# 8th KOREA-SINGAPORE WORKSHOP on Bioinformatics and NLP

Meeting Room 1, 3rd floor, COM1, National University of Singapore
12.00nn-5.30pm, 12 December 2008

*Supported by NUS OLS Bioinformatics Programme,*
*NUS SOC Computational Biology Lab, &*
*MOE T1 Grant R-252-000-274-112*

Co-chairs:
Jong Cheol Park, KAIST
See-Kiong Ng, I2R
Limsoon Wong, NUS

Contact: Limsoon Wong, wongls@comp.nus.edu.sg, 9634-8506

## Programme

*Opening*

12.00nn-12.05pm. Welcome by Limsoon Wong

12.05pm-12.10pm. Welcome by Jong C. Park

*Session I (Gene Expression Analysis)*

12.10pm-12.30pm. **Donny Soh***:* Agreement between Databases of Biological Pathways

12.30pm-12.50pm. **Difeng Dong**: Deciphering Drug Action and Escape Pathways: An Example on Nasopharyngeal Carcinoma

**Lunch Break (60 minutes)**

*Session II (Biomedical Text Analysis)*

01.50pm-02.50pm (Keynote). **Jong C. Park**: On E3Miner, A Text Mining Tool For Ubiquitin-Protein Ligases

02.50pm-03.10pm. **Gong Tianxia**: Information Extraction from Radiology Reports

03.10pm-03.30pm. **Hee-Jin Lee**: Towards Knowledge Discovery through Automatic Inference with Text Mining in Biology and Medicine

**Tea Break (20 minutes)**

*Session III (Medical Visualization Analysis)*

03.50pm-04.50pm (Keynote). **Jinah Park**: Visualization, Simulation, and Interaction.

04.50pm-05.10pm. **Hon Nian Chua**: Automated Mammographic Density Assessment

Closing

05.10pm-05.15pm. Remarks by Jong C. Park

05.15pm-05.20. Remarks by See-Kiong Ng

**Banquet for Visitors**

# Abstracts

## Keynotes

### On E3Miner, A Text Mining Tool For Ubiquitin-Protein Ligases

Jong C. Park
park@cs.kaist.ac.kr
KAIST

In this talk, I will be mainly presenting the following work that was published at Nucleic Acids Research (NAR) this year, conducted primarily by a former student of mine, Hodong Lee, and in cooperation with Prof. Gwansu Yi. I will also be presenting our related work in the field of bioinformatics by the NLP & CL Research Lab at KAIST. Ubiquitination is a regulatory process critically involved in the degradation of more than 80% of cellular proteins, where such proteins are specifically recognized by a key enzyme, or a ubiquitin-protein ligase (E3). Because of this important role of E3s, a rapidly growing body of the published literature in biology and biomedical fields reports novel findings about various E3s and their molecular mechanisms. However, such findings are neither adequately retrieved by general text mining tools nor systematically made available by such protein databases as UniProt alone. E3Miner is a web-based text mining tool that extracts and organizes comprehensive knowledge about E3s from the abstracts of journal articles and the relevant databases, supporting users to have a good grasp of E3s and their related information easily from the available text. The tool analyzes text sentences to identify protein names for E3s, to narrow down target substrates and other ubiquitin-transferring proteins in E3-specific ubiquitination pathways, and to extract molecular features of E3s during ubiquitination. E3Miner also retrieves E3 data about protein functions, other E3-interacting partners, and E3-related human diseases from the protein databases, in order to help facilitate further investigation. E3Miner is freely available through http://e3miner.biopathway.org.

CV: Prof. Jong C. Park received his BE and MSE degrees in Computer Engineering from Seoul National University and earned his PhD in Computer & Information Science from the University of Pennsylvania in 1996. His research interests include biomedical text mining and management, language-enhanced robots, and linguistically motivated computational applications in general, all based on Natural Language Processing (NLP) techniques. He is currently working as founding co-Editor-in-Chief of the Journal of Computing Science and Engineering (JCSE, http://jcse.kiise.org), published by the largest society for computing science in Korea, or Korea Institute for Information Scientists and Engineers (KIISE). He also founded International Symposium on Languages in Biology and Medicine (LBM) in 2005 together with Prof. Limsoon Wong, and worked to coordinate LBM as a sister symposium for Semantic Mining in Biomedicine (SMBM), a primarily European symposium of a similar nature in the range of topics. He is Associate Professor and Associate Department Head of Computer Science Department at Korea Advanced Institute for Science and Technology (KAIST).

