# A NOVEL ALGORITHM FOR *IN-SILICO* EST EXPRESSION PROFILING

**** Leyfer D.*** [1]***, Funari V.*** [1]***, Berwick R., Haverty P., Frith M., Tolan D.***

Boston University, Boston, USA, e-mail: dmitriyl@bu.edu
[*] Presenting author
[1] Both authors contributed equally to this project

***Key words:*** *gene expression, expression profiling, transcriptional profiling, EST, transcriptome, dbEST, biological databases, in-silico, cDNA libraries, genomics, functional genomics*

## Resume

*Motivation:* Various tasks in computational biology, including primer and oligonucleotide array design; epitope mapping, etc. require finding unique regions in gene sequences. We developed a novel algorithm that finds such unique regions based on an alignment of the gene sequence to its paralogs. This algorithm was utilized for *in-silico* expression profiling using EST databases – an accurate low cost alternative to high-throughput "wet bench" expression profiling methods.
*Results:* Analysis of expression patterns of enzymes in glycolytic pathway led to alternative hypothesis of fructose metabolism in the brain. Publicly available database of Expressed Sequence Tags (dbEST) consisting of multiple flat files was mined for crucial information and reformatted into a relational database. A public Internet access to the database is planned.
*Availability:*
-Available as a commercial package through Boston University technology transfer.
-Free availability over the Internet: watch for update at http://zlab.bu.edu/~dmitriyl.

## Introduction

With the completion of the human genome sequencing the focus of biological research has shifted to 'postgenomics': gene expression analysis, signal transduction pathways, modelling biological processes. Various high-throughput methods of expression profiling are now common, but costly and labor-intensive. A low cost, fast alternative to wet bench transcriptional profiling is utilizing information from public and private EST databases.
ESTs (Adams et al., 1991) are single-pass sequenced cDNAs representing expressed genes from a specific cell population. EST library is a collection of ESTs from a single experiment. Database of Expressed Sequence Tags (dbEST, http://www.ncbi.nlm.nih.gov/dbEST/) is a public domain collection of flat files containing information about ESTs. Current number of ESTs in dbEST is close to 9 million and growing exponentially, the number of EST libraries exceeds 8000 and the number of species is 345. Large EST databases have also been compiled in some of the genomics companies. If one compares an EST library to a single gene array type, the benefits of complementing other high throughput expression profiling technologies with EST data become obvious.
Gleaning reliable expression information from EST data is challenging. One of the problems is that commonly used algorithms are designed not for EST expression profiling, but for assembling ESTs into contigs (clusters) for resolving full length cDNAs of novel genes. A by-product of such clustering is a collection of gene-specific ESTs, from which expression information can be derived. Such "misapplication" of algorithm could result in misinformation and errors. High sequence error rate (3.3%), alternatively spliced genes, 2-pass (instead of single pass) sequenced ESTs and contamination of dbEST with vector and wrongly indicated species sequences add to the challenge. Another common problem is the libraries in which the ratio between highly expressed and low abundance genes was altered to find rare transcripts. Such libraries are not suitable for quantitative analysis. If only quantitative libraries are used, then the number of ESTs for a certain gene can quantitatively represent the expression level of this gene in a tissue from which the ESTs originated (Funari et al., 2000). Yet another problem is inconsistent dbEST annotation that complicates data mining.

## Algorithms

*Virtual Northern Blot.* While many of the EST-clustering algorithms are EST-centric, i.e. contig is assembled by "walking" from one EST to another, our approach is gene (cDNA)-centric. VNB starts with a gene sequence that is computationally divided, based on an alignment of this gene to its paralogs, in multiple probes that are unique for this gene. These probes are then computationally "hybridized" with identical sequences in dbEST to find ESTs corresponding to the gene. The number and the length of the probes are optimized based on several parameters that allows for high sensitivity and specificity. Using 100% sequence identity in the probe-EST alignment makes virtually certain that EST is specific for the gene, which eliminates the need for choosing an arbitrary cut-off, which is a major problem with identifying ESTs using BLAST, another gene-centric approach (Peri et al., 2001).

