5th KOREA-SINGAPORE WORKSHOP on Bioinformatics and NLP

Exec Class Room, 5th floor, SOC1, National University of Singapore
9.00am-5.30pm, 22 February 2006

Supported by NUS OLS Bioinformatics Programme,
NUS SOC Computational Biology Lab,
& Institute for Infocomm Research

Co-chairs:
Jong Cheol Park, KAIST
See-Kiong Ng, I^{2}R
Limsoon Wong, NUS

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

Programme

Opening

09.00am. Welcome by Limsoon Wong

09.05am. Address by Jong C. Park

Session I (NUS)

09.15am. **Zeyar Aung**: Discovering Statistically and Biologically Significant Clusters of 3D Protein-Protein Interfaces

09.45am. **Xu Xin**: Mining Top-k Covering Rule Groups for Gene Expression Data

10.15am. **Vipin Narang***: In Silico* Detection of Localized Subtle Motifs in Regulatory Sequences

Tea Break (15 minutes)

Session II (KAIST/ICU)

11.00am. **Ho-Joon Lee**: Augmenting Visualization with Audioization for Enhanced Knowledge Discovery

11.30am. **Hodong Lee**: Explorative search with relational description of biological entities into multiple heterogeneous databases

12.00pm. **Hye-Jin Min**: Construction of Amygdala-related Pathway for Diagnoses of Emotional Disorders using BioIE

Lunch Break (45 minutes)

Session III (NTU)

01.15pm. **Zhang Tianyou**: MRSA Identification System – a novel system for fast identification of MRSA strain by data mining with MLST database

01.45pm. **Deyu Zhou**: Extracting Protein-Protein Interactions from the Literature using the Hidden Vector State Model

Slide Show (15 minutes)

Session IV (I^{2}R)

02.30pm. **Soon-Heng Tan**: Automated Motif Discovery from Protein Interaction Data

03.00pm. **Chen Jin**: Discovering meso-scale network motifs for protein interaction validation

Tea Break (15 minutes)

Session V (KAIST/ICU)

03.45pm. **Hee-Jin Lee**: Towards an efficient CCG parser for RNA secondary structure prediction

04.15pm. **Jin-Bok Lee**: Automatic Extension of Gene Ontology with Induced Prediction and Flexible Validation of Candidate Terms

04.45pm. **Jung-jae Kim**: Term Characterization for Information Extraction with Syntactic Pattern Matching

Closing

05.15pm. Remarks by Jong C. Park

05.20pm. Close by See-Kiong Ng

Banquet for Visitors

Abstracts

Presentations by NUS

Discovering Statistically and Biologically Significant Clusters of 3D Protein-Protein Interfaces

Zeyar Aung
zeyaraun@comp.nus.edu.sg
supervisor: Kian Lee TAN

The study of the structural properties of protein-protein interfaces, which are responsible for interactions of proteins, can give us a better overview of protein functions, as compared to studying individual protein structures separately. In this work, we present a new method to encode, cluster and analyze the similar 3D interface patterns among various protein complexes. We represent the protein-protein interfaces as 2D residue-residue distance matrices, and encode them as multi-dimensional feature vectors. Then, we cluster the interfaces using these feature vectors, and analyze the resultant clusters by various means. Experimental results show that we can discover a number of statistically significant clusters of interfaces. A visual inspection also confirms that the interfaces that fall into the same cluster are visually similar. We can find out the clusters of similar interface patterns in the protein complexes belonging to diverse structural fold types. We can also discover in some clusters the recurring interface patterns associated with biologically important functional motifs. Furthermore, we compare our method with the sequence-only clustering approach, and observe that ours is much better in terms of the statistical significance of the resultant clusters.

Mining Top-k Covering Rule Groups for Gene Expression Data

Xu Xin
xuxin@comp.nus.edu.sg
supervisor: Anthony Tung

We propose a novel algorithm to discover the topk covering rule groups for each row of gene expression problems. Several experiments on real bioinformatics datasets show that the new top-k covering rule mining algorithm is orders of magnitude faster than previous association rule mining algorithms. Furthermore, we propose a new classification method RCBT. RCBT classifier is constructed from the top-k covering rule groups. The rule groups generated for building RCBT are bounded in number. This is in contrast to existing rule-based classification methods like CBA which despite generating excessive number of redundant rules, is still unable to cover some training data with the discovered rules. Experiments show that the RCBT classifier can match or outperform other state-of-the-art classifiers on several benchmark gene expression datasets. Also, the top-k covering rule groups provide insights into the mechanisms responsible for diseases directly.

