Introduction

Proceedings of the Fourth Annual Conference of the MidSouth Computational Biology and Bioinformatics Society

Dawn Wilkins[1], Yuriy Gusev[2], Raja Loganantharaj[3], Susan Bridges[4], Stephen Winters-Hilt[5] and Jonathan D Wren*[6]

Address: [1]Computer & Information Science Department, The University of Mississippi, University, MS, USA, [2]Department of Surgery, Health Sciences Center, The University of Oklahoma, Oklahoma City, Oklahoma 73104, USA, [3]Bioinformatics Research lab, University of Louisiana, Lafayette, LA, USA, [4]Department of Computer Science and Engineering, Mississippi State University, Box 9637, Mississippi State, MS 39762, USA, [5]Department of Computer Science, University of New Orleans, and The Research Institute for Children, 200 Henry Clay Ave., New Orleans, LA 70118, USA and [6]Arthritis and Immunology Research Program, Oklahoma Medical Research Foundation; 825 N.E. 13th Street, Oklahoma City, Oklahoma 73104-5005, USA

Email: Dawn Wilkins - dwilkins@cs.olemiss.edu; Yuriy Gusev - Yuriy-Gusev@ouhsc.edu; Raja Loganantharaj - logan@cacs.louisiana.edu; Susan Bridges - bridges@cse.msstate.edu; Stephen Winters-Hilt - winters@cs.uno.edu; Jonathan D Wren* - Jonathan.Wren@OMRF.org

* Corresponding author

Introduction

The Fourth Annual MidSouth Computational Biology and Bioinformatics Society (MCBIOS-IV) conference was held in New Orleans, Louisiana on February 1st–3rd, 2007. The conference venue included two locations: the Lindy Boggs Conference facility at the University of New Orleans (UNO); and Dinner/Speaker venues in the French Quarter – Broussard's of New Orleans and The House of Blues. The conference featured three days of technical presentations, with the third day partly devoted to a satellite Conference on Nanopore Cheminformatics with an on-site demo presentation of a nanopore detector (from the Research Institute for Children and UNO Nanopore Cheminformatics/Biophysics Labs). The theme of this year's conference was "Computational Frontiers in Biomedicine".

At MCBIOS 2007, awards for outstanding poster presentations were given to the following students: In Poster Session I, 1st place went to 1st: Mutlu Mete of UALR, 2nd place to William Sanders of MSU and 3rd place to Stephanie Byrum of the UALR. In Poster Session II, 1st place went to Matthew Landry of UNO, 2nd place to Terra Colvin, Jr. of UALR, and 3rd place to Iftekhar Amin of UNO. 1st place

for outstanding oral presentation went to Vijayaraj Nagarajan of USM.

Proceedings summary

This year, 24 out of 31 submitted papers were accepted for inclusion in the proceedings (77%), similar to the number published in last year's Proceedings [1-21]. Each of the papers was peer-reviewed by at least two members of the program committee members and/or external experts in the field. Our goal, as in past years, has been to be inclusive yet rigorous in selecting only high-quality papers. The general themes of this year's proceedings papers fall into several categories:

Systems biology

One of the most important challenges of current miRNA research is to decipher the "code" of miRNA regulation – to find the connection between miRNA expression and phenotypic changes. Gusev *et al* [22] report the results of a systems biology based analysis of aberrantly expressed miRNAs in five human cancers. Their findings suggest that co-expressed miRNAs collectively provide a systemic compensatory response to the abnormal phenotypic changes

in cancer cells by targeting a broad range of functional categories and signaling pathways affected in a particular cancer.

One of the things evident from the Gusev *et al* study is that there is a large body of microarray data that is becoming available for analysis. As such, methods to begin inferring regulatory networks from this data are important. In another paper, Peng Li *et al* compare probabilistic Boolean Network (PBNs) and Dynamic Bayesian Network (DBNs) approaches to correctly inferring regulatory networks [24]. They find that PBNs can reduce the computational complexity, false positive and false negative errors significantly, while DBNs can more accurately derive genetic network interactions, but are more time-consuming.