*AutoProbe.* AutoProbe is the core of VNB; it is an algorithm that finds maximally unique regions in a gene sequence based on a multiple alignment of the gene's cDNA and its paralogs. AutoProbe idea is taken directly from experimental molecular biology, where one uses a short nucleic acid probe to hybridize to complementary mRNA sequences in Northern Blot. Unlike in regular Northern blot, multiple probes along the entire cDNA length have to be used in VNB in order to pull down all the ESTs for this cDNA, given that the ESTs can correspond to any cDNA region.
The probes for virtual hybridization have to satisfy the following criteria:
1) Probe length has to be long enough in order NOT to pull random sequences from the database.
2) Probe length has to take into account the EST error rate.
3) The probes have to be maximally gene-specific, i.e. have the least similarity to the paralogs.
The minimum length requirement (1) is satisfied by applying Erdos-Renyi Law (Erdos, R'enyi, 1970). The maximum length (L) of a random sequence in the database of length (D) is defined by

$$L=\log_{1/P}(DM/\square) \quad (1)$$

where M is the number of probes, P is the probability of encountering any one of the nucleotides = 1/4 and $\alpha$ is a desired significance level. Although the total length of human dbEST is 2 GB (2 billion nucleotides), the non-redundant portion of it is only ca. 100 million nucleotides – the overall length of the coding regions. The number of probes for an average cDNA and the window size of 10 is 250. Substituting for D, M and $\alpha$ we get L = 19.43 for $\alpha$= 0.05 and 20.59 for $\alpha$=0.01, i.e. 20 nucleotides should be the lower limit of probe size in order not to extract random sequences from the database. This lower limit varies slightly with increasing probe length and the size of the database in case of organisms with higher than human length of coding regions.
The requirement (2) determines the upper size of the probe: an average EST error rate is 3.3%, therefore one can expect a sequencing error every 30 nucleotides. In order for a probe to be on average between sequencing errors, the probe should be no longer than 30 nucleotides. To satisfy gene-specificity requirement (3), each probe is given a score that reflects the similarity of the probe region to the paralogs. The score for a probe is calculated as follows: each nucleotide position in the multiple alignment is given a numerical value based on the number of matches, mismatches and gaps. The scores for each individual position are summed across the probe length. Matches are given higher score than mismatches while mismatches are higher than gaps, ensuring that the regions of the least similarity have the lowest score. In order to cover the entire length of the cDNA, each probe is chosen inside a fixed length sliding window based on the minimum probe score in this window. The length of the sliding window as well as the length of the probe could vary for each gene family depending on the sequence similarity between family members, and could be determined experimentally. We showed that for Aldolase C sliding window of 12 with probe size of 24 nucleotides are the parameters achieving the most sensitivity at 100% specificity.

## Results

VNB was used to study expression profiles for fructose metabolism specific enzymes: aldolase isozymes and ketohexokinase (KHK). The goal was to obtain information about possible alternative sites of fructose metabolism, important for our understanding of Hereditary Fructose Intolerance (HFI) – a metabolic disorder in which unassimilated fructose-1-phosphate accumulates in liver, eventually shutting down gluconeogenesis and glycogenolysis, resulting in severe hypoglycemia, hepatic failure and eventually death. Although HFI patients have a mutation in one of the essential enzymes in the fructose metabolism, Aldolase B, a fraction of consumed fructose (ca. 40%) is processed by unclear mechanisms. The liver and kidneys, and to a lesser extent the small intestine, were the only organs reported to carry out this process, however in normal metabolism only ca. 60% of fructose is known to be internalized in these organs. Our strategy was to find alternative metabolic sites by identifying tissues that coexpress pathway-specific enzymes. Expression of KHK has been previously assayed by several different techniques, however the results were inconclusive (Table).

**Table.** KHK expression in brain.

| Method | Result |
|---|---|
| Immunohistochemistry (Bergbauer et al., 1996) | - |
| Activity assays (Aldeman et al., 1967, Bais et al., 1985) | -/+ |
| RNase Protection Assay (Hayward et al., 1998) | - |
| Affymetrix GeneChips (Haverty, 2001) | - |
| RT-PCR (Hayward et al., 1998) | +/- |

VNB demonstrated that KHK exhibits a previously unknown expression in brain, colon and mammary gland. To confirm VNB results we performed RNA in-situ hybridization (RISH, data not shown) on brain sections with digoxigenin-labeled KHK antisense probe that confirmed KHK expression in cerebellum and brain stem. These results suggest certain regions of brain as alternative sites of fructose metabolism. Aldolase C, not Aldolase B isozyme was found to be coexpressed with KHK in brain, suggesting that alternative metabolic sites might use alternative isozymes in the same pathway. These hypotheses have yet to be confirmed by other methods, for example, activity assays on specific tissues/primary cell cultures or animal knockout studies.

