RESEARCH

Large-Scale Taxonomic Profiling of Eukaryotic Model Organisms: A Comparison of Orthologous Proteins Encoded by the Human, Fly, Nematode, and Yeast Genomes

Arcady R. Mushegian,^{1,5} James R. Garey,² Jason Martin,^{1,4} and Leo X. Liu³

¹AxyS Pharmaceuticals, Inc., La Jolla, California 92037 USA; ²Department of Biological Sciences, University of South Florida, Tampa, Florida 33620 USA; ³NemaPharm, Inc., Cambridge, Massachusetts 02139 USA

Comparisons of DNA and protein sequences between humans and model organisms, including the yeast *Saccharomyces cerevisiae*, the nematode *Caenorhabditis elegans*, and the fruit fly *Drosophila melanogaster*, are a significant source of information about the function of human genes and proteins in both normal and disease states. Important questions regarding cross-species sequence comparison remain unanswered, including (1) the fraction of the metabolic, signaling, and regulatory pathways that is shared by humans and the various model organisms; and (2) the validity of functional inferences based on sequence homology. We addressed these questions by analyzing the available fractions of human, fly, nematode, and yeast genomes for orthologous protein-coding genes, applying strict criteria to distinguish between candidate orthologous and paralogous proteins. Forty-two quartets of proteins could be identified as candidate orthologs. Twenty-four *Drosophila* protein sequences were more similar to their human orthologs than the corresponding nematode proteins. Analysis of sequence substitutions and evolutionary distances in this data set revealed that most *C. elegans* genes are evolving more rapidly than *Drosophila* genes, suggesting that unequal evolutionary rates may contribute to the differences in similarity to human protein sequences. The available fraction of *Drosophila* proteins appears to lack representatives of many protein families and domains, reflecting the relative paucity of genomic data from this species.

Similarities between novel protein sequences and their better-characterized counterparts in sequence databases are an increasingly important source of hypotheses concerning protein functions. Particular attention has been paid to identifying homologs of medically relevant human proteins in genetically tractable model organisms, such as mice, the fruit fly Drosophila melanogaster, the nematode Caenorhabditis elegans, the yeast Saccharomyces cerevesiae, and bacteria (Banfi et al. 1996; Bassett et al. 1997; Mushegian et al. 1997). Whole-genome comparisons of microbial proteins (Koonin et al. 1997) have emphasized the importance of distinguishing orthologs, that is, proteins in two species that have evolved by vertical descent from a common ancestor and are presumed to have the same function

⁴Present address: Genos Biosciences, Inc., La Jolla, California 92037 USA. ⁵Corresponding author.

E-MAIL arcady@axyspharm.com; FAX (619) 452-6653.

(Fitch 1970), from paralogs, namely proteins derived from lineage-specific duplication and domain shuffling that hence may have more divergent functions. Failure to resolve orthologs and paralogs can lead to misinterpretation of cellular biochemistry (Tatusov et al. 1996; Henikoff et al. 1997) and inaccuracies in molecular evolutionary reconstructions (Doolittle et al. 1996; Feng et al. 1997). This distinction has been addressed in a protein domain analysis of positionally cloned human genes that are mutated in specific diseases ("disease genes") and their counterparts in the yeast and nematode genomes (Mushegian et al. 1997). Although almost equal fractions of the human disease genes had regions of significant similarity to nematode and to yeast proteins, the latter study identified a true ortholog in the complete yeast proteome for only 20% of human proteins. In contrast, 30% of human disease genes had candidate orthologs in the ~50% completed nematode proteome then available.

Drosophila and C. elegans have emerged as attractive model animal systems for studying human gene pathways, because of their genetic tractability, phenotypically well-characterized genes, and progress in whole-genome sequencing (Rubin 1996; Ahringer 1997). Traditional morphology-based phylogenies have placed C. elegans and other nematodes in a basal metazoan clade composed of pseudocoelomate animals, whereas Drosophila and other arthropods have been placed in the more recently derived protostome clade of eucoelomate animals with a shorter phylogenetic distance to vertebrates and other deuterostomes (for review, see Brusca and Brusca 1989). However, the notion that arthropods belong to a later evolutionary branch than nematodes (e.g., Sidow and Thomas 1994) has been challenged by recent studies based on analysis of morphology (Nielsen 1995), of ribosomal RNA sequences (Aguinaldo et al. 1997), and of selected protein sequences (McHugh 1997). Notably, the estimated sizes of the Drosophila and C. elegans genomes and proteomes are quite similar, on the order of 100 MB of genomic DNA and 15,000 genes (Miklos and Rubin 1996; Waterston 1997). Despite this information, the representation of proteins and conserved protein domains in these two proteomes has not been approached systematically.

The present study was designed to identify a substantial set of orthologous protein-coding genes in the eukaryotic model organisms by using strict criteria to define orthologous candidates. We then analyzed this set of proteins to assess the relative similarity of *Drosophila* and *C. elegans* proteins to their human orthologs. As a complementary approach toward the evaluation of the model organisms, we sought to estimate the fractions of conserved domains and to compare the composition of multidomain proteins in the available protein sets of *Drosophila* and *C. elegans*.

RESULTS AND DISCUSSION

Forty-Two Quartets of Candidate Orthologs

To identify all potential orthologous genes in humans, nematode, fly, and yeast, we first searched the complete *S. cerevesiae* proteome (6141 protein sequences) using as queries identified *Drosophila* proteins (2142 available sequences as of March 1, 1997, then excluding peptides shorter than 110 amino acids). The modified BLATAX program was used to tabulate the highest scoring matches for each *Drosophila* protein. The resultant 848 fly proteins were then used to search the nonredundant protein sequence (NR) database and extract the best matches from humans. The human sequences retrieved in this way were examined to remove incomplete proteins and sequences that occurred more than once as a result of being the best match for two or more Drosophila proteins. The remaining set of 480 human proteins was used to search the NR database. The best matches to these human proteins from Drosophila, C. elegans, and S. cerevesiae were then extracted by the modified BLATAX program using the similarity score measure. Next, we removed low-scoring sets and spurious hits representing matches in low-sequence complexity and coiled-coil segments. The resultant set of proteins was filtered to derive candidate orthologs by excluding (1) sequences for which the yeast sequence was closer to a human homolog than either a fly or nematode sequence (the third criterion of orthology, see Methods); and (2) proteins that shared high similarity in one domain but differed in overall domain architecture from the human protein (the second criterion of orthology). We then excluded members of expanded protein families as listed in Methods. The remaining protein quartets were subjected to a final round of filtering based on reciprocal BLASTP searches (the first criterion of orthology).

The resulting set of proteins contained 42 quartets of human, *Drosophila, C. elegans*, and *S. cerevesiae* candidate orthologs (Table 1). The 42 proteins comprising the data set of orthologs varied extensively in length, for example, the human proteins contained 116–2225 amino acids, with a median of 349 residues. This is a significantly broader spectrum of sizes than the set of 64 enzymes (151–935 residues) used in a recent large-scale phylogenetic comparison (Doolittle et al. 1996). Moreover, the present set of candidate orthologs samples many of the functional categories essential in the eukaryotic cell, including genome replication and expression, organelle structural components, and signal transduction (Table 1).

Different Proteins Generate Different Phylogenetic Tree Topologies

Most of *C. elegans* proteins in the databases, and 33 of 42 nematode proteins in current data set, are predicted from the genomic sequence, whereas all 42 of the *Drosophila* orthologs were derived from full-length cDNA sequences. Therefore, additional measures were taken to verify the orthologous candidates. First, the human protein sequences were used as queries to search the database of unfinished

