Journal of
Data Mining in Genomics & Proteomics

Weier, J Data Mining Genomics Proteomics 2013, 4:4
DOI: 10.4172/2153-0602.1000e109

Editorial Open Access

New Algorithms for Genome Characterization, Epigenetic Profiling Analysis, Data Mining and Population-based Studies

Heinz-Ulli G Weier*

Department of Cancer & DNA Damage Responses, Life Sciences Division,University of California-LBNL, Berkeley, CA, USA

Welcome to a new, exciting issue of Journal of Data Mining in Genomics & Proteomics (J Data Mining Genomics Proteomics). This issue (Volume 4: Issue 4) presents seven exquisite examples describing various approaches to the analysis of next generation (nexgen) and high throughput RNA sequencing (RNA-Seq) data, assessment of genomic diversity as well as patient care.

The first Research Article in this issue falls in line with reports in the previous issue of Journal of Data Mining in Genomics & Proteomics (Volume 4: Issue 3) which had the central theme of 'Bioinformatics for High Throughput Sequencing' [1].

Microbial forensics enables attribution of microbial pathogen samples back to a suspected source. In their paper 'Population Analysis of Bacterial Samples for Individual Identification in Forensics Application', JP Jakupciak and colleagues [2] studied caveats and pitfalls of Next Generation Sequencing (NGS) platforms for hypothesis testing in comparative analyses.Specifically, the study described here developed a novel reference-free, bioinformatics strategy to account for and identify genetic diversity in samples. Ultimately, this may be required for NGS use both as an investigative tool and as a tool for attribution in courts of law.

The next article byYinT et al. entitled 'Visual Mining Methods for RNA-Seq Data: Data Structure, Dispersion Estimation and Significance Testing [3] describes the analysis of RNA-Seq data from soybeans and investigates why initial significance testing yields gene lists that differ between software packages used. This type of contradiction can occur generally in high-throughput analyses. The well-written and richly illustrated paper demonstrates how the disparities between the results were investigated and how they might be explained. To explore the model fitting and hypothesis testing, the authors implemented an interactive graphic that allows the exploration of the effect of dispersion estimation on the overall estimation of variance and differential expression tests.

The following two papers focus on algorithm development for genome analysis and development microsatellite markers. The research described by D. Ophir in his paper 'An Analysis of Palindromes and n-nary Tract Frequencies found in a Genomic Sequence' [4] is based on previous work by E. Chargaff and colleagues [5,6], who studied the over-representation of certain DNA binary tracts in the genomes of various species. The research described in this paper examines ternary tracts and the palindromes called 'designated tracts'. As the author shows, the binary tracts are over-represented to the same extent as the ternary tracts.Therefore, he concludes that the binary tracts dominate because they have biological impacts, but the ternary tracts do not contribute to biological impacts [4].

The contribution by Gong and Ge 'Characterization of Polymorphic Microsatellite Markers Isolated from Genomic DNA of *Elaeocarpusdecipiens* Hemsly (Elaeocarpaceae)' focuses on the assessment of genetic diversity among *Elaeocarpusdecipiens*, a broad-leaved, woody species of the Elaeocarpaceae family with a disjunct distribution in south of Chinese mainland, the Ryukyu Archipelago and Taiwan [7]. They describe 18 microsatellite markers that are adequate for detecting and characterizing population genetic structure and genetic diversity in *Elaeocarpusdecipiens*.

Algorithm development and evaluations is also the focus of the paper by GH Lubke et al. 'Gradient Boosting as a SNP Filter: an Evaluation Using Simulated and Hair Morphology Data' [8]. The authors advocate a two-step approach where step one consists of a filter that is sensitive to different types of SNP main and interactions effects. The aim is to substantially reduce the number of SNPs so that more specific modeling becomes feasible in a second step. The paper describes an evaluation of a statistical learning method called "gradient boosting machine" (GBM) that can be used as a filter. GBM does not require an a priori specification of a genetic model, and permits inclusion of large numbers of covariates. GBM can therefore be used to explore multiple interactions, which would not be feasible within the parametric framework used inGenome-wide association (GWA) studies. They show in a simulation that GBM performs well even under conditions favorable to the standard additive regression model commonly used in GWAS, and is sensitive to the detection of interaction effects even if one of the interacting variables has a zero main effect. The latter would not be detected in GWAS.

