

Letter to the Editor

The Closest BLAST Hit Is Often Not the Nearest Neighbor

Liisa B. Koski, G. Brian Golding

Department of Biology, McMaster University, 1280 Main Street West, Hamilton, Ontario Canada, L8S 4K1

Received: 23 January 2001 / Accepted: 20 February 2001

Abstract. It is well known that basing phylogenetic reconstructions on uncorrected genetic distances can lead to errors in their reconstruction. Nevertheless, it is often common practice to report simply the most similar BLAST (Altschul et al. 1997) hit in genomic reports that discuss many genes (Ruepp et al. 2000; Freiberg et al. 1997). This is because BLAST hits can provide a rapid, efficient, and concise analysis of many genes at once. These hits are often interpreted to imply that the gene is most closely related to the gene or protein in the databases that returned the closest BLAST hit. Though these two may coincide, for many genes, particularly genes with few homologs, they may not be the same. There are a number of circumstances that can account for such limitations in accuracy (Eisen 2000). We stress here that genes appearing to be the most similar based on BLAST hits are often not each others closest relative phylogenetically. The extent to which this occurs depends on the availability of close relatives present in the databases. As an example we have chosen the analysis of the genomes of a crenarchaeota species *Aeropyrum pernix*, an organism with few close relatives fully sequenced, and *Escherichia coli*, an organism whose closest relative, *Salmonella typhimurium*, is completely sequenced.

Key words: BLAST hits — Nearest-neighbor

Introduction

The analysis of ribosomal RNA (rRNA) has suggested that all life forms can be grouped into one of three domains: Bacteria, Archaea, or Eukarya (Woese 1987), Archaea and Eukarya being each others closest relative with Bacteria first to diverge. Archaea has been further subdivided into Crenarchaeota and Euryarchaeota. With the increasing number of gene sequences available, more and more phylogenetic incongruencies are observed with this so-called universal tree of life (Ribeiro and Golding 1998). These incongruencies have been attributed to such things as variation in evolutionary rates of genes, gene duplication, gene loss, and horizontal gene transfer. Identification of horizontally transferred genes based on differences in codon bias and base composition, sequence homology, or phylogenetic incongruencies have led to many instances of gene transfer reported to date (Lawrence and Ochman 1998; Nelson et al. 1999). It is known that caution should be taken when using sequence similarity to infer evolutionary relationships and gene functionality (Sicheritz-Ponten and Anderson 2001; Eisen 1998; Eisen and Hanawalt 1999). Yet papers that report on newly sequenced genomes often state the number of ORFs related to a certain species based on BLAST hits and hypothesize that odd similarities are due to horizontal gene transfer. It is therefore important that we quantify the extent and magnitude of the differences between sequence similarity and phylogenetic proximity.

The complete amino acid sequences of the crenarchaeota species *Aeropyrum pernix* K1 and *Escherichia*

Table 1. Comparison of closest BLAST hits and nearest phylogenetic neighbors

	Crenarch.	Euryarch.	Eukaryote	Gram–	Gram+	Total
<i>Aeropyrum pernix</i>						
Closest BLAST hit	39	105	1	13	15	173
Nearest neighbor	42	85	5	30	11	173
<i>Escherichia coli</i>						
Closest BLAST hit	0	9	1	193	28	231
Nearest neighbor	0	9	4	196	22	231

coli MG1655 were obtained from the NCBI database (<ftp://ncbi.nlm.nih.gov/gen-bank/genomes/bacteria>). All protein sequences of each genome were blasted to the NCBI database and only matching sequences with an expect value less than 10^{-20} to at least one species in each of the five groups of life (Crenarchaeota, Euryarchaeota, Eukaryote, Gram-negative, and Gram-positive) were used for further comparison. Genes with representatives in each of the five groups of life were chosen to ensure that none were unique to specific taxa and that all genes were present in all life forms. Mitochondrial, chloroplast, cyanobacterial, *Thermotoga*, and *Aquifex* sequences were excluded. For *A. pernix* there were 173/2693 proteins that matched this criterion and 231/4289 for *E. coli*. The highest BLAST hit for each of these ORFs which aligned over more than 85% of the sequence was recorded and these ORFs were used for further phylogenetic analysis.

Phylogenetic trees were constructed using the program PUZZLE (Strimmer and von Haeseler 1996). This method applies maximum-likelihood tree reconstruction and accounts for rate heterogeneity across sites. Substitution rate is known to vary widely across amino acid sequences of protein-coding genes. This is attributed to selective and functional constraints that vary between different parts of the molecule. The number of substitutions that have occurred may be underestimated if rate variation is ignored (Tourasse and Gouy 1999; Golding

1983). Four gamma rate categories and 10,000 puzzling steps were used to reconstruct each phylogenetic tree. The species adjacent to *A. pernix/E. coli* through the least number of internal nodes (or in the case of a tie, via the shortest branch length) was chosen as its nearest phylogenetic neighbor.

