BMC Bioinformatics

BioMed Central

Introduction

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

Jonathan D Wren*[1], Yuriy Gusev[2], Andrey Ptitsyn[3] and Stephen Winters-Hilt[4]

Address: [1]Advanced Center for Genome Technology, Stephenson Research and Technology Center, Department of Botany and Microbiology, 101 David L. Boren Blvd., The University of Oklahoma, Norman Oklahoma 73019, USA, [2]Department of Surgery, Health Sciences Center, The University of Oklahoma, Oklahoma City, Oklahoma 73104, USA, [3]Department of Microbiology, Immunology and Pathology, Colorado State University, Fort Collins, CO 80523-1619, USA and [4]Department of Computer Science, University of New Orleans, New Orleans, LA, 70148, USA and The Research Institute for Children, 200 Henry Clay Ave., New Orleans, LA 70118, USA

Email: Jonathan D Wren* - Jonathan.Wren@OU.edu; Yuriy Gusev - Yuriy-Gusev@ouhsc.edu; Andrey Ptitsyn - ptitsyaa@pbrc.edu; Stephen Winters-Hilt - winters@cs.uno.edu

* Corresponding author

from The Third Annual Conference of the MidSouth Computational Biology and Bioinformatics Society
Baton Rouge, Louisiana. 2–4 March, 2006

Published: 26 September 2006

BMC Bioinformatics 2006, **7**(Suppl 2):S1 doi:10.1186/1471-2105-7-S2-S1

Introduction

The Third Annual MidSouth Computational Biology and Bioinformatics Society (MCBIOS-III) conference was held in Baton Rouge, Louisiana on March 2nd-4th, 2006, under the banner of "Bioinformatics: A Calculated Discovery". The conference featured three days of scientific platform presentations, posters, and panel discussions, with a Cheminformatics satellite conference on the final day. The conference resulted in the 22 papers shown in this proceedings, almost doubling the number over last year's 12 papers [1-12], which is consistent with the rapid growth of MCBIOS since its inception [13]. The strong growth in the Society's proceeding publications, and the strong showing of over 100 presenters at MCBIOS-III, is all the more impressive given the last-minute change of venue forced by the impact of Hurricane Katrina on the Gulf Coast region.

At MCBIOS 2006, awards for outstanding oral presentations were given to the following students: Yuanyuan Ding of the University of Mississippi (1st place), Stephanie Hebert of the University of Arkansas at Fayetteville (2nd place), and Michael Dyar of Ouachita Baptist University (3rd place). Awards for outstanding poster presentations were given to the following students: Zhijie Jiang of Louisiana State University (1st place), Charles McChesney of the University of New Orleans (2nd place), Vinay Ravindrakumar of UALR/UAMS (3rd place).

Proceedings summary

Papers submitted to these proceedings were peer-reviewed by at least two reviewers, including program committee members and external experts as necessary. The aim of the proceedings was to be inclusive yet rigorous in selecting only high-quality papers, with the final acceptance rate being 73%. The innovative bioinformatics research in the region is reflected in these accepted papers, which fall into several general themes as follows:

Advances in methods for microarray analysis

There are a number of steps in microarray analysis and the papers published here are indicative of the impact of computational methods at each one. The first step is technical – getting good empirical measurements. Tao Han *et al.* present a technical study on the best methods to optimize washing and hybridization conditions for microarrays[14], and showed that optimizing these conditions improved accuracy and reproducibility.

After microarray data has been gathered, the next challenge is to determine the statistical significance of the transcriptional response. Robert Delongchamp *et al.* present a new method for computing the overall statistical significance of a treatment effect among predefined sets of genes (e.g., sets of genes grouped by gene ontology (GO) terms) [15]. Computer simulations demonstrated that ignoring the correlations among genes overstates the significance assigned to GO terms. The authors propose statistical tests based upon meta-analysis methods for combining p-values to correct for gene expression correlations.