---

### Visualization, Simulation, and Interaction

Jinah Park
jinah@icu.ac.kr
ICU

Computer graphics deals with the computational generation of images and image sequences from given data stored in a virtual world, and visualization addresses the issues of casting data to suitable representations. Furthermore, computer haptics allow the users to feel by touch of the virtual objects. In this talk, I will overview the basic concepts of disciplines that transfer the data in virtual world to something that human can perceive, and introduce the related on-going research work at the Computer Graphics and Visualization Laboratory in Information & Communications University, including (1) Automatic 3D heart mesh modeling from CT data, (2) Surface-volume dual deformable model, (3) Dental implant simulation – bone drilling, (4) Haptic grasping, (5) Spray paint simulation, and (6) Virtual anatomy learning book.

CV: Prof. Jinah Park received a B.S. (1988) in Electrical Engineering at the Columbia University in New York, and an M.S.E. (1991) and a Ph.D. (1996) in Computer and Information Science at the University of Pennsylvania in Philadelphia. Her main interests are in Computer Graphics and Computer Vision as applied to Medical Image Analysis and Visualization. Her major contributions to the field include developing a computational technique to analyze cardiac motion based on MR tagging data. Upon coming to Korea in 1999, she worked at EE&CS Department of Korea Advanced Institute of Technology and Science (KAIST) as a research professor. She joined Information and Communications University (ICU) in 2002 as a regular faculty member. She is currently an associate professor of School of Engineering in ICU, an elected member of the faculty council at ICU, and Director of Institute for IT-Gifted Youth.

# Session I

### Deciphering Drug Action and Escape Pathways: An Example on Nasopharyngeal Carcinoma

Dong Difeng
dong.difeng@gmail.com
NUS

Biological pathways have been incorporated into gene expression analysis to understand drug treatment response in disease population. However, existing methods fall short on several issues: 1) these works provide little information on the interplay between selected genes; 2) the collection of pathways that can be used, evaluated and ranked against the observed expression data is limited; 3) the generated hypotheses are still too general to guide further research and treatment.

Recently, we have proposed a drug pathway identification system, which we called Drug Pathway Decipherer (Decipherer) to generate hypotheses of specific treatment responsive pathway. Decipherer takes in both pre- and post-treatment gene expression data, and evaluates known biological pathways on the data. In this talk, I will demonstrate a study case of applying Decipherer to two gene expression datasets of human nasopharyngeal carcinoma (NPC) treated with CYC202. As a result, we find the regulation of cell proliferation RAS-ERK pathway and anti-apoptosis PI3K-NFkB-IAP pathway are closely associated with treatment outcome.

### Agreement between Databases of Biological Pathways

Donny Soh
donnysoh@gmail.com
Imperial College London

There is an increasing need to analyze microarray experiments together with biological information to make better biological inferences. One key component often ignored is the integrity of the biological information used. This is important because using inaccurate or incomplete information may lead to different analytical results. We extracted and analyzed all pathway information from three major pathway repositories: KEGG, Ingenuity and Wikipathways. We found that the percentage overlap between the pathways fall within the range of 51% for genes and 18% for interacting gene pairs.

These three original sources used can be assumed to be reliable in the sense that the interacting gene pairs reported in them are correct because they are curated. However, the lack of concordance between these databases suggests each source has missed out many interacting gene pairs. Therefore it is sensible and critical to aggregate these sources into a more comprehensive pathway database. We have accumulated sufficient data to create such an aggregated resource with the convenience of an API to access its information.