Addenosylhomocysteinase 432 194184 1200230 81 134182 77 730701 74 SERCA Calcium-transporting ATPase 994 1586563 114306 80 K11D9* 79 114301 42 SODC Superoxide dismutase 154 134611 1173472 62 464769 57 134693 55	Adenina nucleotes translocator 2 9.88 1134:99 1800741 79 1914555 00 113452 50 Adenosanetriphosphalase 119 1392745 491294 71 720098 51 ATF profitase beta chain 520 114549 287945 62 116837 01 1015845 62 Catalaxe 520 11702 1708622 76 1078837 73 11550 58 COck12 101 100684 20077 63 738565 68 119533 52 Clanamato dehydroganase 1 procursor 559 119571 71 191700 1127230 02 419935 04 Clausee-barnetophala dehydroganase 515 1070933 130470 05 1321745 00 171745 55 Giveereid-sphale dehydroganase 304 170892 120497 76 107683 77 119370 119376 Civeorderived abrain 241 255051 1070084 78 1091454 <t< th=""><th>Abbreviation</th><th>Protein name</th><th>Length (aa)</th><th>Human gi#</th><th>Drosophila gi#</th><th>fiy-human % identity</th><th>C. elegans gi#</th><th>nematode-hurnan % identity</th><th>S. cerevesiae gl#</th><th>yeast-human % identity</th><th>Tes testent</th></t<>	Abbreviation	Protein name	Length (aa)	Human gi#	Drosophila gi#	fiy-human % identity	C. elegans gi#	nematode-hurnan % identity	S. cerevesiae gl#	yeast-human % identity	Tes testent
Prese Adenosinentiphosophalase 119 1362746 481294 71 780226 71 731086 51 TFB ATFP synthase beta chain 520 111640 287045 62 1168872 61 1015845 82 ATA catalace 527 115702 1706522 76 1078837 73 116703 88 DC42 CD42 101 120584 720077 63 738565 66 115615 52 Elemastion factor 2 566 115172 119770 51 1627820 52 416923 60 NOA Endiase alpha 494 119339 13951 73 113252 70 119787 55 PDA Glucese-Sphosophale deflydrogenase 351 100649 75 12053 77 119277 55 PDA Glucese-Sphosophale deflydrogenase 349 1706476 120540 75 120553 77 119277 55 PDA <td< td=""><td>Adenosinentghosphalase 119 1992/26 481/284 71 780/286 71 731/088 611 ATP synthese beta chain 520 114549 287045 62 1168372 01 1076422 76 1078437 73 115/70 68 CDC-42 101 120584 728077 73 115/70 73 115/70 68 CDC-42 101 120584 728077 73 115/70 74 115/70 74 140/70 74 140/70 <</td><td>энх</td><td>Formaldehyde dehydrogenase</td><td>374</td><td>113408</td><td>1168359</td><td>72</td><td>619738</td><td></td><td>417709</td><td>62</td><td></td></td<>	Adenosinentghosphalase 119 1992/26 481/284 71 780/286 71 731/088 611 ATP synthese beta chain 520 114549 287045 62 1168372 01 1076422 76 1078437 73 115/70 68 CDC-42 101 120584 728077 73 115/70 73 115/70 68 CDC-42 101 120584 728077 73 115/70 74 115/70 74 140/70 74 140/70 <	энх	Formaldehyde dehydrogenase	374	113408	1168359	72	619738		417709	62	
TPB ATF synthase beta shain 520 114540 287045 0.2 1168072 0.1 1016845 ea ATA Catalase 527 115702 1705022 76 1078837 73 116708 58 DC42 CDC43 101 120884 728077 63 738665 66 116515 52 LHC Clashin heavy chain 167 1705616 21811 82 461752 73 116515 52 Elongation factor 2 600 119172 119170 61 1027282 82 48 CPD Glucese-Schooghate dehydrogenase 515 1070433 194670 65 1321745 60 171845 50 ADPTH Glucese-Schooghate dehydrogenase 35 12054 120407 75 105832 74 173026 442105 50 PC Glucese-Schooghate dehydrogenase 35 120244 75 105832 74 173064 74 173025 44 <td>ATP anthase beta shain 520 114549 287945 02 1165872 01 1015845 6.2 Catalase 527 115702 1705622 76 1078337 73 115709 5.8 CDC42 101 120584 720077 9.9 738565 85 119533 80 Clathin heavy chain 1075 1705416 221811 92 461752 72 119505 53 729323 29 1 Clandmath dehydrogenes 1 prouteor 550 119170 51 126282 68 119339 113251 73 113573 119339 64 Glucase-Schotsphala Gehydrogenase 335 120540 76 12053 77 119977 65 Glycersd-Sphotsphala Gehydrogenase 342 1700264 78 1301684 74 1720247 69 Tarastational lisitation factor 2.8 schunit 31 14204 1700584 73 184724 53 11735 44 194252 64 <</td> <td>DT2</td> <td>Adenine nucleotide translocator 2</td> <td>298</td> <td>113459</td> <td>1805741</td> <td>79</td> <td>1914555</td> <td>69</td> <td>113462</td> <td>56</td> <td>ļ</td>	ATP anthase beta shain 520 114549 287945 02 1165872 01 1015845 6.2 Catalase 527 115702 1705622 76 1078337 73 115709 5.8 CDC42 101 120584 720077 9.9 738565 85 119533 80 Clathin heavy chain 1075 1705416 221811 92 461752 72 119505 53 729323 29 1 Clandmath dehydrogenes 1 prouteor 550 119170 51 126282 68 119339 113251 73 113573 119339 64 Glucase-Schotsphala Gehydrogenase 335 120540 76 12053 77 119977 65 Glycersd-Sphotsphala Gehydrogenase 342 1700264 78 1301684 74 1720247 69 Tarastational lisitation factor 2.8 schunit 31 14204 1700584 73 184724 53 11735 44 194252 64 <	DT2	Adenine nucleotide translocator 2	298	113459	1805741	79	1914555	69	113462	56	ļ
ATA Catalase 527 116700 1705622 76 1078627 73 116700 68 DC42 CDC42 101 120634 729077 93 733665 96 115933 80 LH Clathrin heavy chain 197 1705616 231611 62 441762 73 116515 52 LHB Chumante dehydrogenaet preoursor 656 115172 119705 61 1627820 92 416925 68 NOA Endase alpha 434 119333 193201 73 1133263 70 119330 64 DED Eucose-sphosphale dehydrogenase 515 107054 78 120553 77 1169787 65 PDA Eucose-sphosphale dehydrogenase 355 120640 75 120553 77 1169787 65 PDA Eucose-sphosphale dehydrogenase 355 102643 147047 82 1445256 64 ER Eukorni factor 25 eucont	Edatase 527 116700 705622 76 107837 73 116700 58 CDC42 101 120634 720077 93 73866 96 116933 90 Clahrin hewy chain 1675 1706916 231811 62 491752 73 118515 52 Elongation factor 2 656 119521 17070 1 1027820 62 119339 04 Elongation factor 2 656 119172 119170 01 1027820 62 119339 04 Glucese-Schotsothate dehydrogenase 355 120693 73 113252 70 119379 50 Glucese-Schotsothate dehydrogenase 240 1709026 84982 64 462197 50 GTP cyclohydrolase 1 241 285061 1079024 78 1301634 74 170247 69 Transferional initiation factor 4E 217 1362434 179 149175 53 11720247 69	Pase	Adenosinetriphosphatase	119	1362745	481294	71	780226	71	731098	51	<u> </u>
DC42 CDC42 101 12084 72077 63 73866 64 11633 80 LH Clathrin heavy chain 1675 1706916 231611 82 461752 73 116515 52 HE3 Channate dehydrogenee 1 procursor 556 118541 1706402 72 1340025 63 729323 28 F2 Erongelon factor 2 666 118172 11970 61 1627620 62 416933 08 RDP Biocose-Schosphals dehydrogenase 515 1070353 1384070 65 1321745 60 1119330 64 RDP Biocose-Schosphals dehydrogenase 335 120649 75 100653 67 11169767 65 APP Giocose-Schosphals dehydrogenase 349 170026 4462 47 1014316 37 124223 37 AF Fubroshits dehydrogenase 655 472307 126434 47 1014316 37 124223 37 <td>CDC42 191 120634 728077 93 735665 96 11693 90 Clathin heavy chain 1676 1706916 231811 82 461752 73 118515 52 Clatamana dehydrogenese 1 precursor 656 119521 119170 0.1 1267820 82 419935 98 Encreation factor 2 656 119172 119170 0.1 1267820 82 419935 98 Encreation factor 2 655 115339 1133311 73 1132523 70 1189767 65 Ghocenstering 4 531 1070433 3304070 65 462197 50 Ghocenstering 4 641 120640 76 120653 74 1730247 69 Translational initiation factor 4E 217 1362436 1381434 71 1194316 37 1124205 64 Manneyt-ofsoseecharises ether.1.2.mannesideare 692 6924 750019 74 1340505 66 124205<</td> <td>rpb</td> <td>ATP synthase beta chain</td> <td>529</td> <td>114549</td> <td>287945</td> <td>92</td> <td>1168572</td> <td>91</td> <td>1015845</td> <td>82</td> <td>ļ.,</td>	CDC42 191 120634 728077 93 735665 96 11693 90 Clathin heavy chain 1676 1706916 231811 82 461752 73 118515 52 Clatamana dehydrogenese 1 precursor 656 119521 119170 0.1 1267820 82 419935 98 Encreation factor 2 656 119172 119170 0.1 1267820 82 419935 98 Encreation factor 2 655 115339 1133311 73 1132523 70 1189767 65 Ghocenstering 4 531 1070433 3304070 65 462197 50 Ghocenstering 4 641 120640 76 120653 74 1730247 69 Translational initiation factor 4E 217 1362436 1381434 71 1194316 37 1124205 64 Manneyt-ofsoseecharises ether.1.2.mannesideare 692 6924 750019 74 1340505 66 124205<	rpb	ATP synthase beta chain	529	114549	287945	92	1168572	91	1015845	82	ļ.,
LH Clethrin heavy chain 1675 1706916 231611 62 461762 7.3 118615 5.2 MB3 Ghammate dehydrogenese 1 procursor 556 118641 1709402 72 1340025 63 729823 28 F2 Elongation factor 2 690 119172 119170 61 1927820 02 419836 69 NAA Enolese alpha 434 119339 118351 7.3 113253 70 119376 69 ORD Glucces 4-Schoppatis dehydrogenase 351 120491 120640 75 120653 77 119787 55 PDA Glycers/s-thosphate dehydrogenase 333 120401 170647 72 1483252 68 142055 61 2B Translational initiation factor 2.6 subunit 333 124041 170647 72 1483252 68 124205 61 44E Eukaryotio initiation factor 4E 217 1362433 1707084 72 1919459	Clerhrin heavy chain 1678 1706916 231911 82 461782 73 116515 52 Ghuamata dehydrogenses 1 procursor 556 119521 119170 611 192720 62 419995 63 Elongation factor 2 650 119172 119170 61 192720 62 419935 64 Glucose 4-shosphate dehydrogensee 515 1070433 1304070 65 1321745 00 171445 50 Glycerok-shosphate dehydrogensee 340 120803 1304070 65 142055 74 1700247 69 Glycerok-shosphate dehydrogensee 340 1708026 84982 64 142176 61 Eukayotio initiation factor 2 8ubunit 33 132644 76 1914315 37 12423 37 Memosyl-sligosechetids ekhydrogensee 354 40284 780 194255 53 41730647 69 Transteforal initiation factor 4E 217 1352431 138424 71	ATA	Catalase	527	115702	1705622	76	1078837	73	115709	58	ļ
HEB Glutamate detrydrogenese 1 precursor 568 118641 1708402 72 1940025 63 729823 28 P2 Elonastion factor 2 958 119172 119170 61 1927820 62 419935 68 NOA Enclasse alpha 434 119335 133477 51 1132583 70 113356 64 4519 50 ADPH Glucoster-trictostantal detrydrogenase 335 120649 120640 76 120853 77 1169787 65 PDA Glucoster-trictostantal detrydrogenase 340 1708026 84862 64 42118 50 PTC GTP cyclohydrofase 1 241 258061 1070884 72 1301684 74 173025 44 454 tukaryoto initiation factor 2.6 subunit 333 124204 1170497 62 1483252 68 124203 7 12423 7 14223 7 1422 7 141245 144245 146455	GAlternate dehydrogenee 1 procursor 558 119641 170402 72 1340028 63 729923 28 1 Elongation factor 2 950 119172 119170 01 1927820 82 419935 09 Enclases atoma 434 119339 119351 73 1132523 70 119330 04 GRozenidettride Schoolsphate dehydrogenase 335 120649 76 120853 77 1169787 69 GRozenidetride Schoolsphate dehydrogenase 349 1708028 84982 64 462197 50 GROzenide I 217 1952435 195644 47 1700247 69 Translational initiation factor 2.6 subunit 33 14204 117047 2 1480525 68 124223 77 Mannewic-Bigozeschetride ajbra: 1.2.manneridase 652 495253 170899 52 194724 53 417305 44 Methionine adenosyltransferase 952 402645 700019 74 1946325 <td>DC42</td> <td>CDC42</td> <td>191</td> <td>120634</td> <td>729077</td> <td>93</td> <td>733565</td> <td>86</td> <td>115933</td> <td>80</td> <td>ļ</td>	DC42	CDC42	191	120634	729077	93	733565	86	115933	80	ļ
F2 Elongation factor 2 658 119172 119170 61 1027820 62 419930 98 NOA Enolase alpha 434 119339 113331 73 1132523 70 119330 64 6PD Glucose-0-shosphalt dehydrogenase 335 120649 76 120653 77 1169787 65 CR Glycerol 3-phosphalt dehydrogenase 349 170026 64982 64 462156 58 462197 50 TPC Glycerol 3-phosphalt dehydrogenase 349 170026 4982 64 462156 58 462197 50 TPC GTP cyclohydrolase1 28 Translational inflation factor 2.6 subunit 333 124204 1101497 52 149202 64 11914316 37 124223 37 Aff Elongation factor 2.6 subunit 333 124204 17 1940326 64 124203 37 124233 37 Mannewi-bitals dehydrogename resolutin 333 <t< td=""><td>Elongation factor 2 656 119172 119170 01 1027020 02 419955 00 Enclases alpha 434 119339 119351 73 1132523 70 119330 04 Glucose-9-strosphale dehydrogenase 315 120649 76 120853 77 1169787 65 Glycerdi-3-phosphale dehydrogenase 335 120649 76 120853 77 1169787 65 GTP cyclohydrolaae 1 241 255061 107094 78 1301684 74 1730247 69 Translational ivitation factor 2 /t subunit 333 124204 117497 62 1483252 68 42205 61 Fukaryotio initistion factor 4 E 217 1362431 32434 47 1914316 37 124205 64 Cyclohrome P450 reductase 676 247307 1296517 51 69765 46 216453 33 Nucleasitid ethyosphate kinase B 152 137283 128697 22</td><td>Н</td><td>Clathrin heavy chain</td><td>1675</td><td>1705916</td><td>231811</td><td>82</td><td>461752</td><td>73</td><td>118615</td><td>52</td><td>Į</td></t<>	Elongation factor 2 656 119172 119170 01 1027020 02 419955 00 Enclases alpha 434 119339 119351 73 1132523 70 119330 04 Glucose-9-strosphale dehydrogenase 315 120649 76 120853 77 1169787 65 Glycerdi-3-phosphale dehydrogenase 335 120649 76 120853 77 1169787 65 GTP cyclohydrolaae 1 241 255061 107094 78 1301684 74 1730247 69 Translational ivitation factor 2 /t subunit 333 124204 117497 62 1483252 68 42205 61 Fukaryotio initistion factor 4 E 217 1362431 32434 47 1914316 37 124205 64 Cyclohrome P450 reductase 676 247307 1296517 51 69765 46 216453 33 Nucleasitid ethyosphate kinase B 152 137283 128697 22	Н	Clathrin heavy chain	1675	1705916	231811	82	461752	73	118615	52	Į
