Computational analysis of proteomes from parasitic nematodes

James D. Wasmuth

Institute of Evolutionary Biology
University of Edinburgh

Thesis presented in accordance with the requirements for the degree of Doctor of Philosophy

2006

Declaration

I declare that this thesis has been composed by myself and, except where otherwise stated, is entirely my own work.

James D. Wasmuth

February 2006

Abstract

Funding allocation for complete genome sequencing has a severe taxonomic bias; nearly half of the completed or draft stage genomes are vertebrates. The luxury of a full genome sequence is unlikely to be available for the vast majority of organisms, regardless of their importance in terms of evolution, health or ecology. This has lead to the investigation of genomes of an increasing number of species, especially parasites, through generating expressed sequence tags (ESTs).

Over 300,000 ESTs are available for 36 species of parasitic nematodes. These species include whipworm and filarial worms, which currently infect 3 billion people, as well as a large number of plant parasites. The need for a database and analysis suite, which integrates these transcriptomes with associated annotation and metadata, led to the creation of NEMBASE (http://www.nematodes.org). The system enables comparative transcriptomic analyses across the phylum Nematoda.

The focus of this thesis is the comparison of nematode proteomes. I describe my work to identify sequences and sequence features that have patterns of interest to nematode biology. Such patterns, include proteins that are unique to parasitic feeding strategy, and protein domains that have been lost in certain nematode lineages. This involved not only global comparisons of the proteomes, but also investigating the proteins domain complement from each species.

One vital step in the analysis is identifying credible coding regions within the error-prone EST sequences. Robust identification of the coding regions presents an opportunity to perform comparative analysis previously confined to those working with complete genomes.

To achieve this, I built the translation pipeline prot4EST, a hierarchical collection of freely-available algorithms. The benchmarking showed that prot4EST produced coding region predictions that were better than its constituent algorithms. Exploring the effect of sequence composition of both the studied species and the program's training sets, improved the accuracy of prediction. A database of high quality protein translations for all nematodes studied was generated, called NemPep. This was accompanied by a collection of predicted domains (NemDom). The decoration of protein sequences with domain annotation is not trivial, especially given the incomplete nature of ESTs. It was necessary to explore domain model assignment to ensure the most accurate results. The rigorous analysis of NemPep and NemDom has revealed:

1) proteins specific to certain nematode-lineages,
2) the level and potential effects of contamination in the original cDNA libraries,
3) the extent of protein loss and domain modification in the caenorhabditid lineage.

These findings are of particular importance to parasitic nematodes, as they highlight possible candidates for new anthelmintic drug targets. The methods used also offer new approaches for other complex cross-species comparisons.

Acknowledgements

I must thank Mum, Dad and Nick for all their love through the trials and celebrations of the last 28 years. It's all worth it. Dad you were right, I should 'stop being a tart and buck [my] bloody ideas up!' I'm also glad I heeded Mum telling me to 'do whatever makes [me] happy.' Nick, thank you for showing me a whole world that I want to stay well away from. The support from you three has been amazing and I love you all.

The rest of my family deserves my love and thanks. Thanks for supporting me and repeatedly asking me what it was I did. By letting me describe my work to you, I reminded myself how much I enjoyed it.

Big man-hugs go to Chris 'the Boy with the dancing feet' and Gaz. Thanks for always being there (well in Bristol at least) when I needed you. And for keeping my feet firmly on the floor. Thanks for the long, long nights (and occasionally days) dancing to the best of music. Some of those will live very long in my memory. My love goes to Whitney, Dale, Mel, George and Chris for being superb friends for what seems like forever. I've had some amazing times with you lot, and looking forward to more.

None of this work, or what follows (what follows?), would have been possible without the teachers who have graced my presence for 23 years. You are too numerous to mention individually, but special tributes should go to Martin Watson, Terry Brooks, John Evans, James MacPherson, David Perkins, Nick Edwards, Brian 'Dr. K' Kempley, Peter Little, Anne Dell, Gary Warren, Peter Young and Emma Rand. Two people deserve huge thanks yous - Sidney Reid-Smith, you more than anyone is responsible for starting my trip into science and particularly Biology. Mark Blaxter has been my mentor and friend for the past three years, and without his guidance, support and “permission” to go to (overseas) conferences my Ph.D. would have suffered. That said if he didn't feature creep so much, I'd have finished sooner :o)

To the people who I work with, thank you for putting up with me, I hope I've given a fraction back to repay the massive amount I've taken. To Ralf, Ann, John (see you soon), Al, and Martin (no more drafting - I promise) I'm pleased I shared an office with you. Thank you for furthering my programming knowledge by posing interesting problems and then in turn answering my questions. I must thank Claire, Jen, Jenna, Robin, Fran, Habib, Chris, Asher and Katelyn for reminding me where my data came from, and that there is a world of biology away from my computer. Thank you to Karen, Duff, Anjie, Katie and Henry for not groaning too often every time I popped into their office - well if you will bake cakes...

The biggest of hugs goes to Constance, without whom the last 18 months would have not been so much fun. Thank you for looking after me, and reminding me to eat while I have been writing up. I'm glad that I braved that injured knee to go on that date with you. What love I have left is all yours...

Finally I wish to thank the BBSRC (UK) for funding me for the last three years and Tom Hutchinson and Becky Brown at Astra Zeneca, Brixham for providing me with a CASE award for three years. My work would have been impossible without all those wonderful people who make their programs freely available and their open source philosophy. A big thank you and I hope you keep coding...

This thesis is dedicated to:

Sidney Reid-Smith, for starting me on this crazy journey

and the memory of

Dr. Gary 'it's been emotional' Warren

and

Auntie Fiona

"You have made your way from worm to man,
and much in you is still worm."

Friedrich Nietzsche, *Thus Spoke Zarathustra*

Table of Contents

List of Figures

List of Tables

Abbreviations used

ANN – artificial neural network

BLAST – basic local alignment search tool [1]

cDNA – complement DNA

CDS – coding sequence

COG – cluster of orthologous genes

Dme – *Drosophila melanogaster*

E value – Expect value

EST – expressed sequence tag

HGT – horizontal gene transfer

Hsa – *Homo sapiens*

HMM – hidden Markov model

HSP – highest scoring pair segment. The alignment between query and subject sequences in a BLAST search

KOG – eukaryote cluster of orthologous genes

MCL – Markov clustering

mRNA – messenger RNA

Nem – nematodes

poly(A) – poly-adenylated

Nematode species and their codes

Species	Code	Clade	Taxonomic Order
Ancylostoma caninum	ACC	V	Strongyloidea
Ancylostoma ceylanicum	AYC	V	Strongyloidea
Ascaris lumbricoides	ALC	III	Ascaridomorpha
Ascaris suum	ASC	III	Ascaridomorpha
Brugia malayi	BMC	III	Spiruromorpha
Caenorhabditis briggsae		V	Rhabditoidea
Caenorhabditis elegans		V	Rhabditoidea
Dirofilaria immitis	DIC	III	Spiruromorpha
Globodera pallida	GLC	IV	Tylenchomorpha
Globodera rostochiensis	GRC	IV	Tylenchomorpha
Haemonchus contortus	HCC	V	Strongyloidea
Heterodera glycines	HGC	IV	Tylenchomorpha
Heterodera schachtii	HSC	IV	Tylenchomorpha
Litosomosides sigmondontis	LSC	III	Spiruromorpha
Meloidogyne arenaria	MAC	IV	Tylenchomorpha
Meloidogyne chitwoodi	MCC	IV	Tylenchomorpha
Meloidogyne hapla	MHC	IV	Tylenchomorpha
Meloidogyne incognita	MIC	IV	Tylenchomorpha
Meloidogyne javanica	MJC	IV	Tylenchomorpha
Meloidogyne paranaensis	MPC	IV	Tylenchomorpha
Necator americanus	NAC	V	Strongyloidea
Nippostrongylus brasiliensis	NBC	V	Strongyloidea
Onchocerca volvulus	OVC	III	Spiruromorpha
Ostertagia ostertagi	OOC	V	Strongyloidea
Parastrongyloides trichosuri	PTC	IV	Panagrolaimomorpha
Pratylenchus penetrans	PEC	IV	Tylenchomorpha
Pratylenchus vulnus	PVC	IV	Tylenchomorpha
Pristionchus pacificus	PPC	V	Diplogasteromorpha
Rhadopholus similis	RSC	IV	Tylenchomorpha
Strongyloides ratti	SRC	IV	Panagrolaimomorpha
Strongyloides stercoralis	SSC	IV	Panagrolaimomorpha
Teladorsagia circumcincta	TDC	V	Strongyloidea
Toxocara canis	TCC	III	Spiruromorpha
Trichinella spiralis	TSC	II	Trichinellida
Trichuris muris	TMC	II	Trichinellida
Trichuris vulpis	TVC	II	Trichinellida
Wuchereria bancrofti	WBC	III	Spiruromorpha
Xiphinema index	XIC	II	Dorylamida
Zeldia punctata	ZPC	IV	Cephlalobomorpha

Chapter One - Introduction

1.1 The need for more sequence

Complete genome sequencing is a major investment and is unlikely to be applied to the vast majority of organisms, whatever their importance in terms of evolution, health or ecology. Complete genome sequences are available for only a few eukaryote genomes, most of which are model organisms. These species were chosen for the ease with which they can be manipulated in the laboratory, and not their relevance to 'wild' biology. Hence the focus of eukaryote genome sequencing has been on a restricted subset of known diversity, with, for example, nearly half of the completed or draft stage genomes being from vertebrates [2]. While Arthropoda and Nematoda have, respectively, three and two completed genomes, with a dozen others in progress, when compared to predicted diversity (over a million species each) current genome sequence illuminates only small parts of even these phyla. This disparity between sequence data and motivation for biological study is significant. Allied to this bias in genome sequence is a bias in functional annotation for the derived proteomes: a vertebrate gene is more likely to have been assigned a function due to the focus on humans and closely related model species such as mouse [3].

Diversity between organisms can be explored through any number of avenues. One way, a central theme through this thesis, is to study the proteomes of each organism [4]. In multicellular organisms the set of proteins expressed in a cell differs from cell type to cell type and will change with time because gene regulation controls advances in development from the embryonic stage and further on. The proteome considers all these proteins and can be defined as the full complement of the polypeptide sequences encoded by the majority of a species' genes. There is a proportion of genes which are transcribed but not translated; these are classified under the umbrella term RNA genes and include, tRNA (transfer), miRNA (micro), rRNA (ribosomal) and ncRNA (non-coding) [5,6].

The proteomes of all organisms can be thought of occupying proteinspace. The concept of **proteinspace** conveys a different meaning to different people. For this body of work, I define proteinspace as the *composite of properties for all proteins*. Protein properties range from relatively basic features such as amino acid composition and structural (secondary and tertiary) conformation to more complex expression profiles and protein domain (delineation and architecture) and finally the intricate quaternary structure and interaction partners. Proteinspace can be thought of as a multi-dimensional graph where each axis represents one of these properties. The orientation of an organism's proteome in proteinspace can then be described by positioning each of its constituent proteins on this graph. The proteomes of two or more species can be compared with any of the properties used to describe proteinspace.

The well documented phylogenetic deficit [7], led to the term 'neglected genomes' being coined [8]. The term refers purely to the amount of molecular sequence data available for an organism, and is distinct from the more widespread title 'neglected taxa'. The need for more sequence has led many to explore the use of expressed sequence tags (ESTs) or genome survey sequences (GSS) as a tool for investigating the transcriptome and possibly the proteome of the target organisms.

1.2 Expressed sequence tags

Expressed sequence tags (EST) or genome survey sequences (GSS) have proved to be a cost-effective and rapid method of identifying a significant proportion of the genes of a target organism [9]. ESTs provide a valuable adjunct to whole genome sequencing, as they facilitate gene identification. That said, collections of ESTs from one or multiple organisms can be studied in their own right, as part of the transcribed genome.

ESTs are short DNA sequences (usually 300 to 600 nucleotides in length) generated by sequencing either or both ends of the mRNA transcribed from an expressed gene. Typically unedited, they are single-read sequences from cDNAs (Figure 1.1). An advantage ESTs have over a complete genome sequence is that libraries of cDNAs can be prepared from a variety of tissue types, defined developmental stages and environmental challenges, revealing specificity in gene expression.