In Silico Detection of Localized Subtle Motifs in Regulatory Sequences

Vipin Narang
vipinnar@comp.nus.edu.sg
supervisor: Ken SUNG

In silico motif finding algorithms are often used for discovering protein-DNA binding sites in a set of regulatory sequences. Most algorithms can easily detect prominent motifs. Keich and Pevzner (2002), Chin et al. (2004), and others have addressed the problem of detecting weak or subtle motifs. The problem arises frequently in the analysis of metazoan regulatory sequences where binding sites are widely spaced and the DNA sequences to be analyzed are long. Competing random patterns eclipse the real motifs producing false positives. In several situations however regulatory motifs show positional localization in segments of the genome relative to some biological marker. This talk reports a new algorithm called LocalMotif for detecting such subtle localized motifs. A novel scoring function that measures motif strength in a local sequence region is employed in an optimized algorithm to search for localized motifs. Multiple best scoring motifs are simultaneously reported along with their corresponding sequence regions of maximum strength. LocalMotif's performance tested on simulated data shows its advantage in discovering localized subtle motifs as compared to other algorithms. In real regulatory sequences, Localmotif could discover biologically relevant motifs around the transcription start site or a related binding site which were not easily visible to other motif finding programs. In addition, LocalMotif reports the position of localization of a motif, which is useful in its identification and study.

Presentations by NTU

MRSA Identification System – a novel system for fast identification of MRSA strain by data mining with MLST database

Zhang Tianyou
zhangty@pmail.ntu.edu.sg
Supervisor: Kwoh Chee Keong

Emerging Infectious Diseases (EID) such as SARS, Bird Flu and Avian Flu, have alarmed the world of its epidemicity and become serious threats to the health of human society. Staphylococcus aureus (S.aureus) is one of life-threatening EID bacterium, which can cause large variety of diseases such as pneumonia, meningitis, endocarditis and septicemia. During its 50-year history, S.aureus has demonstrated strong capability of acquiring antibiotic resistance by frequent mutations. Methicillin-resistant S.aureus (MRSA) has recently become a major cause of hospital-acquired infections and been recognized with increasing frequency in community acquired infections. The S.aureus family has been growing rapidly due to the rate of mutation. Hence, it is necessary to classify those strains and identify methicillin resistance before any treatment to the S.aureus infections.

In this project, genetic mutation of S.aureus is intensively studied by means of two data mining algorithms: Associate Rule Mining (ARM) and Bayesian Classifier (BC). Multilocus Sequence Typing (MLST) data is first cleaned and then mined by ARM algorithm to generate output of strong patterns correlating with MRSA. DNA sequences of those patterns are investigated by Methicillin Resistant Related (MRR) pattern extraction model, and meanwhile, allele index sequences of patterns are processed to build BC predictor. The extracted MRR patterns and BC predictor are jointly employed to facilitate fast and mutation-tolerated identification of MRSA strains. In addition, we also modeled the strain linkage and mutation rate for better understanding in the MRSA prediction.

Extracting Protein-Protein Interactions from the Literature using the Hidden Vector State Model

Deyu Zhou
Zhou0063@ntu.edu.sg
Supervisor: Yulan He, Chee Keong Kwoh

In the field of bioinformatics in solving biological problems, the huge amount of knowledge is often locked in textual documents such as scientific publications. Hence there is an increasing focus on extracting information from this vast amount of scientific literature. Many approaches, such as pattern matching, shallow and full parsing, have been proposed to automatically extract medical and biological text, for instance protein-protein interactions. In this paper, we present an information extraction system which employs a semantic parser using the Hidden Vector State (HVS) model for protein-protein interactions. Unlike other hierarchical parsing models which require fully annotated Treebank data for training, the HVS model can be trained using only lightly annotated data whilst simultaneously retaining sufficient ability to capture the hierarchical structure needed to robustly extract task domain semantics. When applied in extracting protein-protein interactions information from medical literature, we found that it performed better than other established methods and achieved 61.7% and 71.8% in recall and precision respectively.

