



Comparative genomics, genome cross- referencing and XREFdb

Comparative genomics approaches have had a strong presence in the literature in the past two years, with researchers drawing, for example, upon the wealth of expressed sequence tag (EST) data available in the EST division of GenBank (dbEST¹) to clone human homologs of genes originally cloned and characterized in other organisms². Identification of related genes in multiple organisms has proven to be particularly valuable in the identification and functional analysis of human disease genes, such as the colon cancer genes *MSH2* (Ref. 3) and *MLH1* (Refs 4, 5), and the recently cloned gene for ataxia telangiectasia⁶. As the sequence, protein structure and mapping databases grow, comparative methods will be used with increasing success to solve the mysteries of gene function and genome organization in model organisms and humans.

Historically, new members of a gene family or homologs of a particular gene

Box 1. Establishing an XREFdb account

To obtain further information on this project, and to establish an XREFdb account, access the following World Wide Web URL:
<http://www.ncbi.nlm.nih.gov/XREFdb/>

The WWW is the preferred mode of access to XREFdb. However, if you do not have access to the WWW, send an email message to the following address with 'help' in the message body:
xref@ncbi.nlm.nih.gov

have been identified in other organisms using several methods, including degenerate PCR, cross-hybridization of DNA or antibodies and similarity searching. Similarity searching has the advantage that it is automatable and can be readily applied on a genome-wide scale to identify related genes among various species. Its value in the process of identification and cloning of new genes, however, is tempered by the fact that sequence data for a particular gene must exist in the databases before it can be detected. For human gene discovery, this problem has been all but overcome by the rapidly-growing number of public EST sequences present in dbEST⁷.

A database for cross-referencing genomes

XREFdb is a publicly accessible database that is a component of a research project (the XREF project), which is devoted to cross-referencing the genetics of model organisms with mammalian phenotypes and accelerating the identification of genes mutated in human diseases. XREFdb is accessible via the WWW, and provides similarity search, mapping and relevant mammalian phenotype information (Boxes 1 and 2). The database provides researchers with BLAST⁸ similarity search results that identify significant matches between sequences of model organism proteins and mammalian peptide sequences predicted by conceptual translation of ESTs. In addition, XREFdb tracks EST matches automatically for each account holder and flags those that have not been reported previously, thereby eliminating the need to re-evaluate matches that have already been analyzed. The XREF project is also determining mouse and human map

positions for those ESTs most significantly matched by proteins from the budding yeast *Saccharomyces cerevisiae*. These map data, also available through XREFdb, will systematically establish potential cross-references between genes in model organisms and mammalian phenotypes, via the phenotype-rich mouse and human maps. Such cross-references can be particularly valuable when functional data are present for the gene product in one or more model organisms. Establishing connections in this manner has strong implications for expediting the discovery and characterization of genes mutated in human diseases⁹.

How much sequence data are available?

The success of this cross-referencing project is largely dependent upon two factors, the first of which is the amount of sequence information available for the organisms being cross-referenced. Genome sequencing projects are well underway for several model organisms. For example, the complete genomic sequence of *S. cerevisiae* is estimated to be available in early 1996, and 90% of the ORFs in *Caenorhabditis elegans* are expected to be in the public databases by the end of 1997. Many currently sequenced ORFs in these model organisms are associated with mutant phenotypes and/or other protein functional data.

Mammalian EST sequences are plentiful as well. The Washington University-Merck & Co. EST project, for example, has submitted nearly 150,000 ESTs to dbEST and GenBank at a rate of approximately 1,500 per day. The impact of this effort on the representation of human genes by EST sequences in

Yeast Interest Group

The Yeast Interest Group of the National Institutes of Health offers research information regarding budding yeast, *Saccharomyces cerevisiae*, and fission yeast, *Schizosaccharomyces pombe*, on the WWW at the URL:

<http://www.nih.gov/sigs/yeast/index.html>

We provide a description, with Nomarski pictures, of the different stages in the life cycle of fission yeast on our *S. pombe* page and schematic information on the replication cycle of budding yeast L-A virus on our *S. cerevisiae* page. In addition, research interests, recent publications and contact information is listed for the 14 NIH Bethesda Campus laboratories that use yeast as a model system. Our up-to-date WWW information is intended to facilitate collaborations and to assist postdoctoral fellows who want to join our groups. Contributed by Frans Hochstenbach (fransh@Box-f.nih.gov).

dbEST has been dramatic. The current release of dbEST (6 July 1995) contains 211936 human cDNA sequences. We have analyzed a set of genes mutated in human disease states that were isolated by positional cloning and found that 71% (32/45) are represented by one or more ESTs in dbEST (see http://www.ncbi.nlm.nih.gov/dbEST/dbEST_genes/). For example, multiple ESTs in the current release of dbEST were found to be derived from the recently cloned gene mutated in early-onset Alzheimer's disease¹⁰. Thus, it is probable that a significant proportion of the total human gene repertoire is already represented by ESTs in dbEST.

How related are the genomes?