NOA Enclase alpha 434 119339 119331 73 1132223 70 119330 04 GPD Glucose-9-phosphala dehydrogenase 515 1070433 1304670 05 1321745 00 171545 50 ADPH Glucose-9-phosphale dehydrogenase 335 120649 76 120653 77 1169787 65 PDA Glucose-9-phosphale dehydrogenase 349 1708025 84982 64 462196 58 462197 50 Translational initiation factor 2.6 subunit 333 124204 1171447 82 1483262 66 124205 61 24E Eukaryotic initiation factor 2.6 subunit 333 124204 17 1914316 37 124223 37 A12 Mannozvi-digeseotharide alpha 1.4 mannosidese 655 452563 1706899 62 1994724 53 417505 44 ETK Methonine adenosytransferase 395 400245 730019 74 1706586 62	Enclase alpha 434 119339 119321 73 113223 70 119330 64 Glucose 9-chosphale dehydrogenase 515 1070433 1304070 65 1321745 60 171545 50 Glycercl-3-phosphale dehydrogenase 335 120649 120640 76 120533 77 1169787 55 Glycercl-3-phosphale dehydrogenase 340 1708026 84982 64 462197 50 GTP cyclohydrolase1 231 124204 1708027 1904336 72 14433252 66 124205 61 Eukaryotio initiation factor 2.6 subunit 233 124204 170497 62 1443325 147233 147233 147233 147233 147233 147233 73 14223 77 142625 66 124205 64 145233 33 147264 53 147265 44 14525 64 214523 33 1 Methionine adonosyltraneferase 675 247307 126817	HE3	Glutamate dehydrogenase 1 preoursor	558	118641	1706402	72	1340025	63	729323	28	1
BPD Glucose-5-phosphale dehydrogenase 515 1070433 1304070 65 1321745 60 171545 50 ADPH Glucerol-3-phosphale dehydrogenase 335 120649 120640 76 120653 77 1169787 68 PDA Glucerol-3-phosphale dehydrogenase 340 1708026 84982 64 462195 58 462197 50 TPC GTP cyclohydrolase I 241 255061 1079084 78 1301684 74 1730247 69 288 Translational initiation factor 2.6 subunit 333 124204 117.0497 82 1483262 66 124205 61 44E Eukaryotic initiation factor 4.E 217 126435 1706899 62 1894724 53 417305 44 ETK Methionine adenosytransferase 595 247307 1286517 51 68765 217305 78 152319 69 549341 61 RML NNAMF 2 126807 <td>Glucose-6-shosphale.dethydrogenase 515 1070433 1304070 65 1321745 00 171144 50 Glucose-6-shosphale.dethydrogenase 335 120649 120640 75 120653 77 1169787 65 Glycerol-3-phosphale.dethydrogenase 349 1708026 84982 64 462196 58 462197 50 GTP cyclohydrolase I 241 255061 1070047 82 148362 66 124205 61 Evkaryotic initiation factor 2.6 subunit 33 124204 1170497 82 148362 66 124205 61 Evkaryotic initiation factor 2.6 subunit 335 400245 730019 74 194316 37 124223 37 Methicine adenosytransferase 395 400245 730019 74 196986 76 2184333 3 NUcleositid diphosphale kinase 152 127983 127980 78 122817 51 687765 22 100897 30 1266857</td> <td>2</td> <td>Elongation factor 2</td> <td>858</td> <td>119172</td> <td>119170</td> <td>81</td> <td>1627820</td> <td>82</td> <td>416935</td> <td><u>08</u></td> <td></td>	Glucose-6-shosphale.dethydrogenase 515 1070433 1304070 65 1321745 00 171144 50 Glucose-6-shosphale.dethydrogenase 335 120649 120640 75 120653 77 1169787 65 Glycerol-3-phosphale.dethydrogenase 349 1708026 84982 64 462196 58 462197 50 GTP cyclohydrolase I 241 255061 1070047 82 148362 66 124205 61 Evkaryotic initiation factor 2.6 subunit 33 124204 1170497 82 148362 66 124205 61 Evkaryotic initiation factor 2.6 subunit 335 400245 730019 74 194316 37 124223 37 Methicine adenosytransferase 395 400245 730019 74 196986 76 2184333 3 NUcleositid diphosphale kinase 152 127983 127980 78 122817 51 687765 22 100897 30 1266857	2	Elongation factor 2	858	119172	119170	81	1627820	82	416935	<u>08</u>	
ADP: Giveeral/detuxide 2-ph/osphate. detux/crogenase. 335 120649 76 120653 77 1169787 65 PDA Giveerol-3-phosphate. detux/crogenase. 349 1708026 84982 64 462195 58 462197 50 TTPC GTP cyclohydrolase I 241 255061 1070054 78 1301684 74 1730247 69 28B Tranelational initiation factor 2 B subunit 333 124204 1170497 82 1483252 66 12423 37 M12 Memoart-oligosecoharide alpha-1.2,mannosidase 925 492543 1708999 62 1994724 53 417305 44 ETK Methionine adenosyltransierase 395 400245 730019 74 1708996 74 134652 56 CPD Cytochrome P450 reductase 676 247307 1286517 51 5874 52 136383 32 DKB Nucleocide diphosphate kinase B 152 127980 78 1403171 </td <td>Gikceraldehyde 3-phosphale dehydrogenase 335. 120649 76 120653 77 1169787 65 Gikcerol 3-phosphale dehydrogenase 349 1708026 84982 64 462195 68 462197 60 GTP cyclohydrolase I 241 256061 1079084 78 1301684 74 1730247 69 Translational initiation factor 2 B aubunit 333 124204 1170497 82 1483252 66 124205 61 Eukaryotic initiation factor 4 E 217 1362434 47 1914316 37 124223 37 Methionine adenosyttransferase 395 400245 730019 74 1709986 74 194525 69 Cytochrome P450 reductase 676 247307 1298517 51 697655 46 218433 33 Nucleoside diphosphate kinase B 152 127983 127980 78 122919 69 548541 61 NFAMP 2 504 1325232 1246972 <</td> <td>NOA</td> <td>Enclase alpha</td> <td>434</td> <td>119339</td> <td>119351</td> <td></td> <td>1132523</td> <td></td> <td>119336</td> <td>64</td> <td></td>	Gikceraldehyde 3-phosphale dehydrogenase 335. 120649 76 120653 77 1169787 65 Gikcerol 3-phosphale dehydrogenase 349 1708026 84982 64 462195 68 462197 60 GTP cyclohydrolase I 241 256061 1079084 78 1301684 74 1730247 69 Translational initiation factor 2 B aubunit 333 124204 1170497 82 1483252 66 124205 61 Eukaryotic initiation factor 4 E 217 1362434 47 1914316 37 124223 37 Methionine adenosyttransferase 395 400245 730019 74 1709986 74 194525 69 Cytochrome P450 reductase 676 247307 1298517 51 697655 46 218433 33 Nucleoside diphosphate kinase B 152 127983 127980 78 122919 69 548541 61 NFAMP 2 504 1325232 1246972 <	NOA	Enclase alpha	434	119339	119351		1132523		119336	64	
EDA Giveend-3-phosphate dehydrogenase 949 1708026 84982 64 462195 58 462197 50 TPC GTP cyclohydrolasa I 241 255661 1070084 78 1301684 74 1730247 69 2B Translational initiation factor 2.8 subunit 333 124204 1170497 82 1483252 66 124205 61 4E Eukaryotic initiation factor 2.8 subunit 333 124204 170497 82 1483252 66 124205 61 4E Eukaryotic initiation factor 2.8 subunit 333 124204 170899 62 1994724 53 417305 44 ETK Methionine adenosyltranslerase 676 247307 1286517 51 82785 46 219453 33 DKB Nucleoside diphosphate kinase B 152 127980 78 1403171 65 130860 61 RM2 NTAMF 2 504 132523 134600 730372 78 1	Skycerck3-phosphate dehydrogenase 349 1708026 84982 64 462196 58 462197 50 GTP cyclohydrolase I 241 256061 1070084 78 1301684 74 1720247 69 Translational initiation factor 2 B subunit 333 124204 1170497 92 1483252 66 124205 61 Eukaryotic initiation factor 4E 217 1352434 47 1914316 37 124223 37 Mannexi-sitgesaccharide alpha-1.2,mannoaldase 525 492553 1703899 52 1994724 53 417305 44 Methionine adenosyltranslerase 576 247307 1296517 51 687654 2184533 33 Nucleoside diphosphatia kinase B 152 127980 78 152319 69 548341 61 NTAMP 2 504 1325223 1248509 60 746574 52 136388 32 1 Proteasome component C3 244 120580 730372	6PD	Glucose-8-phosphate dehydrogenase	515	1070433	1304670	65	1321745	60	171545	50	1
PDA Glycerol-3-phosphate dehydrogenase 349 1708026 84982 64 462195 58 462197 50 TPC GTP cyclohydrolasa I 241 285061 1078084 78 1301684 74 1730247 69 2B Translational initiation factor 2.8 aubunit 333 124204 1170497 82 1483252 66 124205 51 4E Eukaryotic initiation factor 2.8 aubunit 333 124204 170497 82 1493252 66 124205 51 4E Eukaryotic initiation factor 2.8 aubunit 335 400245 730019 74 170898 74 134652 66 CPR Cytochrome P450 reductase 676 247307 1286517 51 68765 46 21853 33 DKB Nucleoside diphosphate kinase B 152 127983 127980 78 1523919 69 548341 61 RM2 NTAMP 2 504 135252 1346509 60 746574	Glycerol-3-phosphate dehydrogenase 349 1708026 84982 64 462195 55 462197 50 GTP cyclohydrolasa I 241 285061 1070084 78 1301684 74 1730247 69 Translational initiation factor 2 B subunit 333 124204 1170497 52 1483252 68 124223 37 Mannezvi-oligosaccharids alpha-1.2,mannosidase 595 498253 1708939 52 1894724 53 417305 44 Methionine adenoxyltranslerase 395 400245 730019 74 1708936 74 1945925 69. Cytochrome P450 reductase 676 247307 1298517 51 687865 462 124533 33 Nucleoside diphosphatis kinase B 152 127893 127980 78 1523919 69 548341 61 NTAMF 2 504 1352523 1348509 60 746574 52 136388 32 126687 32 126687 32	ADPH	Givceraldehyde 3-phosphate dehydrogenase	335	120649	120640	76	120653		1169787	65	<u>.</u>
28 Translational initiation factor 2 ß subunit 933 124204 1170497 52 1483252 66 124205 61 44 Eukaryotic initiation factor 4E 217 1352435 1352434 47 1914316 37 124223 37 A12 Mannozyt-oligosaccharide alpha-1.8,mennoridase 926 492563 170899 62 1994724 53 417305 44 ETK Mathionine aderocytransferase 936 400245 780019 74 1706996 74 1346525 66 CPR Cytochrome P450 reductase 676 247307 1296517 51 69765 46 218453 33 DKB Nucleoside diphosphate kinase B 152 127985 1282917 78 1523919 69 548341 61 FMZ NFLAMF 2 504 1325232 134800 60 746574 52 136388 32 FMZ Protessome component C3 294 1300650 730372 78 1403	Translational initiation factor 2 8 subunit 333 124204 1170497 82 1483262 68 124205 61 Eukaryotic initiation factor 4E 217 1352434 1352434 47 1914316 37 124223 37 Mennoeut-oligosescohenise alpha-1.2,mennoeidase 985 .492853 1708899 62 1894724 53 417305 44 Methionine adenosultransferaze .985 .402845 730019 74 1706986 74 134525 66 Cytochrome P450 reductase .676 .247307 1226617 51 .697865 46 218453 33 Nucleositide diphosphate kinase B 162 127883 127880 78 1523914 65 543341 61 INFLAMP 2 504 1352523 1246609 60 746574 52 136388 32 GAP-associated tyrbeine phosphorotein p62 443 420044 1622930 40 1047005 50 1256657 32 1256877 32 <tr< td=""><td>PDA</td><td></td><td>349</td><td>1708026</td><td>84982</td><td>64</td><td>462196</td><td>58</td><td>462197</td><td>50</td><td></td></tr<>	PDA		349	1708026	84982	64	462196	58	462197	50	
Het Eukaryolio initiation factor 4E 217 1962435 1962434 47 1914316 37 124223 37 A12 Mannseyt-bilgosaccheride.siphe-1,8,manneaidase 585 492583 1708899 62 1994724 53 417305 44 ETK Methionine adenoxyltransferase 395 400245 730019 74 1706998 74 1346525 60 CPR Cylochrome P450 reductase 676 247307 1296517 51 697665 46 218453 33 DKB Nucleosidie diphosphate kinase B 152 127983 127880 78 1323919 69 548341 61 RMZ NRAMP 2 504 1352523 1346609 60 746574 52 1363888 32 RG2 GAP-associated tyrosine phosphoprotein pS2 443 420044 1622030 40 1947005 39 1256657 32 RG3 Proteasome component C3 234 130650 730372 78 <td< td=""><td>Eukaryotis initiation factor 4E 217 1352434 47 1914316 37 124223 37 Mannesyl-eligosascheride alpha-1,2,mannesidase 925 492583 1708899 62 1994724 53 417305 44 Methionine adenosyltransferase 395 400245 730019 74 1709396 74 1346525 66 Cytochrome P450 reductase 676 247307 1296517 51 62785 46 218453 33 Nucleosidie diphosphate kinase B 152 127980 78 1523919 69 548341 61 NFAMP 2 504 1352823 1346609 60 746574 52 1363883 32 GAP-associated hyrosine phosphorotein p52 443 420044 1622930 40 194705 91 125667 32 Proteasome cents chain 241 464456 149859 72 152914 65 1709764 65 DNA Primase 49kDa subunit 420 1346722 666989</td><td>TPC</td><td>GTP cyclohydrolase i</td><td>241</td><td>255061</td><td>1079084</td><td>78</td><td>1301684</td><td>74</td><td>1730247</td><td>69</td><td>1</td></td<>	Eukaryotis initiation factor 4E 217 1352434 47 1914316 37 124223 37 Mannesyl-eligosascheride alpha-1,2,mannesidase 925 492583 1708899 62 1994724 53 417305 44 Methionine adenosyltransferase 395 400245 730019 74 1709396 74 1346525 66 Cytochrome P450 reductase 676 247307 1296517 51 62785 46 218453 33 Nucleosidie diphosphate kinase B 152 127980 78 1523919 69 548341 61 NFAMP 2 504 1352823 1346609 60 746574 52 1363883 32 GAP-associated hyrosine phosphorotein p52 443 420044 1622930 40 194705 91 125667 32 Proteasome cents chain 241 464456 149859 72 152914 65 1709764 65 DNA Primase 49kDa subunit 420 1346722 666989	TPC	GTP cyclohydrolase i	241	255061	1079084	78	1301684	74	1730247	69	1
A12 Mannosvi-oligosaccharide.elpha-1.8.mannosidase 685 492583 1708939 62 1994724 53 417305 44 ETK Methionine adenosyliransferase 395 400245 730019 74 1705936 74 1346525 68 CPR Cylochrome P450 reductase 676 247307 1296517 51 697665 46 218453 33 DKB Nucleoside diphosphate kinase B 152 127983 127880 78 1232919 69 548341 61 RMZ NRAMP 2 504 1332523 1346609 60 740705 39 1256857 32 RG2 GAP-associated hyrosine phosphoprotein pS2 443 420044 1622930 40 1047005 39 1256857 32 RG3 Proteasome component C3 234 130650 73072 78 1403171 65 130690 61 RG2 Proteasome zeta chain 241 464456 149859 73 152914 65 170774 65 RI11 DNA Primase 49kDa suburit <	Manneavi-oligosescharide sibha: 1,2,manneaidase 925 492553 1708999 92 1994724 53 417305 44 Methionine adenoxitransferase 395 400245 730019 74 1708996 74 1346525 66 Cvtochrome P450 reductase 676 247307 1296517 51 687865 46 218453 33 Nucleoside diphosphats kinase B 152 127980 127980 78 1523919 69 548341 61 NRAMP 2 504 1352523 1346609 60 746574 52 135388 32 GAP-associated tyrosine phosphoprotein pS2 443 420044 1622930 40 1947005 30 1256657 32 Protessome component C3 234 130860 730372 78 1403171 65 1709764 65 DNA Primase 49kDa subunit 420 1346722 666999 42 464461 39 730382 31 CAD protein 7001 1148	28	Translational initiation factor 2 ß subunit	333	124204	1170497	82	1483252	66	124205	61	
