Korea-Singapore Workshop on Bioinformatics and NLP 2013

School of Computing, National University of Singapore

18 February 2013

VENUE

- Morning sessions will be held at Meeting Room 1, Level 3, COM1 Building.
- Lunch and tea breaks will be held at Multi-Purpose Space, Level 4, COM2 Building.
- Afternoon sessions will be held at Executive Classroom, Level 4, COM2 Building.
- Workshop Banquet will be held at University Club, Level 4, Shaw Foundation House.

SPONSORS

SBI NCRC
시스템 바이오 정보의학 국가핵심연구센터
Systems Biomedical Informatics National Core Research Center

NUS School of Computing
Computational Biology Laboratory

PROGRAMME

9.15am-9.30am: Welcome

Jong Cheol Park (KAIST) & Limsoon Wong (NUS)

9.30am-11.30am: BioNLP

@ 9.30am-10.00am

Biomedical Ontology Alignment For Equivalence and Subsumption Correspondences

Kim Jung Jae (NTU)

Ontology alignment refers to the task of finding correspondences between entities in different ontologies. It facilitates inter-operability between applications using different ontologies as the correspondences allow them to understand one another's data. We present novel methods for the discovery of both equivalence and subsumption correspondences between concepts of different ontologies and use them for knowledge representation and integration in the biomedical domain. For equivalence correspondence discovery, we introduce a novel technique, BOAT. In BOAT, we improve accuracy by combining a word-based comparison with a structural comparison. Given a pair of concepts, we collect difference words and determine if each of them distinguishes the concepts by using the ontology structures. Concept pairs with no distinguishing difference words are considered equivalent. BOAT is also one of the fastest matchers as it reduces the time taken for matching large ontologies using a candidate selection technique that selects only concept pairs with high similarities for comparison. Equivalence correspondences are alone insufficient to fully support inter-operability. Subsumption correspondences complement them by explicitly stating the generalization of a concept in one ontology over other concepts in another ontology. Instance-based techniques can be used to determine whether a subsumption correspondence exists between a pair of concepts based on the common instances they share. However, this technique cannot be applied to ontologies which are not instantiated. We propose a technique for subsumption and equivalence correspondence discovery, which resolves this issue by instantiating ontologies with their annotations on text corpora.

@ 10.00am-10.30am

Annotation of Gene Expression Changes Related to Cancers in the Literature for Gene Classification

Hee-Jin Lee (KAIST)

In order to develop efficient diagnostics and therapy for cancers, it is important to identify various genes that are involved in cancers and understand their roles in cancer development. For the identification of cancer-involved genes, the process of how a gene's expression level changes across normal cells and cancer cells is often investigated and reported in the literature. It is why there are several text mining (TM) systems to effectively collect mentions of gene expression changes related to cancers from the literature. However, these TM systems are not yet designed to recognize and distinguish diverse ways in which gene expression changes are stated about cancers. In this talk, we present a corpus to be made publicly available on gene expression changes as linked to cancer, which consists of 826 sentences, about prostate, breast and ovarian cancers. This corpus is based on a novel annotation scheme with four concepts that are mutually independent. The annotation results enable us to classify genes into three classes, or `oncogenes', `tumor suppressor genes' and `biomarkers'. The mapping between the annotation results and gene classes is summarized into 10 inference rules. We show that the annotation achieved high inter-annotator agreements, together with validated inference rules.

@ 10.30am-11.00am

Use of Clue Word Annotations as the Silver-standard in Training Models for Biological Event Extraction

Seung-Cheol Baek (KAIST)

Current state-of-the-art approaches to biological event extraction train models by reconstructing relevant graphs from training sentences, where labeled nodes correspond to tokens that indicate the presence of events and the relations between nodes correspond to the relations between these events and their participants. Since multi-word expressions may also indicate events, these approaches use heuristic rules to define target graphs to re-construct by mapping various clue words into single tokens. Since training instances define actual problems to solve, the method of deriving graphs must affect the system performance. However, there has not been any related work on this aspect, to the best of our knowledge. In this talk, we present a method to incorporate an EM algorithm into supervised learning to look for training graphs that are more favorable to model construction. We evaluate our algorithm on the development dataset in the 2009 BioNLP shared task and show that this algorithm makes a statistically meaningful improvement on the performance of trained models over a supervised learning algorithm on a fixed set of training graphs.

