

## THE *CAENORHABDITIS ELEGANS* GENOME PROJECT

Sean R. Eddy

MRC Laboratory of Molecular Biology  
Hills Road  
Cambridge CB2 2QH  
England  
sre@mrc-lmb.cam.ac.uk

### INTRODUCTION

In the next five years, molecular biology will get its first look at the complete genetic code of a multicellular animal. The *Caenorhabditis elegans* genome sequencing project, a collaboration between Robert Waterston's group in St. Louis and John Sulston's group in Cambridge, is currently on schedule towards its goal of obtaining the complete sequence of this organism and all its estimated 15,000 to 20,000 genes by 1998 (Sulston et al., 1992). By that time, we should also know the complete genome sequence of a few other organisms as well, including the prokaryote *Escherichia coli* (Daniels et al., 1992; Plunkett et al., 1993) and the single-celled eukaryote *Saccharomyces cerevisiae* (Oliver et al., 1992).

Sydney Brenner became interested in *C. elegans* in the 1960s as a model genetic system for the study of nervous system development and animal behavior (Brenner, 1974). Since that time, a growing community of *C. elegans* researchers has extended Brenner's already ambitious vision, and *C. elegans* has become a major experimental organism for the study of eukaryotic molecular, cellular, and developmental biology (Wood, 1988). The community's goal is to understand the biology of this particular animal in the minutest detail possible. This unashamedly naive philosophy has been one of the driving forces behind a set of large descriptive projects that have supported the *C. elegans* community with a remarkably deep knowledge base. These include the elucidation of the somatic developmental lineage by direct observation of embryos (Sulston et al., 1983), the reconstruction of the connectivity and anatomy of the 302-cell nervous system from serial EM micrographs (White et al., 1986), and the cloning and physical mapping of the genome in cosmid and yeast artificial chromosome (YAC) vectors (Coulson et al., 1986; Coulson et al., 1988; Coulson et al., 1991). The genome sequencing project is the latest endeavor in this program of accumulating basic knowledge about the nematode.

As of this summer (1993), 2.2 million bases of DNA sequence have been obtained

known to be relatively gene-dense, like all the centers of *C. elegans* autosomes. It was also known to contain some quite interesting genetically identified loci such as a cluster of homeobox genes involved in determination of anterior/posterior pattern (Burglin et al., 1991; Kenyon and Wang, 1991; Wang et al., 1993). Although the current data comprise only 2% of the genome, we have a taste of what is to come.

Here I describe the overall strategy and goals of the genome sequencing effort. I emphasize the mechanisms being used to make data and clones available to the rest of the research community, both via the international sequence databases and via the freely distributable *C. elegans* database, ACeDB (R. Durbin and J. Thierry-Mieg, unpublished). Finally, I describe the first exploratory efforts by the computational biology groups at St. Louis and Cambridge to predict genes and other features of the sequence, using both conventional similarity searching techniques and more recent developments in biological sequence pattern recognition.

## STRATEGY

### Physical map

The physical map originally consisted of overlapping 40-50 kb cosmid clones. 17,500 cosmids were mapped into about 700 contigs, covering perhaps 90% of the genome (Coulson et al., 1986; Coulson et al., 1988). The remaining gaps proved impossible to clone into cosmids and are thought probably to be repetitive sequence of some sort. Yeast artificial chromosome (YAC) vectors, which accommodate large inserts of about 200-400 kb (or more; megabase inserts are possible) became available by the late 1980s. YACs have an additional advantage in that DNA that is difficult to maintain stably in cosmids seems more easily cloned in YACs (Coulson et al., 1988; Coulson et al., 1991). Over 2,300 YAC clones have now been mapped, closing most of the remaining gaps on the physical map (Coulson et al., 1988; Coulson et al., 1991).

The physical map is currently eight gaps away from closure (Alan Coulson, personal communication). No special effort has yet been made to clone the ends of the chromosomes. Although the size of the remaining gaps and telomeric regions is unknown, the current physical map is thought to be about 98% complete because 98% of the current cDNA expressed sequence tags (McCombie et al., 1992; Waterston et al., 1992) can be mapped to existing clones (Alan Coulson, personal communication). About 95 million bases have been cloned and mapped.

Most new sequences can be rapidly physically mapped to a resolution of about 100 kilobases by hybridization to YAC grids dubbed polytene filters (Coulson et al., 1988; Coulson et al., 1991). The origin of "polytene" is a play on words, derived from the *Drosophila* technique of physical mapping by *in situ* hybridization to polytene chromosomes. Spotted on each filter are 958 representative YACs, covering most of the physical map with an average coverage of two-fold. Polytene filters are available on request from Cambridge.

### Genome sequencing

The collaboration that was established between Robert Waterston's group in St. Louis and John Sulston's group in Cambridge for physical mapping was continued to sequence the genome (Sulston et al., 1992; Wilson et al., 1993). They began at a point in the middle of the chromosome III and sequenced outwards. The strategy is to sequence

Figure 1: Summary of the current state of the sequencing project. Finished or in progress cosmids are indicated to the left of a linear scale in nucleotides. Gray bars indicate the extent of finished sequence. Positions of predicted genes on each strand are indicated as black bars to the left or right of a central line, and genetically identified genes are named to the right of this. More cosmids appear here than the 77 discussed in the text. The figure was generated from the Cambridge database, which includes Cambridge sequence in progress as well as the finished sequence from both groups.

center of III is completed.

Although there is much interest in new technological developments for large-scale genome sequencing, the strategy currently employed is one of scaling up and refining existing automated sequencing technology (Sulston et al., 1992; Wilson et al., 1993). Overlapping cosmid clones are selected from the physical map, random fragments of each cosmid are subcloned into M13 (1-3 kb insert size) and phagemid (6-9 kb insert size) vectors, and a large number of these “shotgun clones” are sequenced on automated machines using standard fluorescent primer extension methods. These sequence fragments, consisting of about 400 nucleotides per gel read, are assembled into contigs using software developed by Rodger Staden and his group (Dear and Staden, 1991). The remaining gaps are closed by primer walking from custom-synthesized oligonucleotides. A cosmid sequence is finished when it has been sequenced at least once on both strands and any ambiguities resolved by inspection of the original traces or by further sequence reads. Very roughly, 600–800 shotgun reads and about 10–40 custom oligos are used per cosmid, giving about 4- to 5-fold redundancy on average.

Through both improved automation – for instance, robotic DNA template preparation (Watson et al., 1993) – and additional man- and sequencing machine-power, the pace continues to accelerate. The groups plan to finish another ten million bases of sequence in 1994 (John Sulston, personal communication).

## Preliminary computational analysis

The program GENEFINDER (Phil Green and LaDeana Hillier, unpublished) is used to predict coding regions in the sequence. In the nematode genome, which is gene-dense with (usually) short introns and relatively little repetitive DNA, gene-finding is not too difficult. The predictions of GENEFINDER are manually refined using ACeDB’s sequence displays.

The DNA sequence is conceptually translated in all six frames and compared against the protein sequence databases, using the program BLASTX from the BLAST suite of similarity searching programs (Altschul et al., 1990). BLASTX output is filtered to remove hits resulting from biased composition, and to detect and assign higher significance to multiple BLAST matches consistent with a single gapped alignment (Sonnhammer and Durbin, 1993). Significant similarities are recorded and displayed in ACeDB, and become one of the criteria by which GENEFINDER predictions are manually revised.

Other features in the DNA or predicted protein sequence that are noticed by various sorts of programs (large inverted repeats, members of repeat sequence families) are annotated in ACeDB as well.

## Data organization and distribution

Complete cosmid sequences are annotated with predicted and known genes, detected similarities, and other features. The annotated sequences are deposited in the EMBL and GenBank sequence databases (Benson et al., 1993; Rice et al., 1993). Release 35 of the EMBL database (June 1993) contained data for 53 cosmids, 1.6 million bases of slightly overlapping genome sequence.

A great deal of information about *C. elegans* is available through the database program ACeDB (A *C. elegans* Data Base), written by Jean Thierry-Mieg and Richard Durbin (unpublished). Among other things, ACeDB organizes and displays the genetic

with the help of the entire *C. elegans* community. The ACeDB software is organism-independent; different databases can be created for other purposes, and the ACeDB software has been adopted by a number of other genome projects. The program is built to accommodate new extensions and data types flexibly. The versions in use in Cambridge are in constant flux as new features are explored.