ETK Methionine adenosyltransferase 395 400245 730019 74 1706996 74 1345025 06 CPR Cytochrome P450 reductase 676 247307 1296517 51 667665 46 218453 33 DKB Nucleoside diphosphate kinase B 152 127983 127980 78 1523919 69 548341 61 FMZ NFLAMP 2 504 13252523 1346609 60 746574 52 1363888 32 GAP-associated tyroeine phosphoprotein p62 443 420044 1622930 40 1947005 30 1256657 32 RC3 Protessome component C3 234 130650 730372 76 1403171 65 1309680 61 RC2 Protessome component C3 234 13045792 666989 42 464461 39 730382 31 YR1 CAD protein fact Achar 1418 410219 2121267 56 1419476 41	Methionine adenosyltransferase 395 400245 730019 74 1706996 74 1346525 69 Cytochrome P450 reductase 676 247307 1296517 51 697655 46 219153 33 Nucleoside diphosphate kinese B 152 127980 72980 78 1523819 69 548341 61 NFAMP 2 504 1352523 1346609 60 746574 52 1363888 32 GAP-associated tyrosine phosphoprotein p52 443 420044 1622930 40 1947005 30 1256867 32 Proteasome component C3 234 130660 73072 78 1403171 66 1709764 65 DNA Primase 45kDa subunit 420 1346792 666989 42 464461 39 730382 31 CAD protein 2225 1709855 131992 63 1009053 63 173146 60 Beplication factor C large subunit 1148 410218 21212	4E	Eukaryotic initiation factor 4E	217	1362436	1352434	47	1914316	37	124223	37	
ETK Methionine adenosyltransferase 398 400245 730019 74 1708996 74 1946525 66 CPR Cytochrome P450 reductase 676 247307 1296517 51 667665 46 219453 33 DKB Nucleoside diphosphate kinase B 152 127983 127980 78 1523919 69 548341 61 RMZ NFLAMP 2 504 13252523 1346909 60 746574 52 1363888 32 RC3 GAP-associated hyrosine phosphoprotein p62 443 420044 1622930 40 1647005 30 1256857 32 RC3 Proteasome component C3 234 130650 730372 78 1403171 65 130690 61 1709764 68 RC4 Proteasome zeta chain 241 464456 149858 73 1523914 65 1709764 68 R11 DNA Primase 45kDa aubunit 120 1346792 666989 42 <td>Methionine adenosyltransferase 395 400245 730019 74 1706996 74 1346525 66 Cytochrome P450 reductase 676 247307 1296517 51 687665 46 218453 33 1 Nucleoside diphosphate kinase B 152 127983 127980 78 1523919 69 548341 61 NRAMP 2 504 1352523 1346600 60 746574 52 1363888 32 GAP-associated tyrosine phosphoprotein p52 443 420044 1622930 40 1947005 39 1256857 32 Proteasome component C3 234 130650 730372 78 1403171 65 130680 61 Proteasome component C3 234 130650 730372 78 1403171 65 130680 63 1709764 65 DNA Primase 45kDa subunit 420 1346792 665989 42 46461 39 730382 31 64 GOS Fibos</td> <td>A12</td> <td>Mannosyl-oligosaccharide alpha-1.2,mannosidase</td> <td>625</td> <td>462563</td> <td>1708899</td> <td>62</td> <td>1694724</td> <td>53</td> <td>417305</td> <td>44</td> <td></td>	Methionine adenosyltransferase 395 400245 730019 74 1706996 74 1346525 66 Cytochrome P450 reductase 676 247307 1296517 51 687665 46 218453 33 1 Nucleoside diphosphate kinase B 152 127983 127980 78 1523919 69 548341 61 NRAMP 2 504 1352523 1346600 60 746574 52 1363888 32 GAP-associated tyrosine phosphoprotein p52 443 420044 1622930 40 1947005 39 1256857 32 Proteasome component C3 234 130650 730372 78 1403171 65 130680 61 Proteasome component C3 234 130650 730372 78 1403171 65 130680 63 1709764 65 DNA Primase 45kDa subunit 420 1346792 665989 42 46461 39 730382 31 64 GOS Fibos	A12	Mannosyl-oligosaccharide alpha-1.2,mannosidase	625	462563	1708899	62	1694724	53	417305	44	
CPR Cytochrome P450 reductase 676 247307 1296517 51 687665 46 218453 33 DKB Nucleoside diphosphate kinase B 152 127983 127980 78 152919 69 548341 61 FM2 NFAMP 2 504 1352523 1346609 60 746574 52 1363888 32 62 GAP-associated tyrosine phosphoptolein p62 443 420044 1622930 40 1947005 39 1256857 32 RG2 Proteasome component C3 234 130650 730372 78 1403171 65 130680 61 RG2 Proteasome zeta chain 241 464456 1498589 42 464461 38 730382 31 YP1 CAD protein 2225 1709855 131692 63 1008053 63 173146 60 FC Replication factor C large subunit 1148 410218 212267 56 1418479 41	Cytochrome P450 reductase 676 247307 1296517 51 687865 46 218453 33 1 Nucleoside diphosphate kinase B 152 127980 78 1523919 69 548341 61 NFIAMP 2 504 1352523 1246609 60 746574 52 1363888 32 GAP-associated tyrosine phosphoprotein p52 443 420044 1622930 40 1047005 39 1256657 32 Protessome component C3 234 130650 730372 78 1403171 65 130680 61 Protessome component C3 234 130650 730372 78 1403171 65 130680 61 Protessome zeta chain 241 464456 1498689 73 1623914 65 1706764 65 DNA Primese 49kDa subunit 420 1346792 666989 42 464461 99 730382 31 60 Replication factor C large subunit 1148 410218 <td>ЕТК</td> <td></td> <td>395</td> <td>400245</td> <td>730019</td> <td>74</td> <td>1708998</td> <td>74</td> <td>i</td> <td>68</td> <td></td>	ЕТК		395	400245	730019	74	1708998	74	i	68	
DKB Nucleoside diphosphate kinase B 152 127983 127980 78 152319 69 548341 61 RM2 NFAMP 2 504 1352523 1346609 60 746574 52 1363888 32 62 GAP-associated tyrosine phosphoprotein p62 443 420044 1622930 40 1947005 30 1256857 32 RG3 Protessome component C3 234 130650 730372 78 1403171 55 130680 61 RG2 Protessome component C3 241 454456 1498589 73 1523914 65 1709764 65 RIC DNA Primase 49kDa subunit 420 1346792 666989 42 464461 39 730362 31 YR1 CAD protein factor C large subunit 1148 410218 2121267 56 1418476 41 584399 44 L17 065 Ribosomal protein L17 140 269527 1350673 86 1350671 </td <td>Nucleoside diphosphate kinase B 152 127980 78 1523919 69 548341 61 NRAMP 2 504 1352523 1346609 60 746574 52 1363888 32 GAP-associated tyrosine phosphoprotein p62 443 420044 1622930 40 1947005 39 1256857 32 Protessome component C3 234 130850 730372 78 1403171 65 130860 61 Protessome zeta chain 241 464456 1498589 73 152314 65 1709764 65 DNA Primase 49kDa subunit 420 1346792 666989 42 464461 39 730352 31 CAD protein 1actor C large subunit 1148 410218 2121267 56 1418478 41 58499 44 60S Ribosomal protein L17 140 269927 1350673 86 1350671 55 132744 79 60S Ribosomal protein L22 128 464628 1633049</td> <td></td> <td></td> <td>676</td> <td></td> <td></td> <td>51</td> <td>687865</td> <td>46</td> <td>218453</td> <td>33</td> <td>Ī</td>	Nucleoside diphosphate kinase B 152 127980 78 1523919 69 548341 61 NRAMP 2 504 1352523 1346609 60 746574 52 1363888 32 GAP-associated tyrosine phosphoprotein p62 443 420044 1622930 40 1947005 39 1256857 32 Protessome component C3 234 130850 730372 78 1403171 65 130860 61 Protessome zeta chain 241 464456 1498589 73 152314 65 1709764 65 DNA Primase 49kDa subunit 420 1346792 666989 42 464461 39 730352 31 CAD protein 1actor C large subunit 1148 410218 2121267 56 1418478 41 58499 44 60S Ribosomal protein L17 140 269927 1350673 86 1350671 55 132744 79 60S Ribosomal protein L22 128 464628 1633049			676			51	687865	46	218453	33	Ī
RMA NFLAMP 2 504 1352523 1346609 60 746574 52 1363888 32 B2 GAP-associated lyroeine phosphoprotein p62 443 420044 1622930 40 1947005 39 1256857 32 BC3 Proteasome component C3 234 130950 730372 78 1403171 65 130960 61 RC3 Proteasome zeta chain 241 464456 1498589 73 1523814 65 1709764 65 R11 DNA Primese 49kDa subunit 420 1346792 666989 42 464461 39 730382 31 YR1 CAD protein 2225 1709955 131692 63 1009053 83 173146 60 FC Replication factor C large subunit 1148 410218 212267 56 1418475 41 584899 44 L17 605 Filbosomal protein L17 140 265927 1350673 86 1350671 85<	NRAMP 2 504 1352523 1346609 60 746574 52 1363989 32 GAP-associated tyrosine phosphoprotein p62 443 420044 1622930 40 1047005 30 1256857 32 Protessome component C3 234 130650 730372 78 1403171 65 130690 61 Protessome zeta chain 241 464456 1496589 73 1523014 65 1709764 65 DNA Primase 49kDa subunit 420 1346792 666989 42 464461 39 730382 31 CAD protein 2225 1709955 131692 63 1008053 63 173146 60 Replication factor C large subunit 1149 410218 2121267 58 1350671 85 132744 79 60S Ribosomal protein L12 128 464628 1633049 58 1710514 57 171806 55 RNA polymerase II, largest subunit 1970 133224 79 <td></td> <td></td> <td>1</td> <td>Ì</td> <td>ĺ</td> <td></td> <td></td> <td></td> <td>1</td> <td></td> <td>I</td>			1	Ì	ĺ				1		I
B2 GAP-associated tyroeine phosphoprotein p62 443 420044 1622930 40 1947005 39 1256857 32 RG3 Protessome component C3 234 130850 790372 78 1403171 65 130860 61 RG2 Protessome sets chain 241 464456 1408599 73 1522914 65 1709764 65 RI1 DNA Primase 46kDa subunit 420 1346792 666989 42 464461 39 730382 31 YR1 CAD protein 2225 1709955 131692 63 1008053 63 173146 60 FC Replication factor C large subunit 1148 410218 212867 84 1350573 86 1350571 85 132744 79 L22 605 Ribosomal protein L22 128 464528 1633049 58 1710514 57 171905 55 PB1 RNA polymerase II, largest subunit 1970 133226 133324	GAP-associated tyrosine phosphoprotein p52 443 420044 1622930 40 1947005 39 1256857 32 Protessome component G3 234 130850 730372 78 1403171 65 130860 61 Protessome component G3 241 464456 1498589 73 152314 65 1709764 65 DNA Primase 45kDa subunit 420 1346792 666989 42 464461 39 730382 31 CAD protein 2225 1709955 131692 63 1008063 63 173148 60 Replication factor C large subunit 1149 410218 2121267 58 1350671 85 132744 79 60S Ribosomal protein L17 140 269527 1350673 88 1350671 85 132744 79 60S Ribosomal protein L22 128 464528 1633049 58 1710514 57 171805 5 RNA polymerase II, larget subunit 1970 133226			1	1	[1		
RG3 Protessome component C3 234 130850 730372 76 1403171 65 130860 61 RG2 Protessome zeta chain 241 464456 1499599 73 1523914 65 1709764 65 RI1 DNA Primase 49kDa subunit 420 1346792 666989 42 464461 39 730362 31 YR1 CAD protein 2225 1709955 131692 63 1008053 63 173146 60 FC Replication factor C large subunit 1148 410218 2121267 56 1418478 41 584399 44 L17 605 Ribosomal protein L17 140 265827 1350673 86 1350671 85 132744 79 L22 605 Ribosomal protein L22 128 464628 1633049 58 1710514 57 171806 55 PB1 RNA colymerase II, largest subunit 1970 133326 133324 79 1525915 73	Protessome component G3 234 130860 730372 78 1403171 56 130860 61 Protessome cata chain 241 464456 1499589 73 1523914 65 1709764 65 DNA Primase 45kDa subunit 420 1346792 666989 42 464461 39 730382 31 CAD protein 2225 1709955 131692 63 1008053 63 173146 60 Replication factor C large subunit 1144 410218 2121267 56 1418478 41 584999 44 60S Ribosomal protein L17 140 266927 1350873 86 1350671 55 132744 79 60S Ribosomal protein L22 128 464628 1633049 58 1710514 57 1710538 49 60S Ribosomal protein S12 132 133742 113997 69 1523915 67 1710538 49 60S Ribosomal protein S12 132 133742 1173185			-								1
RCZ Proteaseme zeta chain 241 464456 1498589 73 1523014 66 1709764 66 RUI DNA Primase 49kDa subunit 420 1346792 666989 42 464461 39 73362 31 YR1 CAD protein 2225 1709955 131692 63 1006053 63 173146 60 FC Replication factor C large subunit 1148 410218 2121267 58 1418476 41 584939 44 L17 605 Ribosomal protein L17 140 265927 1350673 86 1350671 85 132744 79 L22 605 Ribosomal protein L22 128 464628 1633049 58 1710514 57 171806 55 PB1 RNA polymerase II, largest subunit 1970 13326 133324 70 1255805 73 133330 54 S12 405 Ribosomal protein S12 132 13712 1173185 63 1350025 68	Protessome zeta chain 241 464456 1498589 73 152914 66 1709764 65 DNA Primase 49kDa subunit 420 1346792 666999 42 464461 39 730382 31 CAD protein 2225 1709955 131692 63 1008053 63 173146 60 Replication factor C large subunit 1149 410218 2121267 50 141847B 41 594899 44 605 Filosomal protein L17 140 266927 1350673 86 1350671 85 132744 79 605 Filosomal protein L122 128 464528 153349 58 1710514 57 171805 55 RNA polymerase II, larget subunit 1970 13326 133324 79 1256905 73 133300 54 405 Filosomal protein S12 132 133742 1173185 63 1350927 58 1350929 62 405 Filosomal protein S1 204 1172267 120805<				1	1		1		1		T
PH1 DNA Primese 49kDa subunit 420 1346792 666989 42 464461 39 730382 31 YR1 CAD protein 2225 1709955 131692 63 1008053 63 173146 60 FC Replication factor C large subunit 1148 410218 2121267 58 1418478 41 584899 44 L17 60S Ribosomal protein L17 140 266927 1350673 36 1350671 35 132744 79 L22 60S Ribosomal protein L22 128 464628 1633049 58 1710514 57 1710538 49 LA0 60S Ribosomal protein P0 317 133041 113987 69 1523915 67 171806 55 PB1 RNA polymerase IL largest subunit 1970 133326 133324 79 125895 73 133330 54 S12 40S Ribosomal protein S12 132 133742 1173185 63 1350025 58	DNA Primase 49kDa subunit 420 1346792 666989 42 444461 99 730382 31 CAD protein 2225 1709955 131692 63 1008053 63 173146 60 Replication factor C large subunit 1148 410218 2121267 58 1418478 41 584899 44 605 Ribosomal protein L17 140 266927 1350673 26 1350671 85 132744 79 605 Ribosomal protein L122 128 464628 1633049 58 1710514 57 171806 55 RNA polymerase II, largest subunit 1970 133264 133324 79 1256905 73 133300 64 405 Ribosomal protein S12 132 133742 1173185 63 135092 58 1360929 62 405 Ribosomal protein S1 204 117287 1203905 88 1351001 85 1173269 71 405 Ribosomal protein S5 204 1173267 1											T
