsu T, Ott S (U Tokyo), Int. Conf. Computing and Combinatorics, 15 August.

A gibbs sampling algorithm for numerical sequences: detection of subtle motifs from protein sequences and structures, Akutsu T, Annual Meeting of Korean Society for Bioinformatics, 15 November.

Point matching under non-uniform distortions and protein side chain packing based on an efficient maximum clique algorithm, K.C. Dukka Bahadur, Akutsu T, et al., Int. Conf. Genome Informatics, 18 December.

Grants

Akutsu T, Miyano S, Ueda N, Algorithms for finding common patterns in bioinformatics, Grant-in-Aid for Scientific Research (C) (2), 1 April 2001 - 31 March 2005.

Akutsu T, Genome Information Science (a member of the project), Grant-in-Aid for Scientific Research Priority Areas (C) , 1 April 2000 - 31 March 2005.

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).

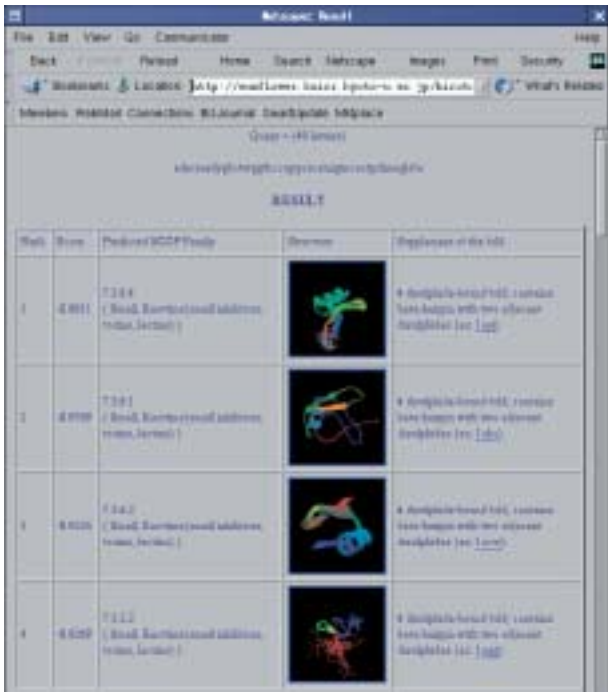

Figure 1. A system for remote homology detection of proteins.

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.

1.Dukka Bahadur K. C. et al., Point matching under non-uniform distortions and protein side chain packing based on an efficient maximum clique algorithm, *Genome Informatics*, **13**, 143-152 (2002).

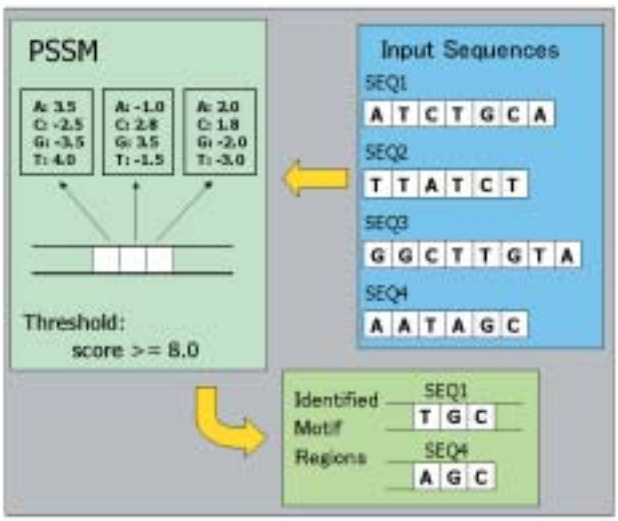

Figure 2. Motif detection using a position specific score matrix.

On the complexity of deriving position specific score matrices from examples

PSSMs (Position-Specfic 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.

1. T. Akutsu et al., On the complexity of deriving position specific score matrices from examples, *Lecture Notes in Computer Science*, **2373**, 168-177 (2002).