@ 10.00am-11.30am

DigSee: Disease Gene Search Engine with Evidence Sentences

Jeong Kyun Kim (Gwangju Institute of Science and Technology)

Biological events such as gene expression, regulation, phosphorylation, localization, and protein catabolism play important roles in the development of diseases. Understanding the association between diseases and genes can be enhanced with the identification of involved biological events in this association. Although biological knowledge has been accumulated in several databases and can be accessed through the Web, there is no specialized Web tool yet allowing for a query into the relationship among diseases, genes, and molecular events. For this, we developed DigSee (Disease oriented search engine with evidence sentences: ver. cancer, http://gcancer.org/digsee) to search Medline abstracts for evidence sent-ences describing that “genes” are involved in the development of “cancer” through &ldq uo;molecular events”. For validation of evidence sentences, we constructed 563 gold standard evidence sentences describing that a gene is involved in cancer through biological events. Then, we built a novel machine learning model based on 10 linguistically- motivated features using the training sentences. Using these features, a Bayesian classifier was evaluated against the test sentences, which achieved 0.823 AUC score. This accuracy was compared to that of the simple bag-of-words approach using a support vector machine as a classifier. Our new method outperformed the baseline.

11.30am-12.30pm: Lunch Break

12.30pm-13.30pm: Biological Networks

@ 12.30pm-13.00pm

Counting motifs in the entire biological network from noisy and incomplete data

Tran Ngoc Hieu (NTU & NUS)

Small over-represented motifs in biological networks are believed to represent essential functional units of biological processes. Given that current high-throughput biotechnology is only able to interrogate a portion of a biological network with non-negligible errors, how to estimate motif occurrences in the entire network from noisy and incomplete data is studied in this work. We develop a powerful method to correct link errors in estimating the number of occurrences of an undirected or directed motif. The proposed estimators are proved mathematically to be asymptotically unbiased and consistent. They are further applied to four eukaryotic protein-protein interaction (PPI) networks and forty-one cell-specific transcription factor (TF) regulatory networks in human. It is found that the number of triangles in the entire human protein interactome is 125x larger than in the S. cerevisiae interactome, 2.5x as large as expected. It is also found that there is a very strong positive linear correlation between the number of occurrences of widely studied triad and quadriad motifs in human cell-specific TF regulatory networks.

@ 13.00pm-13.30pm

Supervised maximum-likelihood weighting of composite protein networks for complex prediction

Yong Chern Han (NUS)

Protein complexes participate in many important cellular functions. Finding the set of existent complexes is crucial for understanding the organization and regulation of processes in the cell. With the availability of large amounts of high-throughput protein-protein interaction (PPI) data, many algorithms have been proposed to discover protein complexes from PPI networks. However, such approaches are hindered by the high rate of noise in high-throughput PPI data, including spurious and missing interactions. Further, many transient interactions are detected between proteins that are not from the same complex, while not all proteins from the same complex may actually interact. As a result, predicted complexes often do not match true complexes well, and many true complexes go undetected. We address these challenges by integrating PPI data with other heterogeneous data sources to construct a composite protein network, and using a supervised maximum-likelihood approach to weight an edge based on its posterior probability of belonging to a complex. We then use six different clustering algorithms, and an aggregative clustering strategy, to discover complexes in the weighted network. We test our method on S. cerevisiae and H. sapiens, and show that complex discovery is improved: compared to previously proposed supervised and unsupervised weighting approaches, our method recalls more known complexes, achieves higher precision at all recall levels, and generates novel complexes of greater functional similarity. Furthermore, our maximum-likelihood approach allows learned parameters to be used to visualize and evaluate the evidence of novel predictions, aiding human judgment of their credibility.

13.30pm-15.00pm: Systems Biology

@ 13.30pm-14.00pm

Model Identifiability of Biochemical Networks

Sridharan Srinath (I^{2}R)

Mathematical modeling has become an integral component in biotechnology, in which these models are frequently used to design and optimize bioprocessses. Canonical models, like power-laws within the Biochemical Systems Theory, offer numerous mathematical and numerical advantages, including built-in flexibility to simulate general nonlinear behavior. Construction of such models relies on the estimation of unknown case-specific model parameters by way of experimental data fitting, also known as inverse modeling. Despite the large number of publications on this topic, this task remains the bottleneck in canonical modeling of biochemical systems. The focus of this paper concerns with the question of identifiability of power-law models from dynamic data, that is, whether the parameter values can be uniquely and accurately identified from time-series data. Existing and new parameter identifiability methods were applied to two power-law models of biochemical systems, and the results pointed to the lack of parametric identifiability as the root cause of the difficulty faced in the inverse modeling. Despite the focus on power-law models, the analyses and conclusions are extendable to other canonical models, and the issue of parameter identifiability is expected to be a common problem in biochemical system modeling.

