Differential Gene Expression in the Auditory System

Irene S. Gabashvili[1], Richard J. Carter[1], Peter Markstein[1], and Anne B.S. Giersch[2]

[1] Hewlett-Packard Labs, Computational Biosciences Research, 1501 Page Mill Road, Palo Alto, CA, 94304, USA
{Irene.Gabashvili, Dick.Carter, Peter.Markstein}@HP.com
http://hpl.hp.com/research/cbsr
[2] Department of Pathology, BWH, Harvard Medical School, 75 Francis Street, 02115 Boston, USA
agiersch@rics.bwh.harvard.edu
http://hearing.bwh.harvard.edu/

Abstract. Hearing disorders affect over 10% of the population and this ratio is dramatically increasing with age. Development of appropriate therapeutic approaches requires understanding of the auditory system, which remains largely incomplete. We have identified hearing-specific genes and pathways by mapping over 15000 cochlear expressed sequence tags (ESTs) to the human genome (NCBI Build 35) and comparing it to other EST clusters (Unigene Build 183). A number of novel potentially cochlear-specific genes discovered in this work are currently being verified by experimental studies. The software tool developed for this task is based on a fast bidirectional multiple pattern search algorithm. Patterns used for scoring and selection of loci include EST subsequences, cloning-process identifiers, and genomic and external contamination determinants. Comparison of our results with other programs and available annotations shows that the software developed provides potentially the fastest, yet reliable mapping of ESTs.

1 Introduction

Personalized medicine in the future will be based on the comparison of individual genetic information to reference gene expression, molecular interactions and pathways in tissues and organs, in health and disease. It will be based on advanced genome sequencing, gene expression, proteomic and metabolomic technologies, as well as efficient computational tools for mapping of genes and pathways.

The reliability of computational approaches and models is improving, as "omic" technologies mature and the accuracy of predictions grows with increasing data input. There is a growing need for fast software tools capable of handling massive amounts of data and reanalyzing the data to discover integrated knowledge and identify broken links and wrong connections between intricate processes in individual datasets.

The first step in comparing genomic information is to align DNA sequences, that is, to map nucleotides of expressed sequence tags (ESTs) or full cDNAs to the genome and sequences of known and predicted genes. Sequence alignment is one of

J.C. Setubal and S. Verjovski-Almeida (Eds.): BSB 2005, LNBI 3594, pp. 1–8, 2005.

the oldest and most successful applications of Computer Science to Biology [1-2]. Many local pairwise alignment methods exist [1-6] and most software tools are freely available. These tools, however, are customized for specific tasks and do not allow enough flexibility for new specialized tasks to external users. The most popular generic programs relevant to EST mapping, BLAST from the National Center for Biotechnology Information [6] and BLAT from U.C. Santa Cruz [4], each have their strengths and weaknesses. The BLAST service offered by NCBI is too slow to use for sets of tens of thousands ESTs. Moreover, it does not handle intron gaps well when used for the whole-genome mappings and works best on expressed sequence databases. The BLAT service offered by UCSC is fast, but its interactive nature and 25-sequence submission limit would prevent its use on a large number of sequences.

To direct and control the process of EST mapping, we needed software with problem-specific intelligence that was not available with existing tools. One of the most important tasks in processing experimental data is estimating the errors and potential sources of errors in measurements [7]. Cloning and sequencing artifacts, for example, could be eliminated using pre-screening procedures. Accordingly, we needed not only to align ESTs, but also check for a number of favorable and detrimental signals, to identify the most likely mapping amongst many possibilities.

In this work, we have analyzed over fifteen thousand ESTs expressed in the human cochlea. The cochlea is one of the smallest organs in the body located in the inner ear and responsible for auditory transduction (conversion of sound into the language of the brain). Hearing impairment is always the result of damage to either the middle ear, the cochlea or its associated auditory nerve. Over one hundred genes responsible for deafness have been discovered, but many more candidates apparently exist. A much smaller fraction of molecular-level auditory pathways have been identified [8-10], mostly due to the lack of knowledge of human biology in general.