Presentations by I^{2}R

Automated Motif Discovery from Protein Interaction Data

Soon-Heng Tan
soonheng@i2r.a-star.edu.sg
manager: See-Kiong NG

Protein motifs are important for the study of protein's function and disease pathways. The biological experiments to find protein motifs are laborious and expensive. On the other hand, the computational approach to discover such motifs is hindered by the inadequate biological knowledge use to group sequences for pattern detection. The talk will present a new problem of using protein interaction data to circumvent current motif discovery's bottleneck caused by the lack of adequate biological knowledge. Existing works grouped sequences binding to a protein or protein group for pattern detection by motif discovery programs. A novel "motif pairs" motif finding approach will be presented which we show is more sensitive and more robust to find protein motifs from interaction data than existing methods.

Biodata:
Soon-Heng Tan is currently a research officer in Institute of Infocomm Research, Singapore. He obtained his B.Sc (Molecular Biology) from National University of Singapore in 2001 and has recently completed a part-time M.Sc (Computer Science) from the same university. Soon-Heng is active in the research of data mining and knowledge discovery from biological networks.

Discovering meso-scale network motifs for protein interaction validation

Chen Jin
chenjin@comp.nus.edu.sg
supervisors: See-Kiong Ng, Wynne Hsu, Mong Li Lee

Network motifs have been shown to be useful in biological applications such as improving the reliability of protein interaction data generated by highly erroneous high-throughput experimental methods. However, existing motif mining algorithms are not scalable enough to find the meso-scale network motifs required for biological applications. This talk describes an efficient network motif discovery algorithm, NeMoFinder, that can find meso-scale repeated and unique network motifs in protein interaction networks. Our comparative evaluations show that NeMoFinder could successfully discover, for the first time, up to size-12 network motifs from whole-genome yeast protein interaction network. It also enables us to employ actual network motifs derived from biological networks, instead of predefined small-sized network motifs, to improve the reliability of currently erroneous protein interaction network obtained from experimental methods.

Presentations by KAIST/ICU

Automatic Extension of Gene Ontology with Induced Prediction and Flexible Validation of Candidate Terms

Jin-Bok Lee
jblee@nlp.kaist.ac.kr
supervisor: Jong C. Park

Gene Ontology (GO) has been manually developed to provide a controlled vocabulary for gene product attributes, and continues to evolve with new concepts that are compiled mostly from existing concepts in a compositional way. In this talk, I present a novel method that automatically predicts more detailed concepts by utilizing syntactic relations among the existing concepts in a compositional manner such that GO evolves. I also present a validation measure for the automatically predicted concepts by matching the concepts to biomedical articles. Finally, I discuss how to find a suitable direction for the extension of a constantly-growing ontology such as GO.

Term Characterization for Information Extraction with Syntactic Pattern Matching

Jung-jae Kim
jjkim@nlp.kaist.ac.kr
supervisor: Jong C. Park

Term characterization is a task that associates terms from the literature with standard terminology. This task is necessary for information extraction that works by matching syntactic patterns to sentences, since the syntactic patterns specify syntactic types, but not semantic types, of their variables. We present methods that semantically characterize terms in the results of information extraction from the biomedical literature by associating the terms with standard biomedical resources such as Swiss-Prot and Gene Ontology. These methods deal with a number of linguistic phenomena, including term variation, anaphoric expression, and coordination. We also show experimental results of the methods on the information of biological interactions and contrastive information extracted from the biomedical literature.

Augmenting Visualization with Audioization for Enhanced Knowledge Discovery

Ho-Joon Lee
hojoon@nlp.kaist.ac.kr
supervisor: Jong C. Park

Visualization provides a method that displays information with graphic elements, allowing users to navigate enormous amount of information. However, it has the difficulty in displaying verbal information, for example summaries of selected objects. To address this problem, we suggest augmenting visualization with audioization, which additionally provides sound information that helps the users to understand visualized results more effectively. We present a method that augments visualization of biological interaction information with audio descriptions of interactions and interacting proteins. As a result we expect a more intuitive and elaborate knowledge discovery system with synergy of audio and visual information