YR1 CAD protein 2225 1709955 131692 63 1008053 63 173146 60 FC Replication factor C large subunit 1148 410218 2121267 56 1418478 41 584899 44 L17 605 Ribosomal protein L17 140 266927 1350673 86 1350671 85 132744 79 L22 605 Ribosomal protein L22 128 464628 1633049 58 1710514 57 1710538 49 LA0 605 Ribosomal protein L22 128 464628 1633049 58 1710514 57 171806 55 PB1 RNA polymerase II, largest subunit 1970 133326 133324 79 125895 73 133330 54 S12 405 Ribosomal protein S12 132 133742 1173185 63 1950925 58 1250029 52 S3 405 Ribosomal protein S3 249 417719 548856 87 1350897 9	CAD protein 2225 1709955 131692 63 1008053 63 173146 60 Replication factor C large subunit 1148 410218 2121267 56 1418475 41 594899 44 605 Ribosomal protein L17 140 266927 1350673 86 1350671 85 132744 79 605 Ribosomal protein L22 128 464528 1633049 58 1710514 57 1710638 49 605 Ribosomal protein L22 128 464528 1633049 58 1710514 57 171806 55 RNA polymerase II, largest subunit 1970 133326 133324 79 1255805 73 133300 54 405 Ribosomal protein S12 132 133742 1173185 63 1350925 58 1360929 52 405 Ribosomal protein S3 243 417719 548856 87 1350989 79 1173255 71 405 Ribosomal protein S5 204 1173267			1		i		1		1		Ī
FC Replication factor C large subunit 1148 410218 2121267 56 1418478 41 584899 44 L17 605 Ribosomal protein L17 140 266927 1350673 86 1350671 85 132744 79 L22 605 Ribosomal protein L22 128 464628 1633049 58 1710514 57 1710538 49 LA0 605 Ribosomal protein L22 128 464628 1633049 58 1710514 57 171806 55 PB1 RNA colvmerase II, largest subunit 1970 133326 133324 79 1256805 73 133330 54 S12 405 Ribosomal protein S12 132 132 13712 1173185 63 1350925 58 1350929 52 S3 405 Ribosomal protein S3 243 417719 548856 87 1350929 62 133 1173265 71 S6 406 Ribosomal protein S5 204 1173267 1203906 <	Replication factor C large subunit 1146 410218 2121267 56 1418476 41 594899 44 605 Ribosomal protein L17 140 266927 1350673 86 1350671 85 132744 79 605 Ribosomal protein L22 128 464628 1633049 58 1710514 57 1710538 49 605 Ribosomal protein L22 128 464628 1633049 58 1710514 57 1710638 49 605 Ribosomal protein L22 128 464628 133041 113997 69 1523915 67 171806 55 605 Ribosomal protein P0 317 13324 113987 63 1350925 58 1350929 52 405 Ribosomal protein S12 132 133742 1173185 63 1350927 1173255 71 405 Ribosomal protein S5 204 1173267 1203905 88 1351001 85 1173256 71 Adencay/homocysteinase 432 194164							1		1		T
L17 90S Ribosomal protein L17 140 269927 1350673 86 1350671 85 132744 79 L22 60S Ribosomal protein L22 128 464628 1633049 58 1710514 57 1710538 49 LA0 60S Ribosomal protein L22 128 464628 1633049 58 1710514 57 1710538 49 LA0 60S Ribosomal protein L22 127 132041 113987 69 1523915 67 171806 55 PB1 RNA polymerase II, largest subunit 1970 133326 133324 79 1256905 73 133330 64 \$12 40S Ribosomal protein S12 132 133742 1173185 63 1350925 68 1350925 71 \$55 40S Ribosomal protein S3 244 417719 548856 87 125099 79 1173265 71 \$55 40S Ribosomal protein S5 204 1172827 1203905 88 1351001 8	60S Ribosomal protein L17 140 269927 1350673 86 1350671 85 132744 79 60S Ribosomal protein L22 128 464628 1633049 58 1710514 57 1710538 49 60S Ribosomal protein L22 128 464628 1633049 58 1710514 57 1710538 49 60S Ribosomal protein P0 317 133041 113987 69 1523915 67 171806 55 RNA polymerase II, largest subunit 1970 13326 133242 79 1255805 73 133300 64 40S Ribosomal protein S1 132 133742 1173185 63 1360925 68 1360929 62 40S Ribosomal protein S3 243 417719 546865 67 1350949 79 1173255 71 40S Ribosomal protein S5 204 1173267 1203905 88 1351001 65 1173259 71 Adencey/homocyateinase 432 134184 12									1		1
L22 60S Filbosomal protein L22 128 464628 1633049 58 1710514 57 1710638 49 LA0 60S Filbosomal protein P0 317 133041 113997 69 1523915 67 171806 55 PB1 RNA polymerase II, largest subunit 1970 133326 133324 70 1255805 73 123330 54 S12 40S Filbosomal protein S12 132 133742 1173185 63 1950825 58 1350929 52 S3 40S Filbosomal protein S12 132 133742 1173185 63 1950825 58 1350929 52 S3 40S Filbosomal protein S3 243 417719 546865 87 1350829 71 132326 71 S5 40S Filbosomal protein S5 204 1173267 1203905 58 1351001 86 1173269 71 AHH Adenosylhomocrysteinase 432 134184 1200230 81 134182 7	60S Ribosomal protein L22 128 464628 1633049 58 1710514 57 1710538 49 60S Ribosomal protein P0 317 133041 113997 69 1523915 67 171806 55 RNA polymerase II, largest subunit 1970 133326 133324 79 1255805 73 133330 64 40S Ribosomal protein S1 132 133742 1173185 63 1950925 58 1360920 62 40S Ribosomal protein S1 243 417718 54866 87 1350989 79 1173255 71 40S Ribosomal protein S5 204 1173267 1203905 88 1351001 65 1173269 71 Adencey/homocyateinase 432 134184 1200230 81 134182 77 730701 74 Calcium-transporting ATPase 994 1586563 114306 80 K11D9* 79 114301 42 Superoxide dismutase 154 134611 1173472<							1				Ť
LAO 60S Filbosomal protein P0 317 132041 113987 69 1523915 67 171806 55 PB1 RNA colymerase II, largest subunit 1970 133326 133324 79 1255805 73 13330 54 S12 40S Filbosomal protein S12 132 133742 1173185 63 1350925 58 1350920 52 S3 40S Filbosomal protein S12 132 133742 1173185 63 1350925 58 1350925 52 S3 40S Filbosomal protein S12 243 417719 548856 67 1350925 58 1350926 52 S4 40S Filbosomal protein S5 204 1173267 1203905 58 1351001 55 1173269 71 Addenosylhomocysteinase 432 134184 1200230 51 134182 77 730701 74 ERCA Calcium-transporting ATPase 994 1586563 114306 80 K1109* 79 1	60S Ribosomal protein P0 317 133041 113987 69 1523915 67 171806 55 RNA polymerase II, largest subunit 1970 133326 133324 79 1256805 73 133320 64 405 Ribosomal protein S12 132 133742 1179185 63 1350925 58 1350920 62 405 Ribosomal protein S1 243 417718 54866 87 1360929 79 1173265 71 405 Ribosomal protein S5 204 1173287 1203905 68 1351001 65 1173269 71 Adenceylhomocysteinase 432 134184 1200230 81 134192 77 730701 74 Calcium-transporting ATPase 994 1586663 114306 80 K1109* 79 114301 42 Superoxide dismutase 154 134611 1173472 62 464769 57 134693 55 Swnaptobrevin 116 135094 436307 <			1		Ì				i		1
PB1 RNA polymerase II, largest subunit 1970 133326 133324 79 1255805 73 133320 54 \$12 40S Ribosomal protein S12 132 13374 1173185 63 1350025 68 1350029 62 \$33 40S Ribosomal protein S12 132 13374 1173185 63 1350025 68 1350029 62 \$33 40S Ribosomal protein S3 243 417719 548856 87 1350929 62 \$35 40S Ribosomal protein S5 204 1172877 1203905 88 1351001 65 1173269 71 AHH Adencaylhomocrysteinase 432 134184 1200230 61 134182 77 730701 74 ERCA Calcium-transporting ATPase 994 1586563 114306 90 K11D9* 79 114301 42 QDC Superoxide dismutase 154 134611 1173472 62 464769 57 134633 55	RNA polymerase II, largest subunit. 1970 133326 133324 79 1255805 73 13330 54 405 Riboscomal protein S12 132 133742 1173185 63 1350925 58 1350929 62 405 Riboscomal protein S12 243 417719 548856 87 1350925 58 1350929 62 405 Riboscomal protein S3 243 417719 548856 87 1350926 79 1173255 71 405 Riboscomal protein S5 204 1173267 1203905 88 1351001 65 1173259 71 Adencaylhomocysteinase 432 194184 1200230 61 134192 77 730701 74 Calcium-transporting ATPase 994 1586563 114306 80 K1109° 79 114301 42 Superoxide dismutase 154 134611 1173472 62 464769 57 134833 55 Swnaptobrevin 116 135094 436307			1				1				Ť
S12 40S Ribosomal protein S12 132 133742 1173185 63 1350925 68 1350929 62 S3 40S Ribosomal protein S3 243 417719 548856 87 1350929 79 1173265 71 S5 40S Ribosomal protein S3 204 1173267 1203905 88 1351001 85 1173268 71 AHH Adenosylhomocysteinase 432 134184 1200230 61 134192 77 730701 74 ERCA Calcium-transporting ATPase 994 1586563 114306 60 K11D9* 79 114301 42 ODC Superoxide dismutase 154 134611 1173472 62 464769 57 134633 55 YB Synaptobrevin 116 135094 436307 72 1405508 79 401107 38 FS-II Transcription elongation factor S-II 301 107907 135660 51 1729914 49 1	405 Ribosomal protein S12 132 133742 1173185 63 1950925 63 1350929 62 405 Ribosomal protein S3 243 417719 548856 87 1350989 79 1173255 71 405 Ribosomal protein S3 204 1172267 1203905 88 1361001 65 1173289 71 405 Ribosomal protein S5 204 1172267 1203905 88 1361001 65 1173289 71 Adencaylhomocysteinase 432 134184 1200230 61 134182 77 730701 74 Calclum-transporting ATPase 994 1586563 114305 80 K11D9''' 114301 42 Superoxide dismutase 154 13401 1173472 62 464789 57 134833 55 Synaptobrevin 116 135094 438307 72 1405508 79 401107 38 Transcription elongation factor S-II 301 107007 126660 51			1				1		1		1
S3 405 Ribosomal protein S3 243 417719 549856 67 1350969 79 1173255 71 S5 405 Ribosomal protein S5 204 1173267 1203905 68 1351001 65 1173269 71 AHH Adenosylhomocytabilitiase 432 134184 1200230 61 134162 77 730701 74 ERCA Calclum-transporting ATPase 994 1586563 114306 80 K11D9* 79 114301 42 ODC Superoxide dismutase 154 134611 1173472 62 494769 57 134633 55 YB Synaptobrevin 116 135094 436307 72 1405508 79 401107 38 FS-II Transcription elongation factor S-II 301 107907 135660 51 1729014 49 1729015 33	40S Filosernal protein S3 243 417719 548856 87 1350989 79 1173255 71 40S Filosernal protein S5 204 1173267 1203905 88 1351001 65 1173289 71 Adencay/homocysteinase 432 134194 1200230 61 134192 77 730701 74 Calclum-transporting ATPase 994 1589593 114306 60 K11D9* 79 114301 42 Superxide dismutase 154 134611 1173472 62 484759 57 134633 55 Synaptiobrevin 116 135094 436307 72 1405508 79 401107 38 Transcription elongation factor S-II 301 107007 136660 61 1720914 49 1720915 33 DNA Topolsomerase I 765 136988 267147 72 1934847 70 135993 57	.	400 Billion and analysis 040	190	400740	1179105		1950005	===	1250020	<u>97</u>	ļ
S5 405 Ribosomal protein S5 204 1173267 1203905 88 1351001 86 1173268 71 AHH Adenosylhomocystelinase 432 134184 1200230 61 134182 77 730701 74 ERCA Calcium-transporting ATPase 994 1586563 114306 80 K11D9* 79 114301 42 ODC Superoxide dismutase 154 134101 1173472 62 494769 57 134633 55 YB Synaptobrevin 116 135094 436307 72 1405508 79 401107 38 FS-II Transcription elongetion factor S-II 301 107907 136660 51 1729014 49 1729015 33	40S Ribosomei protein S5 204 1173267 1203905 68 1351001 65 1173269 71 Adencsylhomocyateinase 432 134184 120030 61 134182 77 730701 74 Calcium-transporting ATPase 994 1586563 114306 80 K11D9* 79 114301 42 Superoxide dismutase 154 134611 1173472 62 464759 57 134533 55 Synaptobrevin 116 135094 436307 72 1405508 79 401107 38 Transcription elongation factor S-II 301 107907 136660 61 1729914 49 1729915 33 DNA Topolocomerase I 765 136988 267147 72 1934847 70 135993 57			049	i	549955	<u></u>	1250020	70	1170066	71	Ť
AHH Adenosylhomocystelnase 432 134194 1200230 81 134192 77 730701 74 ERCA Calcium-transporting ATPase 994 1586563 114306 60 K11D9* 79 114301 42 ODC Superoxide dismutase 154 134611 1173472 52 494769 57 134633 55 YB Synaptobrevin 116 135094 436307 72 1405508 79 401107 38 FS-II Transcription elongation factor S-II 301 107007 135660 51 1729014 49 1729015 33	Adencay/homocysteinase 432 134194 1200230 61 134192 77 730701 74 Calclum-transporting ATPase 994 1588563 114306 60 K11D9* 79 114301 42 Superoxide dismutase 154 134611 1173472 62 484759 57 134633 55 Synaptobrevin 116 135094 436307 72 1405508 79 401107 38 Transcription elongation factor S-II 301 107007 136680 61 1720914 49 1720915 33 DNA Topolsomerase I 765 136988 267147 72 1934847 70 135993 57			1	1	1				1		Ť
ERCA Calcium-transporting ATPase 994 1586563 114308 80 K11D9* 79 114301 42 ODC Superoxide dismutase 154 134611 1173472 62 494769 57 134633 55 YB Synaptobrevin 116 135094 436307 72 1405508 79 401107 38 FS-II Transcription elongation factor S-II 301 107007 135660 51 1729014 49 1729015 33	Calcium-transporting ATPase 994 1586563 114306 80 K11D9* 79 114301 42 Supercoxide dismutase 154 134611 1173472 62 464769 57 134633 55 Synaptobrevin 116 135094 436307 72 1405508 79 401107 38 Transcription elongation factor S-II 301 107007 136660 51 1720914 49 1720915 33 DNA Topolsomerase I 765 136980 267147 72 1934847 70 135993 57					!		ł		1		1
ODC Superoxide dismutase 154 134611 1173472 62 464769 57 134633 55 YB Synaptobrevin 116 135094 436307 72 1405509 79 401107 38 FS-II Transcription elongation factor S-II 301 107007 135660 51 1720014 49 1720015 33	Superoxide dismutase 154 134811 1173472 62 494769 57 134833 55 Swnaptobrevin 116 135094 436307 72 1405508 79 401107 38 Transcription elongation factor S-II 301 107007 125660 61 1720914 49 172915 33 DNA Topolsomerase I 765 136980 267147 72 1934847 70 135993 57		1		1			1				1
YB Synaptobrevin 116 135094 436307 72 1405508 79 401107 38 FS-II Transcription elongation factor S-II 301 107907 135660 51 1729014 40 1729015 33	Synaptobrevin 116 135094 436307 72 1405508 79 401107 38 Transcription elongation factor S-II 201 107007 126660 51 1720914 49 1729915 33 DNA Topolsomerase I 765 136980 267147 72 1934847 70 136993 57			1	1			1				1
FS-II Transcription elongation factor S-II 301 107907 135660 51 1729914 49 1729915 33	Transcription elongation factor S-II 301 107907 135660 61 1729914 49 1729915 33 DNA Topolsomerase I 765 136980 267147 72 1934847 70 135993 57			1	1	1		ł		1		1
	DNA Topolsomerase I 765 136989 267147 72 1934847 70 135993 57			1	1			1				į
<u>OP1 IDNA Topolsomerase I 765 135989 267147 72 1934847 70 135993 57</u>				1	1	1		1		•		4.
PIS Triceephosphate isomerase 249 136066 267156 66 1729998 62 136069 54	Triosephosphate isomerase 249 136066 267156 65 1729998 62 136069 54				Ì	I		1		1		+

