Gene Discovery in dbEST

Mark S. Boguski; Carolyn M. Tolstoshev; Douglas E. Bassett Jr.

Science, New Series, Vol. 265, No. 5181, Genome Issue (Sep. 30, 1994), 1993-1994.

SCIENCE

Published by the **American Association for the Advancement of Science (AAAS),** *Science* serves its readers as a forum for the presentation and discussion of important issues related to the advancement of science, including the presentation of minority or conflicting points of view, rather than by publishing only material on which a consensus has been reached. Accordingly, all articles published in *Science*—including editorials, news and comment, and book reviews—are signed and reflect the individual views of the authors and not official points of view adopted by the AAAS or the institutions with which the authors are affiliated.

The American Association for the Advancement of Science was founded in 1848 and incorporated in 1874. Its objectives are to further the work of scientists, to facilitate cooperation among them, to foster scientific freedom and responsibility, to improve the effectiveness of science in the promotion of human welfare, to advance education in science, and to increase public understanding and appreciation of the importance and promise of the methods of science in human progress.

Membership/Circulation
Director: Michael Spinella
Deputy Director: Marlene Zendell
Member Services: Rebecca Dickerson, *Manager*; Mary Curry, *Supervisor*; Pat Butler, Helen Williams, Laurie Baker, *Representatives*
Marketing: Dee Valencia, *Manager*; Jane Pennington, *Europe Manager*; Hilary Baar, *Associate*; Angela Mumeka, *Coordinator*
Research: Renuka Chander, *Manager*
Business and Finance: Jacquelyn Roberts, *Manager*; Robert Smariga, *Assistant Manager*
Administrative Assistant: Nina Araujo de Kobes
***Science* Member Services**
Marion, Ohio: 800-347-6969;
Washington, DC: 202-326-6417
Other AAAS Programs: 202-326-6400

Advertising and Finance
Associate Publisher: Beth Rosner
Advertising Sales Manager: Susan A. Meredith
Recruitment Advertising Manager: Janis Crowley
Advertising Business Manager: Deborah Rivera-Wienhold
Finance: Randy Yi, *Senior Analyst;* Shawn Williams, *Analyst*
Marketing: John Meyers, *Manager;* Allison Pritchard, *Associate*
Traffic Manager: Tina Turano
Recruitment: Terri Seiter, *Assistant Manager*; Michael Sweet, *Production Associate*; Debbie Cummings, Celeste Wakefield, Rachael Wilson, *Sales*
Reprints Manager: Corrine Harris
Permissions Manager: Arlene Ennis
Sales Associate: Carol Maddox

PRODUCT ADVERTISING SALES: East Coast/E. Canada: Richard Teeling, 201-904-9774, FAX 201-904-9701 • **Southeast:** Mark Anderson, 305-856-8567, FAX 305-856-1056 • **Midwest:** Elizabeth Mosko, 312-665-1150, FAX 312-665-2129 • **West Coast/W. Canada:** Neil Boylan, 415-673-9265, FAX 415-673-9267 • **UK, Scandinavia, France, Italy, Belgium, Netherlands:** Andrew Davies, (44) 457-838-519, FAX (44) 457-838-898 • **Germany/Switzerland/Austria:** Tracey Peers, (44) 270-760-108, FAX (44) 270-759-597 • **Japan:** Mashy Yoshikawa, (3) 3235-5961, FAX (3) 3235-5852
RECRUITMENT ADVERTISING SALES: US: 202-326-6555, FAX 202-682-0816 • **Europe:** Gordon Clark, (44) 0223-302067, FAX (44) 0223-302068 • **Australia/New Zealand:** Keith Sandell, (61) 02-922-2977, FAX (61) 02-922-1100
Send materials to *Science* Advertising, 1333 H Street, NW, Washington, DC 20005.

Information for Contributors appears on pages 37–39 of the 7 January 1994 issue. Editorial correspondence, including requests for permission to reprint and reprint orders, should be sent to 1333 H Street, NW, Washington, DC 20005.
Internet addresses: science_editors@aaas.org (for general editorial queries); science_letters@aaas.org (for letters to the editor); science_reviews@aaas.org (for returning manuscript reviews); membership@aaas.org (for member services); science_classifieds@aaas.org (for submitting classified advertisements)

LETTERS

Gene Discovery in dbEST

The number of public complementary DNA (cDNA) sequences ("expressed sequence tags" or ESTs) has recently exceeded 50,000 (Fig. 1), and we were interested in assessing the usefulness of this resource for gene discovery. We therefore compiled a list of 32 human disease genes that had been cloned by either the positional cloning or positional candidate methods (*1*) and performed sequence homology searching (*2*) against dbEST, the database of expressed sequence tags (*3*). Thirty-eight percent of these human genes had exact and often multiple matches in dbEST, and an additional 47% were represented by homologs in other organisms (*4*). Only five human disease genes had no convincing matches with ESTs. Thus, for 85% of the human disease genes positionally cloned to date, there are homologous partial cDNA sequences in the public domain.

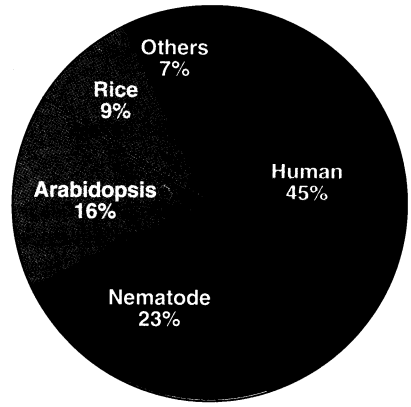

Fig. 1. dbEST contents by organism. dbEST release 2.27 contained 50,214 DNA sequences from 22 different organisms. The top four organisms, represented by more than 1200 sequences each, are shown. Information about the current release is available through the World Wide Web at http://www.ncbi.nlm.nih.gov/dbEST/index.html.