# Session II

### Combining Text Mining and Automatic Inference for Biological Knowledge Discovery

Hee-Jin Lee
heejin@nlp.kaist.ac.kr
KAIST

Field experts in biology and medicine search the literature for state-of-the-art results and occasionally discover knowledge through manual inference on published causal relations. However, the results of such inference cannot be

sufficiently accurate and/or complete, as the domain of published relations is rather huge. In this talk, we introduce an automatic inference system, BioDetective, which works on literature mined qualitative causal information in biology and medicine. BioDetective provides proofs for such qualitative causal information, and predicts the existence of new causal information, if there is any. We also examine the characteristics of natural language descriptions to use them as input data of such an inference system.

---

### Information Extraction from Radiology Reports

Gong Tianxia
gongtian@comp.nus.edu.sg
NUS

Information extraction in medical domain has gained increasing interest in recent years. Radiology reports contain rich information describing radiologist's observations on the patient's medical conditions in the associated medical images. However, as most reports are in free text format, the valuable information contained in those reports cannot be easily accessed and used. We propose a system to apply NLP techniques to extract and use the medical finding information in radiology reports. The information extraction process consists of three phases: concept mapping, syntactic relation extraction, and medical finding extraction. The medical findings extracted are used to index the radiology reports for knowledge-based retrieval.

---

# Session III

### Automated Mammographic Density Assessment

Hon Nian Chua
hnchua@i2r.a-star.edu.sg
$I^2R$

The proportion of dense regions in a breast relative to total breast area, referred to as mammographic density (MD), has been shown to be well correlated to breast cancer risk. Some studies have shown that subjects in the highest quartile MD range is 3 to 5 times more predisposed to developing breast cancer than subjects of similar age in the lowest quartile. Hence MD measurement can provide early warning during screening which can allow early intervention. The current procedure for the measurement of MD involves a clinical expert manually selecting the breast region and dense region on a digital mammogram and computing the ratio between the two. The widely accepted software used in this procedure is the Cumulus interactive threshold software, which can partially automate the procedure. However, the process can still be laborious in large scale mammogram screenings, which provides us with the motivation to develop an approach to reliably automate the process. Here we present our current progress on this work.

