Comparison of gene indexing databases

John Bouck
jbouck@bcm.tmc.edu

Wei Yu
yuw@bcm.tmc.edu

Richard Gibbs
agibbs@bcm.tmc.edu

Kim Worley
kworley@bcm.tmc.edu

Department of Molecular and Human Genetics, Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX 77030 USA.

A significant fraction of the coding regions of the human genome have been identified through the use of expressed sequence tags (ESTs) and complementary DNA sequences (cDNAs)[1,2]. ESTs are highly redundant and both EST and cDNA clones seldom encode the entire gene. The databases therefore contain a large amount of data but identification of complete genes is difficult. This difficulty is compounded because the data representing these genes are from different sources, have inconsistent annotation, and are often of low sequence quality[3]. The problem remains to sift through the existing data and identify the genes that are represented. Equally important is the issue of representing the information from many sources in a rich context, while allowing easy access to pertinent information.

Several public and private organizations have addressed this problem of representing a unique set of all human genes[2,4–6]. The purpose of this article is to examine the merits and applications of three of the publicly available tools that group sequences into clusters. The three datasets examined are Unigene at the National Center for Biotechnology Information (http://www.ncbi.nlm.nih.gov/UniGene/index.html)[6], Sequence Tag Alignment and Consensus Knowledgebase (STACK) at the South African National Bioinformatics Institute (http://ziggy.sanbi.ac.za/stack/stacksearch.htm)[4], and the Human Gene Index (HGI) at the Institute for Genome Research (http://www.tigr.org/tdb/hgi/hgi.html)[2]. All of these systems attempt to obtain as complete a list as possible of all genes. However, the approaches and practical uses of each are very different. This article is divided into two parts; first we describe the parameters used to cluster sequences into gene families and discuss the practical uses for each system, next we test the utility of each database by comparison with 20 recently sequenced cDNAs.

Clustering parameters

The data and parameters used to group sequences greatly influences the usefulness and completeness of the resulting database. The three datasets examined here each have different data sources and different algorithms to group sequences (Table 1). There are a number of different types and sources of data that can be used to create a gene index. Genomic DNA sequence is generally very accurate, but the exons of a gene can be dispersed, complicating the identification of transcripts. Although cDNA sequence is a richer source of transcript information, cDNAs are often incomplete and seldom extend to the 5′ end. Expressed sequence tags (ESTs) are single-pass sequencing reads generated from either the 5′ or 3′ end of a cDNA clone. These sequences are short (less than 500 bases) and were initially intended to simply identify a transcript. Ideally one would like to have a complete cDNA sequence that entirely covers a gene. Due to the scarcity of full length cDNA sequences and the abundance of ESTs, several groups have used these sequences to assemble electronic transcripts.

The relative abundance of the various types of data varies greatly. The amount of genomic DNA identified as coding regions is limited as is the number of full length cDNA insert sequences. The most abundant source of gene information is the EST data. The ESTs have both advantages and disadvantages; there are large numbers of EST sequences derived from different tissues and different disease states, but because ESTs represent a single sequencing read, the probability of sequencing errors is high.

Unigene utilizes ESTs and properly annotated mRNA sequences that are derived from the dbEST database and GenBank[7,8]. Efforts are made to prevent the incorporation of very low quality ESTs and non-gene sequences. Unigene is frequently built and the reader is referred to the Unigene web site for details on the parameters used for the current build. The sequences are compared with each other and all sequences that have a statistically significant overlap are placed into a single group. Sequences are also clustered if the sequencing reads were performed on the same clone. For instance, a particular cDNA clone might have been sequenced at each end, but the resulting EST reads might not overlap. In this case, the sequences are grouped according to the template that was used. As indicated in Table 1, no consensus sequence is determined for each cluster. Due to the non-stringent parameters used to cluster sequences, there is expected to be only a single group that represents each gene.

The STACK data set utilizes EST data obtained from dbEST (but not cDNA sequence data) and categorizes them according to the tissue from which they were derived[4]. The clustering algorithm identifies ESTs that are highly similar (>95% identical over 150 bases) and from the same tissue. These sequences are then aligned to provide a consensus sequence. Similar to Unigene, reads are grouped together if they are derived from the same cDNA clone. The relatively large number of groups in STACK (Table 1) may be a result of dividing nearly identical ESTs from different tissues into separate groups. The presence of the same mRNA in different tissues guarantees that the gene will be contained in multiple groups. Because certain tissues have had many more ESTs sequenced from them than others, the frequency of representation of a transcript in different tissues does not necessarily represent the correct expression levels.