Released stable versions of ACeDB and the *C. elegans* database are freely available by anonymous ftp from a number of sites on the Internet, including `ncbi.nlm.nih.gov` and `cele.mrc-lmb.cam.ac.uk`. The software and the databases are discussed on the Usenet newsgroup `bionet.software.acedb`.

## ANALYSIS

### Gene predictions and database similarities

From the sequence of 77 cosmids (2.2 Mb), GENEFINDER predicts 470 coding regions. The predictions imply an average of 1 gene every 5,000 bp. About 29% of the genome appears to be coding. Seventeen genes which had been previously identified genetically have been localized in the sequence (some of which had already been sequenced individually): *ceh-16*, *dpy-19*, *emb-9*, *egl-5*, *egl-45*, *gst-1*, *glp-1*, *lin-9*, *lin-12*, *lin-36*, *mab-5*, *mig-10*, *ncc-1*, *sup-5*, *tbg-1*, *unc-36*, and *unc-86* (Wilson et al., 1993).

An estimate for the total number of genes in the worm that corrects for biased gene density along the chromosomes can be obtained using data from the expressed sequence tag (cDNA) sequencing projects (McCombie et al., 1992; Waterston et al., 1992). Although the cDNAs are biased towards highly-expressed genes, one assumes that highly expressed genes are not significantly clustered in the genome. 129 out of 4,615 (2.8%) of the current cDNA tags map into this 2.2 Mb interval. The extrapolation 470 divided by 2.8% implies there are about 17,000 total genes.

How many of the predicted genes are similar to something in the existing protein databases? I used the BLAST algorithm (Altschul et al., 1990) to scan the SWISS-PROT and PIR databases (Bairoch and Boeckmann, 1993; Barker et al., 1993) for proteins similar to these 470 coding regions. High-scoring matches were further checked with the Smith/Waterman dynamic programming alignment algorithm (which allows gaps) (Smith and Waterman, 1981) to ascertain how much of the two proteins could be aligned (Appendix).

Of the 470 predicted coding regions, 159 (34%) are significantly similar to one or more proteins in the database. Of these, 102 (22%) have alignments which span more than half of both the predicted and the database sequences, suggesting that they might be true functional homologues. The cutoff of 50% alignment was an arbitrary choice. The remaining 57 coding regions with database similarities may be more divergent homologues, or may share functional domains with each other. They may also reflect errant GENEFINDER predictions; we know that some GENEFINDER predictions are missing exons or are fused to inappropriate exons. The putative homologues and similarities are listed in the Appendix.

There are genes for multiple members of some protein families, such as protein kinases (10 proteins), homeodomain proteins (7 proteins), and RNA helicases (6 proteins). Although extrapolation from these small numbers is dangerous, particularly since related genes may occur in clusters, it appears that there could be hundreds of members of genes encoding these protein families in the nematode genome. The homeodomain estimate is probably biased because the sequence includes a known cluster of

glin et al., 1991; Kenyon and Wang, 1991; Wang et al., 1993). It has been estimated from hybridization and PCR data that there are about 60 homeodomain genes (Chalfie, 1993).

It is the common experience of genome projects in several different organisms that about a third of randomly selected inferred protein sequences match something in the databases (Bork et al., 1992; Green et al., 1993). To some extent, this reflects the incompleteness of the databases, but there is also a deeper lesson. Green *et. al.* performed extensive pairwise comparisons of unselected protein data sets from the human, yeast, and *C. elegans* genome projects to the protein sequence database (Green et al., 1993). About 30-40% of each dataset matched sequences in the database. Green *et. al.* refer to matches between sequences from different phyla as ancient conserved regions (ACRs). ACRs represent fundamental protein components of animal life – proteins that arose before the major radiation of metazoan phyla 580 to 540 million years ago (Knoll, 1992) and have been conserved since then. The surprising result of Green *et. al.* is that few new ACRs are detected by comparing the new datasets to *each other*. In other words, newly sequenced worm proteins that match a newly sequenced human protein are very likely (>90%) to match an existing database sequence as well. This implies that the database already contains examples of most ACR's, and that the number of ACR families is relatively small (Green *et. al.* estimate on the order of 1,000). The remaining 60-70% of sequences have either diverged too far to be detected by sequence-based comparisons, or they have arisen fairly recently in evolution.

In this regard, it is interesting that there are many similarities between pairs of the 470 predicted worm genes with no similarities to existing database sequences. Using approximately the same procedure as Green et. al. (1993), I found 29 ACR (encoding kinases, RNA helicases, collagens, etc.) and 23 non-ACR families in pairwise comparisons between the 470 predicted proteins (119/470 have similarity to at least one other predicted protein in the set). Many of the non-ACR similarities are very highly similar and sometimes nearly identical; one match, between two predicted genes (F22B7.5 and C38C10.4) about 750 kb apart, shows 95% identity over almost 500 deduced amino acids and must be the result of a recent duplication. There are a number of scenarios that could explain the presence of a high proportion of gene families within the nematode that are not already in the database, given the ACR story of Green et. al. (1993). Two likely scenarios are: 1) The matches may be to gene families that emerged late in evolution and are playing functional roles specific to all or part of the phylum *Nematoda* (a strong Darwinist viewpoint). 2) The matches may reflect the fact that the genome is fluidly evolving by gene duplication and divergence, and parts of the genome are being copied around at a high rate regardless of function (a more neutral, evolutionary drift viewpoint). Either way, it appears likely that a significant fraction of nematode genes radiated fairly recently.

## Transposons

Two different sorts of transposons are known in *C. elegans* (Collins et al., 1989; Dreyfus and Emmons, 1991; Levitt and Emmons, 1989; Wood, 1988; Yuan et al., 1991). The first type, exemplified by Tc1, Tc2, and Tc3 elements, are 1.6 to 2.1 kilobases in length with short inverted terminal repeats, and open reading frames that may encode transposases (Collins et al., 1989; Levitt and Emmons, 1989; Wood, 1988). They are members of a large family of eukaryotic transposable elements that includes, for instance, *Drosophila* mariner and P elements (Robertson, 1993). A Tc3 element has

each other and significantly similar to *Drosophila* mariner elements have also been found (in C30A5 and ZK370). The other known type of *C. elegans* transposon, exemplified by Tc4 and Tc6 elements, are composed entirely of long inverted repeats and appear to encode no protein; their structure is analogous to that of *Drosophila* foldback elements (Dreyfus and Emmons, 1991; Yuan et al., 1991). Two Tc4-related elements have been detected in the genome sequence (in C27D11 and ZK686).

Another major class of eukaryotic transposable elements are retrotransposons (Singer and Berg, 1991). Retrotransposons encode proteins with similarity to reverse transcriptases and are thought to transpose through an RNA intermediate. Retrotransposons come in two distinct families. Some have long terminal direct repeats of several hundred nucleotides (LTR's) and have structural similarity to retroviruses. A second class, the non-LTR retrotransposons, lacks the direct repeats and is less understood. Until recently, neither form of retrotransposon had been seen in *C. elegans*, though examples of LTR-containing retrotransposons have been found in the nematodes *Panagrellus redivivus* and *Ascaris lumbricoides* (Wood, 1988).

Six regions with significant similarity to reverse transcriptases have been detected in the genome sequence. None of these appear to be associated with long terminal repeats. Five of the six are clearly related to each other, though divergent (less than 60% identity), and their closest similarities are to insect non-LTR retrotransposons, so they are likely to represent a family of nematode non-LTR retrotransposons. The sixth element (F44E2.1) is very diverged from the other five, and is more closely related to RVT's of the gypsy-like LTR-containing retrotransposons, though I can find no LTR's around it. It is impossible to tell whether any of the elements encode functional reverse transcriptase. One element, in C07A9, almost certainly is nonfunctional because a small inverted repeat element has apparently inserted itself into the retrotransposon and disrupted the RVT reading frame. Much more analysis needs to be done on these regions, but for now it appears that there are at least two families of retrotransposons in the nematode genome sequence, and members of the one family may be quite numerous.

## SENSITIVE FEATURE RECOGNITION

There are many interesting parallels between linguistic problems – for instance, parsing a sentence, or recognizing words spoken by speakers with different accents – and sequence recognition problems – for instance, parsing sequence into exons and introns, or recognizing divergent sequence family members (Searls, 1992). A direction that I have been involved in is to use adaptive statistical models and other theoretical techniques borrowed from the fields of speech recognition and formal linguistics to recognize features in the genome sequence.