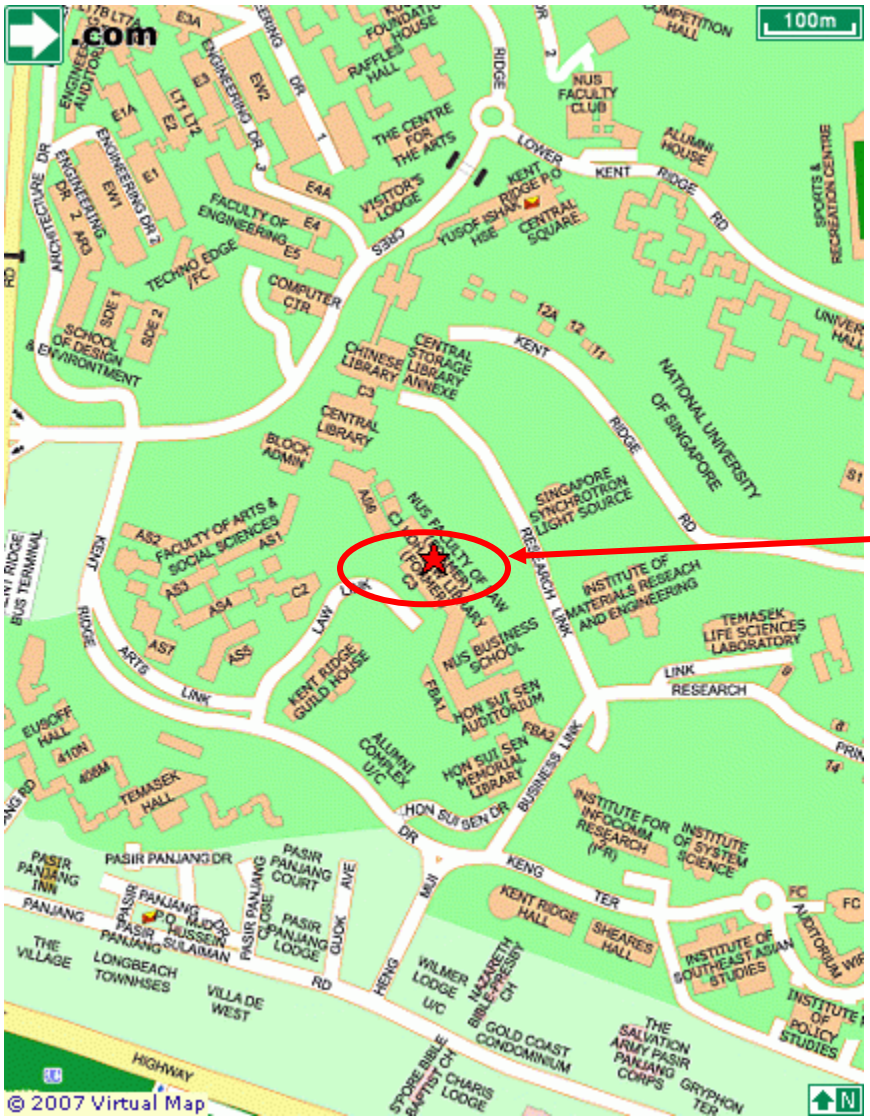
---

## DIRECTIONS TO SCHOOL OF COMPUTING

Venue: Meeting Room 1, Level 3, Blk COM1, NUS School of Computing.

If you are lost, please call Prof. Wong Limsoon at 9634-8506 or Ms Tan Poh Suan at 6516 2726.

### Directions to NUS School of Computing, Blk COM1




By Taxi:

- Destination – NUS (National University of Singapore);
- Please instruct Taxi Driver to use the **second entrance to NUS on Clementi Road**.

- Upon entering NUS (see map below), please take the first right turn into Kent Ridge Drive, then take the first left turn into Law Link.

By Bus:

- Take bus services 95 or 151 towards the direction of Kent Ridge campus
- Alight at the bus stop outside Central Library
- Take the internal shuttle bus A1 or B from the same bus stop
- Alight at the Business Canteen bus stop (which is the third bus stop after the one at Central Library).

Taxi enters here. Then turns right at first junction into Kent ridge Drive. Then turns left at first junction into Law Link.

# Participants

**(If name is in bold font, it means Limsoon has received your "official" registration)**

## NUS

1. **WONG Limsoon, wongls@comp.nus.edu.sg**
2. NG Hwee Tou, **nght@comp.nus.edu.sg**
3. **LIU Guimei, liugm@comp.nus.edu.sg**
4. **ZHANG Jie, zhangjie_gd@hotmail.com**
5. **Huy Hoang DO, huang@comp.nus.edu.sg**
6. **DONG Difeng, dongdife@comp.nus.edu.sg**
7. **Koh Chuan Hock, kohchuanhock@gmail.com**
8. **Donny Soh, donnysoh@gmail.com**
9. **LI Bowen, dcslbw@.nus.edu.sg**
10. **GONG Tianxia, gongtian@comp.nus.edu.sg**
11. **Chandana Tennakoon, drcyber@gmail.com**
12. Li Xiaohui
13. Ngo Thanh Son
14. Liu Bing
15. **FAN Mengyuan, g0801861@nus.edu.sg**
16. **Nim Tri Hieu, g0801981@nus.edu.sg**

## NTU

17. **CHEN Xin, chenxin@ntu.edu.sg**
18. **LIAN Heng, henglian@ntu.edu.sg**

## A*STAR

19. **NG See Kiong,** skng@i2r.a-star.edu.sg
20. **Su Jian, sujian@i2r.a-star.edu.sg**
21. **LI Xiaoli, xlli@i2r.a-star.edu.sg**
22. **Hon Nian Chua, hnchua@i2r.a-star.edu.sg**
23. **James Ma, tlmah@i2r.a-star.edu.sg**
24. **Zeyar Aung,** azeyar@i2r.a-star.edu.sg
25. Stanley Ng
26. Geoffrey Koh
27. **Vipin Narang, narangv@gis.a-star.edu.sg**

## KAIST

28. **Jong Cheol PARK,** park@cs.kaist.ac.kr
29. **Eunyoung Chang,** euny@nlp.kaist.ac.kr
30. **Hee-Jin LEE,** heejin@nlp.kaist.ac.kr

## ICU

31. **Jinah PARK,** jinah@icu.ac.kr