Biology Direct

BioMed Central

Editorial

Opening Pandora's Box: making biological discoveries through computational data exploration

L Aravind

Address: National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA

Email: L Aravind - aravind@ncbi.nlm.nih.gov

Published: 20 November 2007

Biology Direct 2007, **2**:29 doi:10.1186/1745-6150-2-29

Received: 16 November 2007

Accepted: 20 November 2007

This article is available from: http://www.biology-direct.com/content/2/1/29

Editorial

Over two decades ago the first electronic databases which systematically collected protein and nucleic acid sequence and structure data came into existence. This period also saw the emergence of the first algorithms and their implementations to query these databases for sequence and structure similarities, to align sequences, and to identify of compositional features in sequences. A parallel advance was the integration of these disparate data collections into strongly interconnected databases, which included repositories of biomedical literature (e.g. Entrez). It would be no exaggeration to state that these computational developments played a role comparable to that of the polymerase chain reaction in the rise of modern molecular biology.

The essential philosophy of this new movement within biology – computational biology – has been the use of computational methods to explore repositories of biological information to make new scientific discoveries. An early example of the success of these methods was the identification of the helix-turn-helix domain as a determinant of DNA-protein interaction [1]. This allowed the prediction of diverse bacterial and eukaryotic transcription factors, and resulted in testable hypotheses regarding the functions of key developmental regulators and oncogenes [2,3]. Ever since, computational investigations have resulted in discovery of new protein domains and prediction of their biochemical roles [4], discovery of new RNAs [5], identification of subcellular targeting signals in proteins [6] and prediction of transcription factor binding sites [7]. Application of such methodologies has also been at the heart of genomics – being central to the interpretation of genome sequences. Most remarkably it has allowed us to reconstruct the biology of diverse life forms, such as the syphilis pathogen [8], the malarial parasite [9] or the diverse uncultivable microorganisms [10], which were never too amenable to classical experimentation. The successes of genomics have also spawned whole assemblies of new forms of high-throughput data. These include genome-scale collations of data pertaining to gene expression, protein-protein interactions, genetic interactions and intra-population genomic polymorphisms. By adding a new layer of contextual information to that contained in sequences and structures of biomolecules these new datasets greatly add to the power of the computational discovery process.

The principal idea behind announcing the Discovery Notes section of *Biology Direct* is to augment the process of discovery in light of the unprecedented accumulation of biological data. The articles submitted to this section aim to occupy a specific niche in the already rich menagerie of publications. Papers announcing the sequencing of a particular genome or a high-throughput genome-scale analysis hardly do justice to all that can be inferred from the data presented in them. Especially in the case of the high-profile scientific magazines with a space-crunch, much is relegated to supplementary material and may not necessarily hit the target audience. Likewise, comprehensive papers discussing the evolution of particular biological systems might contain many specific findings that are lost in the bulk of the article. In light of this, we feel that key findings that are likely to elucidate previously obscure issues, provide novel connections or spur experimental investigations in new and unexpected directions are best published separately as succinct publications. Such publications would do greater justice to such discoveries that

might not require a full-scale paper by making the key findings more accessible to the relevant audience. The value of such publications can be easily discerned by considering exemplars from the field of protein sequence analysis. For example, the papers announcing the identification of the UBA domain [11], or the BRCT domain [12] or the PAS domain [13,14] have enormously benefited, respectively, the study of the ubiquitin system, DNA repair and the sensing of light/redox stimuli.

The goal of the new 'Discovery Notes' section of Biology Direct is to publish brief reports of specific discoveries made by computational analysis of nucleic acid and/or protein sequences, structures or other data, with novel observations and conclusions about the function, organization, or evolution of proteins, genes or genomes. In format it will be comparable to the now discontinued Protein Sequence Motifs column in *TiBS* [4], and the currently active Genome analysis section in *TIG* and the Discovery Note section of *Bioinformatics*. Beyond the two above-mentioned venues, there are hardly any regular venues for such publications. The above venues employ the conventional peer review model involving three or more anonymous referees providing reports to the authors and confidential recommendations to the editors. As result, the usual vagaries of the process and the concomitant delays affect these publications. Given the swell in the data, and the high significance of some of these computational discoveries, we believe that the research community would benefit both from an additional forum for such publications, as well as a modified system of peer-review that would allow greater speed and openness.

In the original spirit of *Biology Direct*, the Discovery Notes section will follow the open peer-review format. However, there will be some key differences in the peer review process for the Discovery Notes section relative to conventional articles:

• Only two Editorial Board members are required for review of the manuscript.