The growth in the number of sequences in the primary EST repository (dbEST) mirrors that of other sequence databases. The first ESTs were from human and mouse, and dbEST entries are still dominated by these and other model organisms. However there has been a dramatic growth in the number of ESTs deposited that are from species considered as non-traditional model organisms, as many genome initiatives utilise EST and GSS strategies to gain an insight into "wild" biology (see Figure 1.2 and Table 1.1). By 2000, greater than half the sequences in dbEST were from the neglected genomes. Given this increase in partial sequences that were generated to be studied in their own right rather than as a complement to a complete genome sequence, there was a need to consider the problems common to all EST datasets.

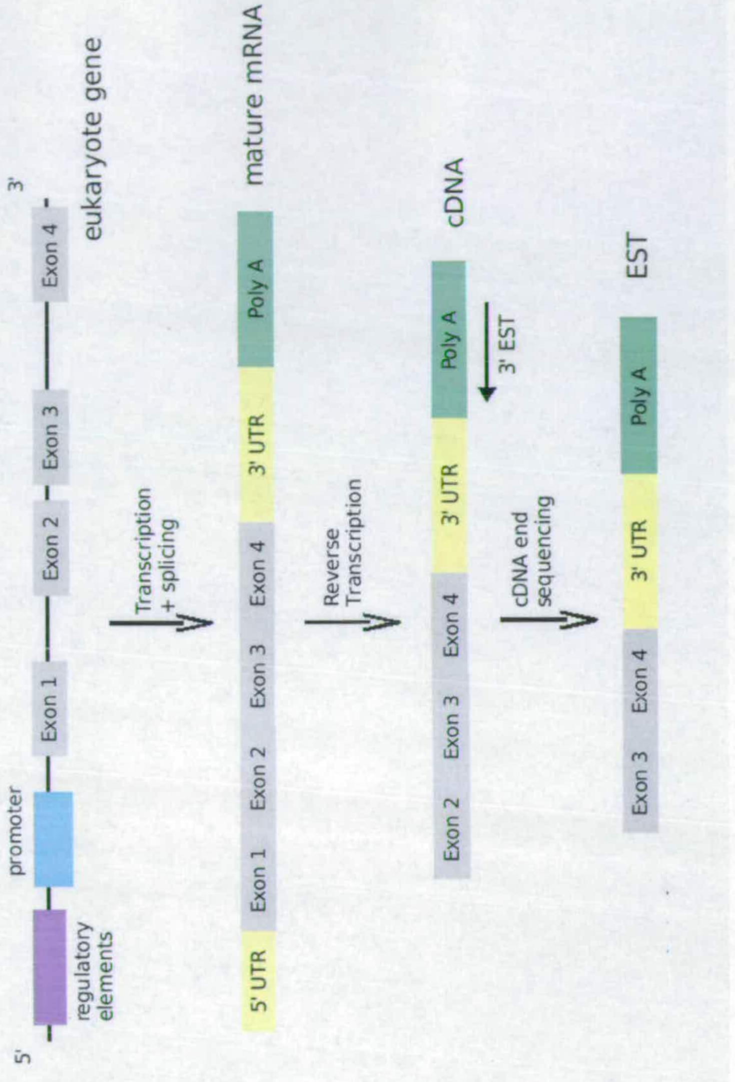


Figure 1.1

Summary of cDNA cloning and expressed sequence tag sequencing (EST).
The generation of the cDNA library through reverse-transcription of RNA rarely results in a full-length clone. The 3' poly(A) tail is often used as the selective tag for mRNA selection, thus the 3' end of the gene is more likely to be represented within the cDNA library. The cDNA can be read from either end, yielding 5' or 3' ESTs. In this figure a PCR primer for the 3' end was used.

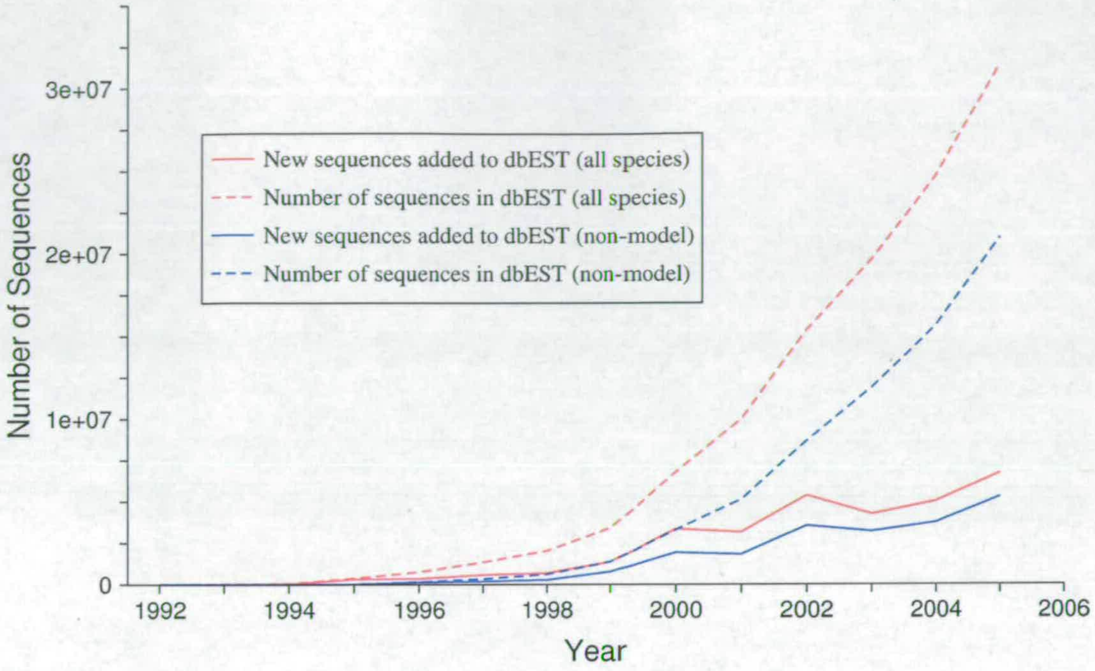


Figure 1.2

Growth of dbEST.

Expressed sequence tags are seen as an excellent way to survey the transcribed regions of the genome. This is particularly true for researchers working with neglected genomes (non-model), where a complete genome sequence is unlikely. Sequences from non-model organisms now constitute over half of all sequences in dbEST.

Phylum	Number of Projects
Streptophyta (VP)	194
Chordata (M)	66
Arthropoda (M)	50
Nematoda (M)	40
Not determined[1]	29
Ascomycota (F)	29
Apicomplexa (A)	16
Basidiomycota (F)	11
Mollusca (M)	8
Platyhelminthes (M)	7
Cnidaria (M)	6
Chlorophyta (VP)	5
Glomeromycota (F)	2
Annelida (M)	2
Echinodermata (M)	2
Bacillariophyta (VP)	2
Phaeophyceae (S)	1
Tardigrada (M)	1
Chytrodiomycota (F)	1
Zygomycota (F)	1
Microsporidia (F)	1

Table 1.1

Taxonomic distribution of EST projects.

VP – Viridiplantae; M – Metazoa; F – Fungi; A – Alveolata; S – Stramenopiles;

(1) species name in dbEST did not correspond to a known phylogenetic lineage in the NCBI taxonomy.

Sequence redundancy and clustering

There is usually significant redundancy in EST datasets, where some genes have been sequenced more than once. This is particularly true for genes that are expressed at high levels. The mapping of multiple ESTs to a gene is useful in confirming the existence of the gene product. However the sheer scale of some EST projects means that the redundancy must be reduced, thus enabling many types of analyses. For the majority of genes it is also the case that a single EST covers only part of the transcribed mRNA. Stochastic clone selection, which produces multiple tags for some genes, does not yield sequence for all the expressed genes of an organism. Some genes may not be expressed under the conditions sampled, and others may be expressed at very low levels and therefore missed through random sampling.

An effective way to overcome redundancy problems is to group the ESTs into clusters that represent putative genes. These clusters can then be annotated. There are a number of clustering methods available for EST projects, however it is beyond the scope of this thesis to describe them all. The majority of methods employ an all-against-all comparison. In StackPACK [10] the clustering is based upon shared word multiplicity; that is do the same words (string of sequence) occur the same number of times in both sequences [11,12]? Other clustering approaches make use of the BLAST algorithm [13] to perform iterative comparisons. Two popular methods which have adopted this approach are the 'Gene Indices' clustering developed by The Institute of Genome Research (TIGR) [14], and CLOBB (Clustering On the Basis of BLAST similarity) which allows incremental updates of the clusters as more ESTs become available [15]. The initial EST is considered as the first cluster in the database. The next sequence (query) available is compared to the database with a BLASTN search. If there is similarity between the query sequence and a (subject) EST in the database that satisfies the stringency thresholds, the query sequence is assigned to the cluster habouring the subject EST. In other cases, the query sequence is assigned to a new cluster. As the process progresses, query sequences will match more than one subject EST, and these may be from separate clusters. A series of checks re-analyse the BLAST output to determine if the clusters should be merged. The process continues until all sequences are assigned to a cluster (Figure 1.3). CLOBB has an advantage over other clustering programs in that incremental updates, commonplace for EST projects, can be readily performed. Annotation can be assigned to each cluster with a confidence not applicable to unclustered ESTs, a step usually aided by the determination of a consensus sequence for the cluster. Once clusters are assembled, the redundancy of the dataset can then be exploited by enumerating the number of ESTs in a given cluster. This provides a relative measure of gene expression and may uncover striking patterns [16].

Sequence quality

Sequence quality describes the faithfulness with which an EST sequence represents the gene sequence from which it was cloned. Therefore a low quality EST is a poor representation of the originating gene and essentially useless for analysis. As the sequences are single-reads of the cDNA, it is anticipated that the quality of the resulting nucleotide string is less accurate than for most genome sequencing projects where multiple coverage of each position is obtained. The inherent low quality of EST sequences may result in shifts in reading frames (missing or inserted bases) or ambiguous bases (called from the chromatographic trace). The Phred program is one of the more popular base-calling systems which determines a quality score of each nucleotide, based on the strength of signal [17,18]. Assessment of the quality of an EST must consider not only the fidelity of the reverse-transcription and sequencing reaction but also a background of incorrect sequence. Worryingly an incorrect sequence has many potential sources, from stretches of vector or polylinker sequence to contaminants from foreign genomes, commonly *Escherichia coli* and the parasites' hosts. Poly(A) tract and

5' and 3' untranslated regions (UTR) also introduce significant amounts of non-protein coding sequence, estimated at over 100 nucleotides each [19]. Clustering ESTs provides some benefit in improving the effective quality of each consensus base. The consensus for overlapping positions can utilise Phred quality scores, if available, or simply use majority rule. Despite these problems, an EST survey is a relatively cheap and accessible approach to investigate the transcribed part of a genome.

1.2.1 Processing and databasing ESTs

There are currently 484 species for which more than 1,000 ESTs have been submitted to dbEST. It has become common for each EST collection, or several related datasets, to be curated through a database system, which is frequently available online (a small selection is available in Table 1.2). As well as the highly redundant practice of inhouse development of bioinformatic resources in each laboratory, there are several stand-alone analysis suites available for download and local installation. Three widely used programs are, AutoFact [20], StackPACK [10], PartiGene [21]. These address the quality, redundancy and partial nature of EST sequences, and some provide a platform for subsequent functional annotation. The EST sequences used in this thesis motivated the genesis of the PartiGene project, an integrated analysis suite that uses freely available public domain software [21]. Users are able to (1) process raw trace chromatograms into sequence objects suitable for submission to dbEST; (2) place these sequences within a genomic context; (3) perform customisable first pass annotation of the data; (4) present the data as an SQL database; and (5) create an online portal for the database allowing external users to query the database. The system has recently expanded to include an array of annotation tools, collectively named Annot8r (Schmid unpublished). The elaboration of the PartiGene system has been motivated by the large number of EST projects carried out either by or in conjunction with the Blaxter group at the University of Edinburgh. The largest project, in terms of EST numbers, resource devotion and collaboration, has been the generation and analysis of ESTs from nematodes. It is this dataset that has motivated much of the work I describe in this thesis.