The Human Gene Index (HGI) uses the EST data from dbEST and proprietary sources as well as GenBank transcript data to group sequences[2]. The goal of this database is to find protein-coding regions and a consensus sequence for each putative protein is generated (named tentative human consensus or THC). Sequences are initially grouped together if they contain at least 40 bases with greater than 95% identity. The groups are then assembled to generate a consensus sequence. During assembly, discordant sequences are identified and removed.

Problem case handling

The alternative clustering systems result in different handling of problem cases. Many difficult issues arise when sequences found in the national databases are clustered. Problems arise with incorrectly annotated sequences or expressed sequences containing introns[9]. Some of the discrepancies in the databases are avoided through the use of sequence comparisons for clustering. All of the databases discussed here rely on redundancy to help eliminate problem sequences. Undoubtedly, some inappropriate sequences appear in all of the databases which will not be detectable until much more genomic and full-length transcript information is obtained. Here we will discuss two problem cases: chimeric clones and alternatively spliced transcripts.

Occasionally, cDNAs are cloned which contain sequences from two different transcripts. These cloning artifacts can cause serious problems for clustering databases and each database has its own system for dealing with these chimeras.

0168-9525/99/$ – see front matter PII: S0168-9525(99)01709-6

TABLE 1. Characteristics of gene indexing databases[a]

	Unigene	STACK	HGI
DNA Searchable	No	Yes	Yes
Consensus sequences	No	Yes	Yes
Alignment	No	Yes	Yes
Redundancy of groups	Low	High	High
Sequences used	730 296	552 013	615 482
Number of groups	46 569	87 905	72 333
Alternative splices	Same clusters	Different clusters	Different clusters
Build/version[a]	48	1	3.3

[a]As of September 1998.

Unigene tries to identify chimerics by searching for a single clone that joins two otherwise unconnected groups. Because both STACK and HGI use alignments in their systems, chimeras are more easily identified during the alignment process.

A second exception for the clustering programs is the problem of alternatively spliced transcripts. Because two alternatively spliced transcripts contain identical regions as well as disparate regions, their representation can be difficult. Different isoforms of a transcript are generally expressed in different tissues or at different stages of development. Because STACK separates transcripts by their origins, alternative transcripts should be significantly limited within a group, and will be represented by different groups. The fairly relaxed criteria used for clustering Unigene results in groups containing transcripts with a small number of exons in common. HGI relies on its alignment program to separate alternative transcripts into different groups. Both Unigene and HGI need to distinguish between chimeric clones and alternatively spliced transcripts because both of these instances raise problems for clustering.

Alternatively spliced transcripts

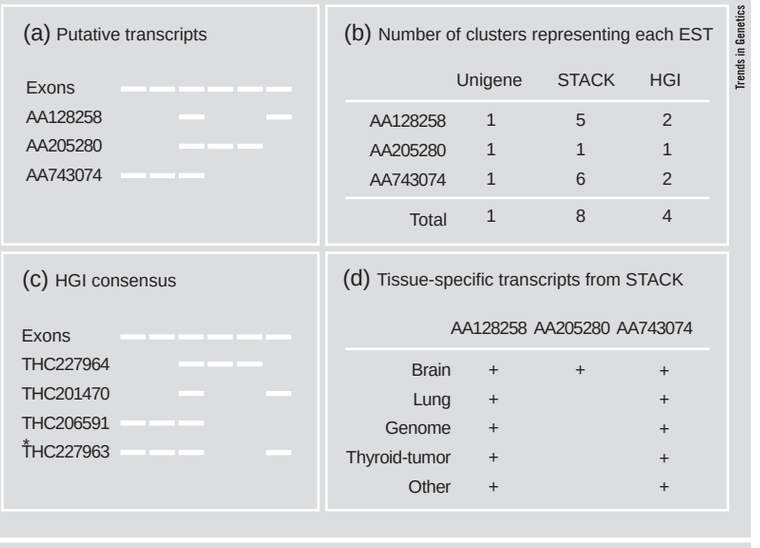
Alternative transcripts are a potential source of important biological information, and therefore we examined the mechanisms for handling a number of alternatively processed sequences. Figure 1 displays the results from one set of alternative transcripts related to a genomic clone sequenced at Baylor College of Medicine, Human Genome Sequencing Center (BCM-HGSC) (AC004106). This sequence contained EST matches from independent but overlapping transcripts indicating that alternative splicing was occurring. The putative alternative transcripts indicated by three representative ESTs are indicated in Fig. 1a.