We have mapped and analyzed genes predominantly expressed in the inner ear and their pathways. We have also studied cochlear genes expressed in low numbers. We show that the vast majority of cochlea-unique genes identified by existing tools and servers are either genomic contaminations or can be also found in other tissues. We have selected a small subset of cochlea-specific genes and they are currently being verified by independent experimental methods.

2 Computational Approach

To speed up alignment of ESTs to the genome and improve the scoring of such mappings, we reduced the problem to that of simultaneous exact matching of multiple motifs within ESTs to localized genome regions. Our approach is illustrated on the example of a particular Morton cochlear EST (Fig. 1).

Mapping and selection of ESTs is realized by dynamic interaction of two in-house programs, *Enhancer2* and *BatchSearch*. *Enhancer2* is a 5000-line C++ program that finds exact matches of a number of input search patterns within a database of sequences (whole genomes, mRNAs, etc). The fast exact string prefix matching algorithm (Dick Carter and Peter Markstein, to be published) was applied to other genome search problems in early stages of its development [11]. Some of the features

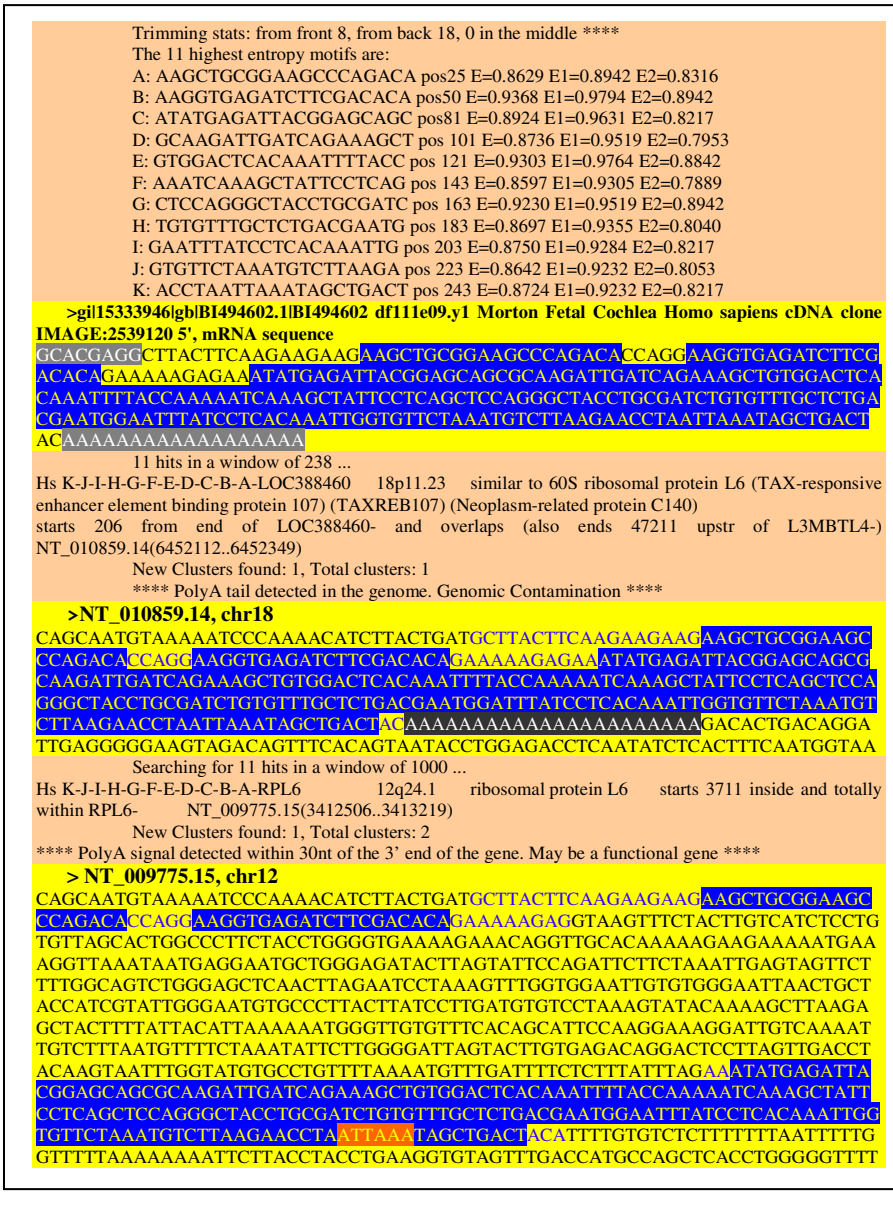

Fig. 1. Our approach to mapping and scoring of results illustrated on the example of a sequence with accession number BI49460. As a first step, we determined detrimental motifs in this sequence (shaded in grey) and trimmed them off. Blue area represents dynamically selected subsequences used for matching to the human genome. The program found two equally well matching regions in chromosomes 12 and 18. A detrimental signal (polyA tail (black shading), in chromosome 18 and a favorable motif in chromosome 12 determined the best mapping. See text for details

of this algorithm are its ability to handle all IUPAC nucleotide codes with little additional overhead and its high parallelization efficiency.