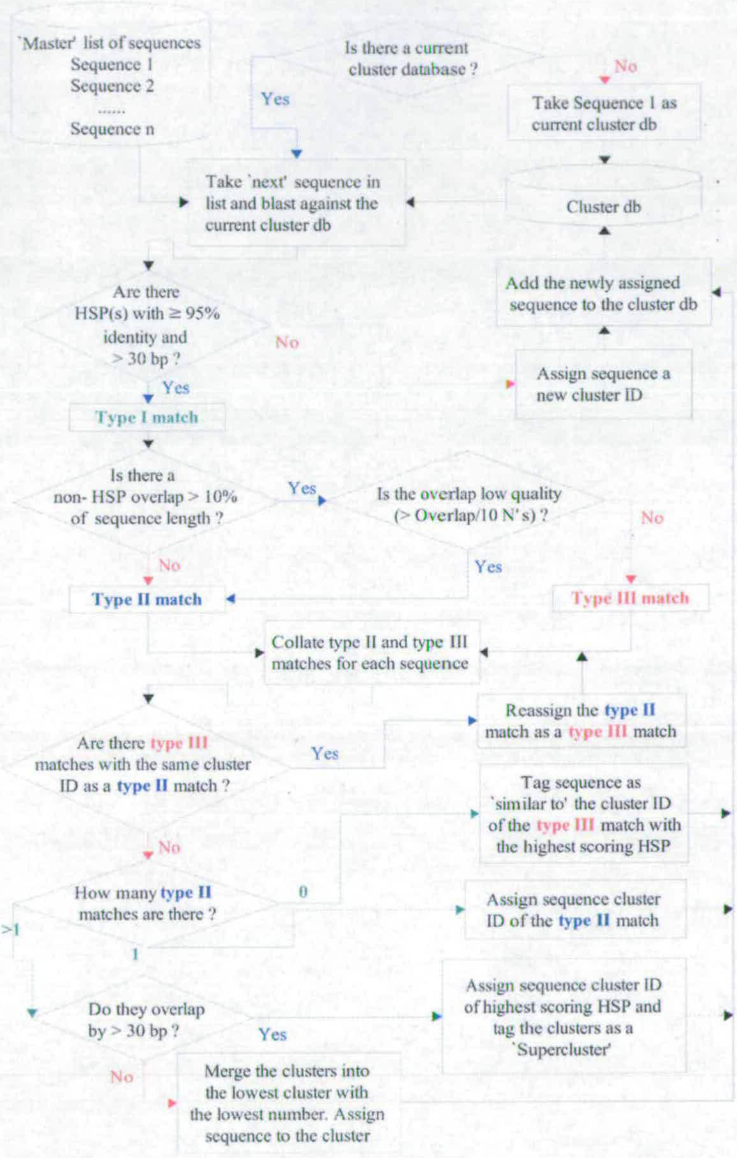

`Master' list of sequences
Sequence 1
Sequence 2
......
Sequence n
Is there a current cluster database ?
Yes
No
Take Sequence 1 as current cluster db
Take `next' sequence in list and blast against the current cluster db
Cluster db
Add the newly assigned sequence to the cluster db
Are there HSP(s) with ≥ 95% identity and > 30 bp ?
No
Assign sequence a new cluster ID
Yes
Type I match
Is there a non- HSP overlap > 10% of sequence length ?
Yes
Is the overlap low quality (> Overlap/10 N's) ?
No
No
Yes
Type II match
Type III match
Collate type II and type III matches for each sequence
Are there type III matches with the same cluster ID as a type II match ?
Yes
Reassign the type II match as a type III match
Tag sequence as `similar to' the cluster ID of the type III match with the highest scoring HSP
No
How many type II matches are there ?
0
>1
1
Assign sequence cluster ID of the type II match
Do they overlap by > 30 bp ?
Yes
Assign sequence cluster ID of highest scoring HSP and tag the clusters as a `Supercluster'
No
Merge the clusters into the lowest cluster with the lowest number. Assign sequence to the cluster

Figure 1.3

The CLOBB clustering algorithm.

Taken from Parkinson et al. 2004c. Used with permission.

Species	Database Name	URL	Reference
Fundulus Heteroclitus (fish)	FunnyBase	http://genomics.rsmas.miami.edu/funnybase/super_craw4/	22
Gallus gallus (chicken)	ChickEST	http://www.chick.umist.ac.uk/	23 24
Many Plants	OpenSputnik	http://sputnik.btk.fi/	25
Nematoda (37 species)	NemBase	http://www.nematodes.org	26
Nematoda (37 species)	NemaGene	http://www.nematode.net	27

Table 1.2

Examples of online resources for EST projects.

1.3 The phylum Nematoda

Nematodes (or round worms) are abundant and diverse in terms of biology and ecology [28]. As individuals, nematodes account for an estimated four out of every five animals on earth [29]. There are approximately 25,000 described nematode species [30,31], and while the estimates for the true species count range from 100,000 to one million [32], it is likely that the upper estimate is too high [33]. Nematodes are ubiquitous members of the meiofauna, found in all but the most arid soils. They can be found in immeasurable numbers in all sorts of sediments [34] and play a core role in the recycling of nutrients. Probably most famously, it was the free-living bacteriovore nematode *Caenorhabditis elegans* that provided the first genome sequence for a multicellular organism [35]. This has since been complemented by a draft sequence for its sister species *Caenorhabditis briggsae* [36]. Understanding the genomic sequence of *C. elegans* has motivated the development of a startling array of analysis tools and the integration of experimental data, much of which is stored and curated at WormBase [37]. Most of what is known about the molecular and developmental biology of nematodes is a consequence of the study of *C. elegans*. However, despite the wealth of information for *C. elegans*, relatively little is known about other species in this phylum.

Another well known feature of the Nematoda is the large diversity of parasitic species which can be found within it. Nematodes infect humans, domestic animals and food crops [38,39,40]. The diseases caused by nematodes are extremely varied including; anaemia (hookworm – *Ancylostoma ceylanicum*), and filariasis related pathology such as African river-blindness (*Onchocerca volvulus*) and elephantiasis (*Brugia malayi*). An estimated 2.9 billion people are infected by nematodes, primarily in tropical regions of Africa, Asia and the Americas [41]. Closely related species of human parasites are also responsible for substantial loss in livestock animals. The root-knot nematodes (*Meloidogyne* spp.) are major pathogens of crop plants throughout the world, impacting both the quantity and quality of marketable yields, causing an estimated $80 billion in damage annually [42]. In addition, root-knot nematodes interact with other plant pathogens, resulting in increased losses due to secondary infections.

1.3.1 Nematode systematics

The field of nematode systematics is in a state of flux. As with many early classifications, nematode relationships were delineated with morphological characteristics. The classification of the multitude of nematode species has been difficult, as most nematodes, whether free-living or parasitic, are small and their morphological distinctness is in structures observed only with higher power microscopy. Nematodes share a similar body plan, although they vary in length from ~100 μm to more than 6 m. The organisation of the phylum has undergone some major adjustments through the work of Blaxter and colleagues [40,28]. On the basis of small subunit ribosomal RNA (SSU rRNA) phylogenetics, the nematodes can be divided into three major clades: Dorylaimia, Enoplia and Chromadorea (Figure 1.4). This framework points to parasitism of both animals and plants arising multiple times during nematode evolution [40, 43].

The reorganisation suggested that to gain an insight into the molecular physiology and evolution of parasitic traits, elaborate strategies would be required. The comparison of a small number of sequences from a few parasitic worms with the genome of a model species, *C. elegans*, is insufficient. The desire for more sequence led to the initiation of two collaborative projects to generate ESTs for nematode parasites spanning the phylogenetic disparity of the phylum [44]. By March 2005, a total of 341,008 sequences had been generated from 37 different species of nematode, the largest collection of ESTs spanning the diversity of a single phylum. Such a collection has provided the opportunity to identify sequence features that are species- and phylum-specific and present them in the context of *C. elegans* biology.

The analyses of the datasets from 39 species raised an important consideration – at what levels of the taxonomy should the nematodes be compared. Species level comparisons are typical, however there is much to be gained from taking a step 'back' in the nematode tree and identifying patterns at various nodes. Ideally combined data from one nematode family should be compared with data from another family. The separation for such taxonomic ranks is often up to the taxonomist and therefore subjective. The taxonomic rankings detailed in Figure 1.4 are not always identical, however it is convenient for this thesis to consider the penultimate ranking (e.g. Spiruromorpha and Rhabditoidea) as orders.

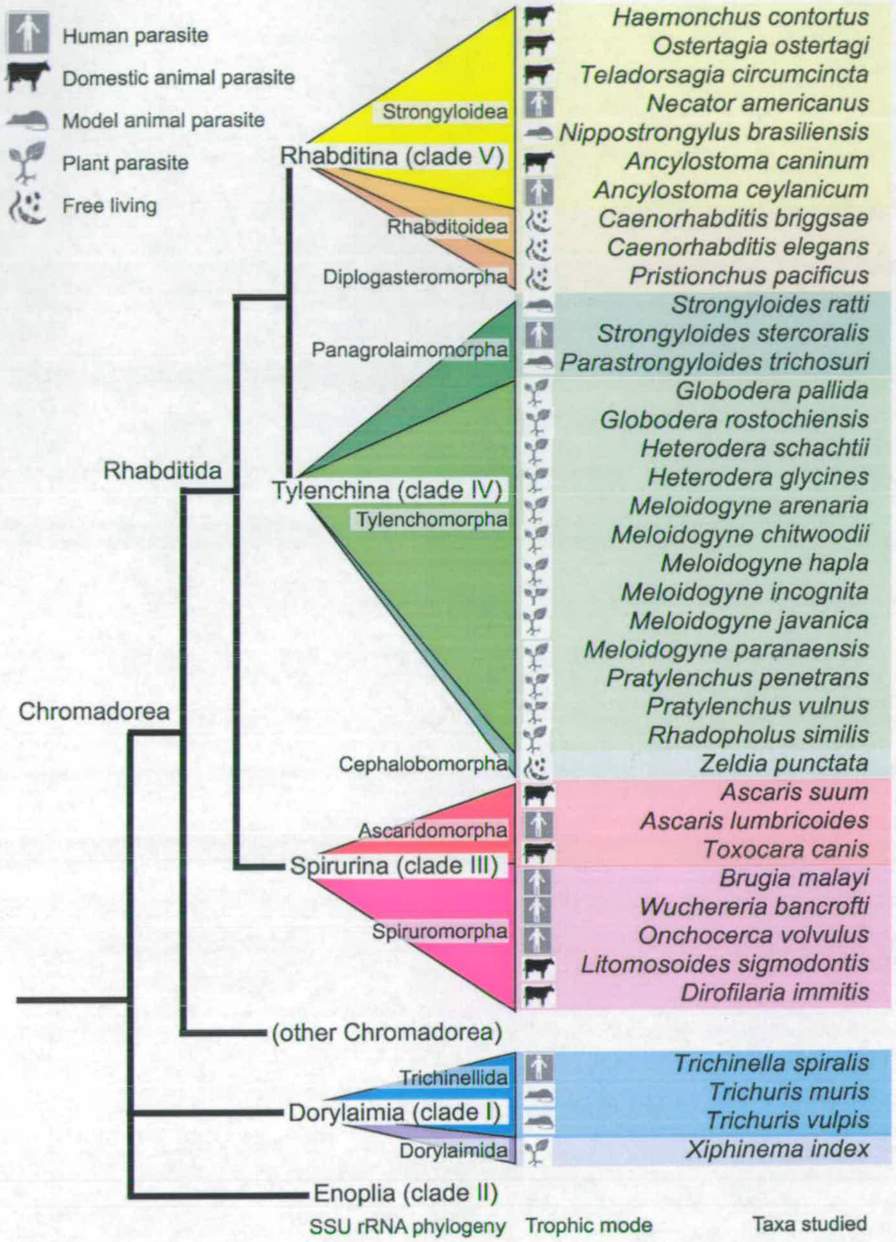


Figure 1.4

The nematode species from which expressed sequence tags were generated.

Species are grouped into major taxonomic groups based on SSU rRNA phylogeny (see 1.3.2). The trophic biology of each targeted species is indicated by a small icon.