In the *A. pernix* genome we found that from 173 phylogenetic reconstructions, 40.5% (70/173) of the ORFs had a BLAST hit different from their nearest phylogenetic neighbor. For 30% (52/173) of the ORFs, the closest BLAST hit was not even to a species in the same domain of life as their nearest phylogenetic neighbor. Table 1 shows the distribution of highest BLAST hits and nearest neighbors among the five groups of life. In the *E. coli* genome 27.3% (63/231) of the ORFs had a blast hit different from their nearest phylogenetic neighbor and 7% (16/231) did not BLAST to a species in the same domain of life as its nearest phylogenetic neighbor. Interestingly, for some genes that BLAST to a species in one particular domain of life, the number with phylogenetic nearest neighbors in this domain may be either over or underestimated (Table 2). This also suggests that the potential for phylogenetic errors of reconstruction, such as long branch attraction, should be carefully evaluated. The high number of ORFs in the *A. pernix* genome that do not have the same BLAST hit and nearest phylogenetic neighbor is not surprising as there are few close relatives present in the databases. More surprising is the

Table 2. ORFs that differ in nearest neighbor versus closest BLAST hit and the taxonomic grouping from which these hits originated

70/173 ORFs from <i>Aeropyrum pernix</i>					
Neighbor/BLAST	Crenarch	Euryarch	Eukaryote	Gram–	Gram+
Crenarch	1	8	0	0	0
Euryarch	2	11	0	5	4
Eukaryote	0	2	0	3	1
Gram–	2	18	0	3	3
Gram+	0	4	0	0	3
63/231 ORFs from <i>Escherichia coli</i>					
Neighbor/BLAST	Crenarch	Euryarch	Eukaryote	Gram–	Gram+
Crenarch	0	0	0	0	0
Euryarch	0	0	0	0	2
Eukaryote	0	1	0	1	1
Gram–	0	0	0	47	7
Gram+	0	1	0	3	0

number of ORFs in the *E. coli* genome (27.3%) that have different BLAST hits and nearest phylogenetic neighbors. Several Gram-negative bacteria have been completely sequenced, including *E. coli*'s close relative *Salmonella typhimurium*. *S. typhimurium* is the nearest phylogenetic neighbor for 20% (48/231) of the *E. coli* ORFs analyzed, 19% (9/48) of which BLAST to a species other than *S. typhimurium*.

These results show that comparisons that rely on the closest BLAST hit alone should be interpreted with great caution as it does not imply phylogenetic proximity. The magnitude of the discrepancy between the two methods of measuring gene relatedness has to be carefully reexamined.

References

- Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 25:3389–3444
- Eisen JA (2000) Horizontal gene transfer among microbial genomes: new insights from complete genome analysis. *Curr Opin Genet Dev* 10:606–611
- Eisen JA (1998) Phylogenomics: improving functional predictions for uncharacterized genes by evolutionary analysis. *Genome Res* 8:163–167
- Eisen JA, Hanawalt PC (1999) A phylogenomic study of DNA repair genes, proteins, and processes. *Mutat Res* 435:171–213
- Freiberg C, Fellay R, Bairoch A, Broughton WJ, Rosenthal A, Perret X (1997) Molecular basis of symbiosis between Rhizobium and legumes. *Nature* 387:394–401
- Golding GB (1983) Estimates of DNA and protein sequence divergence: an examination of some assumptions. *Mol Biol Evol* 1:125–144
- Lawrence JG, Ochman H (1998) Molecular archaeology of the *Escherichia coli* genome. *Proc Natl Acad Sci USA* 95:9413–9417
- Nelson KE, Clayton RA, Gill SR, Gwinn ML, Dodson RJ, Haft DH, Hickey EK, Peterson JD, Nelson WC, Ketchum KA, McDonald L, Utterback TR, Malek JA, Linher KD, Garrett MM, Stewart AM, Cotton MD, Pratt MS, Phillips CA, Richardson D, Heidelberg J, Sutton GG, Fleischmann RD, Eisen JA, Fraser CM, et al (1999) Evidence for lateral gene transfer between Archaea and bacteria from genome sequence of *Thermotoga maritima*. *Nature* 399:323–329
- Ribeiro S, Golding GB (1998) The mosaic nature of the eukaryotic nucleus. *Mol Biol Evol* 15:779–788
- Ruepp A, Graml W, Santos-Martinez ML, Koretke KK, Volker C, Mewes HW, Frishman D, Stocker S, Lupas AN, Baumeister W (2000) The genome sequence of the thermoacidophilic scavenger *Thermoplasma acidophilum*. *Nature* 407:508–511
- Sicheritz-Ponten T, Anderson GE (2001) A phylogenomic approach to microbial evolution. *Nucl Acids Res* 29:545–552
- Strimmer K, von Haeseler A (1996) Quartet puzzling: a quartet maximum-likelihood method for reconstructing tree topologies. *Mol Biol Evol* 13:964–969
- Tourasse NJ, Gouy M (1999) Accounting for evolutionary rate variation among sequence sites consistently changes universal phylogenies deduced from rRNA and protein-coding genes. *Mol Phylogenet Evol* 13:159–168
- Woese CR (1987) Bacterial evolution. *Microbiol Rev* 51:221–271