After defining which genes are differentially expressed, the next step is often grouping or clustering these genes into similar expression profiles. Raja Loganantharaj *et al.* measured the effectiveness of microarray clustering algorithms [16] by calculating inter-cluster cohesiveness and intra-cluster separation for biological processes and molecular functions associated with the genes in the cluster. In a similar effort, Ding and Wilkins introduce a variant of the Recursive Feature Elimination (RFE) method for classifying gene expression data [17]. Their method is implemented using a Support Vector Machine (SVM), and borrows ideas from simulated annealing to reduce what is normally a very computationally intensive task. RFE is a common and well-studied method for reducing the number of attributes used for further analysis or development of prediction models. The goal of the algorithm is to improve the computational performance of recursive feature elimination by eliminating chunks of features at a time with as little effect on the quality of the reduced feature set as possible. The algorithm was tested on several large gene expression data sets and shown to be more time efficient in generating a set of attributes that is very similar to the set produced by RFE.

Data produced in microarray experiments carries a high degree of stochastic variation, and in time series data, this variation can obscure periodic patterns. Furthermore, in many experiments a limited number of replicates covers no more than two complete periods of oscillation. To address this, Andrey Ptitsyn *et al.* developed a new method for identifying periodicity within time-series data and compared its performance versus previous methods of identifying periodicity to show that their new method was more sensitive and precise [18]. The authors applied this method to a study of circadian expression on a large data set, representing three different peripheral murine tissues, and re-analyzed a number of similar time series data sets produced and published independently by other research groups over the past few years. This test is based on a random permutation of time points in order to estimate the non-randomness of a periodogram. This Permutated time, or Pt-test, is able to detect oscillations within a given period in expression profiles dominated by a high degree of stochastic fluctuations or oscillations of different irrelevant frequencies. The software is implemented as a set of C++ programs available from the authors on the open source basis.

Finally, post-response microarray analysis typically consists of identifying functional commonalities among the responding genes. Towards this end, Hongmei Sun *et al.* report a new FDA microarray analysis tool called Gene Ontology For Functional Analysis (GOFFA) [19], which provides an interface for visualization and analysis of GO categories associated with responding genes.

Microarray studies

Bioinformatics entails not just the development of new methods of microarray analysis, but studies on the effectiveness of their application. To this end, several groups report bioinformatics-based microarray analysis using several model systems. Nan Mei *et al.* report their analysis of liver-based gene expression changes of Big Blue transgenic rats when fed comfrey, a perennial plant native to most of North America, Europe, and western Siberia that has been used as a herbal medicine for more than 2000 years [20]. Their study of gene expression profiles helps provide a better understanding of hepatotoxicity induced by comfrey and exerted through pyrrolizidine alkaloid plant components.

Lei Guo *et al.* used microarray analysis to study primary hepatocytes from mice that had been exposed to peroxisome proliferators-activated receptor alpha (PPAR-alpha) agonists [21]. PPAR-alpha agonists lower plasma triglyceride and cholesterol levels, which is a very important pharmacological effect given the rise in cholesterol-related heart deaths as well as obesity in western societies. Their results suggest that PPARα agonist exposure results in increased oxidative stress and increased peroxisome proliferation, which can account for the pleiotropic and sometimes carcinogenic effects of PPAR-alpha agonists.

Tao Han *et al.* report a study of L5178Y mouse lymphoma cells, using large and small colony Thymidine kinase mutants[22]. To gain insight into the underlying mechanisms for formation of large and small colony *thymidine kinase* (*Tk*) mutants due to different growth rates of the mouse lymphoma cells, the authors conducted microarray analysis of gene expression profiles from the two different types of mutants. Their findings suggest that genes in the DNA segment altered by the *Tk* mutations were significantly up-regulated in the small colony mutants, but not in the large colony mutants, leading to differential expression of a set of growth regulation genes related to cell apoptosis and other cellular functions related to the restriction of cell growth.