This figure was created by Mark Blaxter and used with permission.

1.3.2 Comparative studies of the Nematoda

There are currently more than a dozen species or family specific published analyses of the nematode EST datasets. They cover parasites of humans (e.g Hookworms [45], *Brugia malayi* [46] and *Strongyloides stercoralis* [47]), animals (e.g. *Strongyloides ratti* [48]), and plants (e.g. *Meloidogyne incognita* [49] and Tylenchida [50]). The first meta-analysis which considered ESTs from across the phylum Nematoda used 265,000 sequences from 30 species [51]. The ESTs were processed using the PartiGene system [21] and are available from NEMBASE [26]. A total of 93,645 gene clusters were assembled which could be assigned into approximately 60,000 families. Cross-species BLAST comparisons revealed that 30-70% of each species dataset shared no significant similarity (BLAST bit score less than 50) with another sequence either within or outwith the sampled nematodes. A similar comparison between the genus *Caenorhabditis*, between *C. briggsae* and *C. elegans*, for which all the genes are available [36], showed that 10% of their genes were not shared. The data presented by Parkinson and colleagues [51] suggests that this level of novelty may be universal across the phylum.

These findings raise the question: what are these new genes? Of the novel sequences identified in the hookworm *Nippostrongylus brasiliensis*, 32% were predicted to contain a signal peptide [52]. There are reports of genes in other mammalian nematode parasites where a gene has a predicted signal peptide which is absent in the putative *C. elegans* homologue [53,54]. Therefore, there is a suggestion that the conversion to a secretory function for certain gene products is an adaptive strategy embraced by a number of parasitic nematodes. It should be noted that the mechanism of acquisition of the signal peptides is not clear; have they evolved from mutation in the 5' UTR of the gene or been gained from the insertion of a peptide from another gene?

Study of the expression profiles of transcripts has uncovered a number of striking findings. The two largest clusters (greatest number of ESTs) in the EST dataset for *Haemonchus contortus*, an important parasite of sheep and goats, belong to two proteins that shared 68% identity to each other in the amino acid sequence; they were named *nim-1* and *nim-2* [16]. Putative homologues were identified in the free-living *C. elegans* and *C. briggsae* as well as in the EST collections for the parasitic *Parastrongyloides trichosuri* and *Ostertagia ostertagi*. The genes' presence in the nematode orders Strongyloidea (*O. ostertagi*, *H. contortus*), Rhabditoidea (*C. briggsae*, *C. elegans*) and Panagrolaimomorpha (*P. trichosuri*) suggest that the gene emerged before the split of the Rhabditina and Tylenchina. These other *nim* genes have significantly fewer ESTs assigned to them, suggesting that they are expressed at low levels, which was supported by serial analysis of gene expression (SAGE) analysis on *C. elegans* [55]. The protein NIM-1 was localised by immunohistochemistry to the hypodermis in the anterior of the worm. Use of RNA interference (RNAi) on the *C. elegans* homologues produced no visible phenotypes. The striking expression level in *H. contortus* suggests that the *nim* genes have an important role in the life strategy of *H. contortus*.

Such analyses allow the identification of new targets for anthelmintic drugs. A high-throughput molecular technique which has recently been used to study gastrointestinal nematodes is SAGE [56,57]. This method generates data that are both qualitative and quantitative, hence complementing the EST resources being developed [58]. Computational analysis of the EST datasets is one step of a genomic filtering approach that can be supported by scalable functional studies. Such studies, including two-hybrid interactions and RNAi, are currently a focus in *C. elegans* [59]. Use of a large number of BLAST searches and RNAi has already revealed a small number of promising drug targets [59]. It is important to remain cautious when extrapolating findings from *C. elegans* onto parasitic nematodes. While the free-living bacteriovore may possibly be a model for closely-related hookworms (both in the Rhabditina [40]), similarities diminish with evolutionary distance [51,59], thus reducing the relevance of *C. elegans* biology for the Dorylaimia. Indeed great differences between nematodes species exist and include: the presence of the *Wolbachia* endosymbiont in filarial nematodes [60,61]; evidence for horizontal gene transfer between *Rhizobia* bacteria and tylenchid worms [62]; and use of the anaerobic electron transport chain by some strongylid nematodes [63].

1.4 Summary of thesis: ESTs to proteinspace

Alongside the more traditional uses of the ESTs for gene finding, expression studies and SNP identification, there is an opportunity to perform comparative analyses previously confined to those working with complete genomes. The work described in this thesis is the identification of sequences and sequence features that have patterns of interest to nematode biology. Such patterns include proteins that are unique to a parasitic feeding strategy, and protein domains that have been lost in certain nematode lineages. This involves not only global comparisons of the proteomes, but also delineating the protein domain complement of each species. Therefore it is imperative that the proteome is used as the unit for comparison. The polypeptide sequence presents a better template for almost all annotation, including domain determination, as well as construction of more accurate multiple sequence alignments and structural threading and modelling to provide secondary and tertiary structures. A proteome also allows studies into metabolic and other characterised pathways.

In the rest of this section I present a summary of this thesis; more detailed introductions are provided at the start of each chapter.

1.4.1 Identification of coding regions

Before any functional annotation can be sought, the coding region of the EST, or of a clustered consensus, must be identified. Accurate reconstruction of the coding region will permit post-genomic study in a manner similar to that for complete genomes. Prediction of the correct polypeptide from ESTs is not trivial:

1. The inherent low quality of EST sequences may result in shifts in the reading frame (missing or inserted bases) or ambiguous bases. These errors impede the correct recognition of coding regions.
2. ESTs are often partial segments of a mRNA, and as most cloning technology biases representation to the internal parts of the genes, the initiation methionine codon may be missed.

While using consensus EST contigs from a clustered dataset may improve sequence quality, this approach will not address the whole problem. Poor quality EST sequences may not yield high quality consensuses and for smaller volume projects, most genes only have a single EST representative.

There are a number of methods available for coding region identification: based upon direct amino acid similarity, probabilistic characterisation of a sequence composition and rudimentary mRNA structure. Each of these presents benefits and specific limitations. Benchmarking of the approaches led to the development of a hierarchical system combining some of the algorithms into a standalone program, prot4EST.

In **Chapter Two** I describe, in more detail, the problems that need to be overcome when translating ESTs, especially those from neglected genomes. I show the prot4EST translation pipeline to be the most accurate system for identifying coding regions in ESTs. Highlighted in the chapter are important considerations for users to ensure they produce the best translations.

1.4.2 Deriving nematode proteomes

The development of prot4EST and creation of nematode proteomes was tightly interwoven. Translations of the nematode EST contigs were used to identify and overcome problems with the program, which I address in **Chapter Three**. One issue was assembling adequate training sets for the probabilistic models used in part of prot4EST. This involved identifying coding regions based on the sequence composition of known coding regions for a particular species. An internal bootstrapping approach was applied where the most robust translations were used to generate the training set.

The first analyses performed on the proteome collection, NemPep, considered many of the raw features, accuracy of translation, length of coding regions, and the effect of cluster size and method of library construction. A small number of cDNA libraries were identified that provided a disproportionate number of ESTs that could not be robustly translated. These sequences were not used in subsequent analyses.

1.4.3 Phylogenomics of nematodes

Annotation of gene products generally relies upon nearest-neighbour strategies, and it has been estimated that all current annotations are derived from around 5% of proteins, whose function has been experimentally determined [64]. Assignment of annotation is usually performed through BLAST searches where the sequence returned as most significant donates its description line. The opportunity for a snow ball effect miss-annotating sequences is a major concern. The PartiGene developers are helping address this problem by incorporating annotation tools into the set-up that can be readily used by researchers. The additional resources available, named Annot8r, include Gene Ontology, structural predictions and metabolic pathway assignment (Schmid unpublished).

To identify nematode proteins that may be of interest in nematode biology, I performed cross-species BLAST searches, which could then be mapped onto the nematode phylogeny. I describe the findings in **Chapter Four**. The analyses are a progression of the work presented in the first meta-analysis [51], with the inclusion of seven species (including one new order), and the full incorporation of the caenorhabditid proteomes. The rate of new protein discovery has not slowed, even though some of the new species are closely related to previously studied datasets. Global comparisons between proteomes have been used to try to resolve important phylogenetic questions, such as the organisation of deep metazoan branches [65,66,67,68]. These studies involved the generation of protein families by various methods. One was the use of symmetrical BLAST hits to assemble clusters of euKaryotic Orthologous Genes (KOGs) [69]. A major concern with the original analysis was that *C. elegans* proteins may be unassigned to a KOG due to a higher rate of evolution, a trait that has been observed in *C. elegans* [70,71,72,73]. Also, extensive gene loss in the *C. elegans* lineage could therefore misrepresent the true phylogenetic relationship [74]. The availability of NemPep3 allows a larger protein complement from the Nematoda to be considered.

1.4.4 Nematode protein domains

The decoration of polypeptide sequences with protein domains is one of the most popular forms of annotation. There are a number of databases, or libraries, of protein domains which habour a wealth of information, from proposed function to species distribution, about each domain. Assigning domains to the proteins of NemPep3 is an excellent way to identify sequences that are fundamental for nematode survival. Finding domains on EST-derived proteins presents a major problem; only a small section of the domain may be present due to the incomplete nature of ESTs. Using standard domain models the domain would therefore normally go undetected. However, in **Chapter Five** I describe the creation of NemDom3, a protein domain resource to complement the proteome data being generated through the Annot8or modules. To ensure maximum and robust coverage of domain annotation I have explored the effect of using both global and local alignments between the domain model and protein sequences, as well as different scoring thresholds.

A combinatorial approach was adopted to assign Pfam-A domains to NemPep3, creating the NemDom3 collection. Species distribution of previously characterised metazoan-wide and nematode-restricted domains [75] was investigated, identifying domains that have been lost in the caenorhabditid lineage but found throughout the rest of the phylum. Domains unique to *Caenorhabditis elegans* were still identified, suggesting that they have been acquired in that lineage or that the domain models are restricted in their predictive power.

Chapter Two - Predicting coding regions from ESTs

2.1 Abstract

The genomes of an increasing number of species are being investigated through generation of expressed sequence tags (ESTs). However, ESTs are prone to sequencing errors and typically define incomplete transcripts, making downstream annotation difficult. Annotation would be greatly improved with robust polypeptide translations. Many current solutions for EST translation require a large number of full-length gene sequences for training purposes, a resource that is not available for the majority of EST projects.

To aid the investigation of these 'neglected' genomes, I have developed a polypeptide prediction pipeline, prot4EST. It incorporates freely available software to produce final translations that are more accurate than those derived from any single method. I show that this integrated approach goes a long way to overcoming the deficit in training data. prot4EST can, therefore, be usefully applied to > 95% of EST projects to improve downstream annotation.

2.2 Introduction

Nematode species are responsible for a large slice of the expressed sequence tags (ESTs) available in dbEST [9]. However, they represent only 38 out of approximately 430 species for which at least 1,000 ESTs are available. Individual groups or small collaborations motivated many of these projects to generate molecular sequence data for their organism(s) of interest. To date the primary uses of the generated ESTs have been:

1. gene finding in complete genome,
2. expression studies, either microarray or serial analysis of gene expression (SAGE),
3. identification of single nucleotide polymorphisms.

However ESTs also present an opportunity to perform comparative analyses previously confined to those working with complete genomes. Many of these studies use complete proteomes as the unit for comparison. The polypeptide sequences present a better template for almost all annotation, including domain determination with Interpro [76] and Pfam [77], as well as construction of more accurate multiple sequence alignments, the creation of protein-mass fingerprint libraries for proteomic studies and structural threading [78,79] and modelling [80] to provide secondary and tertiary structures. A partial proteome will also allow studies into metabolic and other characterised pathways. This work is of particular relevance to the parasitic nematodes, as it offers a promising identification screen for new anthelmintic drug targets.

Before any functional annotation can be sought, the coding region of the EST must be identified. Accurate reconstruction of the coding region will permit post-genomic study in a manner similar to that for complete genomes. Such a resource would compliment those analyses currently under way, especially expression studies [57,16,81].