The second factor affecting the success of a cross-referencing project is the degree to which the genomes involved are related. The comparison between *S. cerevisiae* and human genomes, which are separated by a large evolutionary distance, provides an excellent example. The human *NF1* gene is homologous to the yeast *IRA2* gene, and the biological significance of this relationship has been demonstrated because the human *NF1* cDNA can complement yeast *ira2* mutations¹¹. Approximately 15% of full-length human cDNA sequences currently present in GenBank (release 89.0) have a counterpart in yeast that are related at least as significantly (by the BLAST P-value) as the human *NF1* and yeast *IRA2* genes. Genes mutated in human diseases are related to yeast genes at about the same frequency; a set of 43 positionally-cloned human disease genes searched against a database of yeast proteins, using the BLAST algorithm, reveals that 20% of these disease genes have correlates in yeast that are related at least as significantly as the *NF1* and *IRA2* pair. Although this project was originally funded by the NCHGR to cross-reference the *S. cerevisiae* and mammalian genomes, XREFdb has recently been expanded to accept protein queries from other model organisms including, but not limited to, *C. elegans*, *Drosophila melanogaster*, *Escherichia coli*, *Mus musculus*, *Rattus norvegicus*, *Schizosaccharomyces pombe* and *Xenopus laevis*. We encourage researchers to take advantage of comparative genomics and organism cross-referencing by submitting protein queries of interest to XREFdb.

Douglas E. Bassett Jr.[¶],
bassett@ncbi.nlm.nih.gov
Mark S. Boguski[¶],
boguski@ncbi.nlm.nih.gov
Forrest Spencer[§],
forrest_spencer@qmail.bs.jhu.edu

Box 2. What are the advantages of querying XREFdb?

- **Queriers will receive assistance in identifying mammalian homologs for proteins of interest.** XREFdb automatically performs monthly searches of dbEST with protein sequences submitted by the queriers and sends updates, via email and the WWW, on any new EST matches.
- **ESTs representing potential human homologs of genes cloned in model organisms will be mapped.** Clones from which EST sequences were derived are being mapped in human and mouse by the XREF project. EST clones are currently prioritized for mapping based upon the statistical significance of *Saccharomyces cerevisiae* protein-EST sequence similarity matches. EST map data generated by the XREF project are provided as a service and made available on a non-collaborative basis.
- **Cross-references between the genetics of model organisms and mammalian phenotypes.** This effort will accelerate the establishment of these connections; ESTs that are related to proteins of known function in model organisms will be placed on disease- and phenotype-rich human and mouse maps. ESTs that map near a locus for a given disease or phenotype can be used as probes for candidate genes for the locus.

Roger Reeves[‡],
reeves@welchlink.welch.jhu.edu
Mark Goebel[¶]
goebel@biochem/iupui.edu
Philip Hieter[¶]
phil_hieter@iupui@mail.bs.jhu.edu

Department of Molecular Biology and Genetics[‡], *Department of Physiology*[¶],
Center for Medical Genetics[§], *The Johns Hopkins University, School of Medicine, 725 N. Wolfe Street, Baltimore, MD 21205-2185, USA.*

The National Center for Biotechnology Information[¶], *National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA.*

Indiana University School of Medicine, Indiana University^{**} *35 Barnhill Drive, Indianapolis, IN 46202, USA.*

References

- 1 Boguski, M.S., Tolstoshev, C.M.

and Bassett, D.E. (1994) *Science* 265, 1993-1994

2 Tugendreich, S., Boguski, M.S., Seldin, M.S. and Hieter, P. (1993) *Proc. Natl. Acad. Sci. USA* 90, 10031-10035

3 Fishel, R. et al. (1993) *Cell* 75, 1027-1038

4 Papadopoulos, N. et al. (1994) *Science* 263, 1625-1629

5 Bronner C.E. et al. (1994) *Nature* 368, 258-261

6 Savitsky, K. et al. (1995) *Science* 268, 1749-1753

7 Boguski, M.S. (1995) *Trends Biochem. Sci.* 20, 295-296

8 Altschul, S.F., Boguski, M.S., Gish, W. and Wootton, J.C. (1994) *Nature Genet.* 6, 119-129

9 Tugendreich, S. et al. (1994) *Hum. Mol. Genet.* 3, 1509-1517

10 Sherrington, R. et al. (1995) *Nature* 375, 754-760

11 Ballester, R. et al. (1990) *Cell* 63, 851-859

Mendel

The Commission on Plant Gene Nomenclature (CPGN) is developing a common nomenclature for sequenced plant genes. Mendel, the CPGN database, is now accessible via the WWW at:

<http://probe.nalusda.gov:8300/cgi-bin/browse/mendel>.

To date, 412 plant-wide gene names have been approved by the CPGN. Work is continuing by authorities in specific areas of research on approximately 1000 additional families of genes, including housekeeping genes and genes encoding enzymes of secondary plant metabolism.

Contributed by Ellen M. Reardon, Waksman Institute, Rutgers University, Piscataway, NJ 08855-0759, (reardon@mbcl.rutgers.edu)