ELSEVIER

Gene 238 (1999) 163–170

GENE
AN INTERNATIONAL JOURNAL ON GENES AND GENOMES

www.elsevier.com/locate/gene

Human and nematode orthologs — lessons from the analysis of 1800 human genes and the proteome of *Caenorhabditis elegans*

Sarah J. Wheelan [a,b], Mark S. Boguski [a], Laurent Duret [c], Wojciech Makałowski [a,*]

[a] *National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA*
[b] *Department of Molecular Biology and Genetics, The Johns Hopkins University School of Medicine, Baltimore, MD 21205, USA*
[c] *Laboratoire BGBP–UMR CNRS 5558, Université Claude Bernard–Lyon 1, 43 Bd du 11 Novembre 1918, F-69622 Villeurbanne Cedex, France*

Received 12 March 1999; received in revised form 22 June 1999; accepted 6 July 1999; Received by G. Bernardi

Abstract

Recently, we have defined and analyzed over 1800 orthologous human and rodent genes. Here we extend this work to compare human and *Caenorhabditis elegans* coding sequences. 1880 human proteins were compared with about 20 000 predicted nematode proteins presumably comprising nearly the complete proteome of *C. elegans*. We found that 44% of human/rodent orthologs have convincing nematode counterparts. On average, the amino acid similarity and identity between aligned human and *C. elegans* orthologous gene products are 69.3% and 49.1% respectively, and the nucleotide identity is 49.8%. Detailed investigation of our results suggests that some nematode gene predictions are incorrect, leading to erroneous pairing with human genes (e.g. calcineurin and polymerase II elongation factor III). Furthermore, other proteins (i.e. homologs of human ribosomal proteins S20 and L41, thymosin) are missing entirely from the nematode proteome, suggesting that it may not be complete. These results underscore the fact that metazoan gene prediction is a very challenging task and that most computer-predicted nematode genes require supporting evidence of their existence from comparative genomics and/or laboratory investigation.

Keywords: Comparative genomics; Molecular evolution

1. Introduction

The nematode *Caenorhabditis elegans* is the first metazoan organism whose genome has been almost completely sequenced (The C. elegans Sequencing Consortium, 1998). The goal of the present study was to define, as accurately as possible, probable orthologs between *C. elegans* and mammals and to assess the average degrees of protein and coding sequence similarity between these two species.

Studies of orthologous sequences, rather than just domains or motifs, are most likely to reveal the most robust clues about gene function in different organisms. It must be remembered, however, that vertebrates are thought to have undergone genome duplications compared with simpler organisms and, therefore, orthologous sequences may be difficult to define because of the possible one-to-many, or even many-to-many relationships, among homologous genes in their genomes.

We have previously identified 1885 full-length human mRNAs that have validated orthologous mouse and/or rat sequences. We have used this data set to identify 819 pairs of orthologous human/nematode genes. We analyzed the conservation of protein and mRNA coding sequences to estimate the range and average degrees of sequence similarity between *H. sapiens* and *C. elegans*. Quantification of the degree of similarity between the coding regions of these genomes creates a more objective scale by which to judge whether a given similarity is significant.

On average, the amino acid similarity and identity between aligned human and nematode orthologs are 69.3% and 49.1% respectively, and the average nucleotide identity is 49.8%. These numbers provide a good standard by which to gauge the significance of putative orthologs identified by other methods; also, using orthologous relationships that have been validated in other organisms (i.e. humans and rodents) may be a useful technique to refine gene predictions.

* Corresponding author. Tel.: +1-301-435-5989; fax: +1-301-480-9241.
E-mail address: makalowski@ncbi.nlm.nih.gov (W. Makałowski)

0378-1119/99/$ – see front matter
PII: S0378-1119(99)00298-X

2. Materials and methods

2.1. Nematode and human protein sequences

We used 1880 human proteins selected in our previous study of human and rodent orthologous genes (Makałowski and Boguski, 1998). Our nematode data set of 19 080 gene products was graciously provided by R. Durbin and S. Jones of the Sanger Centre in November, 1998, with the caveat that the gene predictions and protein translations were unvetted and might contain inaccuracies. Nevertheless, this was the data set used by most authors for the nematode genome issue of *Science* (vol. 282, December 11, 1999) and thus our work may be compared with other analyses reported there. This data set was combined with all *C. elegans* proteins available in GenBank on December 12, 1998 and a non-redundant superset was created using the patdb program (W. Gish, unpublished). Patdb eliminates all identical strings and substrings of sequences, and these strict criteria may result in some residual redundancy in the final data set. Vertebrate sequences were selected from GenBank on November 26, 1998, and, after clustering with patdb, comprised a data set of 82 460 sequences.