One of the surprising results was that VNB was more sensitive in this experiment than Affymetrix GeneChips, which showed no KHK expression in brain. This can be explained by high resolution of VNB on non-quantitative libraries, or, simply, by different methods of tissue preparation.

## VNB Limitations and Scope

VNB is best suited to doing a first, very fast, pilot study prior to confirming the obtained expression data by experimental means. VNB accounts for splice variants only in the regions of alternative splicing. As other expression profiling methods, VNB results depend on the way a tissue was prepared (e.g. microdissected vs. the entire organ). Compare to microarrays VNB has a speed advantage only while using existing EST data. Although some groups do make new EST libraries specifically to derive expression profiles (Bodymap, http://bodymap.ims.u-tokyo.ac.jp/), microarrays allow for higher throughput in new experiments. Dynamic range of the method is dependent on the total number of available ESTs. While current number of ESTs in dbEST does not allow obtaining quantitative profiles for rare transcripts, VNB has a high qualitative resolution owing to normalized libraries. Finally, most ESTs in dbEST were obtained by oligo-dT priming in order to represent the cell's mRNA population. This method misses several recently discovered classes of regulatory RNA (Eddy**,** 2001) that do not have a poly-A tail (although the same is true for cDNA and currently commercially available oligo arrays).

## Implementation

VNB is implemented as PERL script. Time required to obtain all accession numbers of the ESTs corresponding to the cDNA of interest for a high abundance gene (>1000 ESTs in dbEST) is 1-2 minutes on 1 GHz Intel Pentium III processor running LINUX. The time for obtaining expression profiles will be reduced farther after complete automation of the tool. Complete automation includes assembling gene-specific sets of ESTs for every known gene, automatic verification of ESTs that are duplicate reads, parsing for relevant tissue and library construction information, and reformatting dbEST into PostgreSQL object-relational database with web interface (public access is planned for the fall 2002). The database queries will follow basic biological rationales. The complete tool in its initial state will allow obtaining expression profiles for a known genes and novel sequences as well as tissue 'fingerprints'.

## References

1. Adelman R.C., Ballard F.J., Weinhouse S. (1967) J. Biol. Chem. 242, 3360-3365.
2. Bais R., James H.M., Rofe A.M., Conyers R.A. (1985) Biochem. J. 230, 53-60.
3. Bergbauer K. et al. (1996) Dev. Neurosci. 18, 371-379.
4. Eddy S. (2001) Non-coding RNA genes and the modern RNA world. Nature Reviews Genet. 2, 919-929.
5. Erdos P., R'enyi A. (1970) On a new law of large numbers. Jour. Anal. Math. 23:103-111.
6. Funari V. (2001) Novel computational and classical molecular biology approaches to discovering alternative sites of fructose metabolism in mammals, Ph.D. thesis, unpublished.7. Funari V.A., Leyfer D. et al. (2000) Expression Profiling using the Expressed Sequence Tag (EST) Database for Comparative Physiology and Metabolism. In Recent Research Developments in Comparative Biochemistry & Physiology, (S.G.Pandalai, Ed.) Transworld Research Network, Kerala, India. 1, 13-30.
8. Haverty P. (Personal Communication), 2001.
9. Hayward B.E., Bonthron D.T. (1998) Eur. J. Biochem. 257, 85-91.
10. Leyfer D., Funari V. et al. A Novel Algorithm to Derive Unique Regions in Gene Sequences Is Utilized for in-silico EST Expression Profiling. Bioinformatics, manuscript in preparation.
11. Peri S., Ibarrola N. et al. (2001) Common pitfalls in bioinformatics-based analyses: look before you leap. Trends Genet, Issue 9. 1 September 2001. 17, 541-545.