2.3 Translating Expressed Sequence Tags

The structure of mature mRNA is consistent throughout the Eurkaryota. A typical mRNA can be divided into a 5' cap, a 5'-untranslated region (5'UTR), a protein coding region, a 3'UTR and a poly(A) tail. The coding region, almost always, begins with the codon AUG and continues in the same reading frame up to one of three possible stop codons, UAA, UAG or UGA. For reasons discussed in detail in Chapter One, ESTs present low quality copies of mRNA, often only covering part of the sequence. This makes prediction of the correct polypeptide from an EST not a trivial undertaking:

1. The inherent low quality of EST sequences may result in shifts in the reading frame (missing or inserting bases) or ambiguous bases. These errors impede the correct recognition of coding regions. The initiation site may be lost, or an erroneous stop codon introduced to the putative translation.
2. ESTs are often partial segments of mRNA; as most cloning technology biases representation to the internal parts of the genes, the initiation methionine codon may be missed. This is a problem for some of the *de novo* prediction programs that use the initiation methionine to identify the coding region (described below).

A BLAST comparison of ESTs from *C. elegans* against the species coding regions (CDS) showed that the mean number of frame-shifts in ESTs was 1.5 and the error-rate, with respect to the nucleotide sequence, was 1.1% (Wasmuth unpublished). Sequence quality can be improved by clustering the sequences based on identity. For each cluster a consensus can be determined [82]. This however, will not address the whole problem as poor quality EST sequences may not yield high quality consensuses and with smaller volume projects, most genes have a single EST representative. This is true for 67% of nematode clusters. Therefore additional methods must be applied to provide accurate polypeptide predictions. These may be split into three broad categories: similarity-based methods, *de novo* predictors and *ab initio* approaches. Each is now considered in more detail.

2.3.1 Similarity-based methods

A robust method to determine the correctly encoded polypeptide is to map a nucleotide sequence onto a known protein for which there is statistical evidence for homology. This concept is the basis for BLASTX [13], FASTX [83,84] and ProtEST [85]. BLASTX and FASTX use the six-frame translation of a nucleotide sequence to seed a search of a protein database. These programs produce alignments from regions of local similarity, called high scoring segment pair (HSP). The alignment generated for a significant hit provides an accurately translated region of the EST. BLASTX is extremely quick. However the presence of a frame-shift terminates the local alignment, shortening the predicted polypeptide. FASTX is able to identify possible frame-shifts, but its dynamic programming approach is slower than BLASTX.

ProtEST [85] uses a slightly different similarity-based approach. A protein sequence is compared to an EST database. The phrap program [82] is used to construct a consensus sequence from the cluster of ESTs detected to have similarity. The consensus is then compared to the original EST using ESTWISE (Birney unpublished, http://www.ebi.ac.uk/Wise2) giving a maximum likelihood position for possible frame-shifts. The system is accurate but not readily adaptable to the high-throughput approach necessary when dealing with very large numbers of ESTs. More crucially, an EST that does not show significant similarity to a known protein is ignored from the study and not translated. All these methods require that the nucleotide sequence shares detectable similarity with a protein in the selected database. Many genes, from both well-studied and neglected genomes, do not share detectable similarity to known proteins, making these approaches redundant. For example, the latest analysis of the *Caenorhabditis elegans* proteome shows that only ~50% of the 22,000 proteins contains any Pfam-annotation, and 43% share no detectable similarity (using BLAST) with non-nematode proteins in the UniProt database (Wasmuth unpubl.) This feature is not unique to the phylum Nematoda, and is likely to be more extreme for neglected genomes, given the phylogenetic bias of most protein databases (Table 2.1). Finally if the EST contains a known protein domain that has a high copy number in the proteome, so called promiscuous domains, any region of similarity is likely to be localised around these, often short, domains and so excluding the coding potential of the remaining transcript.

Taxonomic Group	fraction of unique clusters
Annelida (phylum)	0.66
Chelicerata (subphylum)	0.54
Heliconius (genus)	0.40
Mollusca (phylum)	0.63
Nematoda (phylum)	0.43
Tardigrada (phylum)	0.45

Table 2.1

Genetic novelty in neglected genomes.

The fraction of clusters in various EST projects for which there is no significant detectable similarity. A BLASTX search against UniProt [86] was performed in all cases except the tardigrade ESTs, where the SWISSProt database [87] was used.

Source: NEMBASE, LumbriBase, MolluscDB, ButterflyBase, ChelDB, TardiBase (all available from http://www.nematodes.org). The rank names are those provided through the NCBI taxonomy server [88].

2.3.2 *de novo* predictions

To overcome the reliance upon sequence similarity, *de novo* approaches based on the recognition of potential coding regions within poor quality sequences have been developed. The programs work by taking known full-length coding regions and characterising their properties in a probabilistic model. To maintain integrity prior data typically comes exclusively from the species under study. The three most widely used methods are DIANA-EST [89], ESTScan2 [90] and DECODER [91]. Each method has a different implementation of characterisation, and are described here with comments upon their relative strengths and perceived weaknesses. This section ends with likely problems the methods share with regard to their training requirements.

ESTScan2

Hidden Markov models (HMM) can describe a sequence of characters in a probabilistic manner [92]. In molecular sequence analyses, HMMs are useful to characterise the defining features of a group of aligned sequences [93]. Variations in the group are interpreted statistically. A sequence of interest can be compared against a number of models, with the model producing the highest likelihood score best describing the input. For a more detailed description see Durbin *et al.* [94] and Krogh [92]. This has been exploited recently in applications to find genes in genomic sequence [95, 96, 97, 98], predict domain composition in protein sequences [77,99] and align multiple sequences [100]

ESTScan2 uses a HMM to predict the most likely coding region in an EST. The probabilities are based on the species-dependent bias in codon usage and amino acid frequencies. All the potential nucleotide strings of size *n* (*n*-tuples) have a particular distribution in coding regions. Therefore, the probability that a region is coding, or its coding potential, can be calculated based on its adherence to this distribution. This approach is used by both GenScan

[95] and GenMark [97], and is formalised as a 3-periodic inhomogeneous fifth-order HMM. This means that the probability of each position in the codon is considered given the five previous positions (nucleotides). The architecture used by ESTScan2 is a modification of these previous methods, allowing write only and read only states in the model that represent the possible insertion/deletion events in the EST.

Using log-odds for hexamers (n-tuples = 6) to generate emission probabilities of the hidden Markov model, ESTScan's algorithm computes a cumulative score for the coding potential along each sequence. The algorithm considers all frames and possible frame-shifts in parallel and determines the most likely path. The first version of ESTScan [101] was able to correctly identify 78% of coding nucleotides, but detection of the boundaries of coding regions was often inaccurate [90]. This led to the attempt to model the complete structure of the mRNA, that is the 5' and 3' UTRs as well as the start and stop codons (Figure 2.1). The most optimal path through the model, the most likely coding region, is calculated with the Viterbi algorithm. These changes were incorporated into ESTScan2.

Training ESTScan

There are two sets of model parameters that need to be calculated.

First, **emission probabilities** are determined by counting the n-tuples in full-length mRNA entries. Typically these are from the EMBL nucleic acid database or, for model organisms, RefSeq [102]. This approach models the coding and untranslated regions. For the start and stop positions there is not enough data for high order dependencies; the environment around these positions is extremely bias so many possible n-tuples do not exist. For these positions therefore, position specific scoring matrices are used.

The second are **transition probabilities**, which are responsible for:

1. where the model moves from a match state to an insertion or deletion state and

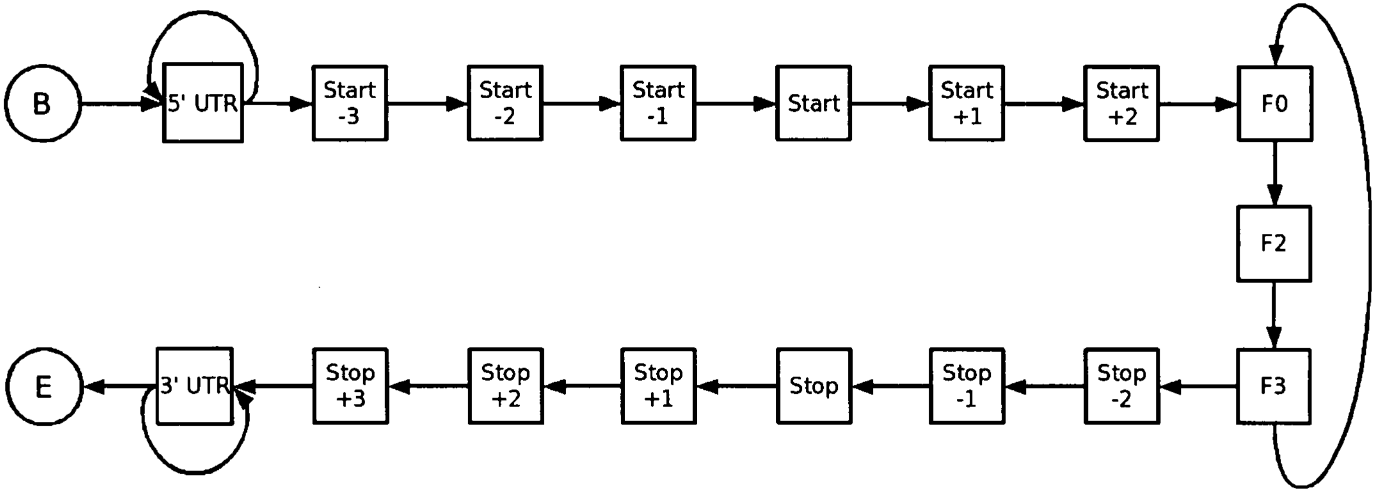


Figure 2.1

A Markov model of mRNA structure.

This is a third order model, therefore the first three nucleotides in the coding region must be explicitly modelled, hence the states 'Start', 'Start +1' and 'Start +2'. A fifth order model would include additional states: Start-5, Start-4, Start+3, Start+4, F4, F5, Stop-3, Stop-4, Stop+4, Stop+5.

This model was adapted for ESTScan2. The coding region section (F0-FN) included insertion and deletion states to overcome frame-shifts in the coding region [90].

B – begin state; **E** – end state.

For the calculation of the emission probabilities the amount of data available for training is the foremost consideration. A tuple size of 6, Markov chain order 5, has been shown a number of times as the optimal trade-off between a sensitivity and selectivity and was shown to be the order which gave the smallest percentage of false positives for ESTScan [104]. The minimum number of nucleotides necessary to train ESTScan is the number of parameters (*n*) in the model, described in Equation 2.1, where the order is of the Markov chain is the length of the tuple minus one.

$n=3*4^{(order+1)}$ **Equation 2.1**

Therefore the absolute minimum number of nucleotides needed for a 5th order Markov chain is over 15,000 (Claudio Lottaz, pers. comms and see Table 2.2). Without this level of training data the order of the Markov chain must be reduced, which will increase the number of nucleotides incorrectly classified as coding (false positives). This problem of a reduction in training data for most EST projects was considered in the construction of ESTScan2. Pseudocounts were added to introduce a priori knowledge into the learning procedure [104]. The authors recognised that while many full-length mRNA entries exist for human and mouse, most other species have much less information available. With small amounts of data, perfectly acceptable tuples that occur in nature may not be observed in the training data. The algorithm used to calculate the emission probabilities would set these instances to zero and so exclude them from the analysis. Using Bayes' rule and applying the Dirichlet distribution *a priori* knowledge derived from single nucleotide probabilities is calculated. When the effect of pseudocounts was evaluated it was shown to weakly influence the discrimination potential of the model. There were some improvements for the performance of very high order models (eight and nine), but as these models suffered from over fitting of the data, the improvement was merely to bring performance in line with other models.

Order[1]	Number of parameters[2]
1	48
2	192
3	768
4	3072
5	12288
6	49152

Table 2.2

Model accuracy against training set requirements.

The order of the Markov model (1) refers to the number of previous positions in the string (here nucleotides) are considered when determining the probability of the current nucleotide occurring. For example, consider the sequence ACTACGTAC in a 5th order Markov model, the probability of the sixth nucleotide, G (underlined), is conditional on the previous ACTAC.

The number of parameters (2) is in essence the **minimum** number of nucleotides required to hold the all possible nucleotide combinations that the model will search over. In practice this number could be significantly larger.