Three parameters were examined for each transcript and each database. First, the number of clusters identified by each EST was determined (Fig. 1b). For Unigene, all three of the ESTs were contained in the same cluster. STACK produced multiple matches for two of the ESTs and a total of eight different clusters were required to describe the three putative transcripts. Four clusters were identified by comparison to the HGI database.

Alternative transcripts are often expressed in different tissue. STACK is designed to identify different tissue types that contain a common gene and so we examined the distribution of tissues that the clusters represent. Figure 1d shows the tissue distribution of the clusters, we note that the EST AA743074 matched two clusters derived from the same tissue (brain). The data suggest that the transcript represented by the EST AA205280 is expressed solely in the brain while the other two putative transcripts are expressed in a wider variety of tissues. This type of analysis is not as straightforward using the other collections. The consensus sequences from HGI were also compared to the exons to determine if the clusters accurately represent the putative transcripts (Fig. 1c). Three of the four clusters faithfully represent the putative transcripts. The fourth, indicated by an asterisk, contains exon 1 in the reverse orientation to exons 3, 4 and 5 indicating an incorrect assembly of this cluster.

The brief analysis here highlights the benefits of the three different datasets for examining alternatively spliced transcripts. Although much of the information can be extracted from any one of the databases, the data are more easily accessed in the different databases. Unigene maintains all of the information in a single cluster, STACK illustrates the different tissues expressing the different transcripts and the HGIs most closely describe the individual sequences.

FIGURE 1. Representations of an alternatively spliced gene in the different databases



(a) The genomic clone AC004106 contains putative alternative transcripts that are minimally represented by three EST sequences, AA128258, AA205280 and AA743074. (b) The number of clusters recognized by each EST is indicated. For STACK and HGI, clusters were only counted if the expectation value was less than 1×10^{-11} or 1×10^{-100} respectively (different *P*-values are generated by the different blast programs employed by the databases). Total indicates the number of unique clusters identified by all three ESTs. (c) Exon organization in the HGI clusters. The asterisk indicates that the last exon is present in an opposing orientation to the other exons. (d) The different tissues that were identified by the individual ESTs are listed. Genome indicates a transcript identified by genomic sequencing and other is a category defined by STACK. The data from this analysis indicates that exons 4 and 5 may represent brain specific expression while the other exons have a broader expression profile.

Clustering of naïve cDNA sequences

To further illustrate the differences between the clustering programs, we examined how well recently submitted cDNA sequences are represented in the three databases. The cDNAs were sequenced at BCM-HGSC and were obtained from the Soares 1NIB infant brain cDNA library[10]. The clones to be sequenced were

randomly selected from the library and sequencing was performed via the concatenation cDNA sequencing method[11,13]. The cDNAs were compared to each database prior to their release from GenBank. A total of 20 cDNAs were examined with an average insert size of 1.4 kb, Table 2.

Unigene does not provide a consensus sequence nor is it possible to search the contents directly using sequence similarity searches. Unigene does provide a compilation of the longest sequence from each cluster that can be retrieved and searched locally. A sequence may fail to match this dataset and yet be contained within Unigene because the longest sequence does not necessarily represent the entire gene. Therefore, the clones that displayed poor matches were compared to the non-redundant sequence database and dbEST using Blast-2 (Ref. 12). If matches were observed (Blast-2 score $\geqslant$ 100), the matched sequences were located in Unigene. Through this analysis, one cDNA produced no significant match to any Unigene cluster, although both a 5′ and 3′ EST read were identified in the EST database. As an estimate of the abundance of similar sequences in GenBank, the number of sequences that were found in each Unigene cluster was determined (Table 2, column 3). Several of the cDNAs matched groups that contained many sequences while others matched groups that contained only a few.