These methods build probabilistic models of sequence families from multiple alignments, similar to sequence “profiles” (Barton, 1990; Gribskov et al., 1990; Krogh et al., 1993). The models can be used for sensitive recognition of more members of a family. Unlike profiles, these models are “adaptive,” meaning that they can be learned automatically from a set of initially unaligned example sequences (Krogh et al., 1993). For analysis of many families in large amounts of genome sequence, it is quite advantageous to have methods which learn on their own, bypassing the laborious step of constructing a trustworthy multiple sequence alignment.

Two kinds of adaptive statistical models are in use. The first, called hidden Markov models (HMM's), model primary sequence information only. They are good for recognition of protein and DNA sequence family members (Krogh et al., 1993). The second,

consensus for RNA sequence families (Eddy and Durbin, 1993; Sakakibara et al., 1993).

Work with both sorts of models is at a preliminary stage. One of our short-term goals is to build a library of models for many known protein, DNA, and RNA families and motifs, and automate the process of screening new sequence using these sensitive models. So far, we have used HMM's and covariance models on some trial problems.

### **An immunoglobulin superfamily member detected by HMMs**

A number of labs use *C. elegans* as a model system for neural development (Wood, 1988). Many nematode genes are known which disrupt proper axonal guidance (Hedgecock et al., 1987; Hedgecock et al., 1990; Wadsworth and Hedgecock, 1992). In the development of mammalian nervous systems, a family of molecules containing multiple repeats of an immunoglobulin (Ig) superfamily domain, including the neural cell adhesion molecule (NCAM), are thought to be some of the key players (Bixby and Harris, 1991; Hynes and Lander, 1992). Indeed, one of the *C. elegans* axon guidance genes, *unc-5*, contains two NCAM-like Ig motifs; mutations in *unc-5* disrupt a subset of dorsal-wards circumferential pioneer axon migrations (Hedgecock et al., 1990). It would be very interesting to detect additional superfamily members in the genome sequence.

Unfortunately, the Ig superfamily is one of the most divergent sequence families. It can be difficult or impossible to detect Ig superfamily members by routine database search techniques. I trained an HMM to recognize NCAM-like Ig domains, using a training set of domains taken from the SWISSPROT protein database, and used that model to search through the 470 predicted proteins. A single protein was detected, ZC262.3, which has one Ig superfamily domain (Figure 2). BLAST searches had missed the similarity. A closer examination of ZC262.3 shows that it has a putative transmembrane region. The Ig domain would presumably be extracellular (although a few examples of intracellular Ig domains are known). The overall structure of the predicted ZC262.3 protein is not similar to any known members of the NCAM family, however, and its function is unclear. Perhaps the next logical step would be to subclone a piece of ZC262.3 and ask whether its expression is tissue-specific, for instance if it is localized to neurons (by either *in situ* RNA hybridization or construction of transgenic animals carrying *ZC262.3::lacZ* fusions).

### **tRNAs detected by covariance models**

Obviously, much of the interest in the genome sequence focuses on the protein coding regions. But there are also RNA genes, such as the well-known tRNAs, snRNAs, and rRNAs, and it will not be at all surprising to find other functional RNA genes and motifs as well. Many RNA sequence families are difficult to detect by conventional primary sequence analysis techniques, because it is in general much more satisfactory to represent RNA families as consensus secondary structures rather than consensus primary sequences. From work by Richard Durbin and myself in Cambridge and work by David Haussler and coworkers at UC-Santa Cruz, it has recently become possible to construct fully probabilistic models of RNA secondary structure consensus, analogous to the HMM's we use for protein and DNA families, and search very sensitively for members of known RNA families (Eddy and Durbin, 1993; Sakakibara et al., 1993). We call these models "covariance models."

There are thought to be about 300 tRNA genes in the nematode (Wood, 1988). Figure 3 shows the results of a search over the genomic sequence using a covariance model trained on tRNA example sequences. Fourteen tRNA genes have been detected.

Figure 2: Alignment of Ig-like domains from three neural cell adhesion molecules with the Ig-like domain detected in ZC262.3. NRG\_DROME is *Drosophila* neuroglian; CAML\_DROME is mouse N-CAM L1; NCA2\_HUMAN is a human N-CAM. The alignment was produced automatically by an HMM trained to recognize N-CAM sequences.

Two of the genes contain introns in their anticodon loop. These genes were detected despite the fact that this particular model was trained entirely on intronless sequences, which is a nice example of the model's flexibility. Three of these tRNA genes turn out to be in introns of predicted protein-coding genes.

## Discussion

The complete cloning and sequencing of *C. elegans* will change how its molecular biology is studied. Much time and effort is still devoted to the cloning of genes, but soon *everything* will be cloned and sequenced from *C. elegans*, and molecular cloning will often be as simple as an electronic database retrieval of the sequence and a letter to Cambridge or St. Louis for the clone. The burden of work can shift to the daunting task of identifying the functions of the estimated 17,000 genes.

Clearly, some hundreds of "interesting" genes will be attacked immediately by conventional molecular biology. But it is neither wise nor possible to devote this kind of effort to every one of the genes. A challenge will be to get at the functions of the rest by scaling up rapid screening techniques, particularly ones that can be done in parallel and/or partially automated. To ask if and where a predicted gene is expressed, one might use *in situ* RNA hybridization to PCR-produced probes, or transgenic animals carrying reporter gene fusions to the predicted promoter (Hope, 1991), or hybridization to tissue-specific or even cell-specific cDNA populations. Gene knockout by PCR-based selection of transposon insertions and systematic determination of null phenotypes is feasible though a bit laborious (Plasterk, 1992). Various sorts of molecular genetic screens could be scaled up, such as enhancer trapping (Bellen et al., 1989; Bier et al., 1989; Wilson et al., 1989) and insertional mutagenesis, since one will be able to immediately identify a candidate gene from a bit of flanking sequence around an insert.

A more computer-reliant style of molecular biology will continue to develop. Computational analysis will become an initial exploratory step in many projects, rather than just the familiar final step of "what is my sequence related to?" For instance, computational probing techniques for particular sequence motifs and sequence families are far more sensitive than molecular hybridization and degenerate PCR cloning approaches, and will rapidly supercede them where genomic sequence is available.

Figure 3: tRNA's detected in the genomic sequence. The tail of non-tRNA scores is shown as well to illustrate the good separation of signals from noise. The scores are in bits, which are log base 2 probabilities.

science. At present, the data are limited and we remain confined to something much like the compilation of a dictionary. But however fascinating the etymology of our “words” may be, we should not be contented with merely cataloging genes and motifs into tidy Linnaean trees with some elaborate nomenclature. The future challenges for computational biology are to find ways to ask questions about the larger context of complex eukaryotic genomes. What genomic changes or new gene families correlate with the relatively explosive major evolutionary radiations, such as the diversification of multicellular eukaryotes about 1,200-1,000 million years ago, or the radiation of the animal phyla about 580-540 million years ago (Knoll, 1992)? What is occupying the spaces between genes – is it functional, or junk? Are there sequences determining higher order chromosome structure? What sequences control the regulation of individual genes and groups of coordinately regulated genes? How much sequence is created by autonomously replicating “selfish” elements? What are the mechanisms responsible for the evolution of genes and whole genomes? These sorts of questions can and will be tentatively explored in the *C. elegans* genome sequence, but they will be attacked most powerfully by comparative analysis of multiple divergent genome sequences.

## Acknowledgements

Many thanks to the members of the nematode genome project and the St. Louis and Cambridge genome informatics groups, particularly John Sulston, Alan Coulson, Richard Durbin, Erik Sonnhammer, and Graeme Mitchison on the Cambridge side, for their continuing help and advice. I am supported by a Human Frontier Science Program long-term postdoctoral fellowship.

