BioData Mining

BioMed Central

Editorial

Open Access

Filling the gap between biology and computer science

Jesús S Aguilar-Ruiz*[1], Jason H Moore[2,3] and Marylyn D Ritchie[4]

Address: [1]School of Engineering, Pablo de Olavide University, Seville, Spain, [2]Norris-Cotton Cancer Center, Lebanon, New Hampshire, USA, [3]Dartmouth Medical School, Lebanon, New Hampshire, USA and [4]Center for Human Genetics Research, Vanderbilt University, Nashville, TN, USA

Email: Jesús S Aguilar-Ruiz* - aguilar@upo.es; Jason H Moore - jason.h.moore@dartmouth.edu; Marylyn D Ritchie - ritchie@chgr.mc.vanderbilt.edu

* Corresponding author

Published: 17 July 2008

BioData Mining 2008, **1**:1 doi:10.1186/1756-0381-1-1

Received: 16 July 2008
Accepted: 17 July 2008

This article is available from: http://www.biodatamining.org/content/1/1/1

Abstract

This editorial introduces *BioData Mining*, a new journal which publishes research articles related to advances in computational methods and techniques for the extraction of useful knowledge from heterogeneous biological data. We outline the aims and scope of the journal, introduce the publishing model and describe the open peer review policy, which fosters interaction within the research community.

Aim and scope

BioData Mining [1] is an open-access, open peer-reviewed, online journal that publishes articles on the development of data mining techniques applied to biological data. The journal stems from the gap between biology and computer science and covers a number of topics in the middle of these fields. One of the main interests of *BioData Mining* is the advance in computational methods or theoretical informatics for the progress in the discovery of new knowledge in biomedical sciences.

Data mining [2] techniques have been traditionally used in many varied contexts. Usually datasets contained many examples (thousands) and some attributes (at most several tens). Algorithms have been developed taking into account these characteristics, and have been validated by means of statistical tests with synthetic and real-world data. Statistics has been the support for any analysis of biological data for many years. However, the biological data has changed over time in size, but above all in structure, and many challenges arise from genetic, transcriptomic, genomic, proteomic and metabolomic data.

The enormous increase of biological data incorporates another element of difficulty because statistics, without losing its relevance, has moved to the background leaving in the foreground a space for complex heuristics. In addition, the curse of dimensionality plays an important role in the design of new data mining algorithms. However, the most important challenge comes from the intrinsic characteristics of new problems to be solved. Due to the high volume of data, optimization and efficiency are key aspects in the design of new heuristics, which many times only provide approximate solutions.

In this sense, *BioData Mining* aims at publishing articles that not only adapt, evaluate or apply traditional data mining techniques, but also that develop, evaluate or apply novel methods from data mining or machine learning fields to the analysis of complex biological data.

Moreover, the situation has substantially changed during the last decade. Nowadays, biological information is distributed and adopts different formats. It is not trivial to consider different types of data, which are located in dif-

ferent databases and present various levels of structure or heterogeneity. In some cases the effort is focused on facilitating the management of biological information, dealing with semantic aspects of the information through the Internet.

In order to promote the advance in science many research groups are making their software development projects publicly available, as open-source software, which encourages researchers to develop extensions of verified software applications, like interfaces, packages or specific services.

BioData Mining aims at publishing articles that design, develop and integrate databases, software and web services for the storage, management and retrieval of complex biological data, with emphasis on open-source software for the application of data mining to the analysis such type of information.

The role of biologists, geneticists, physicians, etc. is critical in the correct interpretation of results obtained by data mining algorithms. In many cases, data needs to be preprocessed for extracting useful knowledge and, in some cases, algorithms produce models that must be post-processed to get an insight of the knowledge that information hides. At the end, experimental validation is crucial to show the research community the quality of the approaches. In this field, statistics offers robust tools that can be applied directly, although new developments are also needed to deal with biological data.

BioData Mining aims at publishing articles that present new methods for pre-processing, post-processing and validation of data mining algorithms for the analysis of genetic, transcriptomic, genomic, proteomic, and metabolomic data.

In the expectation of filling the gap between biology and computer science, we believe that BioData Mining will contribute to the development of theoretical and practical aspects of new methodologies driven by biological data.