2.2. Identification of nematode and human orthologs

1880 human proteins were used as queries in blastp searches against 20 119 nematode sequences. The *E*-value 10^{-5} was used as a threshold of significance in the initial BLAST search. A similar threshold was used in Hovergen clustering (Duret et al., 1994). The top 10 alignments (by *E*-value) were carefully examined for each human query. Only those alignments that comprised at least 70% of both the query and matching sequences were selected for further consideration. By applying this length-consistency rule, we attempted to avoid ortholog candidate selection based merely on strong similarity between highly conserved protein domains (local alignments) that might be present among paralogous genes. It should be noted that the highest-scoring sequence in the blastp results was not always selected as a candidate for the nematode ortholog of a human gene. Whenever more than one nematode sequence met the 70% alignment criterion, the most conserved sequence was selected for further consideration. Our nematode ortholog candidate sequences were then used as queries in blastp searches against a data set of 82 460 vertebrate proteins to provide further evidence of orthology as follows. Only those human/worm sequence pairs for which the blastp score was within 10% of the best score for all nematode/vertebrate alignments were accepted for further analysis. In all, 819 proteins met this criterion.

Most of the human/nematode orthologous pair assignments were further assessed using a phylogenetic approach as implemented in the Hovergen data base (Duret et al., 1994). An experimental database of nematode and vertebrate homologous sequences (Hovercel) was created for this purpose. 55 927 vertebrate proteins and 13 432 *C. elegans* proteins were clustered according to procedures described previously (Duret et al., 1994). The clustering resulted in 1311 gene families containing 14 803 vertebrate and 2432 nematode proteins. The human/nematode pair assignments defined by blastp searching were checked against these families to eliminate pairs that did not reflect orthologous relationships.

2.3. Data acquisition

Sequences (human mRNA and *C. elegans* mRNA or cosmid sequences) were extracted from GenBank using the dump_cds program (J. Zhang, unpublished), which extracts different portions of the GenBank record to different files based on annotation in the GenBank features table.

2.4. Sequence alignment and analysis

Protein sequence alignments were computed using CLUSTAL W 1.7 (Thompson et al., 1994), blastalign (C. Chappey, unpublished) and map (Huang, 1994); BLOSUM 30, BLOSUM62, and BLOSUM80 scoring matrices were tested with all three programs (Henikoff and Henikoff, 1992). All individual alignments were visually inspected and manual adjustments to the alignments were made when necessary. Nucleotide alignments were created based on the protein alignments, placing three gaps in the nucleotide alignment for each gap in the protein alignment.

The value 'percent aligned' was computed by counting the number of amino acids or nucleotides in the given sequence that are aligned respectively with an amino acid or nucleotide of the other sequence (i.e. not aligned with a gap) and dividing by the total length of the given sequence. 'Percent similarity' and 'percent identity' numbers refer to the percentage of aligned sequences (not including gaps) that were considered similar or identical. Two amino acids were considered similar if they had a positive score in a given substitution matrix.

Protein evolutionary distances were estimated according to the Kimura (198)3 empirical formula:

$$\mathrm{Kaa} = -\ln(1-p-p2/5)$$

where p is the fraction of amino acid differences.

3. Results

3.1. Ortholog identification

Out of the original 1880 human sequences examined, 819 were found to have at least one *C. elegans* ortholog

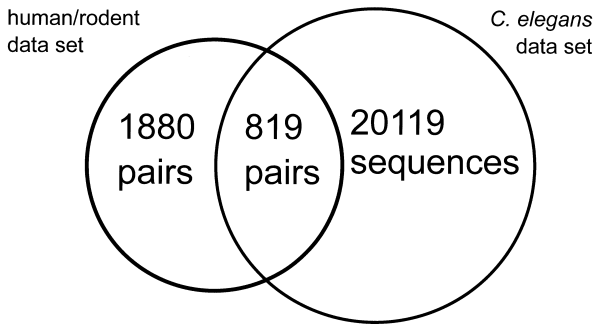

Fig. 1. Venn diagram of the data used in this study.

(Fig. 1). Somewhat surprisingly (given the expectation of gene duplications in vertebrates relative to nematodes) more than half the number of cases (461/819) of the human and worm orthologs in our study appear to represent a one-to-one relationship. This is probably a result of incomplete information on the human proteome. At the other extreme, one nematode gene (C07G1.1) has as many as 10 human orthologs, suggesting multiple duplication events in the evolution of a human lineage.