The other component of our EST-mapping solution is *BatchSearch*, a 2500-line C++ program that interacts with *Enhancer2* by giving it search tasks and dynamically responding to its output. Using the fast exact-matching *Enhancer2* speeds the alignment process since EST-mapping would normally require slower inexact matching to cope with introns and frequent EST sequencing errors or single nucleotide polymorphisms (SNPs). Our idea was to divide an EST into smaller fragments and, using *Enhancer2*, find where some of them occur. Normally the bulk of the fragments would be found clustered within the same locale, thus forming the basis for the reported EST mapping. In the majority of cases, we also observed a very high level of identity, as an entire EST sequence after trimming often exactly matched to a localized region within the genome.

The logic of *BatchSearch* involves a number of steps. First, the input EST is trimmed of bases that are artifacts of the sequencing process (Fig.1). Second, a globally optimal set of high-entropy fragments is chosen from the EST using a dynamic programming algorithm. Then, the formulated exact-match search problem is passed to the waiting *Enhancer2* program. Depending on these results, *BatchSearch* can ask *Enhancer2* to refilter its search results, allowing for more widely dispersed clusters to be reported. In addition, clusters of other detrimental and favorable motifs in the genome are taken into account. Fig.1 demonstrates two such motifs – a polyA tail (*black shading*) that is supposed to be located within 30 nucleotides of the 3' end (larger distance may be allowed in the 5' EST) and a polyA signal (see [12], *orange shading*, not be followed by polyA tail in the genome) Alternatively, *BatchSearch* can redo the genome search with smaller EST subsequences, in an effort to identify the most likely mapping. One search for six 20-nucleotide fragments using *Enhancer2* takes about 2.5 seconds on a 2.8 gHz Xeon CPU with one Giga Byte of RAM. A dual-processor HP XW8000 PC workstation requires 5.5 hours to map the entire library of 15000 cochlear ESTs to the human genome. Datasets with less mapping ambiguity are processed faster.