## REFERENCES

- Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J., 1990, Basic local alignment search tool, *J. Mol. Biol.* 215:403.
- Bairoch, A. and Boeckmann, B., 1993, The SWISS-PROT protein sequence data bank, recent developments, *Nucl. Acids Res.* 21: 3093.
- Barker, W. C., George, D. G., Mewes, H.-W., Pfeiffer, F., and Tsugita, A, 1993, The PIR-international databases, *Nucl. Acids Res.* 21:3089.
- Barton, G. J., 1990, Protein multiple sequence alignment and flexible pattern matching, *Meth. Enzymol.* 183:403.
- Bellen, H. J., O’Kane, C. J., Wilson, C., Grossniklaus, U., Pearson, R. K., and Gehring, W. J., 1989, P-element-mediated enhancer detection: a versatile method to study development in *Drosophila*, *Genes Dev.* 3:1288.
- Benson, D., Lipman, D. J., and Ostell, J., 1993, GenBank, *Nucl. Acids Res.* 21:2963.
- Bier, E., Vaessin, H., Shepherd, S., Lee, K., McCall, K., Barbel, S., Ackerman, L., Carretto, R., Uemura, T., Grell, E., Jan, L. Y., and Jan, Y. N., 1989, Searching for pattern and mutation in the *Drosophila* genome with a P-lacZ vector, *Genes Dev.* 3:1273.
- Bixby, J. L. and Harris, W. A., 1991, Molecular mechanisms of axon growth and guidance, *Ann. Rev. Cell Biol.* 7:117.
- Bork, P., Ouzounis, C., Sander, C., Scharf, M., Schneider, R., and Sonnhammer, E., 1992, Comprehensive sequence analysis of the 182 predicted open reading frames of yeast chromosome III, *Protein Science* 1:1677.
- Brenner, S., 1974, The genetics of *Caenorhabditis elegans*, *Genetics* 77:71.

- F., and Waterston, R., 1991, Nematode homeobox cluster, *Nature* 351:703.
- Chalfie, M., 1993, Homeobox genes in *Caenorhabditis elegans*, *Curr. Opin. Genet. Dev.* 3:275.
- Collins, J., Forbes, E., and Anderson, P., 1989, The Tc3 family of transposable genetic elements in *Caenorhabditis elegans*, *Genetics* 121:47.
- Coulson, A., Kozono, Y., Lutterbach, B., Shownkeen, R., Sulston, J., and Waterston, R., 1991, YACs and the *C. elegans* genome, *BioEssays* 13:413.
- Coulson, A., Sulston, J., Brenner, S., and Karn, J., 1986, Toward a physical map of the genome of the nematode *Caenorhabditis elegans*, *Proc. Natl. Acad. Sci. USA* 83:7821.
- Coulson, A., Waterston, R., Kiff, J., Sulston, J., and Kohara, Y., 1988, Genome linking with yeast artificial chromosomes, *Nature* 335:184.
- Daniels, D. L., Plunkett, G., Burland, V., and Blattner, F. R., 1992, Analysis of the *Escherichia coli* genome: DNA sequence of the region from 84.5 to 86.5 minutes, *Science* 257:771.
- Dear, S. and Staden, R., 1991, A sequence assembly and editing program for efficient management of large projects, *Nucl. Acids Res.* 19:3907.
- Dreyfus, D. H. and Emmons, S. W., 1991, A transposon-related palindromic repetitive sequence from *C. elegans*, *Nucl. Acids Res.* 19:1871.
- Eddy, S. R. and Durbin, R., 1993, Analysis of RNA sequence families using adaptive statistical models, unpublished manuscript.
- Green, P., Lipman, D., Hillier, L., Waterston, R., States, D., and Claverie, J.-M., 1993, Ancient conserved regions in new gene sequences and the protein databases, *Science* 259:1711.
- Gribskov, M., Luthy, R., and Eisenberg, D., 1990, Profile analysis, *Meth. Enzymol.* 183:146.
- Hedgecock, E. M., Culotti, J. G., and Hall, D. H., 1990, The *unc-5*, *unc-6*, and *unc-40* genes guide circumferential migrations of pioneer axons and mesodermal cells on the epidermis in *C. elegans*, *Neuron* 4:61.
- Hedgecock, E. M., Culotti, J. G., Hall, D. H., and Stern, B. D., 1987, Genetics of cell and axon migrations in *Caenorhabditis elegans*, *Development* 100:365.
- Hope, I., 1991, 'Promoter trapping' in *Caenorhabditis elegans*, *Development* 113:399.
- Hynes, R. O. and Lander, A. D., 1992, Contact and adhesive specificities in the associations, migrations, and targeting of cells and axons, *Cell* 68:303.
- Kenyon, C. and Wang, B., 1991, A cluster of antennapedia-class homeobox genes in a nonsegmented animal, *Science* 253:516.
- Knoll, A. H., 1992, The early evolution of eukaryotes: a geological perspective, *Science* 256:622.
- Krogh, A., Brown, M., Mian, I., Sjolander, K., and Haussler, D., 1993, Hidden Markov models in computational biology: applications to protein modeling, unpublished manuscript.
- Levitt, A. and Emmons, S. W., 1989, The Tc2 transposon in *Caenorhabditis elegans*, *Proc. Natl. Acad. Sci. USA* 86:3232.
- McCombie, W. R., Adams, M. D., Kelly, J. M., Fitzgerald, M. G., Utterback, T. R., Khan, M., Dubnick, M., Kerlavage, A. R., Venter, J. C., and Fields, C., 1992, *Caenorhabditis elegans* expressed sequence tags identify gene families and potential disease gene homologues, *Nature Genet.* 1:124.
- Oliver, S., van der Aart, Q., Agostoni-Carbone, M., Aigle, M., et al., 1992, The complete DNA sequence of yeast chromosome III, *Nature* 357:38.
- Plasterk, R. H., 1992, Reverse genetics of *Caenorhabditis elegans*, *BioEssays* 14:629.
- Plunkett, G., Burland, V., Daniels, D. L., and Blattner, F. R., 1993, Analysis of the *Escherichia coli* genome. III. DNA sequence of the region from 87.2 to 89.2 minutes, *Nucl. Acids Res.* 21:3391.
- Rice, C. M., Fuchs, R., Higgins, D. G., Stoehr, P. J., and Cameron, G. N., 1993, The EMBL data library, *Nucl. Acids Res.* 21:2967.
- Robertson, H. M., 1993, The mariner transposable element is widespread in insects, *Nature* 362:241.

- D., 1993, The application of stochastic context-free grammars to folding, aligning and modeling homologous RNA sequences, unpublished manuscript.
- Searls, D. B., 1992, The linguistics of DNA, *American Scientist* 80:579.
- Singer, M. and Berg, P., 1991, *Genes and Genomes*, University Science Books, Mill Valley, CA.
- Smith, T. and Waterman, M., 1981, Identification of common molecular subsequences, *J. Mol. Biol.* 147:195.
- Sonnhammer, E. L. and Durbin, R., 1993, A workbench for large-scale sequence homology analysis, unpublished manuscript.
- Sulston, J., Du, Z., Thomas, K., Wilson, R., Hillier, L., Staden, R., Halloran, N., Green, P., Thierry-Mieg, J., Qiu, L., Dear, S., Coulson, A., Craxton, M., Durbin, R., Berks, M., Metzstein, M., Hawkins, T., Ainscough, R., and Waterston, R., 1992, The *C. elegans* genome sequencing project: a beginning, *Nature* 356:37.
- Sulston, J., Schierenberg, E., White, J., and Thomson, J., 1983, The embryonic cell lineage of the nematode *Caenorhabditis elegans*, *Devel. Biol.* 100:64.
- Wadsworth, W. G. and Hedgecock, E. M., 1992, Guidance of neuroblast migrations and axonal projections in *Caenorhabditis elegans*, *Curr. Opinion Neuro.* 2:36.
- Wang, B. B., Muller-Immergluck, M. M., Austin, J., Robinson, N. T., Chisholm, A., and Kenyon, C., 1993, A homeotic gene cluster patterns the anteroposterior body axis of *C. elegans*, *Cell* 74:29.
- Waterston, R., Martin, C., Craxton, M., Hunyh, C., Coulson, A., Hillier, L., Durbin, R., Green, P., Shownkeen, R., Halloran, N., Metzstein, M., Hawkins, T., Wilson, R., Berks, M., Du, Z., Thomas, K., Thierry-Mieg, J., and Sulston, J., 1992, A survey of expressed genes in *Caenorhabditis elegans*, *Nature Genet.* 1:114.
- Watson, A., Smaldon, N., Lucke, R., and Hawkins, T., 1993, The *Caenorhabditis elegans* genome sequencing project: first steps in automation, *Nature* 362:569.
- White, J., Southgate, E., Thomson, J., and Brenner, S., 1986, The structure of the nervous system of the nematode *Caenorhabditis elegans*, *Phil. Trans. R. Soc. Lond.* 314:1.
- Wilson, C., Pearson, R. K., Bellen, H. J., O’Kane, C. J., Grossniklaus, U., and Gehring, W. J., 1989, P-element-mediated enhancer detection: an efficient method for isolating and characterizing developmentally regulated genes in *Drosophila*, *Genes Dev.* 3:1301.
- Wilson, R., Ainscough, R., Anderson, K., Baynes, C., Berks, M., et al., 1993, The *C. elegans* genome project: nucleotide sequence of over two megabases from chromosome III, unpublished manuscript.
- Wood, W. B., ed., 1988, *The Nematode Caenorhabditis elegans*, Cold Spring Harbor Laboratory, New York, NY.
- Yuan, J., Finney, M., Tsung, N., and Horvitz, H. R., 1991, Tc4, a *Caenorhabditis elegans* transposable element with an unusual foldback structure, *Proc. Natl. Acad. Sci. USA* 88:3334.