One final consideration for estimating emission probabilities is the redundancy in the training data. Redundancy is predominantly caused by: evolutionary mechanisms and the behaviour of researchers. The effect upon hidden Markov models is to introduce a bias towards well known sequences. The consequence is that the model will fail to 'recognise' new coding tuples. The training algorithm for ESTScan reduces redundancy by scanning the training data, remembering tuples of a certain size, typically larger than that used for estimating the emission probabilities. Reoccurring large tuples which overlap are masked from the calculation of emission probabilities. Experimental results of masking redundant stretches have so far shown to have little influence on the models discrimination potential, highlighting the HMM's robust power.

DIANA-EST

To characterise the properties of true mRNA DIANA-EST combines three artificial neural networks (ANN). An ANN is a program that detects patterns and correlations in data [89]. The underlying mechanism is that the ANN learns to recognize a sequence pattern by increasing the emphasis placed upon important information and ignoring irrelevant information. The training of an ANN uses both positive and negative examples. The set-up of ANNs allows high-order correlations in patterns. In comparison to hidden Markov models, this means that correlations are not limited to frequency of information at certain positions.

The three ANNs used by DIANA-EST and their respective training requirements are described here. The training data for each network comes from full-length cDNA.

The *Consensus-ANN* classifies the nucleotides of the translation initiation site (TIS). A twelve nucleotide window surrounding the ATG start codon is used. The training sequences provide one positive example each, with negative information provided from UTR and coding sequence.

The *Coding-ANN* was trained to recognise the coding region. The scanning window is increased to 54 nucleotides. The measure used to classify the coding regions is the codon usage of the window. The frequency of the 64 possible codons is transformed into a vector of 64 units. Positive training is provided by annotated coding regions in their correct frame. There are two sets of negative training, one extracted from non-coding sequence, a second from coding regions which are out of frame (they start with the second or third nucleotides of a codon).

The third network is the *Frame-ANN*, which identifies potential frame-shifts in the nucleotide query. The set-up is the same as the Coding-ANN, except that the negative training data is only coding regions taken out of frame. When the Frame-ANN is applied to the sequence with errors the output will be a high score every third position, representing the first base of the codon. If this periodicity is interrupted then a possible frame-shift has been identified.

By combining these three ANNs, DIANA-EST correctly identified 86.5% of known coding nucleotides in the test set. In a comparison of start site prediction with ESTScan1.0 [101], DIANA-EST correctly predicted 76 (from 107) to 37 by ESTScan.

Similar to ESTScan the performance of DIANA-EST is dependent upon the amount of data available for training. The more data available the better the correlations the ANNs can deduce regarding what makes a coding region.

DECODER

The DECODER program was developed to define start codons and open reading frames in full length cDNA sequences [91]. It uses a rule-based method to identify possible indels in the nucleotides sequence, as well as finding the most suitable initiation site. The program exploits quality scores for the sequence produced from base-calling software, such as phred [17,18] and additional text-based information from the rare sequence. In regions of low sequence quality up to two nucleotides are removed or inserted; these represent possible frame-shifts. A likelihood score is calculated for each possible coding sequence (CDS). The candidate CDS with the lowest score is chosen to represent the sequence. A penalty term limits the number of indel corrections in the CDS. The score is computed from the probability of generating a random sequence with a better Kozak consensus (the nucleotide sequence surrounding the initiation codon of eukaryote mRNA) [105], ATG position and codon usage. DECODER requires a codon usage table, which is used to determine the putative coding regions' optimal codon usage.

Training DECODER

The lone source of training required by DECODER is a codon usage table for the species under study. The codon usage of the potential coding regions proposed by DECODER is compared to that 'known' for the organism and a likeliness score calculated.

Potential problems with training

All of the three methods described in this section make use of full-length cDNA from the species under study. Properties of these sequences are, in some way, characterised to predict the coding region in the error-prone EST. The effectiveness of the training sets is a consequence of primarily two factors. One is the amount of data available, the second how representative it is of the studied species transcriptome.

To make use of ESTScan's optimal architecture of a 5th order Markov model, at least 16,000 nucleotides are required. Even at this level it is unlikely that all possible 6-tuples would present in the dataset, therefore any model would depend upon pseudocounts in the transition probabilities (see 'Training ESTScan'). Pseudocounts have been shown to have little affect on ESTScan's predictive power; relying upon them is inadvisable, meaning substantially more than the theoretical minimum amount of training data is necessary. The ANNs used by DIANA-EST also need as much data as possible. In benchmarking the Consensus-ANN the 325 sequences used to give TIS information were considered "only a relatively small amount of data." This number of sequences is considerable more than is available for the majority of EST projects and all the nematode species, excluding those for which complete genomes are available (Figure 2.2). The amount of data available for EST projects is actually reduced when redundancy is considered. If redundant stretches of sequence are not removed then the models can become over-trained. That is they are very good at recognising a small subset of possibilities but ineffective with a more natural population of sequences.

These problems are expected to influence DECODER's performance, despite its simpler model. DECODER uses a set of rules to identify all potential coding regions and the codon usage of the species to score them. Randomly selecting a reduced set of mRNA to build a codon usage is likely to give a similar codon usage. However the protein membership available for neglected genomes is rarely stochastic. The sequences have been selected for a reason and the set is often redundant therefore the codon usage is unlikely to be suitable to represent the taxa.

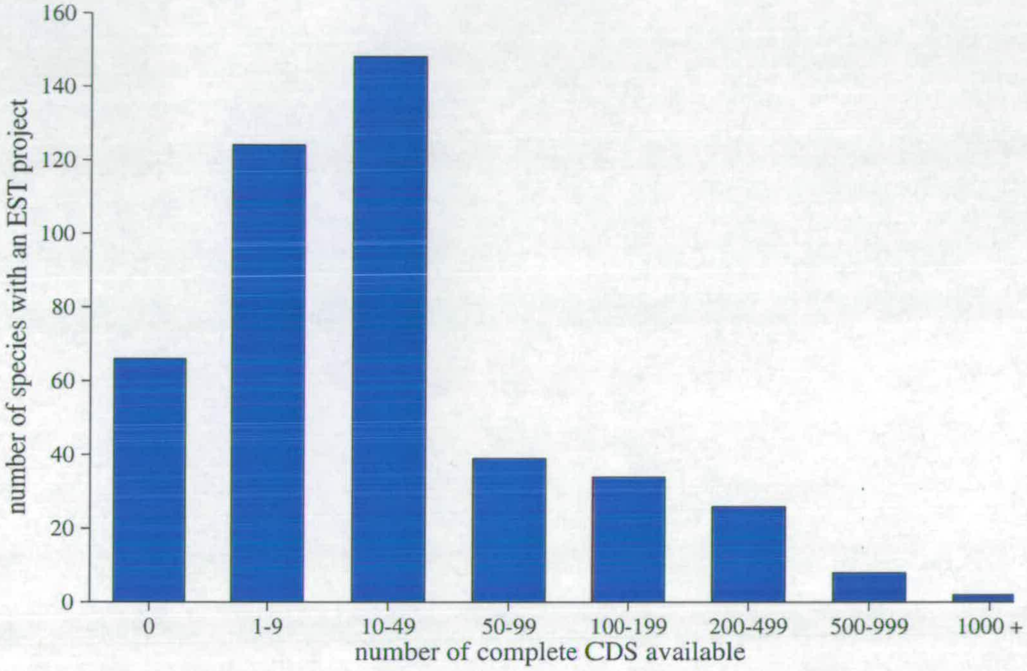


Figure 2.2

The training set deficit for EST projects.

Around 85% of species with representation in dbEST (>500 ESTs) have less than 100 complete CDS entires in the EMBL database. These species comprise ~ 45% of all ESTs. Sixty-six species, with 246,263 dbEST sequence have no full-length CDS. Source dbEST and EMBL database (July 2004).

2.3.3 *ab initio* method

One final method, commonly used, requires nothing more than some of the basic principals of biology; the coding region of the mRNA starts with a methionine and terminates with a stop codon. Applied to this problem, the nucleotide sequence is first translated in all six frames, then the longest open reading frame – the region between a methionine (or start of the sequence) and nearest downstream stop codon (or end of sequence) – is considered as the putative coding region. This approach is naïve as the assumption is made that the nucleotide sequence is error-free. The very nature of obtaining ESTs means that they are prone to frame-shifts and ambiguous or incorrect bases, as explained previously. Therefore erroneous stop codons are present and the true start methionine disguised, if it was every part of the sequence inserted into the vector originally.

2.4 New solution – prot4EST

Prior to this project, nematode ESTs available through NEMBASE had been translated using DECODER. A preliminary study suggested that it outperformed other available methods (DIANA-EST and ESTScan1) (Parkinson pers. comm.). Of the 40,000 or so nucleotide consensuses held in the database at the time, 7,388 were likely to be poorly translated (<30 amino acids in length), and I suspect many more contained considerable errors. This motivated the creation of a solution using several methods to enhance the quality of the polypeptide predictions, exploiting their strengths while recognising their shortcomings. prot4EST is an EST translation pipeline, written in Perl, with a user-friendly interface, that links some of the described methods together. It carries out retrieval and formatting of files from databases for the user. It has been designed as a stand-alone tool, and can be integrated in the PartiGene system.

prot4EST is a hierarchical system using BLAST, ESTScan, DECODER and the longest open reading frame method to predict polypeptide sequences from the error-prone nucleotide sequences. The latter three methods are implemented as described above; how their predictions are assessed is scrutinised below. However a modification to the BLAST output parsing has been adopted to improve its accuracy and is described here.

2.4.1 HSP tiling

The BLAST programs detect local regions of significant sequence similarity. High scoring segment pairs (HSP) are identified that maximise a bit score derived from an amino acid substitution matrix. If an indel is present in the query sequence, causing a frame-shift, the HSP usually terminates around this site. Downstream of this frame-shift the remaining portion of the query may result in another significant HSP to the same protein, perhaps in a different frame from the first HSP. Simple extraction of the best BLAST HSP has been frequently used by those assembling EST datasets and has recently formed the basis for published software [106,20]; however, it will lose additional robustly predicted coding nucleotides. prot4EST implements a rule-based method that considers all the HSPs between the query nucleotide sequence and the protein hit and determines whether a frame-shift can be identified. Where a frame-shift is identified the HSPs are joined. Where two HSPs overlap the sequence with the better score is used (Figure 2.3).

2.4.2 Polypeptide extension

The true polypeptide for the EST may share significant sequence similarity across only part of its length. To rely exclusively on the BLAST report is to miss potential coding nucleotides. An extension process is applied in which a coding region is continued in both 5' and 3' directions. The goal is to identify possible initiation methionine and/or a likely stop codon. A set of restrictions are imposed to ensure this less conventional approach is robust while retrieving as much potential coding sequence as possible. Recently, a similar approach has been used in the program ORFPredictor [106].

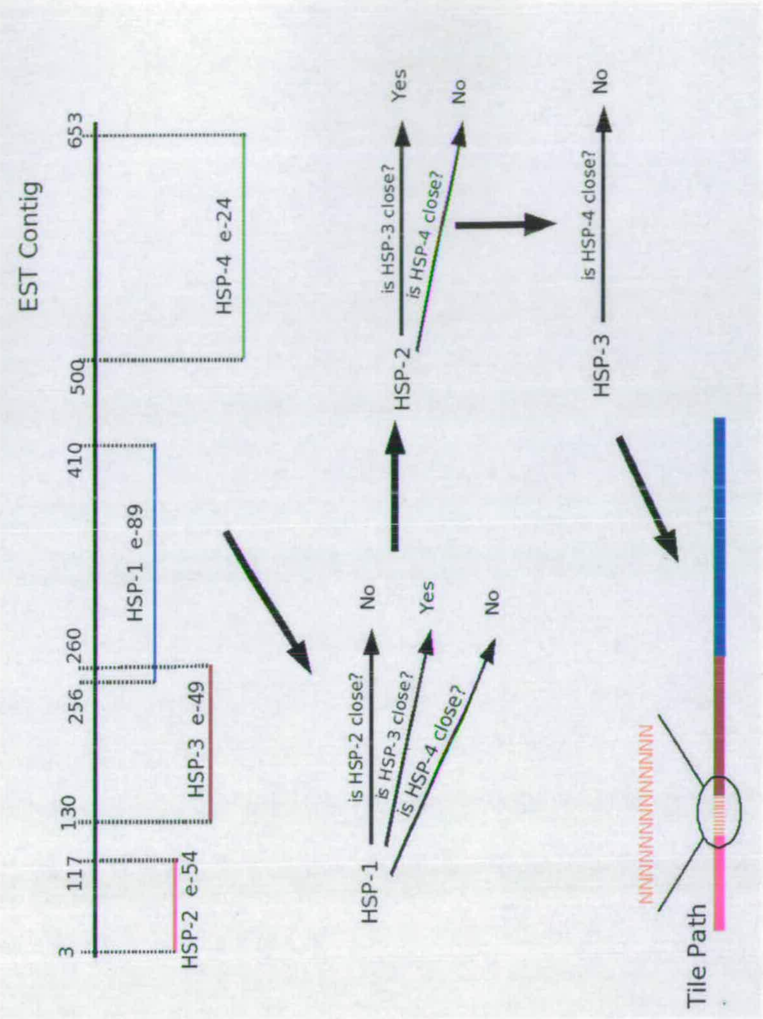


Figure 2.3

HSP tiling method.

