Correspondence

Annotations for all by all - the BioSapiens network

Janet Thornton for the BioSapiens Network

Address: European Bioinformatics Institute, Hinxton CB10 1SD, UK. Email: thornton@ebi.ac.uk.

Published: 10 February 2009

*Genome **Biology*** 2009, **10:**401 (doi:10.1186/gb-2009-10-2-401)

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2009/10/2/401

Abstract

The BioSapiens network has developed a distributed infrastructure for genome and proteome annotation by laboratories anywhere in the world.

Over the last five years, the BioSapiens network has developed a distributed infrastructure to facilitate the combined annotation of genomes and proteomes by laboratories scattered throughout Europe. In a series of four review articles, published in *Genome Biology* [1-4], members of the consortium have collaborated to provide an overview of current methods and challenges for the future.

In total, there are now thousands of completed genomes in the public domain and with the second revolution in DNA sequencing technology, many, many more will be determined. However, DNA sequence is merely a string of letters; it must be interpreted in terms of the RNA and proteins that it encodes and the promoter and regulatory regions that control transcription and translation. Annotation can be described as the process of 'defining the biological role of a molecule in all its complexity' and mapping this knowledge onto the relevant gene products encoded by genomes (Figure 1).

The main objective of BioSapiens, a Network of Excellence funded by the European Commission, is to provide an infrastructure and tools to support a large-scale, concerted effort to annotate genome and proteome data by laboratories distributed around Europe. The Network brought together 26 laboratories in Europe to create a Virtual Institute for Genome Annotation, divided into nodes, each focused on one aspect of genome annotation. The network provides a focus for annotation and through the organization of meetings and workshops encourages cooperation, rather than duplication of effort. The annotations generated are all available in the public domain and easily accessible through a single portal on the web [5].

The review by Harrow *et al.* [1] tackles the challenge of identifying protein-coding genes from genomic sequences. Even the concept of a 'gene' is under revision. The review focuses on the strategies being applied to delineate a number of reference human gene sets - the ones most widely used by researchers in biology - and to assess their quality and completeness. Once the genes are defined, the next challenge is to unravel how regulatory information is encoded in the genome. Gene-expression data has illuminated the consequences of transcriptional activation and propelled the quest to find common regulatory sequences in coexpressed groups of genes. Vingron *et al.* [2] attempt to summarize progress in integrating these approaches for the purpose of identifying regulatory sequence elements and their function. The other two reviews focus on annotating the proteins and their functions. As reviewed by Juncker *et al.* [3], these tasks include identifying functionally important residues, such as those involved in catalysis or binding, and predicting post-translational modifications and cellular localization. Finally, Loewenstein *et al.* [4] show how both sequence and structural data can be used to illuminate the function of the protein by recognizing a homolog. A recent trend is that many prediction tools are combined in complex workflows and pipelines that facilitate the analysis of feature combinations and use a variety of data and methods.

A key to integrated annotation is the ability to combine annotations of different types from different laboratories. Within BioSapiens, the Distributed Annotation System (DAS) is used as a lightweight data-integration infrastructure. Originally developed by Dowell *et al.* [6] for genomic sequences, DAS defines a framework for the annotation of reference