@ 14.00pm-14.30pm

Parameter estimation of bio-pathway models using statistical model checking

Sucheendra K. Palaniappan (NUS)

Building quantitative models that describe the dynamics of bio-pathways is a primary task of computational systems biology. However, these models are often difficult to construct due to the presence of many unknown parameters, which need to be estimated. Once a reliable model has been obtained, numerous analysis tasks can be performed on these models. Traditional methods for parameter estimation on ordinary differential equation (ODE) based models do not take into consideration (i) the uncertainties associated with initial states, the noisiness and the cell-population based nature of experimental data and (ii) qualitative information about the dynamics of the system. Techniques from the domain of formal verification, in particular, model checking can address these limitations. We present a statistical model checking based framework to perform parameter estimation for ODE based bio-pathway models. We combine statistical model checking with standard global search strategy for parameter estimation. We illustrate the usefulness and scalability of the framework with examples.

@ 14.30pm-15.00pm

Bayesian Estimation and Analysis of Bio-Pathway Models using Kernel-enhanced Particle Filters

Benjamin Gyori (NUS)

Quantitative models are essential to understand the dynamics of bio-pathways. However, estimating the parameters of such models remains a challenging task. Bayesian filtering provides a framework for estimateing and representing a posterior probability distribution over the space of parameters. In particular, particle filters can sequentially approximate such distributions in a non-parametric way. However, the performance of particle filters can degrade due to limited sample size, resulting in collapsed distributions. We provide an improved particle filter that guarantees sample diversity, while preserving the parameter posterior. This is achieved by enhancing the particle filter with an additional MCMC kernel, which prevents the set of samples from collapsing. The method is demonstrated on a widely used model of the JAK-STAT signalling pathway, and results are compared to previously proposed particle filtering methods. Using the same number of samples, the proposed method is superior in providing a representative estimation of the parameter posterior. This is critical in making accurate predictions in a Bayesian manner. We show that with the set of particles, we can give more informative model predictions, represented as probability distributions. The particle filter can also be used to select between alternative models with differing structures.

15.00pm-15.30pm: Tea Break

15.30pm-16.30pm: Genome Analysis

@ 15.30pm-16.00pm

Inference of Spatial Organizations of Chromosomes Using Semi-definite Embedding Approach and Hi-C Data

Zhang Zhizhuo (NUS)

For a long period of time, scientists studied genomes assuming they are linear. Recently, chromosome conformation capture (3C) based technologies, such as Hi-C, have been developed that provide the loci contact frequencies among loci pairs in a genome-wide scale. The technology unveiled that two far-apart loci can interact in the tested genome. It indicated that the tested genome forms a 3D chromsomal structure is to model the 3D chromosomal structure from the 3C-derived data computationally. This paper presents a deterministic method called ChromSDE, which applies semi-definite programming techniques to find the best structure fitting the observed data and uses golden section search to find the correct parameter for converting the contact frequency to spatial distance. To the best of our knowledge, ChromSDE is the only method which can guarantee recovering the correct structure in the noise-free case. In addition, we prove that the parameter of conversion from contact frequency to spatial distance will change under different resolutions theoretically and empirically. Using simulation data and real Hi-C data, we show that ChromSDE is much more accurate and robust than existing methods. Finally, we demonstrate that interesting biological findings can be uncovered from our predicted 3D structure.

@ 16.00pm-16.30pm

A computational approach to identifying drug resistance associated mutations in bacterial strains

Michal Wozniak (NUS & Univ of Warsaw)

Drug resistance in bacterial pathogens is an increasing problem, which stimulates research in bioinformatics. However, our current understanding of drug resistance mechanisms remains incomplete.