## APPENDIX

### A. Putative homologues

The name of the predicted protein, its length, the length of the maximal scoring Smith/Waterman alignment to the database sequence, percentage residue identity over that alignment, and the name and description of the most similar database sequence in SWISSPROT and PIR are indicated.

Protein	length	align	id	hits	description
B0303.12	195	135	40.1%	RL11_ECOLI	50S RIBOSOMAL PROTEIN L11.
B0303.4	315	243	23.9%	PNMT_HUMAN	PHENYLETHANOLAMINE-N-METHYLTRANSFERASE ...
B0303.5	497	430	28.1%	THIK_YEAST	3-KETOACYL-COA THIOLASE PEROXISOMAL PRE...

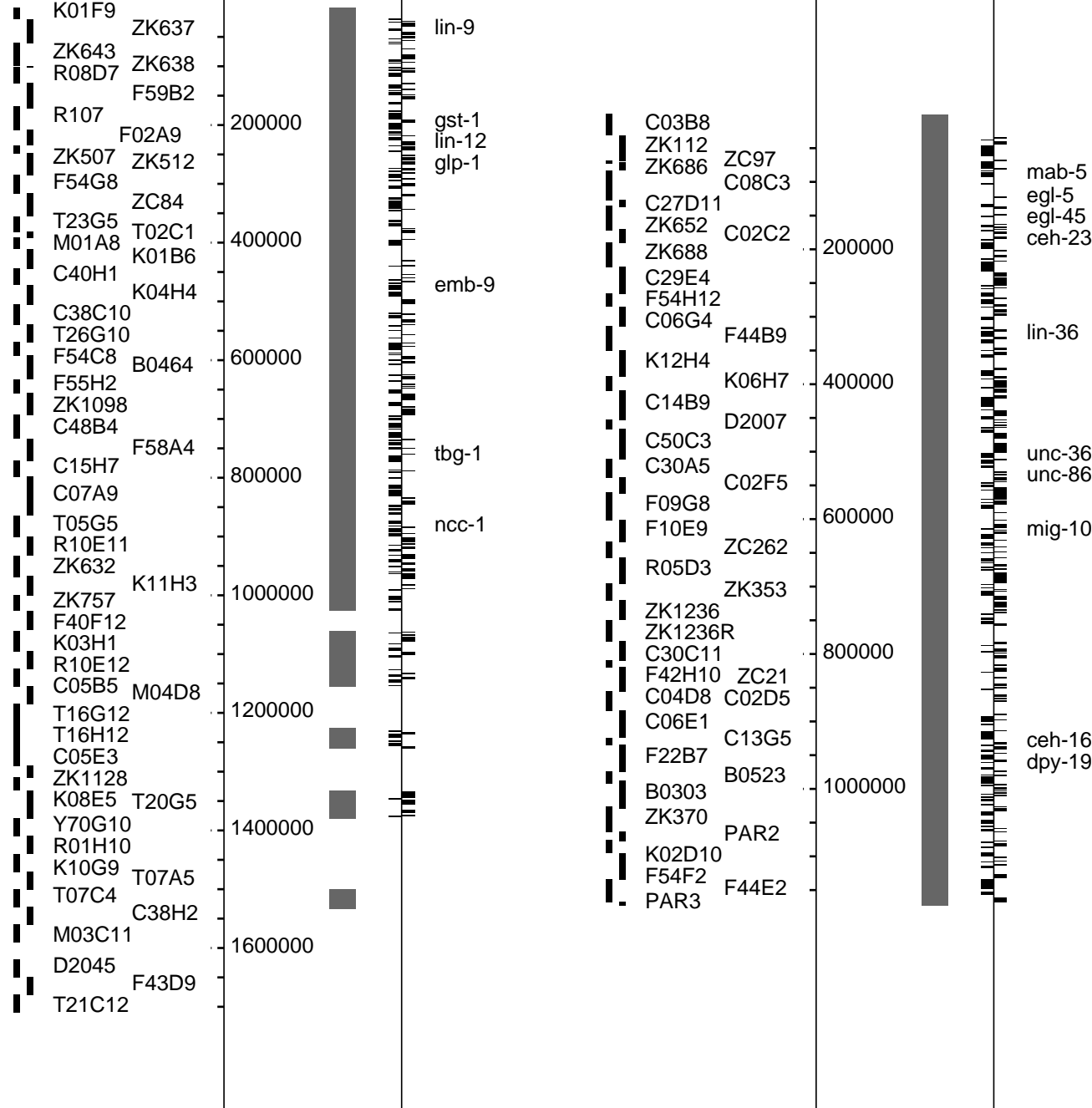
B0303.TC3	329	263	32.8%	YT31_CAEEL	HYPOTHETICAL 31.8 KD PROTEIN FROM TRANS...
B0464.1	531	501	62.4%	SYD2_HUMAN	ASPARTYL-TRNA SYNTHETASE ALPHA-2 SUBUNI...
B0464.5	1087	698	18.4%	S30783	PROTEIN KINASE HOMOLOG KNS1 - YEAST (SA...
B0523.1	363	218	32.5%	KROS_AVISU	ROS TYROSINE KINASE TRANSFORMING PROTEI...
B0523.5	848	748	27.7%	GELS_PIG	GELSOLIN PRECURSOR, PLASMA (ACTIN-DEPOL...
C02C2.3	458	350	20.3%	ACHG_RAT	ACETYLCHOLINE RECEPTOR PROTEIN, GAMMA C...
C02C2.4	568	472	24.4%	S27951	SODIUM/PHOSPHATE TRANSPORT PROTEIN, REN...
C02D5.1	332	321	31.2%	ACDL_RAT	ACYL-COA DEHYDROGENASE PRECURSOR, LONG-...
C02F5.3	573	364	56.7%	JC1349	DEVELOPMENTALLY REGULATED GTP-BINDING P...
C06E1.4	983	913	36.3%	GLRK_LYMST	GLUTAMATE RECEPTOR PRECURSOR.
C08C3.1	218	74	100.0%	HM11_CAEEL	HOMEBOX PROTEIN CEH-11 (FRAGMENT).(egl-5)
C08C3.3	353	210	100.0%	HMMA_CAEEL	HOMEBOX PROTEIN MAB-5 (FRAGMENT).(mab-5)
C14B9.1	110	92	40.2%	CRAB_MESAU	ALPHA CRYSTALLIN B CHAIN (ALPHA(B)-CRYS...
C14B9.2	664	601	48.1%	ER72_MOUSE	PROTEIN DISULFIDE ISOMERASE-RELATED PRO...
C14B9.7	161	160	62.5%	R5RT21	RIBOSOMAL PROTEIN L21 - RAT
C14B9.8	1257	1189	38.2%	KPBA_RABIT	PHOSPHORYLASE KINASE ALPHA CHAIN, SKELE...
C29E4.1	305	266	43.5%	CAC8_CAEEL	CUTICLE COLLAGEN 8.
C29E4.7	250	214	25.0%	S16268	AUXIN-INDUCED PROTEIN (CLONE PGNT35) - ...
C29E4.8	248	233	61.4%	KAD2_RAT	ADENYLATE KINASE ISOENZYME 2, MITOCHOND...
C30A5.3	378	260	30.2%	S30854	PHOSPHOPROTEIN PHOSPHATASE - YEAST (SAC...
C30A5.6	429	428	91.2%	UN86_CAEEL	TRANSCRIPTION FACTOR UNC-86.
C30A5.7	467	466	100.0%	UN86_CAEEL	TRANSCRIPTION FACTOR UNC-86.
C30C11.2	504	497	42.7%	DXA2_MOUSE	PROBABLE DIPHENOL OXIDASE A2 COMPONENT ...
C30C11.4	776	688	36.6%	S30788	HEAT SHOCK PROTEIN HOMOLOG MSI3 - YEAST...
C38C10.1	374	293	36.6%	NK1R_CAVPO	SUBSTANCE-P RECEPTOR (SPR) (NK-1 RECEPT...
C38C10.2	472	448	27.4%	S27951	SODIUM/PHOSPHATE TRANSPORT PROTEIN, REN...
C40H1.4	291	228	26.3%	YCS4_YEAST	HYPOTHETICAL 40.0 KD PROTEIN IN CRY1-RB...
C50C3.11	734	629	25.5%	CIC2_RABIT	DIHYDROPRYRIDINE-SENSITIVE L-TYPE... (unc-36)
F02A9.5	608	484	31.5%	PCCB_RAT	PROPIONYL-COA CARBOXYLASE BETA CHAIN PR...
F09G8.6	278	256	38.3%	A44984	COLLAGEN - NEMATODE (HAEMONCHUS CONTORTUS)
F22B7.7	335	312	22.2%	KCH_ECOLI	PUTATIVE POTASSIUM CHANNEL PROTEIN.
F44B9.8	447	346	38.3%	A45253	ACTIVATOR 1 37 KDA SUBUNIT, A1 37 KDA S...
F44B9.9	220	188	26.3%	S24264	PROTEIN PHOSPHATASE 1A - ARABIDOPSIS TH...
F44E2.1	1746	942	26.0%	POL2_DROME	RETROVIRUS-RELATED POL POLYPROTEIN (PRO...
F44E2.8	308	171	22.8%	YBIA_ECOLI	HYPOTHETICAL 18.7 KD PROTEIN IN RHLE-DI...
F54C8.1	298	286	45.9%	HCDH_PIG	3-HYDROXYACYL-COA DEHYDROGENASE (EC 1.1.1...
F54C8.5	207	189	35.4%	RAS_LENED	RAS-LIKE PROTEIN.
F54F2.1	1226	1194	27.3%	ITAV_HUMAN	VITRONECTIN RECEPTOR ALPHA SUBUNIT PREC...
F55H2.1	184	132	54.5%	SODC_BOVIN	SUPEROXIDE DISMUTASE (CU-ZN) (EC 1.15.1.1...
F55H2.2	257	251	51.0%	S30826	HYPOTHETICAL PROTEIN 11 - YEAST (SACCHA...
F55H2.5	266	243	33.7%	C561_BOVIN	CYTOCHROME B561.
F58A4.10	164	153	50.3%	UBC7_YEAST	UBIQUITIN-CONJUGATING ENZYME E2-18 KD (...)
F58A4.4	410	403	39.3%	PRI1_MOUSE	DNA PRIMASE 49 KD SUBUNIT (EC 2.7.7.-) ...
F58A4.7	292	228	32.0%	A36394	TRANSCRIPTION FACTOR AP-4 - HUMAN (FRAG...
F58A4.8	444	436	43.8%	TBG_XENLA	TUBULIN GAMMA CHAIN. (tbg-1)
F58A4.9	144	129	27.9%	RPC9_YEAST	DNA-DIRECTED RNA POLYMERASES I AND III ...
F59B2.3	418	351	31.8%	NAGA_ECOLI	N-ACETYLGLUCOSAMINE-6-PHOSPHATE DEACETY...
F59B2.7	205	203	77.1%	RAB6_HUMAN	RAS-RELATED PROTEIN RAB-6.
GLP1A.cds	1295	1294	100.0%	GLP1_CAEEL	GLP-1 PROTEIN PRECURSOR.
K04H4.1	1744	1743	92.6%	CA14_CAEEL	COLLAGEN ALPHA 1(IV) CHAIN. (emb-9)
K06H7.4	377	371	40.5%	S24168	HYPOTHETICAL PROTEIN - HUMAN
K06H7.8	283	236	24.2%	HR25_YEAST	CASEIN KINASE I HOMOLOG HRR25 (EC 2.7.1.1...
K11H3.3	374	320	24.5%	A45763	UNCOUPLING PROTEIN, MITOCHONDRIAL - HUMAN
K12H4.4	180	177	48.6%	SPC2_CANFA	MICROSOMAL SIGNAL PEPTIDASE 23 KD SUBUN...
LIN12A.cd	1429	1428	100.0%	LI12_CAEEL	LIN-12 PROTEIN PRECURSOR.
R05D3.7	843	816	46.2%	KINH_LOLPE	KINESIN HEAVY CHAIN.
R08D7.5	173	163	31.7%	CATR_CHLRE	CALTRACTIN (20 KD CALCIUM-BINDING PROTE...
R08D7.6	841	510	36.8%	CNAG_BOVIN	CGMP-DEPENDENT 3',5'-CYCLIC PHOSPHODIES...
R107.2	285	281	25.4%	YJEF_ECOLI	HYPOTHETICAL PROTEIN IN AMIB 5'REGION (...)
R107.7	208	207	100.0%	GTP_CAEEL	GLUTATHIONE S-TRANSFERASE P (EC 2.5.1.18).