3 Genes and Pathways of the Human Cochlea

Only from 60% to 95% of all deposited ESTs in tissue- and organ-specific libraries are classified by Unigene. Fig.2 demonstrates the ratio of classified vs. unclassified sequences for fetal cochlear, eyes and brain libraries and adult bone and stomach datasets. Only 11,913 human cochlear sequences out of fifteen thousand deposited (dbEST Library ID.371 [13,14]) are annotated in Unigene. We mapped over 98% (all but 276 – area 3 in inset of Fig.2 showing sequences not available in Unigene) of the ESTs in the Morton fetal cochlear library to specific regions in the human genome and genomes of laboratory organisms. Of the unmapped sequences, most correspond to highly conserved regions that can be exactly matched to dozens of proteins in a variety of organisms. The remaining unmapped ESTs seem to be formed by nonspecific recombination events and cannot be confidently attributed to a specific

gene or genome. Non-human contaminations in the dataset (259, area 4 in Fig.2) come from laboratory organisms – mainly yeast, E.coli, phages and cloning vectors, but there are also single occurences of such unexpected species as worm and mouse. Among about five thousand genes identified, almost 2000 genes are represented by single ESTs. Less than 200 genes are supported by ten or more sequences. The most abundant mRNAs were for extracellular matrix genes. This can be explained by the importance of structural support in cochlea. We note that this class of proteins acounts for almost half of nonsyndromic deafness genes.

Less than 10% of all our cochlea sequences were deposited with gene-relevant information in their headers, while 41% of the sequences were annotated based on results of BLAST searches against GenBank databases in early 2000s. Almost 80% from this set are annotated in the latest build of Unigene, although about 8% of these annotations remain hypothetical. We selected many different isoforms among ESTs clustered in the same Unigene clusters. In addition to the 4058 Unigene clusters, we determined almost 1000 additional loci, many of which might represent novel genes or isoforms of known genes (areas 1 and 4 in Fig.2). We found about 20% potential genomic contaminations in the dataset and 1% of sequence flips in EST sequences. Many transcripts corresponding to ESTs present in the dataset might not be expressed as proteins, but instead are degraded by nonsense-mediated mRNA decay or other cell surveillance mechanisms. We revealed a number of incomplete, truncated mRNAs in the library, confirming this possibility.

The inset of Figure 2 shows how sequences extracted from the fetal inner ear and not classified by Unigene are mapped to the human genome and genomes of other species (human pathogens and laboratory organisms). Comparison of our mappings to alignments produced by popular tools, such as BLAST [6] and BLAT [4], shows that our solutions are essentially the same. These other tools, however, offer the best solutions among several other top scoring results, thus requiring post-processing of results, often manually. We note that most of our novel genes are also suggested in the AceView database [15] and are being incorporated into the next build of the human genome. On the one hand, we consider it as another confirmation of the reliability of our findings. On the other hand, we note that the subject of this work is analysis of hearing-specific genes and this was not done by the authors of AceView, GeneScan and other global gene-finding programs.

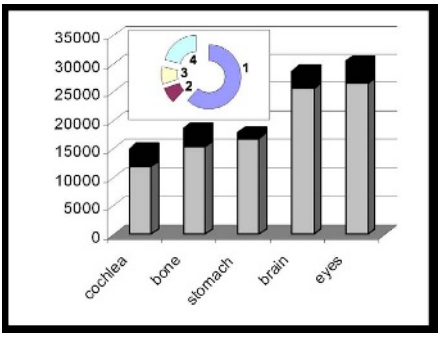

Fig. 2. A bar-chart of sequences of organ-specific libraries classified (white base) and not classified (black top) into Unigene entries. Inset shows our mappings of non-classified cochlear ESTs. Sequences in areas: (1) may be novel isoforms of known genes; (2) are non-human genes; (3) are ambiguous; 4) map to unannotated regions in the human genome

Table 1. The most highly expressed genes and predominant pathways of the human cochlea