We applied very strict and conservative rules for orthology assignment, and these criteria resulted in pairing a relatively low fraction (44%) of human genes with defined nematode orthologs. Surprisingly, not all highly conserved mammalian genes have easily detectable worm counterparts; on the other hand, some less conserved mammalian proteins have clear *C. elegans* orthologs (Fig. 2). In our previously defined set of 1880 human/rodent orthologous proteins, 160 are at least 99% identical. We were able to assign nematode orthologs to only 125 of those highly conserved mammalian genes. The other 35 human genes for which we could not find *C. elegans* counterparts are listed in Table 1. There are at least four factors that could contribute to the 'missing' orthologs phenomena: (1) evolution of new, vertebrate-specific genes since the divergence of the common ancestor of vertebrates and nematodes; (2) specific gene loss from the nematode genome since divergence of the common ancestor; (3) sequence divergence so great as to be undetectable using current methods; (4) inaccurate or incomplete *C. elegans* gene prediction. Human B-cell translocation gene protein 1 (BTG1) is an example of 'missing gene' due to extreme sequence divergence. BTG1 (172 residues, Table 1) has a unique ortholog in *C. elegans* (C03C11.2, 263 residues), but the proteins do not have the same length, conservation is weak and is limited to the N-terminal half of the protein (24% identity, 42% similarity over the first 100 residues). The blastp score is low ($S=44$, $E=8\times10^{-4}$) but homology is confirmed by multiple alignment. This particular case is not due to an error of gene prediction that resulted in incorrect protein lengths: the nematode gene had previously been sequenced as an mRNA (GenBank: AJ011777). However, it seems that erroneous gene predictions have occurred in other cases. Mammalian calcineurin B is 170 amino acids long (GenBank accession number for

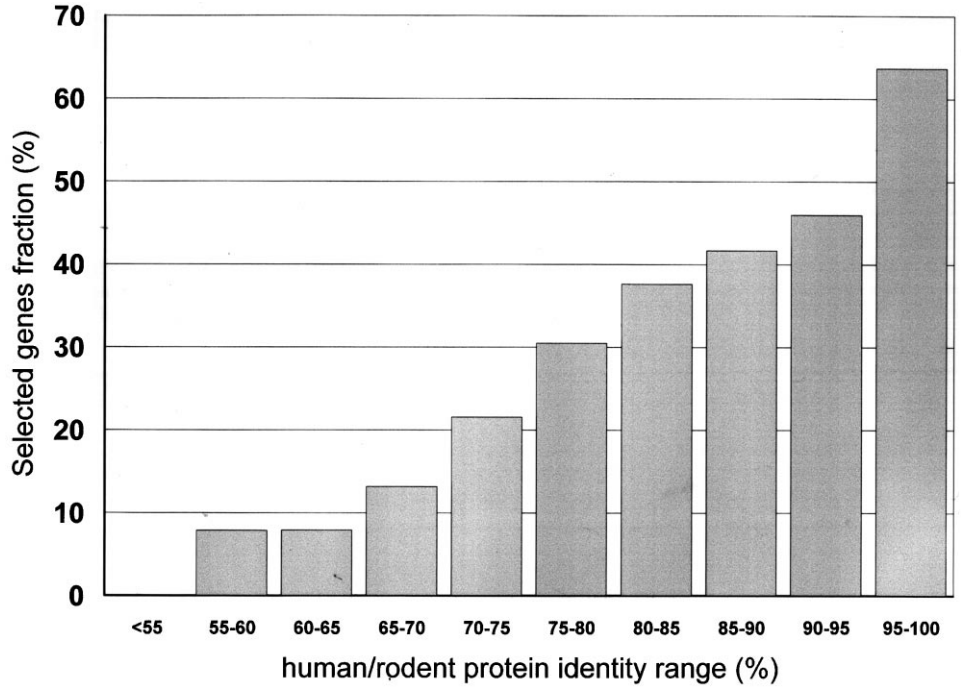

Fig. 2. Fraction of human genes with demonstrable nematode orthologs with different identity for human/rodent sequence pairs. Human/rodent orthologs identity taken from Makałowski and Boguski (1998).