Tao Chen *et al.* contrasted the effect of aristolochic acid (AA) upon different organ systems – liver versus kidney[23]. AA is an active component of herbal drugs and can induce nephropathy and kidney cancer in people and rodents, but does not damage the liver. To evaluate whether microarray analysis can be used for distinguishing the tissue-specific carcinogenicity of AA, they examined gene expression profiles in kidney and liver of rats treated with carcinogenic doses of AA. They found many more significant genes associated with carcinogenesis, defense response, apoptosis, immune response and organic acid metabolism in kidney than in liver due to AA exposure. These differential alterations between kidney and liver could be the underlying mechanisms for the tissue-specific toxicity and carcinogenicity of AA.

Machine learning-based cheminformatics

Stephen Winters-Hilt's group reports several studies in cheminformatics. In Winters-Hilt [24] the author describes Hidden Markov Model (HMM) variants based on hash and/or gap interpolated Markov models, and a novel implementation of HMM-with-Duration is introduced. The new HMM-with-Duration implementation is much simpler than those previously known and has far-reaching application to gene-structure identification and analysis of channel current blockade data. In Winters-Hilt *et al.* [25] they describe SVM implementations for clustering and classification, where novel, information-theoretic, kernels were successfully employed for notably better performance over standard kernels, and where two SVM approaches to multiclass discrimination/classification are described: (i) internal multiclass (with a single optimization), and (ii) external multiclass (using an optimized decision tree). In Iqbal *et al.* [26] they apply Adaboost to circumvent the typical limitations in decision tree approaches, at the expense of requiring an expert to train the classifier (i.e., there is minimal automation in the tuning of key parameters). The approach is based on feature primitives and, once tuned on a given data-type, results in the dramatic reduction in training time on a given data-set to find the best solution.

In Winters-Hilt [27] a nanopore cheminformatics method is described that is able to measure molecular conformational and binding characteristics by use of a reporter molecule that binds to certain molecules, with subsequent distinctive blockade by the bound-molecule complex. A web-interface to many of the machine learning based cheminformatics methods that have been developed is also described. It is hypothesized that reaction histories of *individual* molecules can be observed on model DNA/DNA, DNA/Protein, and Protein/Protein systems, and preliminary results are shown to support this for each case. Nanopore detection capabilities are also described for highly discriminatory biosensing, binding strength characterization, and rapid immunological screening. The author suggests that the heart of chemistry is now accessible to a nanopore-based, single-molecule, observation method that can track both external molecular binding states, and internal conformational states. In Winters-Hilt *et al.* [26] the nanopore detector biophysical advances from [27] and machine learning pattern recognition methods from [24,25] are used to help systematically explore internal DNA dinucleotide flexibility, with particular focus on HIV's highly conserved (and highly flexible/reactive) viral DNA termini. To support this effort a new, HMM-based, filtering method is introduced that amplifies emission variances in the HMM to achieve level-projected filtering for ease of kinetic feature extraction from the observed channel blockade currents. The observed state kinetics of the DNA hairpins containing the CA/TG dinucleotide provides strong evidence for HIV's selection of a peculiarly flexible/interactive DNA terminus.

Jonathan Wren implemented and tested a machine-learning method for automated recognition and extraction of chemical names within text [28]. The method was tested on over 7 million abstracts, which is unusually large as far as most text-based testing datasets go, yet was important to demonstrate the scalability of the approach and show that it might be feasible as a method to automatically identify chemicals within text. The study was also able to pair chemical name variants together and study how these spelling variants affected information retrieval in PubMed and Ovid, demonstrating that document recall for chemical names is sensitive to the exact spelling of the term used.

Databases

Databases are an integral part of modern biomedical research, helping to both locate and analyze categorical data. Thodima *et al.* [29] describe the RiboAptDB, a comprehensive source for sequence information on ribozymes and aptamers. Such 'unnatural' *in vitro* data are not represented in the public 'natural' sequence databases such as GenBank and EMBL. As with the sequence information found in nature, however, the amount of sequence data generated by *in vitro* selection experiments has been accumulating exponentially. In their latest version of RiboaptDB there are 370 artificial ribozyme sequences and 3,842 aptamer sequences. The authors' database also includes numerous functions, such as a general search feature, an individual feature-wise search, and a user submission form for new data through online and also local BLAST search.