While microarray technology is steadily improving, it still suffers from noise; hence experiments are repeated several times to reduce error. To reduce the amount of replication necessary, Dozmorov *et al* [23] used F-tests against system-level noise to identify hypervariable genes from time-course microarray experiments. This novel systems-biology approach to biological network reconstruction investigated urothelial cell response to infection with *Enterococcus* bacteria. A complex response was mapped out involving cytoskeletal rearrangement, immune response, modulation of growth and cellular metabolism, and *Wnt* signaling, as well as responses heretofore unrecognized because they involve poorly annotated genes.

OMICS
Biological analysis spans several different areas from the genome/proteome to the metabolome, collectively referred to as "omics" for the study of different biological bodies. In one study, Schnackenberg *et al* [25] study age-related differences in Sprague-Dawley rats by examining changes in metabolite concentrations in their urine by NMR and UPLC/MS. Their findings are in line with the free-radical theories of aging, as they find a higher concentration of oxidized antioxidants in older rats. They also examine the effects of data normalization procedures and the impact on statistical analyses.

Nagarajan and Elasri [26] use bioinformatics approaches to predict the structure and function of Msa, a novel gene in the human pathogen *S. aureus*. Their combination of methods suggests that Msa is membrane-bound with sites for phosphorylation and protein-binding, suggesting it plays a role in signal transduction, which is consistent with its known activity as a modulator of the protein SarA.

Not all peptide fragments are represented equally in mass spectrometry (MS) experiments. To help predict which peptides might be lost or underrepresented, Sanders *et al.*

[27] use artificial neural networks (ANNs) to predict which proteolytic peptides generated by a protein dataset are likely to be detectable by mass spectrometry. The result is an improved method for calculating protein coverage in proteomics experiments and a mechanism for determining if proteins in specific pathways under study are likely to be detected by mass spectrometry.

Bridges et al [28] develop & describe a system, ProtQuant, to provide relative quantification of proteins in high-throughput proteomics samples (MudPIT) using label-free quantification. ProtQuant differs from existing label-free approaches in that it extrapolates the values of missing data points, where possible, from below-threshold identifications. The Java-based tool has a graphical user interface and accepts multiple file formats.

Cheminformatics
Four papers explore a new transduction-based nanopore detector mechanism. The first [29] introduces the transduction detection method and shows results indicating the applicability to examination of binding in individual molecular complexes in very general circumstances. The next [30] applies this method to the examination of binding for two *DNA-protein binding* interactions: (1) TBP – TATA receptor binding, and (2) HIV Integrase – HIV DNA Terminus binding. The method is also effective at detecting *DNA-DNA binding* interactions that occur with annealing of DNA single-strand overhangs [31] as well as *protein-protein binding* interactions for the medically important case of antibody-antigen interactions [32].

Machine learning based pattern recognition
Pattern recognition is a critical part of making sense of the high-throughput data gathered in modern biomedical experiments. Four papers explore the development of machine learning based pattern recognition methods and their application to resolving complex nanopore-transduction detector signals. The first [33] describes a new Support Vector Machine (SVM) based method for *clustering* (unsupervised learning) – a marked departure from the standard supervised-learning approach to SVMs. The author's objective was to have a powerful, non-parametric, method for phase tracking on nanopore transduction signals, a key requirement for extracting binding kinetics from channel current signals. They also describe a new form of Hidden Markov Model (HMM) that has the strengths of the much more complex HMM-with-Duration (HMMwD) models, but at a computational cost approximating the simpler HMM [34]. The goal is to apply this method in a real-time pattern recognition informed sampling process on the nanopore detector. The third paper, [35], examines learning on exact HMMwD models and their use in two-state signal decomposition. The fourth paper, [36], explores (i) non-standard HMM

implementations for improved feature extraction and SVM classification performance, (ii) SVM classification improvements resulting from introduction of a single "spike density" feature; and (iii) SVM improvements resulting from introducing a huge family of HMM transition probability features subsequently pruned by an Ada-Boost Selection process.

