Microarray and EST database estimates of mRNA expression levels differ: The protein length versus expression curve for *C. elegans*

Enrique T. Muñoz, Leonard D. Bogarad and Michael W. Deem

August 25, 2004

Department of Bioengineering and
Department of Physics & Astronomy,
Rice University, Houston TX 77005–1892

Introduction

The patterns of protein expression, and therefore the estimation of protein expression levels, is of significant interest in the genomics and proteomics fields. Recently, it have also become of great interest within evolutionary studies, in which relationships between expression level and various evolutionary rates have been examined [1–4].

Despite direct measurement of protein expression levels remains non-trivial, many different methods to estimate mRNA or cDNA expression levels have been developed. In this work, we focus on the microarray [5–10] and the abundance within the EST database [11–14] methods for measurement of mRNA expression levels. The microarray, or gene chip, method is perhaps the most popular approach in current use. On the other hand, while growing popularity, the abundance within the EST database method has only recently been proposed for estimation of expression levels [15].

Many possible biases can be hypothesized in both the microarray and the abundance within the EST database methods. The microarray method, however, should generically be more reliable, as microarrays are explicitly intended to quantitatively measure expression levels. In this context, it is important to note that whereas expression data measured with the microarray method arise from a single, large experiment, the ESTs used in the abundance within the EST database method arise from the entire database, which is constructed from many experiments done under different conditions and often examining different subsets of genes of interest [14].

In the *C. elegans* genome, two different correlations between total exon length and expression level have been observed [15,16]. In one work, an estimation of expression level was made from abundance within the EST database for each gene [15]. In the other, gene expression was measured by microarrays [16–18]. Agreement between both approaches reveals a negative correlation between length and expression levels for highly expressed genes, but they disagreed about the trends for moderately and lowly expressed genes. A negative relationship between protein length and expression is expected due to the increased metabolic cost to translate longer genes [19, 20]. Negative correlations between total protein length and expression rate are also expected due to evolutionary reasons [21].

We address these questions about the protein length versus expression curve [22]. The full length versus expression curve was constructed using both the EST abundance and the microarray data. The difference between the two methods of estimating expression rates is shown. Assuming the microarray data to be the more accurate measurement of expression rates, due to reliable internal standards, it is shown that the abundance within the EST database method is biased by coding sequence length, and an explicit form of the length bias is presented. By removing the length bias from the EST database estimation, we achieve agreement between the two sets of data, thus explaining the apparent contradiction. Our results confirm the negative correlation between protein length and expression level expected from both the energetic costs associated with translation and evolutionary theory.

Results

The microarray data [17, 18] show a monotonic decrease of spliced (exonic) gene length with expression level, whereas the abundance within the EST database data [15] show a non-monotonic behaviour. The two corresponding curves differ most significantly in the region of low to moderate expression levels. The abundance within the EST database expression data were normalized according to gene length:

$$\text{Normalized Expression} = \frac{\text{Expression}}{\text{Length}} \; .$$

After removing the length bias from expression rates, the length versus expression curves corresponding to both the microarray data and EST database data agreed in a monotonic decreasing behaviour.

Conclusions

An explicit form of length bias between expression rates measured by microarray and abundance within the EST database methods has been found. If one assumes the microarray data to be more reliable due to internal standards and protocols, this length bias stems from the increased representation of long genes within the EST databases, perhaps because longer genes are more likely to survive the enzymatic conditions within the homogenized samples that lead to the cDNA libraries represented in the EST databases. Normalizing for this bias, we find that both methods for measuring expression agree, and a monotonic decrease of gene length with expression is found, in accord with traditional expectations from genetics and evolutionary biology. The possible presence of a length bias in the microarray data cannot be completely discarded, for example due to decreased accessibility of long transcripts for the microarray surface, and we note that care must be taken to control for length bias in any method that measures expression.

Acknowledgments

It is a pleasure to acknowledge stimulating discussions with Gabriel Marais. This research was supported by the U. S. National Institutes of Health.

References

[1] G. A. Wray, M. W. Hahn, E. Abouheif, J. P. Balhoff, M. Pizer, M. V. Rockman, and L. A. Romano. The evolution of transcriptional regulation in eukaryotes. *Mol. Biol. Evol.*, 20:1377–1419, 2003.

[2] F. A. Kondrashov, A. Y. Ogurtsov, and A. S. Kondrashov. Bioinformatical assay of human gene mobility. *Nucleic Acids Res.*, 32:1731–1737, 2004.

[3] E. P. C. Rocha and A. Danchin. An analysis of determinants of amino acids substitution rates in bacterial proteins. *Mol. Biol. Evol.*, 21:108–116, 2004.