Length in amino acids is given for the human ortholog. (See Fig. 1 for the three possible tree topologies A, B, and C.) (ND) Not determined because yeast-human % identity <35% (see Methods). (*) Nematode SERCA ortholog sequence was determined manually from unfinished sequence of cosmid K11D9. (#) Number. The list of entries including rodent and bacterial orthologs is available at http://www.sequana.com/publications/model_proteomes.

nematode DNA for possible missed exons. Second, the EST databases were searched for the higher scoring sequences in the nematode and fly. Only one nematode sequence (synaptobrevin) was found among the ESTs that was a better orthologous candidate than the sequence in the NR database, and was included in the analysis.

The possible relationships of the human, nematode, and fly sequences can be described by three different tree topologies (Fig. 1): (1) tree A, in which the fly sequence is a sister taxon to the human sequence with the nematode sequence basal to the fly-human clade; (2) tree B, wherein the fly and nematode sequences are sister taxa; and (3) tree C, in which the nematode and human sequences are sister taxa with the fly sequence basal to the nematode-human clade.

Thirty-six protein quartets whose metazoan members contained amino acid identities of $\geq 35\%$ to the yeast sequence were subjected to individual phylogenetic analysis as described in Methods. Neighbor-joining analyses of the 36 orthologous protein sequence alignments (Fig. 2), revealed that 24 quartets generate tree A (with average bootstrap values of $80\% \pm 18$ s.D.), 11 amino acid alignments support tree B (average bootstrap values of $61\% \pm 15$), and 1 alignment produced tree C with a bootstrap value of 45. Results were essentially identical when gamma-corrected distances were used. These data are consis-

COMPARISON OF ORTHOLOGOUS PROTEINS IN MODEL ORGANISMS

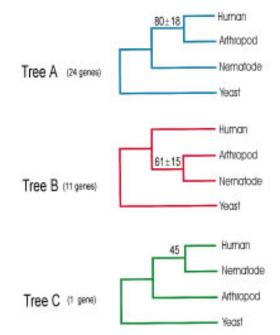


Figure 1 The three possible topologies for a tree describing the evolutionary relationships between nematodes, arthropods, and humans. Tree A (blue) reflects the conventional interpretation of metazoan phylogeny with nematodes as a "protocoelomate" group basal to arthropods and humans. Tree A was supported by neighbor-joining analysis of 24 protein quartets as described in the text. Tree B (red) represents the "Ecdysozoa" phylogeny derived from 18S rRNA gene sequences of a variety of nematodes and arthropods (Aguinaldo et al. 1997), and is supported by 11 protein quartets. Tree C (green) is not expected from any metazoan phylogenetic hypothesis and is supported by a single protein quartet. Average bootstrap values and their standard deviations are shown for each tree.

tent with the distribution of the pairwise similarity scores observed in BLAST searches. Eight of the nine *C. elegans* protein sequences derived from full-length cDNA were less similar to their human orthologs than the corresponding *Drosophila* protein sequences (i.e., supported tree A), suggesting that the prevalence of Tree A was not an artifact of computational prediction of nematode proteins.