The fast-growing number of fully sequenced bacterial strains now enables us to develop new methods to identify mutations associated with drug resistance. We present a new computational method, employing phylogenetic information, to identifying genes and mutations associated with drug resistance mechanisms by comparative analysis of multiple bacterial strains within the same species of bacteria. In order to test our method, we collected genotype and phenoltype data of 100 fully sequenced strains of S. aureus and 10 commonly used drugs. Our computational experiments suggest that the method outperforms the baseline approaches which do not employ the phylogenetic information. Applying our method, we rediscovered the most common genetic determinants of drug resistance and identified some novel putative associations.

16.30pm-17.00pm: 3D Microscopy

@ 16.30pm-17.00pm

Digital Reconstruction of Neuronal Structures from 3D Microscopy Data

Sreetama Basu (IPAL)

Understanding the mechanism of the mammalian brain is a grand challenge in neuroscience. The neurons are known to exhibit an intricate structure-function co-relation. Hence, the neuronal morphology is an important source of information to the biologists. Advances in microscopy techniques produce huge volume of microscopy images where manual reconstruction and analysis is very laborious. We introduce a novel stochastic framework for unsupervised reconstruction of neuronal morphology from 3-dimensional microscopy data. The Marked Point Process aims to extract the neuronal networks by fitting an optimum configuration of objects to the data. We propose an energy function on the configuration of objects for detection of networks of tubular structures. The optimization of the energy function is achieved by a stochastic, discrete-time Multiple Birth and Death dynamics. The performance of the proposed model is demonstrated on axonic arbors from 3-dimensional confocal microscopy data from the DIADEM data set.

17.00pm-17.15pm: Closing

6.30pm-9.30pm: Workshop Banquet

University Club @
Level 4, Shaw Foundation House

Participants

GIST

1. *Hyunju Lee, hyunjulee@gist.ac.kr
2. *Jeong Kyun Kim, jeongkyunkim@gist.ac.kr

KAIST

3. *Jong C. Park, park@nlp.kaist.ac.kr
4. *Hee Jin Lee, heejin@nlp.kaist.ac.kr
5. *Seung Cheol Baek, scbaek@nlp.kaist.ac.kr

I^{2}R

6. *Ng See Kiong, skng@i2r.a-star.edu.sg
7. *Li Xiaoli, xlli@i2r.a-star.edu.sg
8. ^Sridharan Srinath, sridharans@i2r.a-star.edu.sg
9. *Alireza Vazifedoost, alirezav@comp.nus.edu.sg

IPAL

10. *Sreetama Basu, 9sreetama@gmail.com

NTU

11. *Jung-jae Kim, jungjae.kim@ntu.edu.sg
12. ^Han Xu
13. *Tran Ha Nguyen

14. ^Chen Xin, chenxin@ntu.edu.sg
15. *Sun Ruimin, rsun1@e.ntu.edu.sg
16. ^Tran Ngoc Hieu, nhtran@ntu.edu.sg

17. *Pham Nguyen Tuan Anh, pham0070@e.ntu.edu.sg

NUS

18. *Ken Sung, ksung@comp.nus.edu.sg
19. *Zhang Zhizhuo, zzz2010@gmail.com
20. ^Jing Quan Lim, jing.quan.lim@gmail.com
21. *杨瑞杰, A0095650@nus.edu.sg
22. *Narmada Sambaturu, narmada.sambaturu@gmail.com
23. *Peiyong Guan, guan8py@gmail.com
24. Chandana Tennakoon, drcyber@gmail.com

25. *PS Thiagarajan, thiagu@comp.nus.edu.sg
26. *Sucheendra K. Palaniappan, sucheendrak@gmail.com
27. *Benjamin M. Gyori, ben.gyori@gmail.com
28. *Ram
29. *Ratul
30. *Rajarshi

31. ^Wong Limsoon, wongls@comp.nus.edu.sg
32. *Michał Woźniak, mw219725@gmail.com
33. *Yong Chern Han, radiocherny@gmail.com
34. *Kevin Lim, kevinl@comp.nus.edu.sg
35. *Wilson Goh, gohwils@gmail.com

36. ^Tan Chew Lim, tankl@comp.nus.edu.sg
37. *Abhinit Kumar Ambastha, abhinit-kumar.ambastha@stud.ki.se

38. ^Wynne Hsu, whsu@comp.nus.edu.sg
39. ^Lee Mong Li, leeml@comp.nus.edu.sg