These results underscore the utility of "single pass," tag-survey cDNA sequencing (*5*) and demonstrate that much valuable information is already present in the public databases if one knows how to find it (*2*). Also underscored is the value of "model organisms" for accelerating progress in the identification of human genes by homology—an explicit goal of the U.S. Genome Program (*6*). If one is searching for exons in human genomic DNA, a statistically significant match to a cDNA—whether it be from humans, nematodes, rice, maize, or yeast—is the best proof (apart from an experiment) that an exon has been found.

dbEST may be searched by using the BLAST (*2*) e-mail or network services and full reports on individual ESTs may be obtained through the National Center for Biotechnology Information's (NCBI's) retrieve e-mail server (*7*). The capability of retrieving ESTs on the basis of their chromosome assignment and map location has recently been implemented. Instructions for submitting new sequence and mapping data are available (*7*). World Wide Web access is also provided at http://www.ncbi.nlm.nih.gov/. A National Center for Supercomputing Applications (NCSA) Mosaic interface (*4*) allows complex (Boolean) queries of dbEST to be performed.

Mark S. Boguski
Carolyn M. Tolstoshev
National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA
Douglas E. Bassett Jr.
Department of Molecular Biology and Genetics, Johns Hopkins University School of Medicine, Baltimore, MD 21205, USA

References and Notes

1. A. Ballabio, *Nature Genet.* **3**, 277 (1993).
2. S. F. Altschul, M. S. Boguski, W. Gish, J. C. Wootton, *ibid.* **6**, 119 (1994). The TBLASTN program is essential for EST homology searching. TBLASTN takes a protein query sequence and compares it with conceptual translations of DNA sequences in all six reading frames. This kind of comparison is much more sensitive than comparing nucleotides for detecting more distant, cross-phylum relationships [D. J. States and S. F. Altschul, *Methods* **3**, 66 (1991)]. Most homologs representing inexact matches would not have been detected by searching GenBank for nucleotide sequence similarities alone.
3. M. S. Boguski, T. M. J. Lowe, C. M. Tolstoshev, *Nature Genet.* **4**, 332 (1993). Although all dbEST sequences are also present in the EST Division of GenBank [D. Benson, D. J. Lipman, J. Ostell, *Nucleic Acids Res.* **13**, 2963 (1993)], dbEST contains additional value-added annotation such as the latest homologies, mapping data, and contact information for obtaining physical DNA clones. In addition to cDNA data, dbEST contains some genomic sequences that have been isolated by exon "trapping" or "amplification" [for example, A. J. Buckler *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **88**, 4005 (1991)].
4. A detailed summary of these results is available through NCSA Mosaic [B. R. Schatz and J. B. Hardin, *Science* **265**, 895 (1994)] on the World Wide Web. The Uniform Resource Locator (URL) is http://

www.ncbi.nlm.nih.gov/dbEST/index.html.

5. M. D. Adams *et al.*, *Science* **252**, 1651 (1991); A. S. Kahn *et al.*, *Nature Genet.* **2**, 180 (1992); K. Okubo *et al.*, *ibid.*, p. 173; R. Waterston *et al.*, *ibid.* **1**, 114 (1992).
6. F. Collins and D. Galas, *Science* **262**, 43 (1993).
7. The e-mail address for BLAST is blast@ncbi.nlm.nih.gov, and the address for database records is retrieve@ncbi.nlm.nih.gov. To receive documentation, send a message containing the word 'help' (unquoted) in the body of the message. For specific information on dbEST, place the instruction 'datalib dbest' (unquoted) on a line preceding 'help.' For information on the BLAST network service, send e-mail to blast-help@ncbi.nlm.nih.gov. For information on submitting data send e-mail to info@ncbi.nlm.nih.gov. For general information, or if you do not have access to e-mail, telephone 301-496-2475 and ask for the service desk.

■

Adaptive Mutation

The report "Recombination in adaptive mutation" by Reuben S. Harris *et al.* (8 Apr., p. 258) demonstrates the role of biochemical machinery for homologous recombination in adaptive reversion of a *lacZ* gene frameshift mutation. The accompanying Perspective by David S. Thaler "The evolution of genetic intelligence" (p. 224) describes the flow of information between the environment, the cellular activities that can reorganize DNA molecules, and the genome.

Our knowledge of the cellular basis of mutation was revolutionized by Barbara McClintock's discovery of transposable elements in maize and her demonstration of their ability to generate chromosome rearrangements and new alleles at individual genetic loci (*1*). An early example of adaptive mutation in bacteria involved the ability of a transposable element, phage Mu, to form araB-lacZ hybrid protein coding sequences with kinetics that were incompatible with the Luria-Delbruck concept of stochastic mutation (*2*). The importance of transposable elements has been relatively neglected in the debate about adaptive mutation because point mutations have been considered to be more relevant to evolutionary change. Examination of sequence databases, however, has shown that cut-and-splice processes must have been a part of the evolution of loci encoding multidomain proteins and of 5' regulatory regions, which are mosaic composites of many repetitive elements that specify the binding of transcription factors. As transposable elements encode precisely the kind of cleavage and ligation activities that can mediate the required DNA rearrangements, and as their movements frequently create new regulatory configurations, their functions could serve as models for certain evolutionary processes.

The basic similarity between the role of transposable elements in mediating DNA