Similarity of Fly and Nematode Proteins to Human Orthologs May Be Influenced by Unequal Evolutionary Rate Effects

Phylogenetic hypotheses based on molecular sequence data are affected by two important factors governing evolutionary rate. The first factor is geneto-gene variation, where different genes have different evolutionary rates among a given pair of taxa, usually as a result of functional constraints on the encoded protein. The second factor is evolutionary rate heterogeneity within a gene, where the evolutionary rate can vary among different lineages of a tree. A homogeneously evolving gene evolves at the same rate per unit time among all branches of a tree,

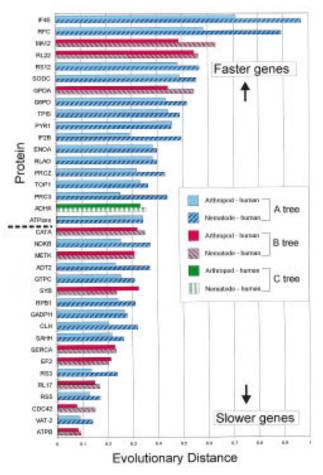


Figure 2 Relative evolutionary rates of the 36 protein quartets subjected to phylogenetic analysis. Protein quartets supporting tree A are shown in blue, those supporting tree B are shown in red, and the quartet supporting tree C is shown in green. The protein name abbreviation is shown along the y-axis, and the proteins are plotted in order of the mean evolutionary distance of nematode to human and arthropod to human where nematode is C. elegans, and arthropod is D. melanogaster. Proteins with the highest number of pairwise substitutions (fast evolving) are at the top and those with the lowest number of pairwise substitutions (slow evolving) are at the bottom. The evolutionary distances along the x-axis were determined from amino acid alignments using a Poisson correction as described in Methods. The broken line (middle left) represents the midway point where 18 proteins are above the line and 18 are below the line. The key to the bars is shown.

whereas a heterogeneously evolving gene may evolve more rapidly in some lineages than others, causing unequal rate effects that are known to produce tree-building artifacts (Hillis et al. 1994; Lyons-Weiler and Hoelzer 1997; Maley and Marshall 1998). For example, the 18S rRNA gene has evolved at a rapid rate of nucleotide substitution in C. elegans compared to other animals (Winnepenninckx et al. 1995; Garey et al. 1996), causing unequal rate effects that artificially place C. elegans as a basal animal, as in tree A of Figure 1. However, 18S rRNA genes from most nonrhabditid nematode taxa appear to have evolved at a slower rate than in C. elegans (Blaxter et al. 1998), and when nematode 18S rRNA sequences with lower substitution rates are analyzed, nematodes emerge as a sister taxon to arthropods (Aguinaldo et al. 1997), as in tree B of Figure 1.

The finding that only one quartet supported tree C was expected because the possibility that arthropods are basal to both nematodes and humans is not supported by any hypothesis of metazoan phylogeny of which we are aware. The finding that 24 quartets support tree A, whereas 11 quartets support tree B has several possible explanations. One is that tree A reflects the correct historical phylogeny, and that the quartets supporting trees B and C represent random noise. An alternative explanation is that the finding of the majority of quartets supporting tree A is caused by unequal evolutinary rate effects, as in the 18S rRNA gene of C. elegans (Aguinaldo et al. 1997). To assess gene-to-gene variation and evolutionary rate heterogeneity among these 36 orthologous proteins of humans, nematodes, and arthropods, the pairwise evolutionary distances from human to nematode were compared with those of human to arthropod for each quartet. These evolutionary distances are shown as a bar graph in Figure 2, with the quartets ordered from the fastest to slowest evolving proteins. There are fewer pairwise substitutions between arthropod and human than between nematode and human except in five quartets (EF2, SYB, METK, ATPase, PYR1). Figure 2 shows that 8 of 11 quartets that support tree B fall among the slower half of the quartets, whereas only three of the quartets supporting tree B fall among the faster half of the quartets. Five of the eight slowest evolving quartets (ATPB, CDC42, RL17, EF2, SERCA) support tree B. These five proteins have fewer overall substitutions, thus unequal evolutionary rates are less likely to develop and tree B is favored. The faster evolving quartets that support tree B (GPDA, RL22, MA12) likely represent genes that have a high number of substitutions, but where the number of substitutions are homogeneous between the taxa within the quartet.

To visualize the degree of evolutionary rate heterogeneity among all four taxa, we plotted the fourway relative rates for the 36 proteins in Figure 3. The *y*-axis in Figure 3 displays the ratio of the evolutionary rates of nematode and arthropod relative to yeast, which should equal one if a protein evolved homogenously in the two lineages, as yeast is undoubtedly an outgroup to nematode and arthropod. However, most of the proteins have a *y*-value greater than one, indicating that they have evolved more quickly in the lineage from yeast to nematode than

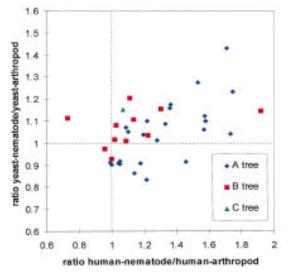


Figure 3 Four-way relative rate plot of evolutionary distances for 36 proteins. The ratio of evolutionary distances from (human-nematode)/(human-arthropod) for each protein is plotted on the x-axis, where a ratio of 1 would be expected if the proteins were evolving homogeneously in those branches assuming that arthropods and nematodes are sister taxa. The ratios of evolutionary distance from (yeast-nematode)/(yeastarthropod) are plotted on the y-axis, which should equal 1 if a protein evolved homogenously in the nematode and arthropod lineages. The position where the x-axis and y-axis both equal one represents the region where genes would fall if they evolved homogeneously in all four taxa, if Tree B is correct. Proteins to the right of the vertical line at x = 1 should favor tree A, proteins to the left should favor tree C, whereas proteins falling near the diagonal line should favor tree B. The distribution of the 36 orthologous proteins is skewed, with those that yield tree B (red squares) scattered uniformly around the diagonal line (with one exception, CDC42 supports tree B but falls to the extreme right of the graph), whereas all of the proteins that yield tree A (blue diamonds) are scattered to the right of the diagonal. The quartet favoring tree C is shown in green (triangle).

from yeast to arthropod. Similarly, the majority of the proteins have an x value >1, indicating that most have evolved more quickly in the lineage from human to nematode than from human to arthropod (assuming that nematodes and arthropods are sister taxa). The distribution of data points along the x-axis is highly skewed, with most of the proteins between 0.9 and 1.3, and the remainder with higher values. Of note, proteins supporting tree B are clustered more closely to the point where the *x*and y-axes both equal one (representing proteins that evolved homogeneously in all four taxa), but proteins supporting tree A are all to the right with most far to the right (representing proteins that evolved more heterogeneously). Thus, these orthologous proteins can be divided into two populations: homogeneously evolving proteins that support tree B, and heterogeneously evolving proteins that support tree A. This four-way relative rate plot suggests that the preponderance of proteins supporting tree A is largely attributable to unequal evolutionary rates. Given that the orthologous proteins selected for this study were extracted from total available genome data, it is reasonable to infer that approximately two-thirds of protein-coding genes in C. elegans have evolved more rapidly than in Drosophila.

Highly Conserved Protein Domains and Diversity of Multidomain Proteins in *Drosophila* and *C. elegans*

C. elegans and *Drosophila*, as the most highly developed genetically tractable model animals, are attractive systems for studying human disease pathways (Miklos and Rubin 1996; Ahringer 1997). In practice, the ability to extract functional inferences depends more on the actual content and complexity of the protein repertoire in different model organisms than on their deduced taxonomic positions per se. To address the question of protein domain diversity in these two proteomes, we first masked putative nonglobular segments and coiled coils in the *Drosophila* and *C. elegans* protein data sets. These two classes of sequences often serve as hinges between globular domains and tend to produce spurious hits with database searches (Altschul et al. 1994). After deleting redundant domains in each database, we constructed *Drosophila* and *C. elegans* libraries of unmasked protein domains >50 amino acids and compared these libraries to each other, to the complete yeast proteome, and to the publicly available human ESTs. Matches with a similarity score of >90 were counted, a cutoff that virtually ensures evolutionary relation and functional relevance under the applied conditions (Koonin et al. 1997; A.R. Mushegian, unpubl.).

The results of this analysis are summarized in Table 2. The most conspicuous result is that the available portion of the Drosophila proteome is strongly enriched in conserved domains as compared to that of C. elegans. It seems unlikely that Drosophila has retained a larger fraction of proteins descended from an ancestral unicellular eukaryote, while also becoming enriched in protein domains shared with humans. Rather, we suspect that this difference is largely attributable to overrepresentation of certain classes of Drosophila proteins in current databases, given that only a small fraction of the fly genome has been sequenced. A substantial increase in Drosophila genomic DNA sequence will clearly be required before the question of domain repertoire in this organism can be addressed in a more definitive way.

A hallmark of eukaryotic genome evolution is the increased number of multidomain proteins thought to have originated largely by domain shuffling (Doolittle 1995). Because a comprehensive comparison of protein sets of *C. elegans* and *D. melanogaster* is limited by the extent of whole-genome sequencing, we wished to analyze a representative set of multidomain proteins in both available proteomes. Toward this end, we extended an earlier

Table 2.	Table 2. Protein Domain Conservation in Model Organisms							
	No. of	Domains conserved in						
Species	domains ^a	C. elegans	Drosophila	yeast	human ESTs			
C. elegans	13169	100%	3133 (24%)	3741 (28%)	5305 (40%)			
Drosophila	2611	1905 (73%)	100%	1264 (48%)	2042 (78%)			
^a The FASTA f			0	nan 90 are reported on-line at http://w				

analysis of 77 proteins encoded by human positionally cloned genes specifically mutated in hereditary diseases, a set consisting largely of complex multidomain proteins (Mushegian et al. 1997). Among 84 proteins in the updated disease gene database (XREFdb as of July 1, 1997; see Methods), 68 (81%) shared similarity with nematode proteins, but in the majority of these cases the similarity was limited to individual domains within larger proteins with a different overall domain architecture. By our criteria of orthology, which requires similarity along the entire length of multidomain proteins and not just individual domains (see Methods), 25 human disease proteins (37% of all detected similarities) had candidate orthologs in the sequenced portion of C. elegans proteome (Table 3). Of the 34 human proteins with similarity matches in the fly proteome, 13 proteins (38%) had candidate orthologs in D. melanogaster. Thus, the likelihood of nematode and fly proteins possessing the same domain architecture as human disease gene products was remarkably similar.

The available portion of the fly proteome contains a high proportion of protein sequences obtained through positional cloning of phenotypically significant genes as well as genes specifically cloned by homology to mammalian proteins. Therefore, one might expect that protein families

	Orthologs			
Data sets (no. of human proteins)	C. elegans	D. melanogaster		
Positionally cloned genes mutated in specific				
human diseases (84) Biosynthetic enzymes ^a	25	13		
(30) purine biosynthesis ^b	24	18		
(6) arginine and proline	5	4		
biosynthesis (8) sterol biosynthesis	6	6		
(11)	8	5		
folate biosynthesis/5 Leukocyte surface	5	3		
antigens (151)	15	5		

^aThe list of proteins is available online at http://www. sequana.com/publications/model_organisms. ^bThree human enzyme sequences in this category are unavailable, so the yeast orthologs were used. would be unevenly represented among available Drosophila proteins. We addressed this issue by querying the Drosophila and C. elegans proteomes with additional sequences of biological interest. In one query using 30 human enzymes belonging to 4 disparate central metabolic pathways (Table 3), 24 had candidate orthologs in C. elegans and 18 in Drosophila. In another search, using a set of 151 human leukocyte surface (CD) antigens, 78 shared similarity with Drosophila sequences and 89 with C. elegans sequences, although not unexpectedly most of these similarity matches were to portable modules (such as immunoglobulin-like or epidermal growth factor (EGF)-like domains) within a nonorthologous protein. Interestingly, there appear to be three times as many orthologs of human CD antigens in C. elegans as in Drosophila (Table 3). Inspection of the similarizies showed that this difference in the number of orthologs is explained by the almost total absence of certain classes of proteins related to CD antigens among the available Drosophila sequences, including large metalloproteases, the type II (4TM) transmembrane receptors, aminopeptidases, and apyrases.