Open access and open peer review publishing model

The time interval between the date an article is written and the date an article is read should be as short as possible. Long intervals are mainly due to slow reviewing process and limited access to articles. *BioData Mining* will put much effort into reducing the reviewing process to several weeks, and will avoid the other aspect due to the open access nature of the journal, i.e., articles will be fully accessible online to any reader immediately upon publication.

In order to make the peer review process transparent *BioData Mining* has adopted an open-review policy. Reviewers' names are included on the peer review reports and are made publically available upon acceptance of an article. We believe that this will foster constructive reviews, and therefore enrich the criticism. This policy will contribute greatly in driving young researchers to improve the quality of their articles.

During the last years, many journals have adopted the open-access policy. Nowadays the success is unquestionable. We expect that the open peer review policy will follow a similar path in the near future, and some experiences show enthusiasm for the concept, such as PLoS ONE[3], that strongly urge reviewers to relinquish the anonymity to promote open decision-making.

Finally, to facilitate the search for topics or related research in articles published in *BioData Mining*, the readers will find all the articles archived in PubMed Central [4].

Editorial Board

The journal is run by two Editors-in-Chief, who subscribe this editorial. Marylyn D. Ritchie acts as Managing Editor. The members of the Editorial Board cover a wide range of research fields related to Biology and Computer Science. To mention some expertise, it ranges from Biomedical Informatics (M. Ramoni [5], F. Azuaje [6]) to Structural Bioinformatics (R. Casadio [7], D. Jones [8], J. M. Carazo [9]), from Soft Computing (O. Cordon [10], I. Zwir [11]) to Clinical Research (M. Eppstein [12]), from Machine Learning (E. Marchiori [13], K. Cios [14]) to Evolutionary Computation (P. Larrañaga [15]), from Cancer Research (S. Volinia [16]) to Data Mining (J. Aguilar-Ruiz [17], D. Simovici [18], D. Gamberger [19], H. Toivonen [20]), from Biostatistics (J. Rahnenfuhrer [21]) to High–performance Technologies (R. Schneider [22]), from Immunology (B.A. McKinney [23]) to Computational Genetics (J.H Moore [24], M.D. Ritchie [25]), from Database Integration (M. Kanehisa [26]) to Functional Genomics (S. Kasif [27]), from Software Technologies (A. Omicini [28]) to Stem Cells (B. Soria [29]).

References

1. **BioData mining** [http://www.biodatamining.org]
2. Hand DJ, Mannila H, Smyth P: *Principles of Data Mining* The MIT Press; 2001.
3. **PLoS ONE** [http://www.plosone.org]
4. **PubMed Central** [http://www.pubmedcentral.nih.gov]
5. Schachter A, Ramoni M: **Clinical forecasting in drug development.** *Nat Rev Drug Disc* 2007, **6:**107-108.
6. Wang H, Zheng H, Simpson D, Azuaje F: **Machine learning approaches to supporting the identification of photoreceptor-enriched genes based on expression data.** *BMC Bioinformatics* 2006, **7:**116.
7. Bartoli L, Calabrese R, Fariselli P, Mita D, Casadio R: **A computational approach for detecting peptidases and their specific inhibitors at the genome level.** *BMC Bioinformatics* 2007, **8:**S3.