Microarray studies

Rather than focus on refinement of methods, several authors report the effectiveness and utility of existing bioinformatics approaches to better understand a biological system. For example, Nan Mei *et al* [37] studied gene expression changes in the livers of riddelline-treated Big Blue rats. Standard analysis methods and popular pathway analysis software was used to determine that the genes differentially expressed with significance were mainly involved in cancer, tissue development, apoptosis, cellular growth and proliferation, and others. The study helped elucidate the mechanisms involved in toxicity and carcinogenesis due to exposure to riddelline.

Guo *et al*. [38] used microarrays in conjunction with pathway analysis software to test the hypothesis that Pyrrolizidine alkaloids (PAs), common in many plants, cause liver toxicity and/or cancer in experimental animals. They found that genes within carcinogenic pathways were disproportionately altered, supporting their hypothesis.

Circadian rhythms are generally associated with the sleep/wake cycle, but also regulate many activities and affect the expression pattern of practically all genes. Time-course microarrays can potentially detect this baseline oscillation, which Ptitsyn and Gimble [39] use in an interesting study on the leptin signaling system. This system is a major regulator of energy metabolism, responsible to the sensation of satiety after a meal. They observe tissue-specific alternative polyadenylation of SOCS3 transcripts, whereby alternative transcripts different by the length of a 3' UTR oscillate in counter-phase. This study suggests a mechanism that can provide a constant abundance of transcript and volume of cytokine signal transduction regardless of circadian time.

Genomic analysis

Identification of DNA binding sites for transcription factors (motifs) is important for a complete understanding of co-regulation of gene expression, but has proven to be quite challenging. Das and Dai [40] review previously published algorithms for DNA motif finding. The algorithms reviewed are string-based, probabilistic, and machine learning techniques that fall into three major classes: Those that use promoter sequences of co-regulated genes from single genome, phylogenetic footprinting, and a hybrid of the two. Although there has been

substantial progress in this area within recent years and algorithms work reasonably well for prokaryotic organisms, success with motif finding in eukaryotes has been more elusive.

Loganantharaj *et al* [41] have proposed a general methodology for validating the effectiveness of phylogenetic profiling, using the Gene Ontology as the gold standard for validating functional similarity among the genes in each cluster. They demonstrated that phylogenetic profiling technique showed poor performance in functional prediction in human and mouse. However, their empirical study shows strong support for few cohesive functional groups in each phylogenetic cluster. They concluded that phylogenetic profiling is still a very useful technique for predicting function of an unknown protein sequence.

Pirooznia *et al* [42] report the results of a large-scale EST sequencing project for the earthworm, *Eisenia fetida*, which is often used in toxicology studies. They describe the sequencing and analysis of 3,144 new ESTs.

Miscellaneous

The insertion of new or altered genes into genomes is a key step in many functional analysis studies, and it is important to determine how many copies of each of these transgenes are present. Yuan *et al* [43] report the development of a new statistical approach that facilitates a more accurate transgene copy number estimation.

Ding *et al* [44] propose a new algorithm for divisive clustering, which is similar to bisecting *k*-median, but which uses statistical spatial depth to identify the "center" of a cluster. A new subcluster selection rule, Relative Average Depth, is also introduced. In data sets that are noisy or have high dimension and low sample size, which is common in gene expression data sets, the bisecting *k*-spatial median algorithm does well compared to the component-wise bisecting *k*-median algorithm.

Cancer diagnosis usually begins with histopathological examinations of tissue biopsies. These evaluations are usually somewhat subjective and the growing number of such images has provides an opportunity to test automated approaches to tissue sample categorization. Mete *et al* [45] report a method for automated analysis of squamous cell carcinomas using a SVM. They report a classification accuracy of 96% on their test set, which is quite promising for the future of histopathology.