Concluding Remarks

In this study we evaluated the nematode C. elegans and the fruit fly D. melanogaster as model systems for studying human proteins using protein sequence comparison techniques. By applying strict and reproducible criteria for identifying orthologous proteins, we could extract numerous proteincoding genes for phylogenetic analysis. Our simultaneous analysis of multiple orthologous proteins shows that different proteins can generate different apparent phylogenetic tree topologies, strongly suggesting that historical phylogenies should not be inferred based on a single protein-coding gene. Unequal evolutionary rates are an important factor in calculating phylogenetic trees, and indeed it appears that the majority of *C. elegans* genes are evolving more rapidly than their Drosophila counterparts. The approaches of ortholog extraction used in this work can be used to better define data sets for phylogenetic analysis among a broader range of representative animal phyla. The available portion of the fly proteome appears to be comparatively enriched in conserved protein domains because of abundant representation of phenotypically defined genes, while missing numerous protein families. The ortholog-to-paralog ratio with regard to human proteins is very similar in the two model animals, indicating that the domain architecture in fly and nematode proteins approximates that of their human homologs to the same extent.

METHODS

Databases

The NR database at the National Center for Biotechnology Information (Bethesda, MD) was used as the source of sequences and for most of the database searches. Speciesspecific sets of proteins were extracted from NR using Nentrez network tools and the species names as queries. Unfinished genome sequences from the *C. elegans* genome project (http:// www.sanger.ac.uk/Projects/C elegans) and database of C. elegans ESTs (http://www.ddbj.nig.ac.jp/c-elegans/html/ CE INDEX.html) were used to verify the protein sequences predicted from genomic DNA. The database of human disease genes and their homologs is available at http://www.ncbi. nlm.nih.gov/XREFdb, and partial data on orthologs in the nematode and yeast are at http://www.ncbi.nlm.nih.gov/ Disease Genes. A nonredundant list of human leukocyte surface (CD) antigens was constructed by modifying the list available at http://www.expasy.ch/cgi-bin/lists?cdlist.txt. Information on biochemical pathways was obtained in part from http://www.genome.ad.jp/kegg.

Sequence Database Searching, Cross-Referencing, and Ortholog Identification

Database searches were performed using the BLAST2 algorithm (Altschul and Gish 1996), with gap width 256 and no filtering. The BLASTP program was used to search protein databases. The TBLASTN program was used to search the nucleotide sequence databases. In all searches, matches with similarity scores <75 were removed. This cutoff eliminates many spurious hits and virtually never eliminates orthologs for medium-sized proteins (A.R. Mushegian, unpubl.). To count the fraction of conserved and unique protein domains, a more restrictive similarity score cutoff, s > 90, was used. BLASTP results were automatically processed using the BLATAX program (Koonin et al. 1996) to extract the best matches in the given species.

Two measures were applied to distinguish candidate orthologs from likely paralogs based on sequence similarity, the BLASTP similarity score and the percentage of amino acid identity in the aligned segments. Criteria used to define candidate orthologs (Tatusov et al. 1996) were as follows. First, protein A in proteome a is a candidate ortholog of protein B in proteome b, if protein B is the best match when sequence A is searched against proteome b, and, conversely, protein A is the best match when sequence B is searched against proteome a. Second, A and B share similarity along their whole lengths. Third, no homolog in a taxonomic outgroup (S. cerevesiae in the present analysis) is closer to A than B, or closer to B than A. Sequences that belong to large, diverged protein families were not considered because of limitations in applying the classic definition of orthology in such cases (Gehring et al. 1994; Tatusov et al. 1997). The following families were thus excluded from analysis: protein kinases, protein phosphatases, RAS-like GTPases and their regulators, chaperones of the HSP60, HSP70, and HSP90 families, and RNA-binding proteins containing RNA recognition motifs.

Phylogenetic Methods

Amino acid sequence data sets for each of 42 protein-coding genes (see Results and Discussion) included orthologs as defined above from S. cerevisiae, C. elegans, D. melanogaster, and Homo sapiens. Orthologous protein quartets were aligned using a star alignment procedure (Myers and Miller 1988), as implemented in Align Plus software version 3 (Scientific and Educational Software Co.) using the yeast sequence as a guide. This method was chosen because it does not invoke phylogenetic assumptions to carry out the alignment. Each quartet alignment was adjusted interactively using the MACAW program (Schuler et al. 1991) to correct alignment errors, and regions where amino acid similarity were too low to be certain of the alignment were deleted. The alignments are available at http://chuma.cas.usf.edu/~garey/alignments/alignment. html. Phylogenetic analysis was carried out only with quartet alignments where amino acid identity was >35% among all members. Sequence sites in an alignment with gaps in any single taxon sequence were excluded from phylogenetic analysis. Maximum parsimony trees were produced with the PHYLIP package (Felsenstein 1993). Evolutionary distances and neighbor-joining trees were calculated using both a Poisson distribution of amino acid substitutions and a γ correction (shape parameter = 2) using the MEGA program (Kumar et al. 1994). All trees were tested by the analysis of 100 bootstrap replicates.

ACKNOWLEDGMENTS

We thank Carl Johnson for his keen interest in this project and valuable suggestions, Greg Ederer, Maurice Leysens, and Douglas Wood for programming support.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES

Aguinaldo, A.M.A., J.M. Turbeville, L.S. Linford, L. Hebshi, M.C. Rivera, J.R. Garey, R.A. Raff, and J.A. Lake. 1997. Evidence for a clade of nematodes, arthropods and other moulting animals. *Nature* **387**: 489–493.

Ahringer J. 1997. Turn to the worm. *Curr. Opin. Genet. Dev.* **7:** 410–415.

Altschul, S.F. and W. Gish. 1996. Local alignment statistics. *Methods Enzymol.* **266:** 460–480.

Altschul S.F., M.S. Boguski, W. Gish, and J.C. Wootton. 1994. Issues in searching molecular sequence databases. *Nature Genet.* **6:** 119–129.

Banfi, S., G. Borsani, E. Rossi, L. Bernard, A. Guffanti, F. Rubboli, A. Marchitiello, S. Giglio, E. Coluccia, M. Zollo, O. Zuffardi, and A. Ballabio. 1996. Identification and mapping of human cDNA homologous to *Drosophila* mutant genes through EST database screening. *Nature Genet.* **13**: 167–174.

Bassett, D.E., Jr., M.S. Boguski, F. Spencer, R. Reeves, S. Kim, T. Weaver, and P. Hieter. 1997. Genome cross-referencing

MUSHEGIAN ET AL.

and XREFdb: Implications for the identification and analysis of genes mutated in human disease. *Nature Genet.* **15:** 339–344.

Blaxter, M., P. DeLey, J.R. Garey, L.X. Liu, P. Scheldeman, J. Vanfleteren, L.Y. Mackey, M. Dorris, L.M. Frisse, J.T. Vida, and K.T. Thomas. 1998. A molecular phylogenetic framework for the phylum Nematoda. *Nature* **392**: 71–75.

Brusca, R.C. and G.J. Brusca. 1990. *Invertebrates*. Sinauer, Sunderland, MA.

Doolittle, R.F. 1995. The multiplicity of domains in proteins. *Annu. Rev. Biochem.* **64:** 287–314.

Doolittle, R.F., D.F. Feng, S. Tsang, G. Cho, and E. Little. 1996. Determining divergence times of the major kingdoms of living organisms with a protein clock. *Science* **271:** 470–477.

Felsenstein, J. 1993. *PHYLIP: Phylogeny inference package,* version 3.5. University of Washington, Seattle, WA.

Feng, D.F., G. Cho, and R.F. Doolittle. 1997. Determining divergence times with a protein clock: Update and reevaluation. *Proc. Natl. Acad. Sci.* **94:** 13028–13033.

Fitch, W.M. 1970. Distinguishing homologous from analogous proteins. *Syst. Zool.* **19**: 99–106.

Garey, J.R., M. Krotec, D.R. Nelson, and J. Brooks. 1996. Molecular analysis supports a targigrade-arthropod association. *Invert. Biol.* **115**: 79–88.

Gehring, W.J., M. Affolter, and T. Burglin. 1994. Homeodomain proteins. *Annu. Rev. Biochem.* **63:** 487–526.

Henikoff, S., E.A. Greene, S. Pietrokovski, P. Bork, T.K. Attwood, and L. Hood. 1997. Gene families: The taxonomy of protein paralogs and chimeras. *Science* **278**: 609–614.

Hillis, D.M., J.P. Huelsenbeck, and C.W. Cunningham. 1994. Application and accuracy of molecular phylogenies. *Science* **264:** 671–677.

Koonin, E.V., R.L. Tatusov, and K.E. Rudd. 1996. Genome-scale comparison of protein sequences. *Methods Enzymol.* **266**: 295–322.

Koonin, E.V., A.R. Mushegian, M.Y. Galperin, and D.R. Walker. 1997. Common and distinctive features of bacterial and archaeal genomes revealed by computer analysis of protein sequences. *Mol. Microbiol.* **25:** 619–637.

Kumar, S., K. Tamura, and M. Nei. 1994. MEGA: Molecular evolutionary genetics analysis software for microcomputers. *Comp. Appl. Biosci.* **10**: 189–191.

Lyons-Weiler, J. and G.A. Hoelzer. 1997. Escaping from the Felsenstein zone by detecting long branches in phylogenetic data. *Mol. Phylogenet. Evol.* **8**: 375–384.

Lupas, A. 1997. Prediction and analysis of coiled-coil structures. *Methods Enzymol.* **266**: 513–525.

Maley, L.E. and C.R. Marshall. 1998. The coming of age of molecular systematics. *Science* **279**: 505–506.

McHugh, D. 1997. Molecular evidence that echiurans and pogonophorans are derived annelids. *Proc. Natl. Acad. Sci.* **94:** 8006–8009.

Miklos, G.L.G. and G.M. Rubin 1996. The role of the genome project in determining gene function: Insights from model organisms. *Cell* **86**: 521–529.

Mushegian, A.R., D.E. Bassett Jr., M.S. Boguski, P. Bork, and E.V. Koonin. 1997. Positionally cloned human disease genes: Patterns of evolutionary conservation and functional motifs. *Proc. Natl. Acad. Sci.* **94**: 5831–5836.

Myers, E.W. and W. Miller. 1988. Optimal alignments in linear space. *CABIOS* **4:** 11–17.

Nielsen, C. 1995. Animal evolution. Interrelationships of the living phyla. Oxford University Press, Oxford, UK.

Rubin, G.M. 1996. Around the genomes: The *Drosophila* genome project. *Genome Res.* **6**: 71–79.

Schuler, G.D., S.F. Altschul, and D.J. Lipman. 1991. A workbench for multiple alignment construction and analysis. *Proteins Struct. Funct. Genet.* **9**: 180–190.

Sidow, A. and W. K. Thomas. 1994. A molecular evolutionary framework for eukaryotic model organisms. *Curr. Biol.* **4**: 593–603.

Tatusov, R.L., A.R. Mushegian, P. Bork, N.P. Brown, M. Borodovsky, W.S. Hayes, K.E. Rudd, and E.V. Koonin. 1996. Metabolism and evolution of *Haemophilus influenzae* deduced from a whole genome sequence comparison to *Escherichia coli. Curr. Biol.* **6**: 279–291.

Tatusov, R.L., E.V. Koonin, and D.J. Lipman. 1997. A genomic perspective on protein families. *Science* **278:** 631–637.

Walker, D.R. and E.V. Koonin. 1997. SEALS: A system for easy analysis of lots of sequences. *ISMB* **5:** 333–339.

Waterston, R.H. 1997. The Genome. In C. elegans II (ed. D.L. Riddle, T. Blumenthal, B.J. Meyer, and J.R. Priess), pp. 23–46. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Winnepenninckx, B., T. Backeljau, L.Y. Mackey, J.M. Brooks, R. De Wachter, S. Kumar, and J.R. Garey. 1995. 18S rRNA data indicate that the aschelminthes are polyphyletic and consist of at least three distinct clades. *Mol. Biol. Evol.* **12**: 1132–1137.

Wootton, J.C. and S. Federhen. 1996. Analysis of compositionally biased regions in sequence databases. *Methods Enzymol.* **266:** 554–573.

Received December 15, 1997; accepted in revised form April 21, 1998.