PublisherInfo		
PublisherName	:	BioMed Central
PublisherLocation	:	London
PublisherImprintName	:	BioMed Central

Transcript mining

ArticleInfo		
ArticleID	:	3604
ArticleDOI	:	10.1186/gb-2000-1-1-reports230
ArticleCitationID	:	reports230
ArticleSequenceNumber	:	95
ArticleCategory	:	Web report
ArticleFirstPage	:	1
ArticleLastPage	:	4
ArticleHistory	:	RegistrationDate : 1999–12–14 Received : 1999–12–14 OnlineDate : 2000–3–17
ArticleCopyright	:	BioMed Central Ltd2000
ArticleGrants	:	
ArticleContext	:	130591111

Colin Semple

Abstract

UniGene, a gene indexing database, is at present the most substantial repository of transcript information from human, rat, mouse and zebrafish.

Content

UniGene, a gene indexing database, is at present the most substantial repository of transcript information from human, rat, mouse and zebrafish. Expressed sequence tags (ESTs) and annotated mRNA sequences from GenBank are automatically partitioned into a non-redundant set of clusters, each of which represents unique genes. Sequences are clustered together when they share a statistically significant overlap, or when they originate from different sequencing reads of the same cDNA clone. Because 5' and 3' reads from the same cDNA clone do not always overlap and clusters may contain splicing variants (different transcripts from the same gene that share exons), no attempt is made to produce contigs or consensus sequences for UniGene clusters. Expressed pseudogenes are also present in the database. There are abundant cross-references to other resources at the National Center for Bioinformatics (NCBI), such as mapping information in the Online Mendelian Inheritance in Man (OMIM) and GeneMap'98 databases, sequence and expression data derived from GenBank via the Entrez retrieval system, and literature referenced in the PubMed database. Separate divisions for each organism covered can be searched by GenBank accession numbers for EST or mRNA sequences and by UniGene cluster accession numbers. Statistics are available for each UniGene build, detailing the numbers of sequences used and the total number of clusters produced. When first launched on the web in August 1996, build #1 of the human UniGene database contained 48,000 clusters. At the time of writing this report, build #103 (released 3 December 1999) contains 1,386,458 sequences that are present in the form of 92,497 clusters.

Navigation

Like most NCBI resources, the site is well documented and designed. Data transfer is rapid, particularly before North America wakes up. It is possible to bookmark particular cluster pages, but given the frequent updates (which can involve 'retiring' and reassigning clusters) links may not remain stable indefinitely. Each cluster page has a clickable button for FTP transfer of sequence data to your account.

Reporter's comments

Timeliness

UniGene is automatically rebuilt every week with new EST sequences and bi-monthly with annotated mRNA sequences.

Best feature

The strength of UniGene lies in the extensive documentation of clusters. It is also expected to be the most comprehensive of the gene-indexing databases; because of its relatively non-stringent criteria for clustering, genes should be present only as one cluster.

Worst feature

The disadvantage of the permissive clustering in UniGene is that clusters often contain undefined splice variants and, occasionally, chimeric clusters are produced. Chimeric clusters are a consequence of sequencing from chimeric clones - artifactual cDNAs that contain sequences from two different genes.

Wish list

It would be helpful to be able to search UniGene sequence clusters with BLASTn or FASTA. Presently, a file from each UniGene build is produced that contains the longest sequence from each cluster. This is available by FTP from UniGene and can be searched locally using your preferred sequence-similarity-search algorithm.

Related websites

Three other EST cluster databases use more stringent criteria for clustering sequences and this allows them to produce searchable consensus sequences for clusters. These are: the Gene Index databases maintained by the The Institute of Genomic Research (TIGR); EuroGeneIndexes, maintained by the European Bioinformatics Institute (EBI); and STACK, maintained by the South African National

Bioinformatics Institute (SANBI). STACK also seeks to produce tissue-specific clusters where there are sufficient data. Internet-accessible resources in this area can also be found through National Center for Bioinformatics.

Table of links

UniGene

FTP from UniGene

The Institute of Genomic Research

EuroGeneIndexes

STACK

National Center for Bioinformatics

Online Mendelian Inheritance in Man

GeneMap'98

GenBank

European Bioinformatics Institute

South African National Bioinformatics Institute

References

1. UniGene.