Table 1
Conserved mammalian proteins without a demonstrable nematode ortholog

GenBank accession number	Protein name	Probable cause of missing ortholog
D14838	fibroblast growth factor (FGF-9)	domain similarity only
J04991	stathmin (p18 protein)	domain similarity only
L13385	Miller–Dieker lissencephaly protein (LIS1)	domain similarity only
L26494	transcriptional repressor (oct-6)	domain similarity only
L29277	DNA-binding protein (APRF)	domain similarity only
M16938	homeo box c8 protein	domain similarity only
M17733	thymosin beta-4	domain similarity only
M24899	triiodothyronine (ear7)	domain similarity only
M27691	transactivator protein (CREB)	domain similarity only
M28372	sterol regulatory element-binding protein (CNBP)	domain similarity only
M30773	calcineurin B	domain similarity only
M64497	apolipoprotein AI regulatory protein (ARP-1)	domain similarity only
M77698	GLI-Krupple related protein (YY1)	domain similarity only
M93415	activin type II receptor	domain similarity only
M96944	B-cell specific transcription factor (BSAP)	domain similarity only
S70721	isl-1 homeobox	domain similarity only
U04241	homolog of *Drosophila* enhancer of split m9/m10	domain similarity only
U33428	K+ channel beta 1a subunit	domain similarity only
X07495	cp19 homeobox from HOX-3 locus	domain similarity only
X55545	cAMP response element binding protein (CREB1)	domain similarity only
X59268	general transcription factor IIB	domain similarity only
X61118	cysteine-rich protein with LIM motif	domain similarity only
X61123	B-cell translocation gene protein 1 (BTG1)	domain similarity only
X75918	immediate–early response protein NOT	domain similarity only
X95404	non-muscle type cofilin	domain similarity only
Z11933	brain-2 POU-domain protein	domain similarity only
Z12962	ribosomal protein L41	not annotated in AF039712
X59711	CAAT-box DNA binding protein subunit A	not annotated in HTGS CEY105E8
L34587	RNA polymerase II elongation factor SIII, p15 subunit	not annotated in HTGS AC006713
L06498	ribosomal protein S20 (RPS20)	not annotated in HTGS CEY105E8
J02645	translational initiation factor (eIF-2), alpha subunit	represented by EST only (GenBank accession number M89149)
J04173	phosphoglycerate mutase (PGAM-B)	no BLAST hit[a]
M54927	myelin proteolipid protein	no BLAST hit[a]
S54005	thymosin beta-10	no BLAST hit[a]
U35100	complexin II	no BLAST hit[a]

[a] BLAST searches were performed under default parameters as implemented on the NCBI web server.

human mRNA is M30773). Using our strict criteria we initially failed to identify a nematode ortholog of this gene. Subsequent examination of blastp results, however, revealed a very similar protein, F55C10.1, in the worm genome. The genefinder prediction for this protein resulted in a sequence 369 amino acids in length. Interestingly, the 169 C-terminal residues show 79% identity and 91% similarity to human calcineurin B. The 199 N-terminal amino acids of the worm protein do not show similarity to any known protein using the blastp program (as of February 25, 1999), indicating possible fusion of exons deriving from the actual gene with putative coding sequences from another hypothetical transcription unit.

To estimate the number of 'missing genes' that have been sequenced but not correctly predicted, we compared the 35 proteins from Table 1 to all *C. elegans* sequences (GenBank: genomic + mRNA = 153 Mb; ESTs = 24 Mb; GenBank release 111 plus daily updates as of May 5, 1999). There are 26 proteins that have good BLAST matches, but the alignments are limited to a single protein domain or for which the nematode and human proteins have very different lengths (e.g. calcineurin). We found one protein that has an obvious ortholog in genomic sequences that is confirmed by ESTs but that has not been yet annotated: Z12962 ribosomal protein L41 (accession nos AF039712 and C09422). We found three proteins that have obvious orthologs in genomic cosmids for which sequence is preliminary and unannotated (HTGS phase 1: 'Sequencing in progress'): X59268 general transcription factor IIB, in AC006704; L34587 RNA polymerase II elongation factor SIII, p15 subunit, in AC006713; L06498 ribosomal protein S20 (RPS20), in CEY105E8. Additionally, we found one protein that has an obvious ortholog among ESTs, but no match in genomic sequences: the human alpha subunit of translational initiation factor (GenBank accession no. J02645) is an

Table 2
Basic statistics of nematode and human orthologous genes comparison

	Protein identity	Protein similarity	Protein evolutionary distance	DNA identity
Sequence pairs analyzed	819	819	819	758[a]
Minimum	18.3%	39.3%	0.016	20.3%
Maximum	98.4%	99.5%	3	82.5%
Mean	49.1%	69.3%	0.957	49.8%
Median	46.5%	68.5%	0.896	48.2%
SD	17.1%	12.3%	0.513	11.0%
Variance	2.9%	1.5%	0.263	1.2%
Std error	0.6%	0.4%	0.018	0.4%

[a] Not all *C. elegans* DNA sequences were available at the time of analysis.

ortholog of a protein represented in GenBank by an EST (accession no. M89149). Thus at least five of the 35 "missig genes (14%) are due unfinished, unannotated or incorrectly annotated sequences. There remain four highly conserved mammalian proteins from our data set that have no obvious nematode homologs: complexin II (U35100), myelin proteolipid protein (M54927), phosphoglycerate mutase (J04173), and thymosin beta-10 (S54005).