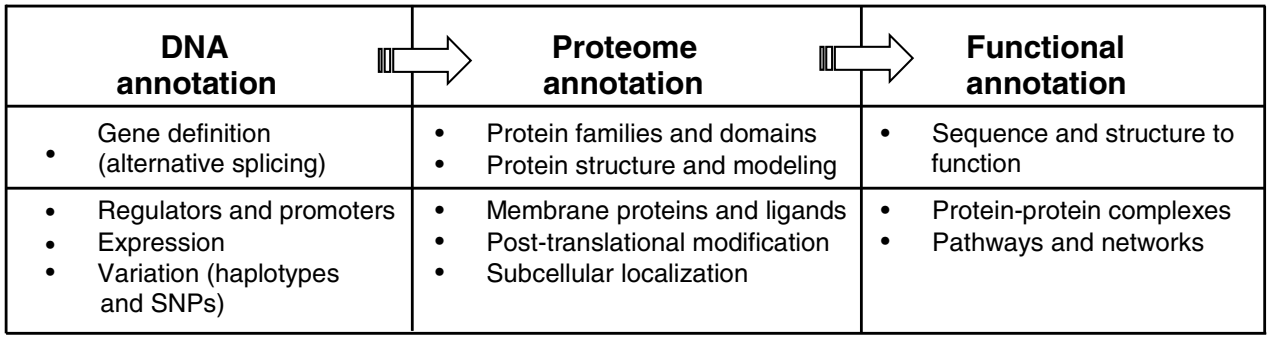

Figure 1
Steps in the analysis and annotation of genomes.

sequences by multiple independent sites. The DAS concept was extended [7] from genomic sequences to protein sequences, structures, and protein interactions. DAS clients such as DASTY [8,9] now visualize the results of many different approaches for functional protein annotation in a consistent framework. One consequence of this was the need to develop an ontology for annotating sequences [10], so that annotations from different laboratories are consistent.

This infrastructure is open to all, allowing any laboratory to generate its own annotations for proteins or genes, and to view their results in the light of other annotations, derived in other laboratories. More detail is available in a book, written by the consortium [11].

Author information

Members of the BioSapiens Network: Janet Thornton, Ewan Birney, Alvis Brazma, Rolf Apweiler, Kim Henrick, European Bioinformatics Institute, Hinxton CB10 1SD, UK; Peer Bork, European Molecular Biology Laboratory, D-69117 Heidelberg, Germany; Jacques van Helden, BiGRe - Université Libre de Bruxelles, Campus Plaine, Bvd du Triomphe - CP263, B-1050 Bruxelles, Belgium; Alfonso Valencia, Structural Biology and Biocomputing Programme, Spanish National Cancer Research Centre (CNIO), Melchor Fernández Almagro, 3, E-28029, Madrid, Spain; Roderic Guigó, Centre de Regulació Genòmica, Institut Municipal d'Investigació Mèdica, Universitat Pompeu Fabra, E-08003 Barcelona, Catalonia, Spain; Richard Durbin, Tim Hubbard, Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SA, UK; Thomas Lengauer, Max-Planck-Institut für Informatik, 66123 Saarbrücken, Germany; Martin Vingron, Computational Molecular Biology, Max-Planck-Institut für molekulare Genetik, Ihnestrasse 73, D-14195 Berlin, Germany; Dmitrij Frishman, Helmholtz Zentrum, German Research Center for Environmental Health, Munich 85764, Germany; Michal Linial, Department of Biological Chemistry, The Hebrew University of Jerusalem, Sudarsky Center, Jerusalem 91904, Israel; Anna Tramontano, Department of Biochemical Sciences, University of Rome "La Sapienza", Rome 00185, Italy; Gunnar von Heijne, Center for Biomembrane Research and Stockholm Bioinformatics Center, Department of Biochemistry and Biophysics, Stockholm University, SE-106 91 Stockholm, Sweden; Richard Mott, Bioinformatics and Statistical Genetics, University of Oxford, Wellcome Trust Centre for Human Genetics, Roosevelt Drive, Oxford OX3 7BN, UK; Christine Orengo, Research Department of Structural and Molecular Biology, University College, London WC1E, UK; Gert Vriend, Radboud University Medical Centre, 6500 HB Nijmegen, The Netherlands; Christos Ouzounis, Centre for Research and Technology, Hellas (CERTH), Thermi Road, Thessaloniki, Greece; Anne-Lise Veuthey, Swiss Institute of Bioinformatics, rue Michel Servet, CH-1211 Geneva, Switzerland; Søren Brunak, Center for Biological Sequence Analysis, Department of Systems Biology, Technical University of Denmark, DK-2800 Lyngby, Denmark; Esko Ukkonen, Helsinki Institute for Information Technology, Helsinki University of Technology and University of Helsinki, 00014 Helsinki, Finland; Stylianos Antonarakis, Department of Genetic Medicine and Development, University of Geneva Medical School and University Hospitals of Geneva, Geneva 1211, Switzerland; László Patthy, Institute of Enzymology, Biological Research Center, Hungarian Academy of Sciences, H-1113 Budapest, Hungary; Dietmar Schomburg, Department of Bioinformatics and Biochemistry, Institute for Biochemistry and Biotechnology, Technical University of Braunschweig, Langer Kamp, D-38106 Braunschweig, Germany; Antoine Danchin, Institut Pasteur, rue du Docteur Roux, Paris CEDEX 15, France; Leszek Rychlewski, BioInfoBank Institute, Poznañ Limanowskiego 24A16 60-744, Poland; Vincent Schachter, Genoscope Centre National de Sequencage Institut de genomique, Direction des Sciences du vivant, rue Gaston Cremieux, CP5706 91 057 Evry Cedex, France.