The STACK database was also queried with each of the cDNA sequences. Eighteen sequences produced strong matches to clusters (Blast-1.4.9 $P > 1\times10^{-11}$) and two sequences had no significant matches. The number of sequences that comprise the cluster with the strongest match is listed in Table 2, Column 4 and the length of the consensus sequence generated is shown in Column 6. Lastly, the cDNA sequences were compared to the HGI database. All of the sequences produced significant matches (Blast-2 $P < 1\times10^{-100}$). The length of the consensus sequence of the cluster with the best match is indicated in the last column of Table 2.

Although the majority of the data in these databases is derived from the EST database, there is significant variation between the datasets. For instance, the cDNA AF070615 is well represented in the Unigene database with 78 clones but the corresponding entries in STACK and HGI have only one and six clones respectively. Conversely, the cDNA AF070620 did not have a significant match to any Unigene cluster but demonstrated strong matches to both STACK and HGI. The size of the consensus sequences also varies between STACK and HGI, although this may be due to the different goals of these two systems. In just over half of the occurrences examined, the STACK consensus is longer than the HGI consensus.

TABLE 2. Database comparisons using naïve cDNAs

cDNA name	cDNA length	Unigene size of group[b]	STACK size of group[b]	HGI size of group[b]	STACK consensus length	HGI consensus length
AF070614	1753	22	18	12	2297	897
AF070615	1501	78	1	6	373	896
AF070616	1148	24	8	16	1004	1533
AF070617	1326	25	13	8	2400	1106
AF070618[a]	1211	22	5	7	878	1001
AF070619[a]	1006	7	6	2	830	448
AF070620[a]	1608	0	2	3	802	897
AF070621[a]	1694	5	2	11	702	715
AF070622	1115	78	15	18	1431	941
AF070623	1249	12	64	12	3970	730
AF070624	1052	5	0	2	710	303
AF070625	1728	2	12	5	938	466
AF070626	1193	82	19	52	2272	1920
AF070627	1320	74	11	39	1729	5270
AF070628	1487	4	0	2	350	443
AF070629	1487	128	39	18	2667	1292
AF070630	849	137	6	54	883	1668
AF070631	1477	43	6	5	942	2701
AF070632	1446	15	7	4	1467	882
AF070633	1481	8	2	7	887	1293

[a]Indicates the presence of an Alu repetitive element.
[b]Data derived from group with the strongest match to the sequence.

We also note that 9 out of 20 of the cDNAs we sequenced were longer than both the STACK and HGI entries, confirming the usefulness of full length cDNA insert sequencing[13].

In addition to the lengths of the consensus sequences, we also examined the composition of the STACK and HGI consensus sequences. As expected, when a consensus has been derived from a previously sequenced cDNA insert, there is extremely good concordance between the consensus and the test sequence with very few base differences. In contrast, when a cluster is composed solely of EST data the consensus sequence contains many insertions and deletions, and is often aligned such that less than half of the test sequence is matched. This result demonstrates the need for high quality cDNA sequencing despite the presence of abundant EST data.

The comparison of the cDNAs described here demonstrates the different uses of the three gene indexing databases. The Unigene groups contain more sequences than the other groups, but some genes are not represented. The STACK database is often successful in stitching together ESTs to produce a long consensus sequence and highlights variable tissue expression, although the sole use of ESTs ignores the abundant mRNA data available. The HGI database produces strong matches to more of the test sequences, perhaps due to the larger EST set utilized. It should be noted that the sequences that did not match to any Unigene or STACK group do have strong matches to ESTs in the public database. This indicates that those ESTs were excluded by the clustering algorithms.

Conclusions

We have described three databases that are currently available to examine genes. Each has strengths and weaknesses. The decision of which database to use depends on the type of information one wishes to obtain. The strength of the Unigene database lies in the non-stringent clustering parameters that allow alternatively spliced transcripts to be incorporated into the same cluster, thereby simplifying their identification. The benefit of STACK comes from the segregation of sequences based on tissue types whereby different tissues expressing different isoforms can be readily identified. Lastly, the HGI incorporates a larger set of ESTs as well as other transcript data to produce a consensus sequence and may therefore be more complete. As we have demonstrated here, not all of the clustering databases contain the same information and furthermore, the consensus sequences generated often differ. For these reasons it is recommended that the database user does not rely upon a single dataset when searching for gene information.

Acknowledgements

We thank W. Hide, O. Litcharge, and I.E. Holt for critical review and helpful comments. J.B., W.Y. and R.G. are supported by the National Human Genome Research Institute (HG01459) and the National Cancer Institute (R01 CA80200-01) at the National Institutes of Health, and the Human Genome Project at the Department of Energy (DE-FG03-97FR62375). K.W. is supported by the Human Genome Project at the Department of Energy (DE-FG03-95ER62097).