3.2. Statistical properties of the data

The human protein sequences ranged from 51 to 4544 amino acids in length [these extremes are represented by ribosomal protein L39 (U57846) and LDL-receptor-related precursor (X13916) respectively]. The mean length of analyzed proteins was 456 (standard deviation, SD = 343) and the median value was 390 residues. Half of the analyzed proteins fall between 245 and 540 residues, and 90% of the proteins were shorter than 814 residues. All together, over 800 000 amino acids were analyzed. Because of our strict method of orthologous sequence selection (see Section 2.1), the statistical properties of the analyzed nematode proteins were very similar.

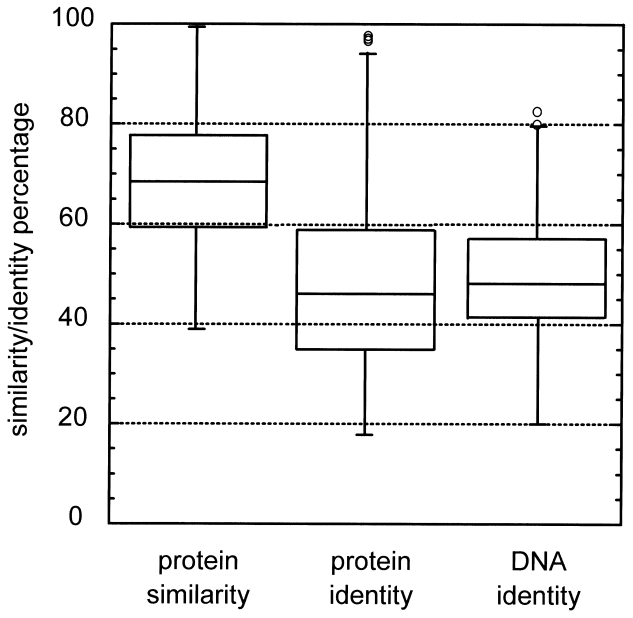

Fig. 3. Box plots of protein and DNA sequence conservation in aligned worm and human genes. For each category, the central box depicts the middle 50% of the data between the 25th and 75th percentiles and the enclosed horizontal line represents the median value of the distribution. Extreme values are indicated by circles that occur outside the main bodies of data.

Some basic features of the nematode/human orthologous sequence alignments are presented in Table 2 and the distributions of protein and coding sequence identities, along with the distribution of protein similarities, is presented in Figs. 3 and 4. On average, human and nematode orthologs were found to be 49.1% (SD = 17.1) identical and 69.3% (SD = 12.3) similar, with protein identity ranging between 18.3% (protein similarity 39.3%) for transient axonal glycoprotein (human accession no. X67734 and worm protein symbol C47E12.8) and 98.4% for ubiquitin (human accession no. M26880 and worm protein symbol F25B5.4).

At the time our analysis was performed (December, 1998) not all coding sequences for *C. elegans* proteins were available. Therefore, we compared 758 human and nematode CDSs comprising over 2 000 000 nucleotides. The average aligned identity of human/nematode CDS pairs was 49.2% (SD = 6.78) with median value of 48.2%. The most diverged coding sequences are about 20% identical, which indicates multiple nucleotide substitutions in each position. Alignment of such divergent coding sequences would be impossible without using aligned proteins as a guide (see Section 2). At the other extreme, some of the human/nematode coding sequences are surprisingly conserved: coding sequences for three human actin genes present in our data set (accession nos: X16940, J00068, and X04098) are over 80% identical with their nematode ortholog, M03F4.2. Additional statistics on coding sequence identity are presented in Table 2.

3.3. Evolutionary analysis

Nematodes diverged at some very early time during metazoan evolution, over 600 million years ago (Feng