[4] L. Zhang and W.-H. Li. Mammalian housekeeping genes evolve more slowly than tissue-specific genes. *Mol. Biol. Evol.*, 21:236–239, 2004.

[5] A. C. Pease, D. Solas, E. J. Sullivan, M. T. Cronin, C. P. Holmes, and S. P. A. Holmes. Light-generated oligonucleotide arrays for rapid dna-sequence analysi. *Proc. Natl. Acad. Sci. USA*, 91:5022–5026, 1994.

[6] P. O. Brown and D. Botstein. Exploring the new world of the genome with dna microarrays. *Nat. Genetics*, 21:33–37, 1999.

[7] D. J. Duggan, M. Bitter, Y. Chen, P. Meltzer, and J. M. Trent. Expression profiling using cDNA microarrays. *Nucleic Acids Res.*, 27:1300–1307, 1999.

[8] S. Nagpal, M. W. Karaman, M. M. Timmerman, V. V. Ho, B. L. Pike, and J. G. Hacia. Improving the sensitivity and specificity of gene expression analysis in highly related organisms through the use of electronic masks. *Nucleic Acids Res.*, 32:e51, 2004.

[9] C.Romualdi, S. Trevisan, B. Celegato, G. Costa, and G. Lanfranchi. Improved detection of differentially expressed genes in microarray experiments through multiple scanning and image integration. *Nucleic Acids Res.*, 31:e149, 2003.

[10] M. H. Asyali, M. M. Shoukri, O. Demirkaya, and K. S. A. Khabar. Assessment of reliability of microarray data and estimation of signal thresholds using mixture modeling. *Nucleic Acids Res.*, 32:2323–2335, 2004.

[11] Y. Gitton, N. Dahmane, S. Balk, A. Ruiz, L. Neidhardt, M. Schoize, B. G. Hermann, P. Kahlem, A. Benkahla, S. Schrinner, R. Yildirimman, R. Herwig, H. Lehrach, and M-L Yaspo. A gene expression map of human chromosome 21 orthologues in the mouse. *Nature*, 420:586–590, 2002.

[12] L. Skrabanek and F. Campagne. Tissueinfo: high-throughput identification of tissue expression profiles and specificity. *Nucleic Acids Res.*, 29:e102, 2001.

[13] X. Mu, S. Zhao, R. Pershad, T.F. Hsieh, A. Scarpa, S.W. Wang, R.A. White, P.D. Beremand, T.L. Thomas, L. Gan, and W.H. Klein. Gene expression in the developping mouse retina by est sequencing and microarray analysis. *Nucleic Acids Res.*, 29:e102, 2001.

[14] R. Sorek and H. M. Safer. A novel algorithm for computational identification of contaminated est libraries. *Nucleic Acids Res.*, 31:1067–1074, 2003.

[15] G. Marais and G. Piganeau. Hill-robertson interference is a minor determinant of variations in codon bias across Drosophila melanogaster and Caenorhabditis elegans genomes. *Mol. Biol. Evol.*, 19:1399–1406, 2002.

[16] C. I. Castillo, S. L. Mekhedov, D. L. Hartl, E. V. Koonin, and F. A. Kondrashov. Selection for short introns in highly expressed genes. *Nat. Genetics*, 31:415–418, 2002.

[17] A. A. Hill, C. P. Hunter, B. T. Tsung, G. Tucker-Kellogg, and E. L. Brown. Genomic analysis of gene expression in C. elegans. *Science*, 290:809–812, 2000.

[18] Genomic analysis of gene expression in C. elegans supplemental data files [http://mcb.harvard.edu/hunter/Publications/ 1053496_supplemental.zip].

[19] E. P. C. Rocha and A. Danchin. An analysis of determinants of amino acids substitution rates in bacterial proteins. *Mol. Biol. Evol.*, 21:108–116, 2004.

[20] L. Duret and D. Mouchiroud. Expression pattern and, surprisingly, gene length shape codon usage in caenorhabditis, drosophila, and arabidopsis. *Proc. Natl. Acad. Sci. USA*, 96:4482–4487, 1999.

[21] T. Tan, L. D. Bogarad, and M. W. Deem. Modulation of base specific mutation and recombination rates enables functional adaptation within the context of the genetic code. *J. Mol. Evol.*, 0, 2004. in press.

[22] E. T. Munoz, L. D. Bogarad, and M. W. Deem. Microarray and EST database estimates of mRNA expression levels differ: The protein length versus expression curve for C. elegans. *BMC Genomics*, 5:30, 2004.