Mark Boguski, M.D., Ph.D.

Computational Biology Branch
National Center for Biotechnology
Information
Building 38A, Room 5N-503
8600 Rockville Pike
Bethesda, MD 20894
USA

Mark Boguski is a senior investigator in the Computational Biology Branch of the National Center for Biotechnology Information (NCBI) and has written and lectured extensively on bioinformatics and genomics. He received his M.D. and Ph.D. degrees from the Medical Scientist Training Program at Washington University in St. Louis, Missouri. Following residency training in pathology, Dr. Boguski was a medical staff fellow with David Lipman at the National Institute of Diabetes, Digestive and Kidney Disease and then became one of the original staff members of NCBI. There he has worked on protein sequence analysis, comparative genomics and the development of the EST database and its applications including large-scale transcript maps of the human genome, in silico polymorphism analysis and the design of gene chips for expression profiling. He also developed the first publicly available database system for expression array data. Dr. Boguski's current research interests include the analysis of data from large-scale expression studies and pharmacogenomics. He is an organizer of the Cold Spring Harbor Symposium on Genome Sequencing and Biology and has served on grant review and advisory panels for a number of government and private funding agencies and as a consultant to industry. He has received the Regents' Award from the National Library of Medicine and the NIH Director's Award. Dr. Boguski holds an adjunct faculty position in the Department of Molecular Biology and Genetics at the Johns Hopkins University School of Medicine and is a former editor of the journal Genome Research.

Interpretation bottlenecks and annotation inversion in functional genomics

The most informative annotation of genes and proteins classically derives from small-scale, biological experimentation, often resulting in the cloning of a molecular sequence of known function or phenotype. In the genome era, massive cloning and sequencing come first, followed by computational annotation based on similarity to classically annotated sequences. Both types of annotation are subject to continual revision based on new experimental and computational results, resulting in an unstable foundation on which to build interpretations of large-scale gene expression profiles. Furthermore, one's ability to efficiently interpret the data is severely limited by the necessity of manually exploring the annotation spaces. We have been experimenting with semi-automated methods to annotate clusters of genes based on data mining and summarization of information in textual databases using document 'neighbouring' and other techniques. The potential power of this approach may be limited by the structure and consistency of information in archival databases. Expression profiles (and other types of functional genomics data) may themselves be the best hope for a new type of holistic, relational annotation in the post-genome era.