Last-but-not least, the final 2 papers in this issue of the Journal have direct relevance to patient care in families and clinics. In her paper 'The Family Knowledge about the Disease and Complications Risk among Diabetic Patients-in Poland', A. Abramczyk summarizes results from the analysis of 1366 questionnaires from families/ caregivers of diabetic patients randomly chosen from 61 primary healthcare centers in Poland [9]. The high significant results demonstrate that family knowledge about the disease is a significant factor that diversifies a medical condition of diabetic patients and a higher level of knowledge among family members about the disease improves patients' medical condition and reduces the risk of diabetes complications.

The progression renal disease is the focus of the Research Article by M. Ghattas and coworkers entitled 'The Methylation Profile of IFN-g, SOCS1 and SOCS3 Promoter Regions in End-Stage Renal Disease' [10]. The study described here was aimed at profiling the methylation status of promoter regions of three modulators of inflammation (IFN-γ, SOCS1 and SOCS3) in DNA isolated from peripheral blood of end-stage

***Corresponding author:** Heinz-Ulli GWeier, Department of Cancer & DNA Damage Responses, Life Sciences Division, E.O. Lawrence Berkeley National Lab, 1 Cyclotron Road, MS 977, Berkeley, CA 94720, USA; Tel: (001)510-486-5363; Fax: (001)510-486-5343; E-mail: ugweier@lbl.gov

Received November 25, 2013; **Accepted** November 27, 2013; **Published** November 30, 2013

Citation: Weier HUG (2013) New Algorithms for Genome Characterization, Epigenetic Profiling Analysis, Data Mining and Population-based Studies. J Data Mining Genomics Proteomics 4: e109. doi:10.4172/2153-0602.1000e109

renal disease (ESRD) patients and controls. The authors found that the methylation profiles of IFN-γ and SOCS1 promoters were significant different between patients and controls, and conclude that promoter region methylation plays an important role in the pathogenesis of ESRD.

The seven articles in this issue describe mostly the specialized research focus of their teams of investigators. Present efforts at the Journal of Data Mining in Genomics & Proteomics are underway to publish a further volume with broader contributions on 'Bioinformatics for High Throughput Sequencing' before the end of the year. Please see the JDMGP's Special Issue web site for the timeline and further information.

Acknowledgement

This work was supported in part by NIH grants CA168345 carried out at the Earnest Orlando Lawrence Berkeley National Laboratory under contract DE-AC02-05CH11231.

Conflict of Interest

The author declares no conflict of interest.

References

1. Weier HUG (2013) Bioinformatics for High Throughput Sequencing. J Data Mining Genomics Proteomics 4: e108.
2. Jakupciak JP, Wells JM, Lin JS and Feldman AB (2013) Population Analysis of Bacterial Samples for Individual Identification in Forensics Application. J Data Mining Genomics Proteomics 4: 138.
3. Yin T, Majumder M, Chowdhury NR, Cook D, Shoemaker R, et al. (2013) Visual Mining Methods for RNA-Seq Data: Data Structure, Dispersion Estimation and Significance Testing. J Data Mining Genomics Proteomics 4: 139.
4. Ophir D (2013) An Analysis of Palindromes and n-nary Tract Frequencies found in a Genomic Sequence. J Data Mining Genomics Proteomics 4: 140.
5. Tamm C, Shapiro HS, Lipshitz R and Chargaff E (1952) Distribution density of nucleotides within a deoxyribonucleic acid chain. J Biol Chem 203: 673-698.
6. Chargaff E (1963) Essays in Nucleic Acids. Elsevier Publishing Corporation, Amsterdam.
7. Gong X, Ge G (2013) Characterization of Polymorphic Microsatellite Markers Isolated from Genomic DNA of ElaeocarpusdecipiensHemsly (Elaeocarpaceae). J Data Mining Genomics Proteomics 4: 141.
8. Lubke GH, Laurin C, Walters R, Eriksson N, Hysi P, et al. (2013) Gradient Boosting as a SNP Filter: an Evaluation Using Simulated and Hair Morphology Data. J Data Mining Genomics Proteomics 4: 143.
9. Abramczyk A (2013) The Family Knowledge about the Disease and Complications Risk among Diabetic Patients-in Poland. J Data Mining Genomics Proteomics 4: 142.
10. Ghattas M, El-shaarawy F, Mesbah N, Abo-Elmatty D (2013) The Methylation Profile of IFN-g, SOCS1 and SOCS3 Promoter Regions in End-Stage Renal Disease. J Data Mining Genomics Proteomics 4: 144.