Explorative search with relational description of biological entities into multiple heterogeneous databases

Hodong Lee
hdlee@nlp.kaist.ac.kr
supervisor: Jong C. Park

Biological data is complex and dispersed in multiple heterogeneous databases. In order to explore such databases effectively, most of biological information systems are developed to provide interfaces for an efficient and expressive search, such as keyword-based, form-based, and graph-based interfaces. However, they are neither efficient to search for structured, diverse, and complex data, nor expressive enough to describe target data that are associated with other data. In this talk, we present a novel method that searches for the biological databases with natural language queries. This method analyzes syntactic relations and biological information in the natural language queries and generates formal database queries by associating the syntactic relations with appropriate databases. We show a practical example for explorative knowledge discovery from over 13 biological databases.

Construction of Amygdala-related Pathway for Diagnoses of Emotional Disorders using BioIE

Hye-Jin Min
hjmin@nlp.kaist.ac.kr
Supervisor: Jong C. Park

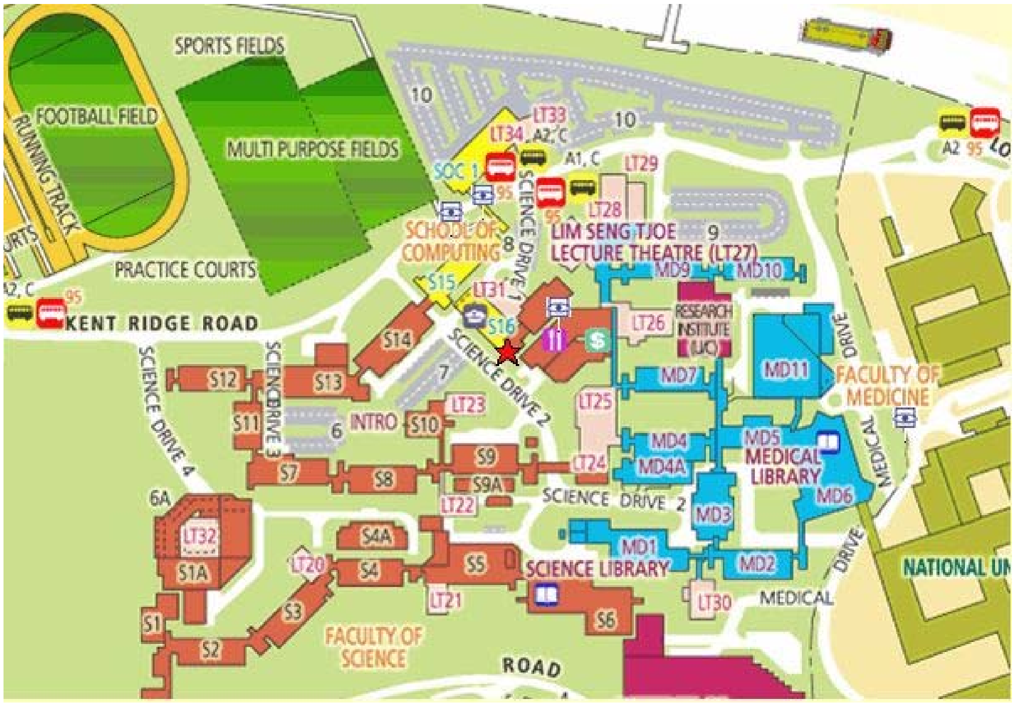
The amygdala in limbic system is critically involved in the control of emotions. It is thus important to be able to consult the neural and genetic pathway of amygdale when making diagnoses on various emotional disorders. In this talk, we present a method that automatically constructs the pathway of amygdala by utilizing 'BioIE', which is a system that extracts interaction information between biological entities from biomedical literature. This pathway can be utilized to develop new drug targets for the disorders as well as to make related diagnoses.

Towards an efficient CCG parser for RNA secondary structure prediction

Hee-Jin Lee
hjmin@nlp.kaist.ac.kr
Supervisor: Jong C. Park

In this talk, I will discuss ways of modeling and predicting RNA secondary structures using a grammar framework. In particular, I will discuss the possibility of using a combinatory categorial grammar (CCG), which is capable of characterizing a number of RNA secondary structures, including stem-loop and pseudo-knot structures. Nevertheless, the CCG covers massively ambiguous structures, rendering its naïve parser to consume much time to examine all the possible results. I will show the possible causes that slow down the parser, and propose ways to overcome them.