Name	EST count, Uni-gene	EST count, this work	PATHWAYS		
			Ion Transport	Cell Shape Maintenance	Housekeeping
Collagen, type I, alpha 2	314	343		Collagen matrix	
Collagen, type III, alpha 1	153	159		Collagen matrix	
Secreted protein, acidic, cysteine-rich(osteonectin)	125	162		Binds Collagen	
Eukaryotic translation elongation factor 1 alpha 1	81	130		Binds Actin	Protein Synthesis
Vimentin	80	84		Intermediate Filament	Structure and Motiilty
Collagen, type I, alpha 1	70	83		Collagen matrix	
Myristoylated alanine-rich protein kinase C substrate	63	67		Binds Actin	Structure and Motiilty
KIAA1040 protein	55	56	Proton transport		
Tumor protein, translationally controlled	51	53		Extracellular matrix	Structure and Motiilty
Chromosome 5 open reading frame 13	50	56		Cell junctions	
Actin, beta	46	54		Actin filament	Structure and Motiilty
Potassium channel tetramerisation domain containing 12	44	49	Potassium transport		
Actin, gamma 1	42	47		Actin Filament	Structure and Motiilty
Ribosomal protein S20	38	39			Protein Synthesis
Cyclin I	36	37			Cell Cycle Regulation

Comparison of our results to available Unigene assignments shows a very good correspondence as well. Our "new gene" mappings often correspond to "transcribed loci" and most discrepancies in gene names are solely due to different naming of the same genes. For example, "ecotropic viral integration site 2A" is the same gene as "neurofibromin 1 (neurofibromatosis, von Recklinghausen disease, Watson disease)", and ALEX2 is the same as ARMCX2. Less than 1% of our EST mappings do not correspond to Unigene assignments. In half of these cases our results might be better. In several cases old Unigene assignments seem to be better than the latest ones.

In order to normalize the cochlear library to find crucial components of hearing transduction, all housekeeping and cell structure maintenance genes have to be subtracted from the set. This task is not trivial, as many proteins have multiple functions and the difference between cochlear and other existing libraries is statistically significant only for a very small number of relatively highly expressed genes. These are collagens (col1a2, col3a1) and osteonectin (if compared to fetal brain, structural tissues or whole embryo). Comparison with libraries from other tissues points additionally to several other candidates. For example, a protein potentially involved in the assembly of potassium channels is known to be implied in the hearing process ("potassium channel tetramerisation domain containing 12"). Table 1 shows fifteen genes of the human cochlea with the highest level of expression. We note that some of the ESTs appear as genomic contaminations (data not shown) and might not be expressed in the cell. Many such sequences, however, are annotated as legitamate genes in public databases.

We identified a number of pathways including abundant transcripts of the dataset, not-directly related to hearing. They describe cell proliferation, maintenance of ion balance, protein synthesis, splicing, transcription, regulation of actin cytoskeleton, etc. The table shows that certain cellular shape maintenance pathways (extracellular junction and matrix-related) are hearing related, rather than for housekeeping (see [16-17] for lists of housekeeping genes). This can be explained by the importance of maintenance of acoustic resonator structures (on the level of cell assemblies) in the ear.

For genes present in a small number of copies, we can employ a bottom-up approach by focusing on potentially novel genes that seem to be solely or predominantly expressed in the cochlea, then reconstructing pathways involving products of these genes. We selected about 200 clusters of ESTs potentially representing novel genes not classified by Unigene. We have further narrowed this list down by filtering out genomic contaminations and highly repetitive sequences. The candidate genes include possible transcription factors (gene-regulatory pathways), a motor protein (cell shape maintenance), an isoform of collagen (cell shape maintenance) and a transmembrane protein (ion transport). The findings are currently being verified by RT-PCR and other laboratory tests.

4 Concluding Remarks

Crucial processes of life, hearing being one of them, are only partially understood at the molecular level. Important but low-abundant proteins remain elusive. Large-scale sequencing of tissue-specific genes and fast yet reliable mapping of sequences will help to identify the key components of sensory sound transduction pathways. Eventually, this will bring a cure and better treatment to now-incurable deafness and age-related hearing loss.