Nahum *et al.* presented EGenBio, a web-based data management system for studies in Evolutionary Genomics and Biodiversity [30]. It includes managed access to curated data from external databases, rapid manipulation of

sequences, alignments, and trees, and integration and visualization of outputs. EGenBio was developed around their research program on comparative genomics and a pilot mithochondrial genome database. It is freely available at http://egenbio.lsu.edu/.

Genomic Analysis

Evolution often proceeds through gene duplication and subsequent functional divergence, and identifying the relationships between gene families is an important part of studying and understanding the difference between species. To aid in this, Ron Frank *et al.* report an automated method of identifying gene family members in plants through a battery of software tools [31], making the process simpler and more rapid.

Alexander Kel *et al.* tackle an important problem in identifying common transcription factor binding sites from microarray data [32]. The authors developed a novel computational approach for revealing key transcription factors by knowledge-based analysis of gene expression data with the help of databases on gene regulatory networks. They demonstrate that promoters of genes encoding components of many known signal transduction pathways are enriched with binding sites of those transcription factors that are endpoints of the considered pathways. Application of the approach to microarray gene expression data on TNF-alpha stimulated primary human endothelial cells helped to reveal key transcription factors that may explain concerted expression changes in signal transduction networks. The corresponding software and databases (TRANSFAC® and TRANSPATH®) are available at http://www.gene-regulation.com.

Miscellaneous

Smolinski *et al.* [33] report an independent component analysis-motivated approach to classify cortical evoked potentials. They analyze data deriving from experiments based on measuring neural activity in a controlled setting (i.e., normal) as well as under exposure to some external perturbation (nicotine exposure). They describe a practical implementation of the method, based on hybridization of independent component analysis, multi-objective evolutionary algorithms, and rough sets.

Nagarajan *et al*. [34] report a novel method to identify antimicrobial activity among peptide motifs. They use a Fourier transformation based method to mine the peptide space for potential antimicrobial activity. They also demonstrate that a property-based coding approach can significantly boost the characteristic properties of biomolecules.

Future Meetings

The fourth annual MCBIOS Conference will be held in New Orleans, Louisiana, February 1–3, at the University of New Orleans' Lindy Boggs International Conference Center. 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.

Authors' contributions

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

Acknowledgements

We thank the Conference Committee: Andrey Ptitsyn and Stephen Winters-Hilt, and the Program Committee: Steve Jennings, Dawn Wilkins, William Slikker Jr., Stephen Winters-Hilt, and Jonathan Wren. We also thank our MCBIOS members for their dedication and efforts to peer review the manuscripts submitted by the attendees, as well as Isobel Peters and Enitan Sawyer for their help with the publication process.