R10E11.4	289	212	30.1%	A24148	N-ACETYLACTOSAMINE SYNTHASE - BOVINE (...)
T05G5.3	332	331	100.0%	S26572	P34 CDC2-LIKE PROTEIN (ncc-1)
T05G5.5	196	162	30.9%	S27735	HYPOTHETICAL PROTEIN A - THERMUS AQUATICUS
T05G5.6	288	273	60.2%	ECHM_RAT	ENOYL-COA HYDRATASE, MITOCHONDRIAL PREC...
T16H12.7	193	175	25.0%	MIPP_MOUSE	MIPP PROTEIN (MURINE IAP-PROMOTED PLACE...
T23G5.1	788	780	73.4%	A24050	RIBONUCLEOSIDE-DIPHOSPHATE REDUCTASE CH...
T23G5.5	514	494	44.7%	NTNO_HUMAN	SODIUM-DEPENDENT NORADRENALINE TRANSPOR...
T26G10.1	489	430	33.8%	DEAD_ECOLI	PUTATIVE ATP-DEPENDENT RNA HELICASE DEAD.
T26G10.3	91	89	24.0%	RS24_HUMAN	P16632 40S RIBOSOMAL PROTEIN S24 (S19).
ZC21.2	823	751	36.4%	JH0588	CALMODULIN-BINDING PROTEIN TRPL - FRUIT...
ZC84.2	772	562	51.4%	CGCC_HUMAN	CGMP-GATED CATION CHANNEL PROTEIN (CYCL...
ZC84.4	425	291	19.7%	SSRB_MOUSE	SOMATOSTATIN RECEPTOR TYPE 2B.
ZK1098.4	305	304	35.3%	GCN3_YEAST	TRANSCRIPTION ACTIVATOR GCN3.
ZK1236.1	581	578	38.2%	LEPA_ECOLI	LEPA PROTEIN.
ZK1236.3	1364	715	22.3%	B34751	HYPOTHETICAL PROTEIN - AFRICAN MALARIA ...
ZK353.6	491	366	30.5%	AMPA_RICPR	ASPARTATE AMINOPEPTIDASE (EC 3.4.11.7) ...
ZK370.5	401	346	30.1%	BCKD_RAT	[3-METHYL-2-OXOBUTANOATE DEHYDROGENASE ...
ZK507.6	409	231	36.3%	CG2A_PATVU	G2/MITOTIC-SPECIFIC CYCLIN A.
ZK512.2	578	499	22.4%	MS16_YEAST	ATP-DEPENDENT RNA HELICASE MSS116.
ZK512.4	76	75	50.7%	A34731	SIGNAL RECOGNITION PARTICLE 9K CHAIN - DOG
ZK512.6	466	399	25.4%	S27951	SODIUM/PHOSPHATE TRANSPORT PROTEIN, REN...
ZK632.1	810	658	30.3%	CD46_YEAST	CELL DIVISION CONTROL PROTEIN CDC46 (MI...
ZK632.4	411	391	33.3%	MANA_EMENI	MANNOSE-6-PHOSPHATE ISOMERASE (EC 5.3.1...
ZK632.6	619	612	40.4%	A37273	CALNEXIN PRECURSOR - DOG
ZK632.8	178	177	48.6%	S28875	GTP-BINDING PROTEIN - ARABIDOPSIS THALIANA
ZK637.1	456	391	20.7%	A43267	SYNAPTIC VESICLE PROTEIN 2, SV2=82.7 KD...
ZK637.10	499	465	33.8%	GSHR_ECOLI	GLUTATHIONE REDUCTASE (EC 1.6.4.2) (GR).
ZK637.13	159	158	44.7%	GLBH_TRICO	GLOBIN-LIKE HOST-PROTECTIVE ANTIGEN PRE...
ZK637.7	610	609	99.0%	LIN9_CAEEL	LIN-9 PROTEIN.
ZK637.8	935	908	55.1%	VPP1_RAT	CLATHRIN-COATED VESICLE/SYNAPTIC VESICL...
ZK643.3	482	418	25.3%	CALR_PIG	CALCITONIN RECEPTOR PRECURSOR (CT-R).
ZK652.10	420	317	25.5%	S27951	SODIUM/PHOSPHATE TRANSPORT PROTEIN, REN...
ZK652.4	123	122	68.0%	R5RT35	RIBOSOMAL PROTEIN L35 - RAT
ZK652.5	305	291	26.6%	HMES_DROME	EMPTY SPIRACLES HOMEOTIC PROTEIN. (ceh-23)
ZK652.6	580	390	23.4%	S29962	REF(2)PERECTA PROTEIN - FRUIT FLY (DROS...
ZK686.2	696	649	23.8%	S31248	PROBABLE RNA HELICASE, ATP-DEPENDENT - ...