• Referees should be selected from the Editorial Board that is specifically set up for discovery notes.

Reviewers would primarily assess the validity of the findings in the article, and have a right to veto publication if the findings are incorrect or trivial. The reviewers need not provide full comments/details of revisions, but are welcome to do so whenever they deem it fit. If reviewers see no requirement for additional comments, their report would merely indicate support for the publication of the manuscript. If either or both referees veto publication, the author should consider the manuscript rejected.

Given the nature of the peer-review of these articles, we believe it should not tax the reviewer excessively and he/she could return a review relatively quickly (~2 weeks). All other aspects of the peer review process remain the same as for research articles submitted to *Biology Direct*. Central to the success of such a process is an Editorial Board with a strong track record in computational analysis of biological data. Approximately 40 excellent researchers have accepted our invitation to join the Editorial Board of Discovery Notes at *Biology Direct*. While such a starting base is encouraging, it is clear that the ultimate success of this section would depend on the authors submitting their interesting findings for publication. We do hope that this section provides a venue for fostering an active community of explorers seeking biological discoveries *in silico*.

References

1. Sauer RT, Yocum RR, Doolittle RF, Lewis M, Pabo CO: **Homology among DNA-binding proteins suggests use of a conserved super-secondary structure.** *Nature* 1982, **298(5873):**447-451.
2. Laughon A, Scott MP: **Sequence of a Drosophila segmentation gene: protein structure homology with DNA-binding proteins.** *Nature* 1984, **310(5972):**25-31.
3. Frampton J, Leutz A, Gibson T, Graf T: **DNA-binding domain ancestry.** *Nature* 1989, **342(6246):**134.
4. McEntyre JR, Gibson TJ: **Patterns and clusters within the PSM column in TiBS, 1992-2004.** *Trends Biochem Sci* 2004, **29(12):**627-633.
5. Eddy SR: **Computational genomics of noncoding RNA genes.** *Cell* 2002, **109(2):**137-140.
6. Eisenhaber B, Eisenhaber F: **Posttranslational modifications and subcellular localization signals: indicators of sequence regions without inherent 3D structure?** *Curr Protein Pept Sci* 2007, **8(2):**197-203.
7. Elemento O, Slonim N, Tavazoie S: **A universal framework for regulatory element discovery across all genomes and data types.** *Mol Cell* 2007, **28(2):**337-350.
8. Subramanian G, Koonin EV, Aravind L: **Comparative genome analysis of the pathogenic spirochetes Borrelia burgdorferi and Treponema pallidum.** *Infect Immun* 2000, **68(3):**1633-1648.
9. Gardner MJ, Hall N, Fung E, White O, Berriman M, Hyman RW, Carlton JM, Pain A, Nelson KE, Bowman S, Paulsen IT, James K, Eisen JA, Rutherford K, Salzberg SL, Craig A, Kyes S, Chan MS, Nene V, Shallom SJ, Suh B, Peterson J, Angiuoli S, Pertea M, Allen J, Selengut J, Haft D, Mather MW, Vaidya AB, Martin DM, Fairlamb AH, Fraunholz MJ, Roos DS, Ralph SA, McFadden GI, Cummings LM, Subramanian GM, Mungall C, Venter JC, Carucci DJ, Hoffman SL, Newbold C, Davis RW, Fraser CM, Barrell B: **Genome sequence of the human malaria parasite Plasmodium falciparum.** *Nature* 2002, **419(6906):**498-511.
10. Sabehi G, Loy A, Jung KH, Partha R, Spudich JL, Isaacson T, Hirschberg J, Wagner M, Beja O: **New insights into metabolic properties of marine bacteria encoding proteorhodopsins.** *PLoS Biol* 2005, **3(8):**e273.
11. Hofmann K, Bucher P: **The UBA domain: a sequence motif present in multiple enzyme classes of the ubiquitination pathway.** *Trends Biochem Sci* 1996, **21(5):**172-173.
12. Bork P, Hofmann K, Bucher P, Neuwald AF, Altschul SF, Koonin EV: **A superfamily of conserved domains in DNA damage-responsive cell cycle checkpoint proteins.** *Faseb J* 1997, **11(1):**68-76.
13. Ponting CP, Aravind L: **PAS: a multifunctional domain family comes to light.** *Curr Biol* 1997, **7(11):**R674-7.
14. Zhulin IB, Taylor BL, Dixon R: **PAS domain S-boxes in Archaea, Bacteria and sensors for oxygen and redox.** *Trends Biochem Sci* 1997, **22(9):**331-333.