# Bioinformatics Center
## - Biological Information Network -

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

Prof
AKUTSU, Tatsuya
(D Eng)

Assist Prof
UEDA, Nobuhisa
(D Eng)

PD
NACHER, Jose C.
(Ph D)

PD
MATSUDA, Setsuro
(D Eng)

PD
OCHIAI, Tomoshiro
(D Sc)

### Students

HAYASHIDA, Morihiro (D3)
FUKAGAWA, Daiji (D2)
K. C., Dukka Bahadur (D2)
SAIGO, Hiroto (D2)
TAMADA, Yoshinori (D2)
MEIRELES, Lidio (M1)
BROWN, John (RS)

### Visitors

| | |
|---|---|
| Mr FENG, Eric | The University of Hong Kong, 2 September - 29 October 2004 |
| Prof NG, Michael Kwak-Po | The University of Hong Kong, 21 October 2004 |
| Dr CHING, Wai Ki | The University of Hong Kong, 21 October 2004 |
| Dr VERT, Jean-Philippe | Ecole des Mines de Paris, 4 - 17 November 2004 |
| Mr MAHÉ, Pierre | Ecole des Mines de Paris, 27 November - 21 December 2004 |

## Scope of Research

Due to rapid progress of the genome projects, whole genome sequences of organisms ranging from bacteria to human have become available. In order to understand the meaning behind the genetic code, we have been developing algorithms and software tools for analyzing biological data based on advanced information technologies such as theory of algorithms, artificial intelligence, and machine learning. We are recently studying the following topics: systems biology, scale-free networks, protein structure prediction, inference of biological networks, chemo-informatics, discrete and stochastic methods for bioinformatics.

## Research Activities (Year 2004)

### Presentations

Fast algorithms for comparison of similar unordered trees, Fukagawa D, Akutsu T, The 15th Annual Int'l Symp. on Algorithms and Computation, 21 December.

Algorithms for point set matching with *k*-differences, Akutsu T, The 10th Int'l Computing and Combinatorics Conference, 18 August.

Extensions of marginalized graph kernels, Mahé P, Ueda N, Akutsu T, Perret J-L and Vert J-P, The 21st Int'l Conf. on Machine Learning, 4 July.

A simple method for inferring strengths of protein-protein interactions, Hayashida M, Ueda N, Akutsu T, The 4th Int'l Workshop on Bioinformatics and Systems Biology, 2 June.

Clustering of database sequences for fast homology search using upper bounds on alignment score, Itoh M, Akutsu T, Kanehisa M, The 4th Int. Workshop on Bioinformatics and Systems Biology, 1 June.

Protein threading with profiles and constraints, Akutsu T, Hayashida M, Tomita E, Suzuki J, Horimoto K, IEEE 4th Symp. on Bioinformatics and Bioengineering, 21 May.

Protein side-chain packing problem: a maximum edge-weight clique algorithmic approach, K.C. D., Akutsu T, Tomita E, Seki T, The 2nd Asia-Pacific Bioinformatics Conference, 20 January.

### Grants

Akutsu T, Miyano S, Maruyama O, Ueda N, Algorithms for extracting common patterns from structured biological data, Grant-in-Aid for Scientific Research (B), 1 April 2004 - 31 March 2008.

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

# Topics

## Analysis on Two Complementary Scale-free Networks

In a wide variety of real-world networks such as the World Wide Web and biological networks, the probability that a node with degree $k$ (the number of edges connected to the node) appears in a graph is proportional to $k^{-r}$, where $r$ is a constant. Such a network is called a scale-free network with exponent $-r$, and metabolic pathways are an interesting instance of the scale-free networks which can be represented by each of two complementary networks. In one of the networks, each node corresponds to a chemical compound, and an edge between nodes represents a reaction from the chemical compound of one node to that of another node. In the other network, a reaction and an enzyme which catalyzes the reaction are placed at a node, and two nodes are connected when the same chemical compound appears in the reactions of the nodes. The latter network can be constructed with the line graph transformation (that is, an edge in the former network is transformed into a node) from the former network, but properties of networks generated with the line graph transformation were not well investigated.

We then showed that, given a scale-free network $G$ with exponent $-r$, its transformed network $L(G)$ is also a scale-free network with exponent $-r+1$. We also experimentally confirmed that $L(H)$ formed a scale-free network, and its exponent was increased by the transformation, where $H$ is a metabolic network stored in the KEGG database.

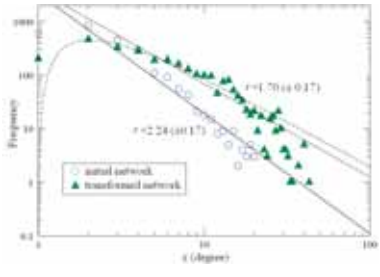

**Figure 1.** The distribution of the degree of nodes in the initial metabolic network from the KEGG database and its transformed networks.

Nacher J C, Yamada T, Goto S, Kanehisa M, Akutsu T, Two complementary representation of a scale-free network, *Physica A,* in press (2004).

## Extensions of Graph Kernels for Classifying Chemical Compounds

Recent theoretical advances and experimental results have drawn considerable attention to the use of kernel methods in computational analysis and classification of data sets. The marginalized graph kernels have recently been proposed to measure a similarity between labeled graphs such as chemical compounds.

These graph kernels, however, are subject to several limitations. First, the marginalized graph kernels decompose a graph into an infinite set of possible paths based on a random walk model for computational efficiency. This may lose structural information of graphs such as subtrees and subgraphs, which can be more relevant features in classification than paths. Moreover, the random walk model sometimes generates paths in which it comes to a node and instantly goes back to its previous node. Such a path only holds information on a local structure of the decomposed graph. Second, the marginalized graph kernel requires much computational cost, which results in slow implementation for real-world problems.

We then proposed two extensions of the marginalized graph kernels, which try to address these issues. The first extension is to relabel each node automatically in order to insert information about the environment of each vertex in its label by an iterative process called the Morgan index. This is effective in terms of feature relevance, because label paths contain information about their environment as well, and computation time, because less number of labeled paths match between graphs. Second, we showed how to modify the random walk model in order to remove irrelevant paths. Each method was validated on benchmark data sets, which are called the MUTAG and the PTC data sets, of classification of chemical compounds.

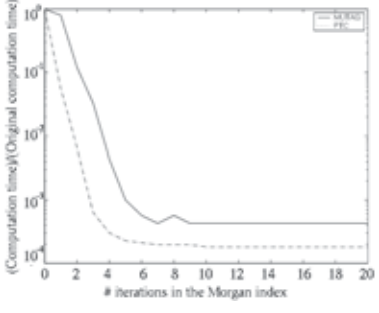


**Figure 2.** Computation time using different number of iterations of the Morgan index for the MUTAG and PTC data sets.

Mahé P, Ueda N, Akutsu T, Perret J-L, Vert J-P: Extensions of marginalized graph kernels, *Proc. 21st Int'l Conf. on Machine Learning,* 552-559 (2004).

Ueda N, Statistical language models that generate a pair of sequences for sequence analysis, Grant-in-Aid for Encouragement of Young Scientists, 1 April 2003 - 31 March 2006.