References

1. Delongchamp RR, Velasco C, Dial S, Harris AJ: **Genome-wide estimation of gender differences in the gene expression of human livers: statistical design and analysis.** *BMC Bioinformatics* 2005, **6(Suppl 2):**S13.
2. Fang H, Tong W, Perkins R, Shi L, Hong H, Cao X, Xie Q, Yim SH, Ward JM, Pitot HC, *et al.*: **Bioinformatics approaches for cross-species liver cancer analysis based on microarray gene expression profiling.** *BMC Bioinformatics* 2005, **6(Suppl 2):**S6.
3. Frank RL, Ercal F: **Evaluation of Glycine max mRNA clusters.** *BMC Bioinformatics* 2005, **6(Suppl 2):**S7.
4. Garge NR, Page GP, Sprague AP, Gorman BS, Allison DB: **Reproducible clusters from microarray research: whither?** *BMC Bioinformatics* 2005, **6(Suppl 2):**S10.
5. Hong H, Dragan Y, Epstein J, Teitel C, Chen B, Xie Q, Fang H, Shi L, Perkins R, Tong W: **Quality control and quality assessment of data from surface-enhanced laser desorption/ionization (SELDI) time-of flight (TOF) mass spectrometry (MS).** *BMC Bioinformatics* 2005, **6(Suppl 2):**S5.
6. Ptitsyn A, Hide W: **CLU: a new algorithm for EST clustering.** *BMC Bioinformatics* 2005, **6(Suppl 2):**S3.
7. Shi L, Tong W, Fang H, Scherf U, Han J, Puri RK, Frueh FW, Goodsaid FM, Guo L, Su Z, *et al.*: **Cross-platform comparability of microarray technology: intra-platform consistency and appropriate data analysis procedures are essential.** *BMC Bioinformatics* 2005, **6(Suppl 2):**S12.
8. Shi L, Tong W, Su Z, Han T, Han J, Puri RK, Fang H, Frueh FW, Goodsaid FM, Guo L, *et al.*: **Microarray scanner calibration curves: characteristics and implications.** *BMC Bioinformatics* 2005, **6(Suppl 2):**S11.
9. Wren JD, Johnson D, Gruenwald L: **Automating genomic data mining via a sequence-based matrix format and associative rule set.** *BMC Bioinformatics* 2005, **6(Suppl 2):**S2.
10. Wren JD, Slikker W: **Proceedings of the Midsouth Computational Biology and Bioinformatics Society 2004 Conference.** *BMC Bioinformatics* 2005, **6(Suppl 2):**S1-13.
11. Xie Q, Ratnasinghe LD, Hong H, Perkins R, Tang ZZ, Hu N, Taylor PR, Tong W: **Decision forest analysis of 61 single nucleotide polymorphisms in a case-control study of esophageal cancer; a novel method.** *BMC Bioinformatics* 2005, **6(Suppl 2):**S4.
12. Xu Z, Patterson TA, Wren JD, Han T, Shi L, Duhart H, Ali SF, Slikker W Jr: **A microarray study of MPP+-treated PC12 Cells: Mechanisms of toxicity (MOT) analysis using bioinformatics tools.** *BMC Bioinformatics* 2005, **6(Suppl 2):**S8.
13. Jennings SF, Ptitsyn AA, Wilkins D, Bruhn RE, Slikker W Jr, Wren JD: **Regional societies: fostering competitive research through virtual infrastructures.** *PLoS Biol* 2004, **2(12):**e372.

14. 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.
15. 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.
16. Logananthara R, Cheepala S, Clifford J: **Metric for Measuring the Effectiveness of Clustering of DNA Microarray Expression.** *BMC Bioinformatics* 2006, **7(Suppl 2):**S5.
17. Ding Y, Wilkins D: **Improving the Performance of SVM-RFE to Select Genes in Microarray Data.** *BMC Bioinformatics* 2006, **7(Suppl 2):**S12.
18. Ptitsyn A, Zvonic S, Gimble JM: **Permutation test for periodicity in short time series data.** *BMC Bioinformatics* 2006, **7(Suppl 2):**S10.
19. 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.
20. 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.
21. 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.
22. 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.
23. 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.
24. Winters-Hilt S: **Hidden Markov Model Variants and their Application.** *BMC Bioinformatics* 2006, **7(Suppl 2):**S14.
25. Winters-Hilt S, Yelundur A, McChesney C, Landry M: **Support Vector Machine Implementations for Classification & Clustering.** *BMC Bioinformatics* 2006, **7(Suppl 2):**S4.
26. Iqbal RT, Winters-Hilt S, Landry M: **DNA Molecule Classification Using Feature Primitives.** *BMC Bioinformatics* 2006, **7(Suppl 2):**S15.
27. 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.
28. Wren JD: **A scalable machine-learning approach to recognize chemical names within large text databases.** *BMC Bioinformatics* 2006, **7(Suppl 2):**S3.
29. Thodima V, Pirooznia M, Deng Y: **RiboaptDB: A Comprehensive Database of Ribozymes and Aptamers.** *BMC Bioinformatics* 2006, **7(Suppl 2):**S6.
30. 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.
31. 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.
32. 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.
33. 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.
34. 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.