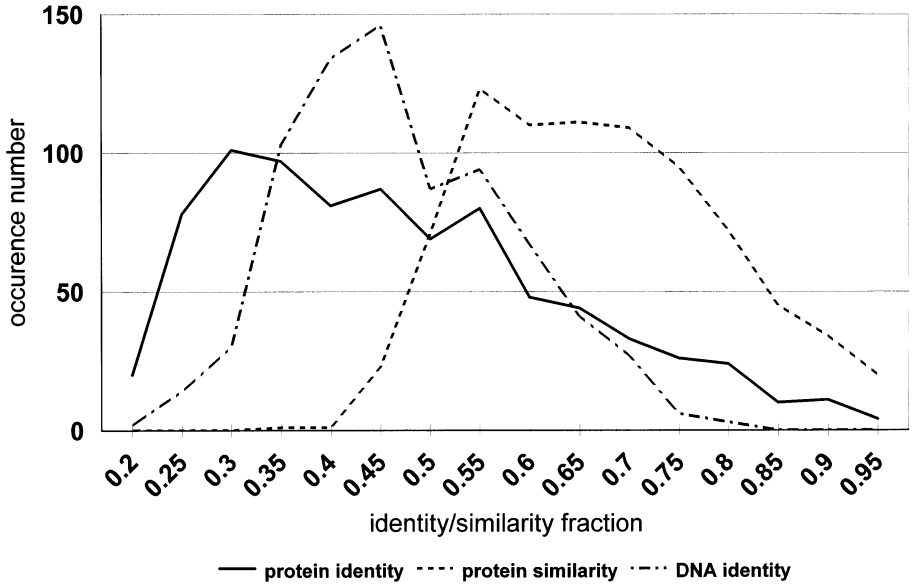

Fig. 4. Distribution of sequence conservation in aligned worm and human genes.

et al., 1997; Futuyma, 1997). During the long period of time since the last common ancestor of *H. sapiens* and *C. elegans*, the coding sequences of orthologous genes have accumulated an extremely large number of changes. It is expected that many of these changes occurred multiple times at the same sequence position during the evolution of primate and nematode lineages, especially in synonymous positions. Unfortunately, there is no method that can reliably estimate the number of sequence changes over such a long period of time. Kimura (1983) provided an empirical formula to estimate the number of probable amino acid changes based on protein identity. Thus we used his formula to estimate evolutionary distance between our human and worm proteins. The mean evolutionary distance is 0.957 (SD = 0.513) with a median value of 0.896 (see Table 2). This means that, on average, almost every amino acid in each protein has changed once during metazoan evolution. Evolutionary distances are widely distributed, from 0.016 in ubiquitin protein, to 3.0 in transient axonal glycoprotein. The latter number is especially interesting; it shows that orthology assignment is possible even for proteins with multiple amino acid substitution in almost every position.

Wolfe and Sharp (1993) examined the relationship between amino acid and nucleotide sequence divergence in rodents and have noted that, for highly conserved genes, amino acid identity exceeds nucleotide identity (because silent mutations are permitted in codons, but the protein sequence is constrained) and that, for less conserved genes, nucleotide identity exceeds amino acid identity (because only one or two random nucleotide changes are required to change an amino acid, but some nucleotide identity will still be conserved). For different data sets, a crossover point that marks the transition from highly conserved to less conserved genes occurs at different levels of identity. For example, for a distribution of mouse and rat orthologs, this crossover point is approximately 93% (Wolfe and Sharp, 1993) and for human/mouse orthologs it is 85% (Makałowski et al., 1996). For comparisons between bacterial sequences, the extrapolated crossover point is closer to 65% (Wolfe

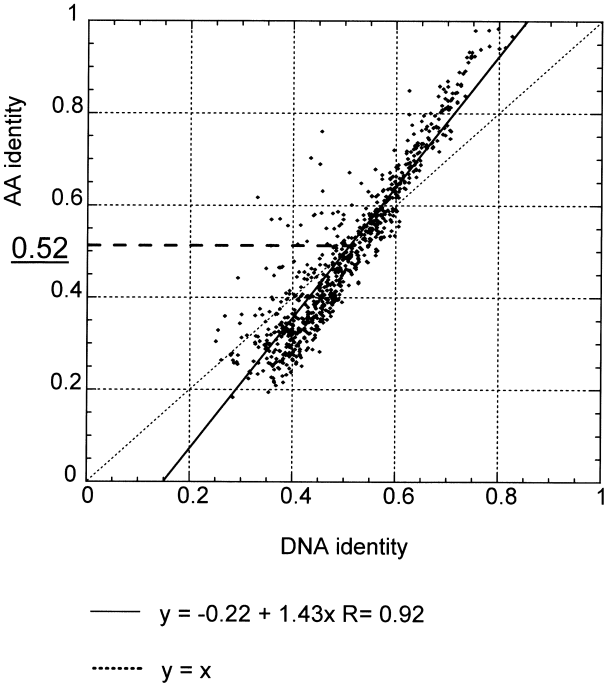

Fig. 5. Scatter plot of protein versus nucleotide sequence conservation. The solid line represents a linear regression model of the data. The dashed line (representing equal nucleotide and protein identity) allows one to estimate the 'crossover' point between nucleotide and protein sequence identity.