Directions to SOC1



By Bus: Take bus service 95 and alight in front of block SoC1.
By Taxi: Tell taxi driver you are going to “NUS”. Ask him to use the “NUH Entrance”. In this case, you will be coming in at the top-left corner of the map above, driving along Kent Ridge Rd. After you pass the small round about, you will see a car park to your left and right. Then you will see a blue/grey building to your right, having an overhead bridge across the road. That is SOC1, marked in yellow on the map.
In case you are lost, please call Limsoon at 9634-8506.

Participants

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

I^{2}R-SOC Joint Lab

1. **WONG Limsoon,** wongls@comp.nus.edu.sg
2. **Wynne HSU,** whsu@comp.nus.edu.sg
3. LEE Mong Li, leeml@comp.nus.edu.sg
4. **SUNG Wing-Kin**, ksung@comp.nus.edu.sg
5. **ZHENG Yun,** zhengy@comp.nus.edu.sg
6. **Zeyar AUNG**, zeyaraun@comp.nus.edu.sg
7. **CHUA Hon Nian**, g0306417@nus.edu.sg
8. **DONG Difeng**, dongdife@comp.nus.edu.sg
9. HOU Yuna, houyuna@comp.nus.edu.sg
10. **Vipin NARANG**, vipinnar@comp.nus.edu.sg
11. WONG Swee Seong, wongss@comp.nus.edu.sg
12. **XU Xin,** xuxin@comp.nus.edu.sg
13. **Hugo WILLY**, hugowill@comp.nus.edu.sg

NUS

14. David HSU, dyhsy@comp.nus.edu.sg
15. **LEONG Hon Wai,** leonghw@comp.nus.edu.sg
16. TAN Kian Lee, tankl@comp.nus.edu.sg
17. OOI Beng Chin, ooibc@comp.nus.edu.sg
18. **P. S. THIAGARAJAN,** thiagu@comp.nus.edu.sg
19. Anthony TUNG, atung@comp.nus.edu.sg
20. **NG Hwee Tou,** nght@comp.nus.edu.sg
21. **TAN Chew Lim,** tancl@comp.nus.edu.sg
22. **Koh Yeow Nam, Geoffrey,** geoffkoh@yahoo.com.sg
23. **Chiang Tsung Han,** th_chiang@yahoo.com
24. **Leong Wai Kay,** leongwai@comp.nus.edu.sg
25. **Liu Bing,** liubing@comp.nus.edu.sg
26. **Melvin Zhang ZhiYong,** zhangzh3@comp.nus.edu.sg

NTU

27. **Kwoh Chee Keong,** asckkwoh@ntu.edu.sg
28. **He Yulan,** ASYLHe@ntu.edu.sg
29. **Stephanus Daniel Handoko,** danielhandoko@pmail.ntu.edu.sg
30. **Zhou Deyu,** ZHOU0063@ntu.edu.sg
31. **Zhang Tianyou,** TYZhang@ntu.edu.sg

I^{2}R

32. **NG See Kiong,** skng@i2r.a-star.edu.sg
33. **LI Jinyan,** jinyan@i2r.a-star.edu.sg
34. **TAN Soon Heng,** soonheng@i2r.a-star.edu.sg
35. **CHEN Jin**, chenjin@comp.nus.edu.sg
36. **Mohit Kumar Sharma,** stumks@i2r.a-star.edu.sg

BTI

26. **LEE Dong Yup**, cheld@nus.edu.sg

KAIST

37. **Jong Cheol PARK,** park@cs.kaist.ac.kr
38. **Jin-Bok LEE,** jblee@nlp.kaist.ac.kr
39. **Jung-Jae KIM,** jjkim@nlp.kaist.ac.kr
40. **Hye-Jin MIN,** hjmin@nlp.kaist.ac.kr
41. **Ho-Joon LEE,** hojoon@nlp.kaist.ac.kr
42. **Hodong LEE,** hdlee@nlp.kaist.ac.kr
43. **Hee-Jin LEE,** heejin@nlp.kaist.ac.kr

ICU

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