Future meetings

The fifth annual MCBIOS Conference will be held in Oklahoma City, Oklahoma in the Cox Convention Center in downtown Oklahoma City on February 23rd and 24th, 2008. Our web site, http://www.MCBIOS.org, contains

further information on the society and future meetings. MCBIOS is a regional affiliate of the International Society for Computational Biology http://www.ISCB.org.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

All authors served as co-editors for these proceedings, with JDW serving as Senior Editor. All authors helped write this editorial.

Acknowledgements

References

1. Chen T, Guo L, Zhang L, Shi L, Fang H, Sun Y, Fuscoe JC, Mei N: **Gene Expression Profiles Distinguish the Carcinogenic Effects of Aristolochic Acid in Target (Kidney) and Non-target (Liver) Tissues in Rats.** *BMC Bioinformatics* 2006, **7(Suppl 2):**S20.
2. Delongchamp R, Lee T, Velasco C: **A method for computing the overall statistical significance of a treatment effect among a group of genes.** *BMC Bioinformatics* 2006, **7(Suppl 2):**S11.
3. Ding Y, Wilkins D: **Improving the Performance of SVM-RFE to Select Genes in Microarray Data.** *BMC Bioinformatics* 2006, **7(Suppl 2):**S12.
4. Frank RL, Mane A, Ercal F: **An Automated Method for Rapid Identification of Putative Gene Family Members in Plants.** *BMC Bioinformatics* 2006, **7(Suppl 2):**S19.
5. Guo L, Fang H, Collins J, Fan X, Dial S, Wong A, Mehta K, Blann E, Tong W, Dragan YP: **Differential gene expression in mouse primary hepatocytes exposed to the peroxisome proliferator-activated receptor alpha agonists.** *BMC Bioinformatics* 2006, **7(Suppl 2):**S18.
6. Han T, Melvin CD, Shi L, Branham WS, Moland CL, Pine PS, Thompson KL, Fuscoe JC: **Improvement in the Reproducibility and Accuracy of DNA Microarray Quantification by Optimizing Hybridization Conditions.** *BMC Bioinformatics* 2006, **7(Suppl 2):**S17.
7. Han T, Wang J, Tong W, Moore MM, Fuscoe JC, Chen T: **Microarray analysis distinguishes differential gene expression patterns from large and small colony Thymidine kinase mutants of L5178Y mouse lymphoma cells.** *BMC Bioinformatics* 2006, **7(Suppl 2):**S9.
8. Iqbal RT, Winters-Hilt S, Landry M: **DNA Molecule Classification Using Feature Primitives.** *BMC Bioinformatics* 2006, **7(Suppl 2):**S15.
9. Kel A, Voss N, Jauregui R, Kel-Margoulis O, Wingender E: **Beyond microarrays: Find key transcription factors controlling signal transduction pathways.** *BMC Bioinformatics* 2006, **7(Suppl 2):**S13.
10. Loganantharaj R, Cheepala S, Clifford J: **Metric for Measuring the Effectiveness of Clustering of DNA Microarray Expression.** *BMC Bioinformatics* 2006, **7(Suppl 2):**S5.
11. Mei N, Guo L, Zhang L, Shi L, Sun Y, Moland CL, Dial SL, Fuscoe JC, Chen T: **Analysis of gene expression changes in relation to toxicity and tumorigenesis in the livers of Big Blue transgenic rats fed comfrey (Symphytum officinale).** *BMC Bioinformatics* 2006, **7(Suppl 2):**S16.
12. Nagarajan V, Kaushik N, Murali B, Zhang C, Lakhera S, Elasri MO, Deng Y: **A Fourier Transformation based Method to Mine Peptide Space for Antimicrobial Activity.** *BMC Bioinformatics* 2006, **7(Suppl 2):**S2.
13. Nahum LA, Reynolds MT, Wang ZO, Faith JJ, Jonna R, Jiang ZJ, Meyer TJ, Pollock DD: **EGenBio: A Data Management System for Evolutionary Genomics and Biodiversity.** *BMC Bioinformatics* 2006, **7(Suppl 2):**S7.
14. Ptitsyn A, Zvonic S, Gimble JM: **Permutation test for periodicity in short time series data.** *BMC Bioinformatics* 2006, **7(Suppl 2):**S10.