A BLASTX search is performed with the EST contigs compared to a database of protein sequences (SWISSPROT / TrEMBL). prot4EST parses the BLAST report considering each high scoring pair segment (HSP). If two HSPs overlap or reside close to one another on the query sequence then a frame-shift event is introduced, and the HSPs joined. The user can define the distance between two HSPs that is considered as a frame-shift event. If there is an overlap between HSPs then the HSP with the higher (worse) E value is trimmed, as seen with HSP-1 and HSP-3. The gap between two, acceptably, close HSPs is filled with the ambiguous nucleotide (N), which is the circumstance for joining HSP-2 and HSP-3.

2.5 The prot4EST pipeline

The architecture of prot4EST was finalised after the analysis described later in this chapter. ESTScan2 and DECODER were selected for incorporation into the pipeline, as they are available as stand-alone programs. This allows retraining of the models for different species. Unfortunately DIANA-EST is not available for download and training the various ANN seems a considerable undertaking.

Here I shall describe the implementation of each individual step. Figure 2.4 presents one particular arrangement that was considered.

2.5.1 Step 1: Identification of ribosomal RNA (rRNA) genes

The protein databases contain (probably spurious) translations of rRNA genes and gene fragments (Blaxter, unpublished). Thus it is important to identify and remove putative rRNA-derived sequences before further processing. A BLASTN search is performed against a database of rRNA sequences obtained from the European Ribosomal RNA database [107]. A BLAST expect value cut off of e-65 is used to identify matches. The cut off was chosen according results of preliminary testing and is conservative to reduce the number of false positives. Those nucleotide sequences with significant matches are annotated as rRNA genes and take no further part in the translation process.

2.5.2 Steps 2 and 3: Similarity searches

The second and third stages are closely affiliated. A BLASTX [13] search is performed against proteins encoded by mitochondrial genomes (see 2.6.1 for details on construction of the database). Any sequences with significant BLAST hits are annotated as mitochondrial-encoded genes for the remainder of the process, permitting the use of a mitochondrial genetic code for translation. Sequences that do not have significant similarity to mitochondrial proteins are compared using BLASTX to a nuclear protein database. Sequences that continue to yield no similarity to known proteins are moved onto step 4 of the process.

For those sequences that show significant similarity to a protein sequence from either database a HSP tile path is constructed (2.4.1). prot4EST then considers whether the nascent polypeptide can be extended at either end in the same reading frame (2.4.2).

2.5.3 Steps 4 and 5: *de novo* methods

The final arrangement of ESTScan2 and DECODER is described in section 2.7.3

ESTScan2 prediction

The training sets, used to generate the emission and transition probabilities for the HMM, are constructed from full-length CDSs present in the EMBL database. The entires for the organism of study are downloaded by prot4EST. The available data is then used as the input for the 'build_models' program, part of the ESTScan distribution. The emission probabilities are estimated as described above (2.3.2). The transition probabilities are left as the default minimum. This is necessary because it is unlikely that the data will be available for these probabilities to be reliably calculated. ESTs must be aligned to mRNA and the transitions determined from here. This was primarily designed for UniGene clusters [104], which are not available for neglected taxa. The step is also considerably slow.

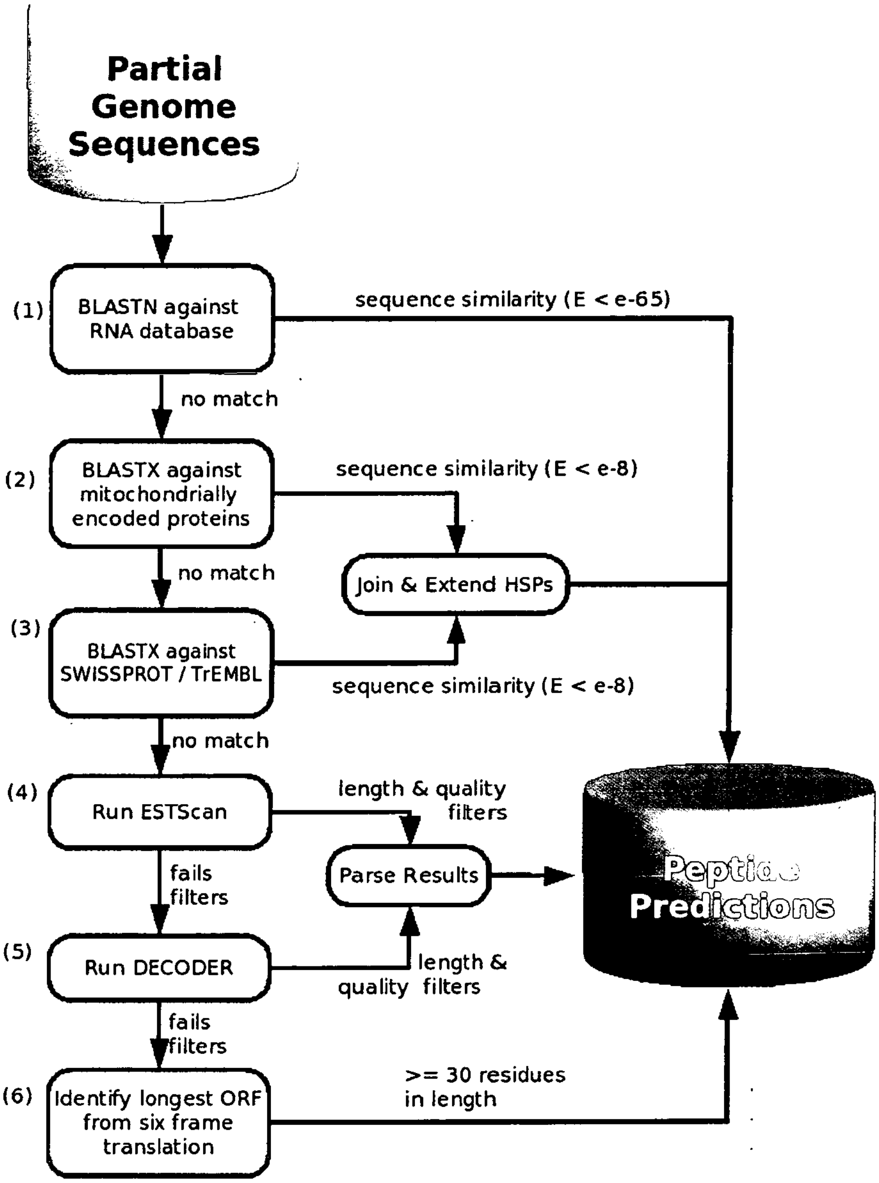

Partial Genome Sequences
(1)
BLASTN against RNA database
sequence similarity (E < e-65)
no match
(2)
BLASTX against mitochondrially encoded proteins
sequence similarity (E < e-8)
no match
Join & Extend HSPs
(3)
BLASTX against SWISSPROT / TrEMBL
sequence similarity (E < e-8)
no match
(4)
Run ESTScan
length & quality filters
Parse Results
fails filters
Peptide Predictions
(5)
Run DECODER
length & quality filters
fails filters
(6)
Identify longest ORF from six frame translation
>= 30 residues in length

Figure 2.4

Schematic of the prot4EST pipeline.

All the nucleotide sequences that have yet to be translated are passed through ESTScan. The predicted coding regions are then subject to checks to ensure their robustness. A pair of length threshold criteria is applied to each putative polypeptide before it is accepted. The polypeptide must be at least 30 amino acids in length and cover at least 10% of the input sequence. Finally the number of potential frame-shift corrections is limited to 5% of the length of the sequence. Polypeptides that satisfy these controls undergo the extension process described above; sequences that fail any of the criteria are passed onto the next step. It is possible that ESTScan predicts a possible coding region on both the positive and negative strand of the nucleotide. This is rare, but in such a circumstance both translations are accepted, if they pass checks described above.

DECODER predictions

The only readily altered parameters for DECODER are the codon usage table and sequence quality scores. Within the environment of prot4EST, the codon usage table can either be constructed by the user or downloaded from CUTG, the codon usage table database [108]. The quality files associated with each sequence are not always available. In such circumstances a uniform quality score is used. By default DECODER only processes the forward strand of each sequence. To improve the performance of this component the reverse complement of each sequence is taken and processed through DECODER. Two putative polypeptides are generated for each nucleotide sequence. The longer polypeptide is selected as the more probable translation. The resulting polypeptides are checked using the same length threshold criteria applied to ESTScan (above).

2.5.4 Step 6: Longest ORF

This last attempt to provide a putative polypeptide translation determines the longest string of amino acids uninterrupted by stop codons from a six-frame translation of the sequence. If a methionine is present in this string it is flagged as a potential initiation site.

2.5.5 Output

The primary output from prot4EST consists of the putative polypeptides in FASTA format. These are complemented with files containing information describing the translated sequences. This information includes: position of the translation with respect to the nucleotide sequence, the genetic code used for translation, position relative to query sequence and BLAST statistics of HSPs used in the tile path. All this information is stored in two CSV format files, permitting parsing and simple insertion into a database.

2.6 Benchmarking EST translation methods

I have compared six translation methods to test their relative and absolute performance. DECODER is designed to consider only the forward strand of the nucleotide sequence, as it was originally designed for full-length CDSs. When applied to ESTs it is imperative that both strands are analysed. Therefore the reverse complement of each nucleotide consensus was also searched. DECODER_default (1) considers only the prediction from the forward strand, whilst DECODER_best (2) considers both predicts accepting the longest. ESTScan (3) considers both strands of the nucleotide sequence and was run as a stand-alone process with default settings. To identify the longest open reading frame (Longest_ORF - 4), each consensus was translated in all six frames and the longest string of amino acids uninterrupted by a stop codon was considered the correct coding region.

Two arrangements of components within prot4EST were tested. prot4EST_ed (5) implements ESTScan before using DECODER on remaining untranslated sequences. Conversely, prot4EST_de (6) uses DECODER first followed by ESTScan. The DECODER module in prot4EST considers potential translations on both strands of the query sequence.

2.6.1 Data Sets

All the datasets and BLAST databases described here are available on the project web-site: www.nematodes.org/thesis/james/supp .

Test EST dataset for translation

I randomly selected 4,000 *Caenorhabditis elegans* ESTs from dbEST. To reduce redundancy, the ESTs were clustered using CLOBB. phrap [82] was used to derive consensus sequences for each cluster. This resulted in 2,899 consensuses. To ensure that the consensuses corresponded to a known coding region, I carried out a BLASTN search for each consensus against the complete *C. elegans* cDNA dataset available from WormBase (version 117). A match was considered significant if the HSP covered ~75% of the consensus and there was 90% percentage identity between the two regions of the HSP. Significant matches were found for 2,372 consensuses. Finally, this set was compared (BLASTX E value cut off e-8) to the *C. elegans* protein dataset (WormPep version 117), thus associating each nucleotide sequence with a corresponding reference polypeptide. A final test set of 2,316 consensus sequences was produced.