References

1. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ.: Basic local alignment search tool. *J. Mol. Biol*. (1990), **215**, 403-410.
2. Batzoglou S.: The many faces of sequence alignment *Brief. Bioinform.* (2005), **6**, 6-22.
3. Gemund, C., Ramu, C., Altenberg-Greulich, B., Gibson, T.J.:Gene2EST: a BLAST2 server for searching expressed sequence tag (EST) databases with eukaryotic gene-sized queries. *Nucleic Acids Res* ((2001) , **29,** 1272–1277
4. Kent, W.J.: BLAT—the BLAST-like alignment tool. *Genome Res (2002)*, **12,** 656–664
5. Krüger, J., Sczyrba, A., Kurtz, S., Giegeri, R.: e2g: an interactive web-based server for efficiently mapping large EST and cDNA sets to genomic sequences. *Nucleic Acids Res.* (2004), **32** (Web Server issue), W301-4
6. Altschul, S.F., Madden T.L., Schoffer A.A., Zhang J., Zhang Z., Miller W., Lipman, D.J.: Gapped BLAST and PSI–BLAST: a new generation of protein database search programs. *Nucleic Acids Research* **25**:3389–3402, 1997.
7. Whirl-Carrillo, M., Gabashvili, I.S., Banatao, D.R., Bada, M., Altman, R.B.: Mining biochemical information: lessons taught by the ribosome. *RNA* (2002) **8**, 279-289.
8. Chen, Z.-Y., Corey, D.P.: Understanding Inner Ear Development with Gene Expression Profiling. *Journal of Neurobiology* (2002), **53**, 276-285.
9. Lin, J., Ozeki, M., Javel, E., Zhao, Z., Pan,W., Schlentz, E., Levine, S.: Identification of gene expression profiles in rat ears with cDNA microarrays. *Hear Res*. (2003), **175**, 2-13.
10. McGuire, J.F. ., Casado, B.: Proteomics: a primer for otologists, *Otol Neurotol*. (2004), **25**, 842-849.
11. Markstein, M., Markstein, P., Markstein, V., Levine, M.S.: Genome-wide analysis of clustered Dorsal binding sites identifies putative target genes in the Drosophila embryo. *Proc Natl Acad Sci U S A* (2002), **99**, 763-768.
12. Beaudoing, E., Freier, S., Wyatt, J.R., Claverie, J.-M., Gautheret, D.: Patterns of Variant Polyadenylation Signal Usage in Human Genes, Gen.Res. (2000), 10, 1001-1010.
13. Robertson, N.G., Khetarpal, U., Gutierrez-Espelata, G.A. , Bieber, F.R., Morton, C.C.: Isolation of novel and known genes from a human fetal cochlear cDNA library using subtractive hybridization and differential screening. Genomics (1994), 23, 42-50.
14. Skvorak, A.B., Weng, Z., Yee, A.J., Robertson, N.G., Morton, C.C.: Human cochlear expressed sequence tags provide insight into cochlear gene expression and identify candidate genes for deafness", *Hum Mol Genet*. (1999), **8**, 439-452.
15. Thierry-Mieg, D., Thierry-Mieg, J-T., Potdevin, M., Sienkiewicz, M.: Identification and functional annotation of cDNA-supported genes in higher organisms using AceView, unpublished. http://www.aceview.org/
16. Hsiao, L.L., Dangond, F., Yoshida, T., Hong, R., Jensen, R.V., Misra, J., Dillon, W., Lee, K.F., Clark, K.E., Haverty, P., Weng, Z., Mutter, G.L., Frosch, M.P., Macdonald, M.E., Milford, E.L., Crum, C.P., Bueno, R., Pratt, R.E., Mahadevappa, M., Warrington, J.A., Stephanopoulos, G., Stephanopoulos, G., Gullans, S.R. A Compendium of Gene Expression in Normal Human Tissues**.** Physiol Genomics. (2001), 21, 97-104
17. Eisenberg, E., Levanon, E.Y.: Human housekeeping genes are compact. Trends Genet. (2003), 19, 362-365.