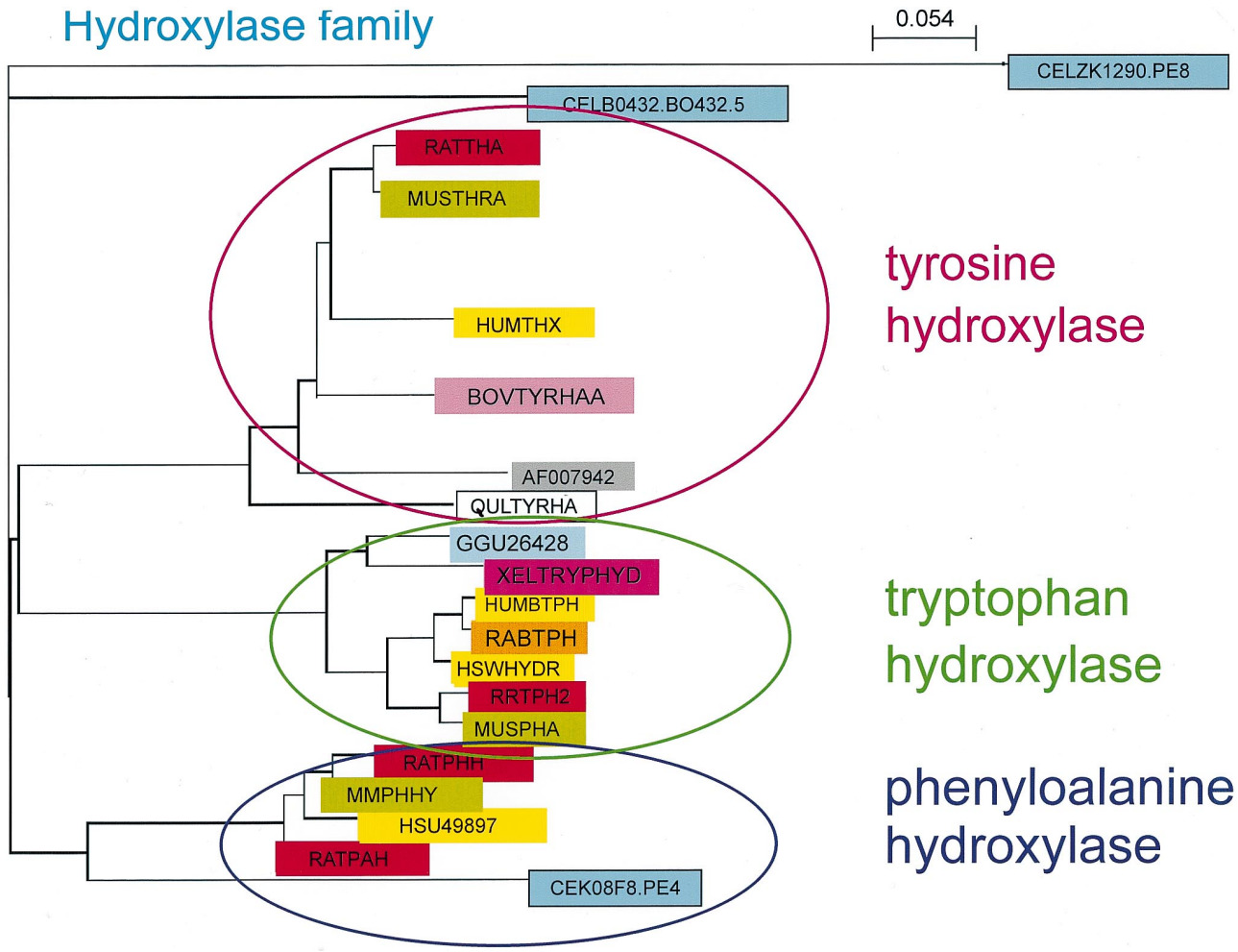

Fig. 6. Phylogenetic tree of sequences of vertebrate and nematode hydroxylases as represented in Hovercel. Genes from different species are color coded as follow: *C. elegans*, blue with black border; rat, red; mouse, green; human, yellow; bovine, pink; chicken, blue; rabbit, orange; frog, pink; perch, gray; quail, white. Names in boxes represent GenBank locus name for a given gene.

and Sharp, 1993). Among human and nematode orthologs, protein and DNA identities are strongly correlated, with a correlation coefficient factor of 0.92. The crossover point for the human/nematode data is 52% (Fig. 5). It is worthwhile noting that, for all animal genes compared, the crossover point is very close to the mean value of both protein and DNA identities. In contrast, the extrapolated crossover point for bacterial genes is much lower than the average protein and DNA identity.