## B. Other significant similarities

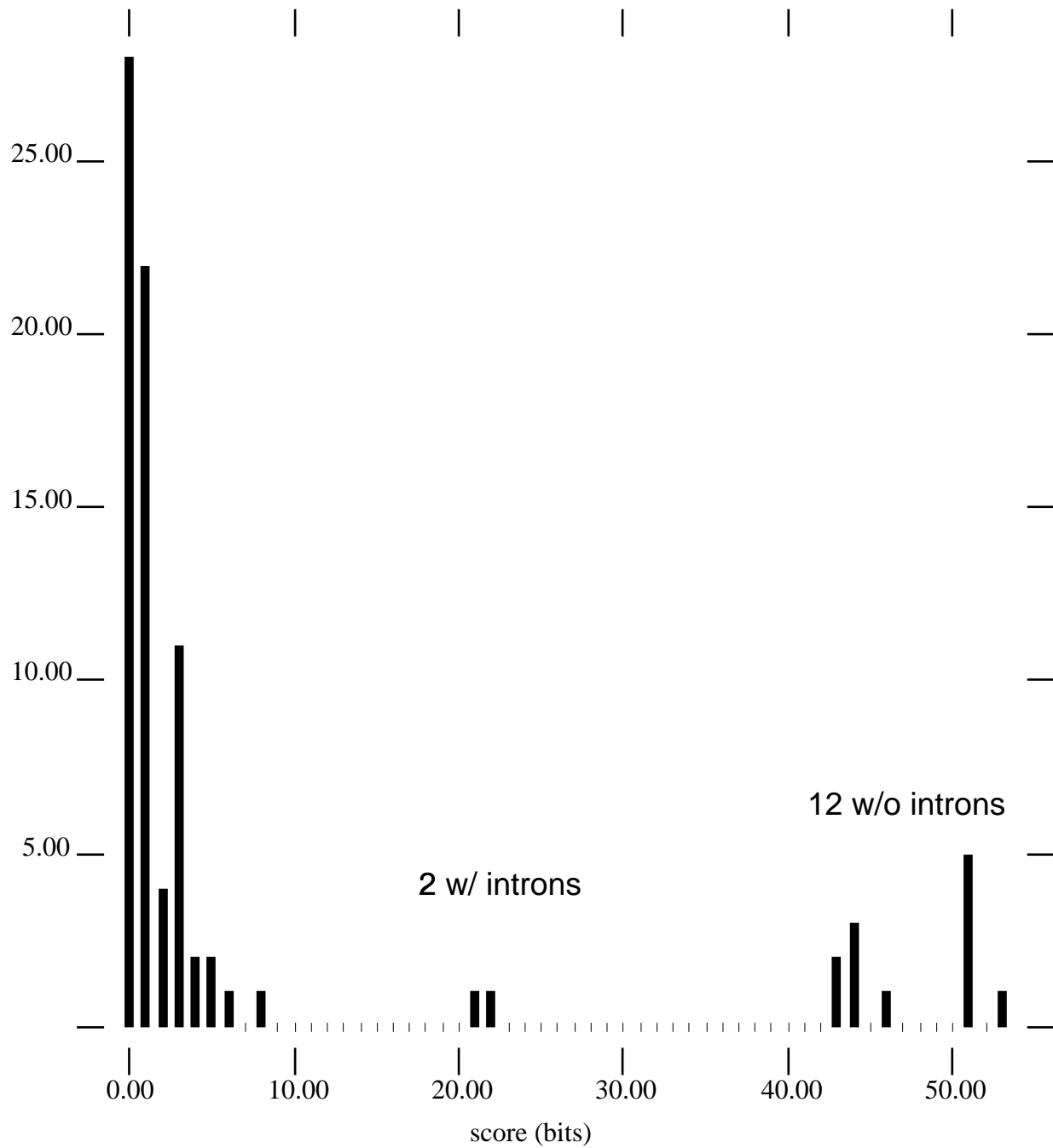
Protein	length	align	id	hits	description
B0303.3	424	182	31.9%	S25770	P33 RSP-1 PROTEIN - MOUSE
B0303.7	359	156	27.1%	NCF2_HUMAN	NEUTROPHIL NADPH OXIDASE FACTOR (P67-PH...
C02C2.1	484	229	25.1%	A40957	MONOPHENOL MONOOXYGENASE PRECURSOR - HU...
C02F5.7	489	362	22.1%	GRR1_YEAST	GRR1 PROTEIN.
C02F5.9	564	227	41.0%	PRC5_HUMAN	PROTEASOME COMPONENT C5 (EC 3.4.99.46) ...
C05B5.5	585	259	25.2%	TENA_HUMAN	P24821 TENASCIN PRECURSOR (TN)...
C06E1.10	1152	407	39.5%	S22609	HYPOTHETICAL PROTEIN - FRUIT FLY (DROS...
C08C3.2	274	128	26.9%	VA55_VACCC	PROTEIN A55.
C13G5.1	240	70	60.0%	HME6_APIME	HOMEBOX PROTEIN E60 (FRAGMENT). (ceh-16)
C15H7.2	266	233	24.5%	KFPS_DROME	DFPS TYROSINE KINASE (EC 2.7.1.112).
C30A5.4	102	83	30.6%	SYB_DROME	SYNAPTOBREVIN.
C38C10.5	1112	279	22.8%	RGR1_YEAST	GLUCOSE REPRESSION REGULATORY PROTEIN R...
C40H1.1	372	307	28.3%	S24577	OVARIAN PROTEIN - FRUIT FLY (DROSOPHILA...
C50C3.2	1009	950	20.6%	SPCA_DROME	SPECTRIN ALPHA CHAIN.
C50C3.5	178	82	34.1%	CALM_ACHKL	CALMODULIN.
C50C3.7	398	241	23.0%	IT5P_HUMAN	75 KD INOSITOL-1,4,5-TRISPHOSPHATE 5-PH...
F22B7.5	943	344	37.9%	DNAJ_BACSU	DNAJ PROTEIN.
F42H10.3	209	88	36.0%	SRC8_CHICK	SRC SUBSTRATE P80/85 PROTEINS (CORTACTIN).
F42H10.4	221	59	39.0%	TSF3_HELAN	POLLEN SPECIFIC PROTEIN SF3.

