Contributed Chair
- Proteome Informatics (SGI Japan) -

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

Vis Assoc Prof
MAMITSUKA, Hiroshi
(D Sc)

Vis Assist Prof
YAMAGUCHI, Atsuko
(D Inf)

Vis Assist Prof
AOKI, Kiyoko F.
(Ph D)

PD
WAN, Raymond
(Ph D)

PD
ZHU, Shanfeng
(Ph D)

Scope of Research

The mission of proteome informatics is to develop information technologies to draw a picture of the complicated relationships among biological components, mainly proteins, from a vast amount of accumulated biological data. The objective of this laboratory is to undergo research to develop new technologies based on computer science that tackle a variety of issues in proteome informatics, and consequently, to acquire new biological knowledge that contributes to molecular biology as well as pharmacology and the medical sciences. Our particular emphasis has been placed on the following three topics: 1) new probabilistic models and methods for estimating parameters for learning/mining, 2) new efficient algorithms for searching similar chemical compounds, and 3) new methods and models for matching and aligning glycans based on statistical techniques.

Research Activities (Year 2004)

Presentations

A General Probabilistic Framework for Mining Labeled Ordered Trees, Ueda, N, Aoki, K F, Mamitsuka, H, Fourth SIAM International Conference on Data Mining (SDM 2004), Orlando, FL, USA, 23 April.

Application of Machine Learning to Bioinformatics, Mamitsuka H, Summer School of Bioinformatics (Sponcered by Japanese Society of Bioinformatics), Wajima, Japan, 21 July.

Application of a New Probabilistic Model for Recognizing Complex Patterns in Glycans, Aoki K F et al., 12th International Conference on Intelligent Systems for Molecular Biology (ISMB/ECCB 2004), Glasgow, UK, 3 August.

A Hierarchical Mixture of Markov Models for Finding Biologically Active Metabolic Paths using Gene Expression and Protein Classes, Mamitsuka H, 2004 IEEE Computational Systems Bioinformatics Conference (CSB 2004), Stanford, CA, USA, 19 August.

Glycan Tree Alignment and Substitution Matrix for Finding Relationships between Glycan Linkages, Aoki K F, US/Japan Glyco 2004 (Joint Meeting of the Society for Glycobiology and the Japanese Society of Carbohydrate Research), Honolulu, HI, USA, 20 November.

Analyzing Metabolic Pathways using Expression Profiles, Mamitsuka H, Workshop, 27th Annual National Meeting of the Molecular Biology Society of Japan, Kobe, Japan, 9 December.

Grant

Mamitsuka, H., Developing Algorithms for Searching and Finding Small Chemical Compounds Binding to Large Biological Molecules, Grant-in-Aid for Scientific Research on Priority Areas (C), 1 April 2004 - 31 March 2005.

Topics

Mining and Predicting Protein-Protein Interactions

Protein-protein interactions play a number of central roles in many cellular functions, including DNA replication, transcription and translation, signal transduction and metabolic pathways. A recent increase in the number of protein-protein interactions has made predicting unknown protein-protein interactions important for the understanding of living cells. However, the protein-protein interactions experimentally obtained so far are often incomplete and contradictory, and consequently existing computational prediction methods have integrated evidence (latent knowledge of proteins) from different and more reliable sources. Analyzing the relationships between proteins and the latent knowledge is important to understanding the cellular processes. For this analysis, we have proposed a new probabilistic model for protein-protein interactions by considering the latent knowledge of proteins. We have further presented an efficient learning algorithm for this model, based on an EM (Expectation-Maximization) algorithm. Experimental results have shown that in a supervised test setting, the proposed method outperformed five other competing methods by a statistically significant factor in all cases. Using the probability parameters of a trained model, we have further shown the latent knowledge that is essential to predicting protein-protein interactions. Overall, our experimental results confirmed that the proposed model is especially effective for analyzing protein-protein interactions from a viewpoint of the latent knowledge of proteins.

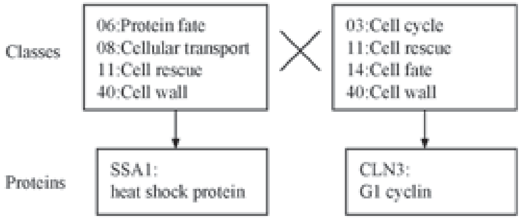
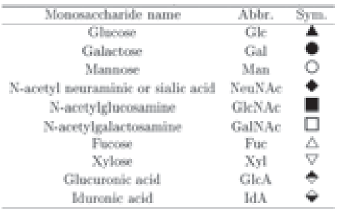


Figure 1. Example of two interacting proteins and the latent knowledge (protein classes).

Managing and Analyzing Carbohydrate Data

One of the most vital molecules in multicellular organisms is the carbohydrate, as it is structurally important in the construction of such organisms. In fact, all cells in nature carry carbohydrate sugar chains, or glycans, that help modulate various cell-cell events for the development of the organism. Unfortunately, informatics research on glycans has been slow in comparison to DNA and proteins, largely due to difficulties in the biological analysis of glycan structures. Our work consists of data engineering approaches in order to glean some understanding of the current glycan data that is publicly available. In particular, by modeling glycans as lableled unordered trees, we have implemented a tree-matching algorithm for measuring tree similarity. Our algorithm utilizes proven efficient methodologies in computer science that has been extended and developed for glycan data. Moreover, since glycans are recognized by various agents in multicellular organisms, in order to capture the patterns that might be recognized, we needed to somehow capture the dependencies that seem to range beyond the directly connected nodes in a tree. Therefore, by defining glycans as labeled ordered trees, we were able to develop a new probabilistic tree model such that sibling patterns across a tree could be mined. In fact, we have developed a new algorithm for estimating the parameters of this model from given training data. The model and algorithm are the cutting edges even in computer science. We provide promising results from our methodologies that could prove useful for the future of glycome informatics.



Monosaccharide name	Abbr.	Sym.
Glucose	Glc	▲
Galactose	Gal	●
Mannose	Man	○
N-acetyl neuraminic or sialic acid	NeuNAc	◆
N-acetylglucosamine	GlcNAc	■
N-acetylgalactosamine	GalNAc	□
Fucose	Fuc	△
Xylose	Xyl	▽
Glucuronic acid	GlcA	⬘
Iduronic acid	IdA	⬙

Figure 2. Common monosaccharide names, their abbreviations and symbols.

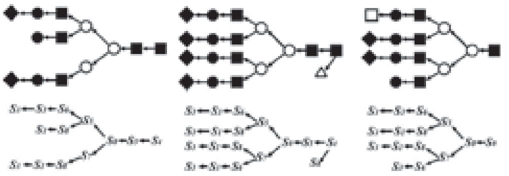


Figure 3. Example of multiple tree alignment obtained by the proposed probabilistic model.