3.4. Gene function assignment

It is commonly believed that orthologs carry out the same function (Chervitz et al., 1998). Although it may be true in most cases for unicellular organisms, which are characterized by limited number of duplicated (paralogous) genes, the situation is more complicated in the case of multicellular organisms. Three decades ago, Ohno (1970) formulated the hypothesis of genome evolution by gene duplication. Recently, Wolfe and Shields (1997) showed that unicellular organisms may also evolve by genome duplications. Remotely related genomes are characterized by one-to-many and many-to-many orthologous genes relationship. Some of the duplicated genes, after species divergence, maintain the primordial function, others, liberated from evolutionary constraint, gain a new function. Therefore, it is difficult to predict confidently the function of a protein when multiple homologs are observed. Eisen (1998) proposed 'phylogenomics' as a solution to this problem, but, unfortunately, this approach is limited to cases with a broad phylogenetic spectrum of available data. Therefore, functional inferences about nematode proteins based on the functions of human orthologs may, in many cases, not be possible. Fig. 6 shows a phylogenetic tree of hydroxylases as represented in Hovercel. There are three groups of vertebrate enzymes, each having different a specific substrate: tyrosine, tryptophan, and phenylalanine. There are also three nematode

proteins in this group. One of them (CEK08F8.PE4; wormpep name — K08F8.4) clearly clusters with vertebrate phenyloalanine hydroxylase and probably serves the same function in *C. elegans*. Two other nematode proteins serve as an outgroup to the rest of the proteins in the tree (Fig. 6). Therefore, their specific roles may not be assigned with confidence.

4. Conclusions

(1) *C. elegans* proteins and their human orthologs are, on average, 49.1% identical and 69.3% similar. Their cognate coding sequences are, on average, 49.8% identical. On average, each amino acid in surveyed proteins has changed almost once (protein evolutionary distance: 0.957) during metazoan evolution.

(2) Although most of nematode proteins we studied show some similarity to known proteins in blast searches, we were able to establish an orthology relationship in only about 40% of cases.

(3) Although this set represents well-defined orthologs of characterized human proteins, the functions of the nematode orthologs are not completely clear in some cases.

References

Chervitz, S.A., Aravind, L., et al., 1998. Comparison of the complete protein sets of worm and yeast: orthology and divergence. Science 282 (5396), 2022–2028.

Duret, L., Mouchiroud, D., et al., 1994. HOVERGEN: a database of homologous vertebrate genes. Nucleic Acids Res. 22 (12), 2360–2365.

Eisen, J.A., 1998. Phylogenomics: improving functional predictions for uncharacterized genes by evolutionary analysis. Genome Res. 8 (3), 163–167.

Feng, D.F., Cho, G., et al., 1997. Determining divergence times with a protein clock: update and reevaluation see comments. Proc. Natl. Acad. Sci. U. S. A. 94 (24), 13 028–13 033.

Futuyma, D.J., 1997. Evolutionary Biology. Sinauer Associates, Sunderland, MA.

Henikoff, S., Henikoff, J.G., 1992. Amino acid substitution matrices from protein blocks. Proc. Natl. Acad. Sci. USA 89 (22), 10 915–10 919.

Huang, X., 1994. On global sequence alignment. Comput. Appl. Biosci. 10 (3), 227–235.

Kimura, M., 1983. The Neutral Theory of Molecular Evolution. Cambridge University Press, Cambridge.

Makałowski, W., Boguski, M.S., 1998. Evolutionary parameters of the transcribed mammalian genome: an analysis of 2,820 orthologous rodent and human sequences. Proc. Natl. Acad. Sci. USA 95 (16), 9407–9412.

Makałowski, W., Zhang, J., et al., 1996. Comparative analysis of 1196 orthologous mouse and human full-length mRNA and protein sequences. Genome Res. 6 (9), 846–857.

Ohno, S., 1970. Evolution by gene duplication. Allen & Unwin/Springer, London/New York.

The C. elegans Sequencing Consortium, 1998. Genome sequence of the nematode *C. elegans*: a platform for investigating biology. The C. elegans Sequencing Consortium. Science 282 (5396), 2012–2018.

Thompson, J.D., Higgins, D.G., et al., 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 22 (22), 4673–4680.

Wolfe, K.H., Sharp, P.M., 1993. Mammalian gene evolution: nucleotide sequence divergence between mouse and rat. J. Mol. Evol. 37 (4), 441–456.

Wolfe, K.H., Shields, D.C., 1997. Molecular evidence for an ancient duplication of the entire yeast genome. Nature 387 (6634), 708–713.