15. Smolinski TG, Buchanan R, Boratyn GM, Milanova M, Prinz AA: **Independent Component Analysis-motivated Approach to Classificatory Decomposition of Cortical Evoked Potentials.** *BMC Bioinformatics* 2006, **7(Suppl 2):**S8.
16. Sun H, Fang H, Chen T, Perkins R, Tong W: **GOFFA: Gene Ontology For Functional Analysis – A FDA Gene Ontology Tool for Analysis of Genomic and Proteomic Data.** *BMC Bioinformatics* 2006, **7(Suppl 2):**S23.
17. Thodima V, Pirooznia M, Deng Y: **RiboaptDB: A Comprehensive Database of Ribozymes and Aptamers.** *BMC Bioinformatics* 2006, **7(Suppl 2):**S6.
18. Winters-Hilt S: **Hidden Markov Model Variants and their Application.** *BMC Bioinformatics* 2006, **7(Suppl 2):**S14.
19. Winters-Hilt S, Landry M, Akeson M, Tanase M, Amin I, Coombs A, Morales E, Millet J, Baribault C, Sendamangalam S: **Cheminformatics Methods for Novel Nanopore analysis of HIV DNA termini.** *BMC Bioinformatics* 2006, **7(Suppl 2):**S22.
20. Winters-Hilt S, Yelundur A, McChesney C, Landry M: **Support Vector Machine Implementations for Classification & Clustering.** *BMC Bioinformatics* 2006, **7(Suppl 2):**S4.
21. Wren JD: **A scalable machine-learning approach to recognize chemical names within large text databases.** *BMC Bioinformatics* 2006, **7(Suppl 2):**S3.
22. Gusev Y, Schmittgen TD, Lerner M, Postier R, Brackett D: **Computational Analysis of Biological Functions and Pathways Collectively Targeted by Co-expressed microRNAs in Cancer.** *BMC Bioinformatics* 2007, **8(Suppl 7):**S16.
23. Dozmorov MG, Kyker KD, Saban R, Shankar N, Baghdayan AS, Centola MB, Hurst RE: **Systems biology approach for mapping the response of human urothelial cells to infection by Enterococcus faecalis.** *BMC Bioinformatics* 2007, **8(Suppl 7):**S2.
24. Li P, Zhang C, Perkins EJ, Gong P, Deng Y: **Comparison of probabilistic Boolean network and Dynamic Bayesian network approaches for inferring gene regulatory networks.** *BMC Bioinformatics* 2007, **8(Suppl 7):**S13.
25. Schnackenberg LK, Sun J, Esperandi P, Holland RD, Hanig J, Beger RD: **Metabonomics Evaluations of Age-Related Changes in the Urinary Compositions of Male Sprague Dawley Rats and Effects of Data Normalization Methods on Statistical and Quantitative Analysis.** *BMC Bioinformatics* 2007, **8(Suppl 7):**S3.
26. Nagarajan V, Elasri MO: **Structure and Function Predictions of the Msa Protein in Staphylococcus aureus.** *BMC Bioinformatics* 2007, **8(Suppl 7):**S5.
27. Sanders WS, Bridges SM, McCarthy FM, Nanduri B, Burgess SC: **Prediction of peptides observable by mass spectrometry.** *BMC Bioinformatics* 2007, **8(Suppl 7):**S23.
28. Bridges SM, Magee GB, Wang N, Williams WP, Burgess SC, Nanduri B: **ProtQuant: A Tool for the Label-Free Quantification of MudPIT Proteomics Data.** *BMC Bioinformatics* 2007, **8(Suppl 7):**S24.
29. Winters-Hilt S: **The a-Hemolysin Nanopore Transduction Detector – single-molecule binding studies and immunological screening of antibodies and aptamers.** *BMC Bioinformatics* 2007, **8(Suppl 7):**S9.
30. Winters-Hilt S, Davis A, Amin I, Morales E: **Preliminary Nanopore Cheminformatics of Protein Binding to Non-Terminal and Terminal DNA regions: Analysis of Transcription Factor Binding and Retroviral DNA Terminus Dynamics.** *BMC Bioinformatics* 2007, **8(Suppl 7):**S10.
31. Thomson K, Amin I, Morales E, Winters-Hilt S: **Preliminary Nanopore Cheminformatics Analysis of Aptamer-Target Binding Strength.** *BMC Bioinformatics* 2007, **8(Suppl 7):**S11.