F44E2.3	244	205	25.5%	RU17_XENLA	U1 SMALL NUCLEAR RIBONUCLEOPROTEIN 70 KD.
F44E2.4	1609	209	22.8%	S03430	LOW DENSITY LIPOPROTEIN RECEPTOR PRECUR...
F44E2.6	152	105	41.0%	PILB_NEIGO	PILB PROTEIN.
F54C8.2	261	123	46.3%	JQ1343	HISTONE H3.3 - FRUIT FLY (DROSOPHILA ME...
F54C8.4	359	171	36.3%	A40781	ORF1 PROTEIN - AUTOGRAPHA CALIFORNICA N...
F54G8.2	827	231	51.0%	KDGL_DROME	PUTATIVE DIACYLGLYCEROL KINASE (EC 2.7....
F54G8.3	1139	144	32.9%	A41543	INTEGRIN ALPHA-6B CHAIN - HUMAN (FRAGMENT)
F54G8.4	932	354	19.9%	KRET_HUMAN	RET PROTO-ONCOGENE TYROSINE KINASE (EC ...
F54G8.5	413	399	19.1%	PATC_DROME	MEMBRANE PROTEIN PATCHED.
F58A4.1	258	170	30.1%	A43932	MUCIN - HUMAN (FRAGMENT)
F58A4.3	288	119	50.4%	H3_PEA	HISTONE H3.
F58A4.5	1222	594	27.4%	B34751	HYPOTHETICAL PROTEIN - AFRICAN MALARIA ...
F59B2.11	337	218	14.9%	AAC2_DICDI	AAC-RICH MRNA CLONE AAC11 PROTEIN (FRAG...
K02D10.1	786	212	28.0%	SN25_MOUSE	SYNAPTOSOMAL ASSOCIATED PROTEIN 25.
K06H7.1	547	278	55.2%	S22127	PROTEIN KINASE - FRUIT FLY (DROSOPHILA ...
K06H7.3	831	245	36.5%	IPPI_YEAST	ISOPENTENYL-DIPHOSPHATE DELTA-ISOMERASE...
K12H4.1	586	538	30.8%	PRO_DROME	PROTEIN PROSPERO.
K12H4.8	1822	323	23.3%	A31922	ATP-DEPENDENT RNA HELICASE HOMOLOG - FR...
M01A8.4	69	47	48.9%	BIK1_YEAST	NUCLEAR FUSION PROTEIN BIK1.
R05D3.1	2434	1139	41.5%	TOP2_SCHPO	DNA TOPOISOMERASE II (EC 5.99.1.3).
R10E11.1	2015	257	24.7%	FSH_DROME	FEMALE STERILE HOMEOTIC PROTEIN (FRAGIL...
R10E11.3	408	295	19.8%	S22158	TRANSFORMING PROTEIN (CLONE 213) - HUMAN
T02C1.1	160	88	29.5%	RA18_YEAST	DNA REPAIR PROTEIN RAD18.
T05G5.1	346	228	16.7%	S27770	HYPOTHETICAL PROTEIN 1 - AFRICAN MALARI...
T23G5.2	470	213	31.3%	SC14_YEAST	SEC14 CYTOSOLIC FACTOR.
ZC21.4	733	199	30.2%	S29956	BETA-CHIMAERIN - RAT
ZC262.6	466	158	51.5%	S24603	KINESIN HEAVY CHAIN - HUMAN
ZC84.1	2885	53	45.3%	ISHP_STOHE	KUNITZ-TYPE PROTEINASE INHIBITOR SHPI.
ZC97.2	296	278	23.7%	S31248	PROBABLE RNA HELICASE, ATP-DEPENDENT - ...
ZK112.7	3343	n.d.	n.d.	A41087	CADHERIN-RELATED TUMOR SUPPRESSOR PRECU...
ZK370.3	923	699	22.4%	TALI_MOUSE	TALIN.
ZK507.1	251	166	32.8%	HR25_YEAST	CASEIN KINASE I HOMOLOG HRR25 (EC 2.7.1...
ZK632.3	510	175	35.2%	S26727	HYPOTHETICAL PROTEIN 186 (RPOA2 3' REGI...
ZK637.11	316	183	33.5%	TWIN_DROME	CDC25-LIKE PROTEIN PHOSPHATASE TWINE (E...
ZK637.5	342	293	27.4%	A25937	ARSA PROTEIN - ESCHERICHIA COLI R-FACTO...
ZK652.9	568	255	38.8%	YIGO_ECOLI	HYPOTHETICAL 28.1 KD PROTEIN IN UDP-RFA...
ZK757.2	292	109	30.3%	TPCL_HUMAN	PROTEIN-TYROSINE PHOSPHATASE CL100 (EC ...



NRG_DROME	DNPFIIEC	EADGQP	EPE	YSWIKN	GKKFDWQAYDNRLRQP	GRGTLVITIPKDEDR	GHYQCFASNEFG
NRG_DROME	GEPFMLKCAAPDGFP	SPT	VNWMIQ	ESIDGSIKINNSRMTLD		PEGNLWFSNVTREDASSDFYYACSATSVFR	
NRG_DROME	GKRMELFC	IYGGTP	LPQ	TVWSKD	GQRIQWSDRITQG H	YGKSLVIRQTNFDDA	GTYTCDVSNVGV
NRG_DROME	DEEVVFEC	RAAGVP	EPK	ISWIHN	GKPIEQSTPNRRTV	TDNTIRIINLVKGD	GNYGCNATNSLG
NRG_DROME	GRNVTIKC	RVNGSP	KPL	VKWLRA	SNWLTGGRYNVQ	ANGDLEIQDVTFSDA	GKYTCYAQNKFG
NRG_DROME	GQSATFRC	NEAHDDTLEIE	IDWVKD	GQSIDFEAQPFRVKT		NDNSLTIAKTMELDS	GEYTCVARTRLD
CAML_MOUSE	TDDISLKC	EARGRP	QVE	FRWTKD	GIHFKPKEELGVVVHEAP	YSGSFTIEGNNSFAQRFQGIYRCYASNKLG	
CAML_MOUSE	GESVVLPCNPPPSAA	PPR	IYW MN	SKIFDIKQDERVSMG		QNGDLYFANVLTSDNH	SDYICNAHFPGT
CAML_MOUSE	GQSLILEC	IAEGFP	TPT	IKWLHP	SDPMPTRVYQN	HNKTLQLLNVGEDD	GEYTCCLAENSLG
CAML_MOUSE	GETARLDC	QVQGRP	QPE	ITWRIN	GMSMETVNKDQKYRI	EQGSLILSNVQPTDT	MVTQCEARNQHG
CAML_MOUSE	GSTAYLLC	KAFGAP	VPS	VQWLDEE	GTTVLQDERFFPY	ANGTLSIRDQANDT	GRYFCQAANDQN
CAML_MOUSE	GARVTFTC	QASFDPQLQAS	ITWRGD	GRDLQERGDSDKYFI		EDGKLVIQSLDYSDQ	GNYSVCVASTELD
NCA2_HUMAN	GESKFFLC	QVAGDA	KDKDISWFSFN	GEKLTPNQQRISVWVWDD		SSSTLTIYNANIDDA	GIYKCVVTGEDG
NCA2_HUMAN	GEDAVIVC	DVVSSL	PPT	LIWKHK	GRDVILKKDVRFIVL	SNNYLQIRGIKKTDE	GTYRCEGRILAR
NCA2_HUMAN	GQSVTLVC	DAEGFP	EPT	MSWTKD	GEQIEQEEDEKEYIFSD	DSSQLTIKKVDKND	AEYICIAENKAG
NCA2_HUMAN	EEQVTLTC	EASGDP	IPS	ITWRTS	TRNISSEKTLDGHMVVRSHA	RVSSLTLKSIQYTD	GEYICTASNTIG
NCA2_HUMAN	GNQVNITC	EVFAYP	SAT	ISWFRD	GQLLPSNYSNIKIYNT	SASYLEVTDPSEND	GNYNCTAVNRIG
ZC262.3	EKPTAISC	FSYGVP	SPK	ISWRRFRPAEKLGSYDPTDEISYTNVSETMKESYEIQSGGSLLRSPNRSHV			ERYVCCVENEY

number of hits



KEYWORDS: genome sequencing / *C. elegans*

The *C. elegans* genome project expects to finish sequencing the 100 megabase genome by 1998. Over two megabases have now been completed, covering 470 predicted genes. A third of these genes are similar to sequences in the databases. Sophisticated pattern recognition methods, using ideas borrowed from speech recognition and formal linguistic theory, are being applied to detect subtler similarities. The sequence data are available in the EMBL and GenBank sequence databases. The sequence, gene predictions, detected homologies, genetic and physical maps, and other information about *C. elegans* are organized and integrated by a freely distributed graphical database program, ACeDB.