Acknowledgements

The BioSapiens project is funded by the European Commission within its FP6 Programme, under the thematic area 'Life sciences, genomics and biotechnology for health', contract number LSHG-CT-2003-503265.

References

1. Harrow J, Nagy A, Reymond A, Alioto T, Patthy L, Antonarakis SE, Guigó R: **Identifying protein-coding genes in genomic sequences.** *Genome Biol* 2009, **10:**201.
2. Vingron M, Brazma A, Coulson R, Helden Jv, Manke T, Palin K, Sand O, Ukkonen E: **Integrating sequence, evolution and functional genomics in regulatory genomics.** *Genome Biol* 2009, **10:**202.
3. Juncker AS, Jensen LJ, Pierleoni A, Bernsel A, Tress ML, Bork P, Heijne Gv, Valencia A, Ouzounis CA, Casadio R, Brunak S: **Sequence-based feature prediction and annotation of proteins.** *Genome Biol* 2009, **10:**206.
4. Loewenstein Y, Raimondo D, Redfern OC, Watson J, Frishman D, Linial M, Orengo C, Thornton J, Tramontano A: **Protein function annotation by homology-based inference.** *Genome Biol* 2009, **10:**207.
5. **A European virtual institute for genome annotation** [http://www.biosapiens.info/]
6. Dowell RD, Jokerst RM, Day A, Eddy SR, Stein L: **The distributed annotation system.** *BMC Bioinf* 2001, **2:**7.
7. Jenkinson AM, Albrecht M, Birney E, Blankenburg H, Down T, Finn RD, Hermjakob H, Hubbard TJ, Jimenez RC, Jones P, Kähäri A, Kulesha E, Macías JR, Reeves GA, Prlic A: **Integrating biological data - the Distributed Annotation System.** *BMC Bioinf* 2008, **9(Suppl 8):**S3.
8. Jimenez RC, Quinn AF, Garcia A, Labarga A, O'Neill K, Martinez F, Salazar GA, Hermjakob H: **Dasty2, an Ajax protein DAS client.** *Bioinformatics* 2008, **24:**2119-2121.
9. **Dasty2** [http://www.ebi.ac.uk/dasty]
10. Reeves GA, Eilbeck K, Magrane M, O'Donovan C, Montecchi-Palazzi L, Harris MA, Orchard S, Jimenez RC, Prlic A, Hubbard TJP, Hermjakob H, Thornton JM: **The Protein Feature Ontology: A Tool for**

the Unification of Protein Feature Annotations. *Bioinformatics* 2008, **24:**2767-2772.
11. Frishman D, Valencia A (Eds): *Modern Genome Annotation. The BioSapiens Network.* New York: Springer; 2009.