32. Winters-Hilt S, Morales E, Amin I, Stoyanov A: **Nanopore-based Kinetics Analysis of Individual Antibody-Channel and Antibody-Antigen Interactions.** *BMC Bioinformatics* 2007, **8(Suppl 7):**S20.
33. Winters-Hilt S, Merat S: **SVM Clustering.** *BMC Bioinformatics* 2007, **8(Suppl 7):**S18.
34. Winters-Hilt S, Baribault C: **A Novel, Fast, HMM-with-Duration Implementation, with Application to Pattern Recognition Informed Sampling Control of a Nanopore Detector.** *BMC Bioinformatics* 2007, **8(Suppl 7):**S19.
35. Churbanov A, Baribault C, Winters-Hilt S: **Duration learning for nanopore ionic flow blockade analysis.** *BMC Bioinformatics* 2007, **8(Suppl 7):**S14.
36. Landry M, Winters-Hilt S: **Analysis of Nanopore Detector Measurements using Machine Learning Methods, with Application to singlemolecule Kinetic Analysis.** *BMC Bioinformatics* 2007, **8(Suppl 7):**S12.
37. Mei N, Guo L, Liu R, Fuscoe JC, Chen T: **Gene expression changes induced by the tumorigenic pyrrolizidine alkaloid riddelliine in liver of Big Blue rats.** *BMC Bioinformatics* 2007, **8(Suppl 7):**S4.
38. Guo L, Mei N, Dial S, Fuscoe J, Chen T: **Determination of the Active Components for Carcinogenicity of Comfrey by Comparison of Gene Expression Profiles Altered by Comfrey and Riddelliine in Rat Liver.** *BMC Bioinformatics* 2007, **8(Suppl 7):**S22.
39. Ptitsyn AA, Gimble JM: **Analysis of circadian pattern reveals tissue-specific alternative transcription in leptin signaling pathway.** *BMC Bioinformatics* 2007, **8(Suppl 7):**S15.
40. Das MK, Dai HK: **A survey of DNA motif finding algorithms.** *BMC Bioinformatics* 2007, **8(Suppl 7):**S21.
41. Loganantharaj R, Atwi M: **Towards Validating the Hypothesis of Phylogenetic Profiling.** *BMC Bioinformatics* 2007, **8(Suppl 7):**S25.
42. Pirooznia M, Gong P, Guan X, Inouye LS, Yang K, Perkins EJ, Deng Y: **Cloning, Analysis and Functional Annotation of Expressed Sequence Tags from the Earthworm *Eisenia fetida*.** *BMC Bioinformatics* 2007, **8(Suppl 7):**S7.
43. Yuan JS, Burris J, Stewart NR, Mentewab A, Stewart CN: **Statistical Tools for Transgene Copy Number Estimation Based on Real-time PCR.** *BMC Bioinformatics* 2007, **8(Suppl 7):**S6.
44. Ding Y, Dang X, Peng H, Wilkins D: **Robust Clustering in High Dimensional Data Using Statistical Depth.** *BMC Bioinformatics* 2007, **8(Suppl 7):**S8.
45. Mete M, Xu X, Fan C-Y, Shafirstein G: **Automatic Delineation of Malignancy in Histopathological Head and Neck Slides.** *BMC Bioinformatics* 2007, **8(Suppl 7):**S17.