8. Lise S, Walker-Taylor A, Jones DT: **Docking protein domains in contact space.** *BMC Bioinformatics* 2006, **7:**310.
9. Scheres S, Gao H, Valle M, Herman G, Eggermont P, Frank J, Carazo J: **Disentangling conformational states of macromolecules in 3D-EM through likelihood optimization.** *Nature Methods* 2007, **4:**27-29.
10. Romero R, Rubio C, Cordon O, Cobb J, Herrera F, Zwir I: **A multi-objective evolutionary conceptual clustering methodology for gene annotation within structural databases: A case of study on the Gene Ontology database.** *IEEE Transactions on Evolutionary Computation* [http://ieeexplore.ieee.org/Xplorlogin.jsp?url=/iel5/4235/4358751/04469888_.pdf?tp=&arnum ber=4469888&isnum-ber=4358751].
11. Zwir I, Shin D, Kato A, Nishino K, Latifi T, Solomon F, Hare JM, Huang H, Groisman EA: **Dissecting the PhoP regulatory network of Escherichia coli and Salmonella enterica.** *PNAS* 2005, **102(8):**2862-2867.
12. Eppstein M, Molofsky J: **Invasiveness in plant communities with feedbacks.** *Ecology Letters* 2007, **10:**253-263.
13. Vanhoutte K, Laarakkers C, Marchiori E, Pickkers P, Wetzels J, Willems J, Heuvel L van den, Russel F, Masereeuw R: **Biomarker discovery with SELDI-TOF MS in human urine associated with early renal injury: evaluation with computational analytical tools.** *Nephrol Dial Transplant* 2007, **22(10):**2932-2943.
14. Swiercz W, Cios KJ, Staley K, Kurgan LA, Accurso F, Sagel S: **A new synaptic plasticity rule for networks of spiking neurons.** *IEEE Transactions on Neural Networks* 2006, **17:**94-105.
15. Larrañaga P, Lozano J: *Estimation of Distribution Algorithms. A New Tool for Evolutionary Computation* 2002 [http://www.springer.com/computer/artificial/book/978-0-7923-7466-4]. Kluwer Academic Publishers
16. Petrocca F, Visone R, Onelli M, Shah M, Nicoloso M, de Martino I, Iliopoulos D, Pilozzi E, Liu C, Negrini M, Cavazzini L, Volinia S, Alder H, Ruco L, Baldassarre G, Croce C, Vecchione A: **E2F1-regulated microRNAs impair TGFbeta-dependent cell-cycle arrest and apoptosis in gastric cancer.** *Cancer Cell* 2008, **13(3):**272-286.
17. Aguilar-Ruiz JS: **Shifting and scaling patterns from gene expression data.** *Bioinformatics* 2005, **21(20):**3840-3845.
18. Jaroszewicz S, Simovici D, Kuo W, Ohno-Machado L: **The Goodman-Kruskal coefficient and its applications in genetic diagnosis of cancer.** *IEEE Trans Biomed Eng* 2004, **51(7):**1095-1102.
19. Gamberger D, Lavrac N, Krstacic A, Krstacic G: **Clinical data analysis based on iterative subgroup discovery: Experiments in brain ischaemia data analysis.** *Applied Intelligence* 2007, **27:**205-217.
20. Landwehr N, Mielikainen T, Eronen L, Toivonen H, Mannila H: **Constrained Hidden Markov Models for Population-based Haplotyping.** *Bioinformatics* 2007, **8:**S9.
21. Schlicker A, Rahnenfuhrer J, Albrecht M, Lengauer T, Domingues FS: **GOTax: investigating biological processes and biochemical activities along the taxonomic tree.** *Genome Biology* 2007, **8(3):**R33.
22. Barbosa-Silva A, Satagopam VP, Schneider R, Ortega JM: **Clustering of cognate proteins among distinct proteomes derived from multiple links to a single seed sequence.** *BMC Bioinformatics* 2008, **9:**141.
23. Kallewaard N, McKinney B, Gu Y, Chen A, Venkataram B, JE Crowe J: **Functional maturation of the human antibody response to rotavirus.** *The Journal of Immunology* 2008, **180:**3980-3989.
24. Moore JH, Barney N, Tsai C, Chiang F, Gui J, White B: **Symbolic modeling of epistasis.** *Hum Hered* 2007, **63(2):**120-33.
25. Schwarz U, Ritchie M, Bradford Y, Li C, Dudek S, Frye-Anderson A, Kim R, Roden D, Stein C: **Genetic determinants of response to warfarin during initial anticoagulation.** *The New England Journal of Medicine* 2008, **358(10):**999-1008.
26. Kanehisa M, Araki M, Goto S, Hattori M, Hirakawa M, Itoh M, Katayama T, Kawashima S, Okuda S, Tokimatsu T, Yamanishi Y: **KEGG for linking genomes to life and the environment.** *Nucleic Acids Res* 2008, **36:**D480-D484.
27. Murali T, Wu C, Kasif S: **The art of gene function prediction.** *Nat Biotechnol* 2006, **24(12):**1474-5.
28. Viroli M, Ricci A, Omicini A: **Operating Instructions for Intelligent Agent Coordination.** *The Knowledge Engineering Review* 2006, **21:**49-69.
29. Vaca P, Martin F, Vegara-Meseguer J, Rovira J, Berna G, Soria B: **Induction of differentiation of embryonic stem cells into insulin-secreting cells by fetal soluble factors.** *Stem Cells* 2006, **24(2):**258-65.