References
1 Collins, F. and Galas, D. (1993) A new five-year plan for the U.S. Human Genome Project. *Science* 262, 43–46
2 Adams, M.D. *et al.* (1995) Initial assessment of human gene diversity and expression patterns based upon 83 million nucleotides of cDNA sequence. *Nature* 377, 3–174
3 Aaronson, J.S. *et al.* (1996) Toward the development of a gene index to the human genome: an assessment of the nature of high-throughput EST sequence data. *Genome Res.* 6, 829–845
4 Burke, J. *et al.* (1998) Alternative gene form discovery and candidate gene selection from gene indexing projects. *Genome Res.* 8, 276–290
5 Houlgatte, R. *et al.* (1995) The Genexpress Index: a resource for gene discovery and the genic map of the human genome. *Genome Res.* 5, 272–304
6 Schuler, G.D. (1997) Pieces of the puzzle: expressed sequence tags and the catalog of human genes. *J. Mol. Med.* 75, 694–688
7 Boguski, M.S. *et al.* (1993) dbEST – database for 'expressed sequence tags'. *Nat. Genet.* 4, 332–333
8 Benson, D.A. *et al.* (1996) GenBank. *Nucleic Acids Res.* 24, 1–5
9 Wolfsberg, T.G. and Landsman, D. (1997) A comparison of expressed sequence tags (ESTs) to human genomic sequences. *Nucleic Acids Res.* 25, 1626–1632
10 Soares, M.B. *et al.* (1994) Construction and characterization of a normalized cDNA library. *Proc. Natl. Acad. Sci. U. S. A.* 91, 9228–9232
11 Andersson, B. *et al.* (1997) Simultaneous shotgun sequencing of multiple cDNA clones. *DNA Seq.* 7, 63–70
12 Altschul, S.F. *et al.* (1990) Basic local alignment search tool. *J. Mol. Biol.* 215, 403–410
13 Yu, W. *et al.* (1997) Large-scale concatenation cDNA sequencing. *Genome Res.* 7, 353–358

User-friendly genome guide

The Human Genome: A User's Guide
by R. Scott Hawley and Catherine A. Mori
Academic Press, 1999. $39.95 pbk (xix + 415 pages) ISBN 0 12 333460 8

Norman A. Johnson
njohnson@ent.umass.edu

Dept of Entomology and Graduate Program in Organismic and Evolutionary Biology, University of Massachusetts, Amherst, MA 01003, USA.

It is difficult enough for most geneticists to keep up with the fast and accelerating pace of breakthroughs in human genetics. It is even more difficult for lay people. Yet these biomedical findings have a great impact on public policy and even our daily lives. How, then, can we better prepare students, especially those not majoring in biology, to be able to assimilate and evaluate the information the media provides about biomedical discoveries?

Scott Hawley and Catherine Mori (respectively, a *Drosophila* cytogeneticist and a health education writer), appear to have thought deeply about this issue. The present book, an expansion of Hawley's human genetics lecture notes, is an excellent guide to the advances made in human genetics and their relevance to daily life. It is both reasonably accessible to general audiences and unlikely to cause specialist readers to cringe. The engaging, informal style of this book is most apparent in its unusual use of asides, set off in italics. In these asides, the authors 'chat' with each other and the reader – explaining why they became interested in a particular topic, their personal opinions about controversies, and the joys and the drudgeries of life in a research lab. In the last half of the book, Julia Richards (a human geneticist) contributes to the dialog. Far from being distracting, the asides are an unexpected bonus.

The book begins with a series of chapters that introduce how heredity works. Chapter 5, which tackles the fundamental but pedagogically challenging topic of the relationship between genotype and phenotype, is particularly good. After first discussing different types of pedigree patterns (X-linked versus autosomal, dominant versus recessive), the authors make distinctions between two types of mutational effects: those that disrupt the normal function of a gene ('monkey wrench') versus those that result in the loss of function. Hawley and Mori point out that while most monkey-wrench mutations appear dominant and most loss-of-function mutations appear recessive, the correlation is not perfect. Loss-of-function mutations can be dominant if half the normal amount of gene product is insufficient for the normal