Training datasets

Caenorhabditis elegans:

Both ESTScan and DECODER require previously annotated full-length coding sequence. Properties of these sequences are used to build the models implemented by these programs. The *C. elegans* RefSeq collection contained 21,033 entries (December 2003). A Perl script built random training sets ranging in size from 5,000 to 350,000 coding nucleotides. Four sets were assembled for each level to allow replication of performance. The build_model script (part of the ESTScan2 package) was used to build the emission probabilities used by ESTScan's HMM.

The same training sets were used to build the codon usage tables required by DECODER. The EMBOSS program CUSP was used to build the tables, and a separate Perl script written to convert the output to that required by DECODER. For each run of prot4EST the codon usage tables were built from the same partition of RefSeq.

Prokaryote genomes:

GenBank entries from 167 complete genomes were obtained (May2004). A Perl script was written to extract the CDS entries and construct a RefSeq-style resource for each prokaryote species. If a taxon's genome consisted of more than one megaplasmid the sequences were combined. CDS annotation was not available for 11 genomes. We used the CDS collections for 156 taxa to determine AT content and construct codon usage tables for ESTScan's HMMs.

Arabidopsis thaliana:

28960 complete CDS entries for *A. thaliana* were obtained from the RefSeq database.

Spirurida (Nematoda):

I queried GenBank for all complete CDS entries from species in the Nematoda order Spirurida.

BLAST databases

Nuclear proteins:

SwissProt (release 42.7) and TrEMBL (release 25.7) were combined to give a well annotated protein database. To recreate the situation facing neglected genome analysis, the accession numbers for all proteins from species in the nematode order Rhabditida were retrieved from the NEWT taxonomy database [109]. These entries (~23,000) were removed from the larger database, leaving sequences in size.

Mitochondrial proteins:

this protocol was provided by Martin Jones (Edinburgh) -

1. Download metazoan mitochondrial genome sequences from GenBank using the search terms "`txid33208[Organism:exp] AND mitochondrion`".
2. For each GenBank entry extract the list of features, then for each feature,

```
if (feature type == CDS AND feature has a 'gene' tag)      {
        then extract the gene name;
        extract the translation;
}
```

Ribosomal RNA:

The rRNA genes were supplied by the European Ribosomal RNA database [107].

2.6.2 Data collection and analysis

Comparison of predicted polypeptides to the 'true' polypeptide

I compared each putative polypeptide predicted from the *C. elegans* test dataset to its cognate reference protein using bl2seq (NCBI distribution). Default parameters were used except for the theoretical database size (-d), which was set to 1,300,000, the size of SwissProt. The blast reports were parsed using BioPerl modules. Each reference *C. elegans* reference protein was also compared to itself, with bl2seq. This gave a maximum bit score for each protein.

Calculation of comparison statistics

The score used to compare the methods was normalised for length and theoretical maximum using Equation 2.2, where: *normBits* is the normalise bit score; *BITlocal* is the bit score of the BLAST alignment between the predicted polypeptide and its WormPep reference protein;

BITmax is the bit score for the alignment between the reference protein and itself; *WPlength* is the length of the WormPep reference protein; and *ESTlength* is the length of the nucleotide consensus that has been translated.

The fraction coverage (*fracCov*) of the reference protein was calculated directory from the blast report file (Equation 2.3), where *lenOfAln* is the length of the HSP alignment between the sequences and *lenOfWP* is the length of the WormPep protein matched.

$$normBits = \left(\frac{BITlocal}{BITmax}\right)\left(\frac{3\text{WPlength}}{ESTlength}\right) \qquad \textbf{Equation 2.2}$$

$$fracCov = \frac{lenOfAln}{lenOfWP} \qquad \textbf{Equation 2.3}$$

2.7 Results and Discussion

To measure the accuracy of translation two statistics were derived from the comparison of the predicted and reference polypeptides. The **coverage** is the fraction of the peptide that aligns with the reference. This is a relative measure and gives an indication of the methods' performance in identifying frame-shifts. If a frame-shift causes the polypeptide to be prematurely terminated then the coverage is reduced. The **bit score** represents the total of the alignment's pair-wise scores, normalised with respect to the substitution matrix used to calculate these scores. In this study the bit score was itself normalised to compensate for EST length and maximum possible bit score for each comparison (see Methods, Equation, 2.2). The number of consensuses translated that had a significant match (E value < e-3) to their cognate *C. elegans* protein was also recorded for each run.

2.7.1 The influence of number of training codons

Both variants of DECODER were unable to produce robust translations of over half the nucleotide sequences no matter how many nucleotides were in the training set (Figure 2.5). As expected, the inclusion of the reverse complement in the DECODER analysis improved its performance. The more accurate translation occurred on the negative strand in almost 20% of consensuses. The inability of DECODER to translate more than 50% of the polypeptides can be traced to its core assumptions. One criterion used is the identification of the most likely initiation methionine. While this is almost always present in full-length cDNAs, the occurrence of any ATG codon is less certain. I noted that DECODER will try any ATG codon to start its prediction, even if this results in a polypeptide of two amino acids in length.

The size of the training set used to build the codon usage table had no influence upon the accuracy of the translations produced by DECODER. This is a consequence of the stochastic nature used to build the training sets. Random sampling from the complete proteome of *C. elegans* is likely to give mRNA whose average codon use is similar to that of the entire proteome. If the codon usage used was considerably different from that of *C. elegans* I would expect its performance to decline. However unless the consensus nucleotide query contained alternative features, such as Kozak sequence and initiation methionine, the codon usage is irrelevant as there is only one candidate translation. Another influence upon coding region recognition is the identification of the most likely Kozak sequence. The translation initiation site is well characterised in vertebrates [105]; however, it has been shown to differ significantly from other taxonomic groups [110]. Unfortunately DECODER implements a regular expression devised from the vertebrate distribution, a situation explained by the mammalian-based research of the RIKEN Institute. The differences may contribute towards the low number consensuses for which a coding region was predicted (Figure 2.5a).

Compounding the lack of discriminatory power is that the original sequence quality files are not available, the ones used are produced by phrap in the consensus building step. That said, this is a situation facing many people who would wish to use prot4EST, so I consider the analysis relevant.

The effect of the number of training nucleotides on ESTScan performance is pronounced. For the majority of training data sets the fraction of predictions that have significant matches to their reference protein was 75-90%. The number of translations fell significantly when the number of nucleotides in the training set dropped below 100,000. Models trained with 10,000 or less nucleotides produced at most two translations from 2,316 input sequences. This is the result of undertraining the HMM used by ESTScan. There are 4,096 possible six-tuples, so in a three-periodic inhomogeneous HMM at least 12,288 nucleotides are required. Even with this number of nucleotides not all six-tuples will be present, even those that appear within the natural distribution for *C. elegans* complete mRNA complement. The HMM will not recognise real coding six-tuples in the consensuses and therefore penalise their occurrence. This also explains the obvious differences between replicates for training sets smaller that 100,000. Random selection of mRNA gives some sets that do not accurately represent the frequency of six-tuples for *C. elegans*. As mentioned in section describing the training of ESTScan (Methods 2.3.2) pseudocounts are used to try and overcome this problem, but their effect is limited with small amounts of data [104].

There is no further improvement in ESTScan's performance once the training sets reach 150,000 nucleotides. This is equivalent to ~150 full-length coding sequences. In a genuine situation, when a small number of full-length CDS exist in the public databases it is probable that they were generated by a few surveys, designed to focus upon a particular subset of the proteome. Often these CDS will be from a few gene families and / or highly expressed genes with atypical codon usage and structure. The method used here for training set construction, random selection of CDS from a complete proteome, will provide a more natural frequency matrix of coding signatures. Therefore I suspect that the performance of ESTScan trained with datasets less than 250,000 nucleotides to be inflated in this study.

When the training sets contained a large number of non-redundant nucleotides (>150,000), prot4EST_ed and ESTScan performed equally well (Figure 2.5a). When the number of coding nucleotides available for training and codon usage determination were reduced, prot4EST translations still showed significant similarity to their reference protein in at least 80% of instances.

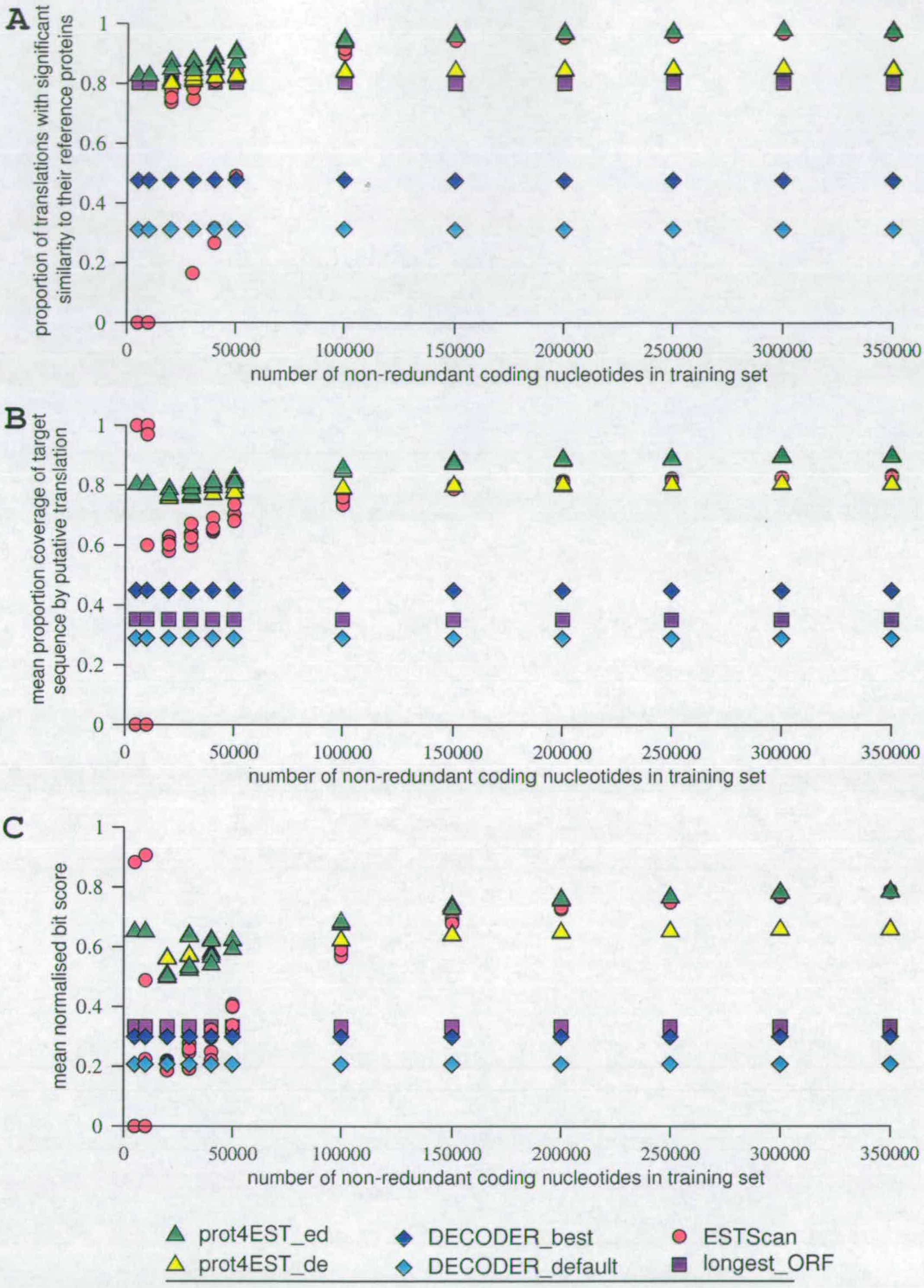


Figure 2.5

Performance of polypeptide prediction methods under different training regimes.

Legend overleaf.

Figure 2.5

Performance of polypeptide prediction methods under different training regimes.

Predicted polypeptides were compared to their reference protein from WormPep. The methods tested were ESTScan, DECODER_default (standard settings), DECODER_best (considers both frames), longest_ORF (open reading frame) and two architectures of prot4EST. See section 2.6 for more details.

Four independent replicates of each training set size were used. (**A**) Proportion of predicted polypeptides having a significant BLASTP match to their reference protein. (**B**) The mean proportion of each sequence covered by a predicted polypeptide. (**C**) The mean relative bit score of each predicted polypeptide compared to its reference protein.

The scores in **B** and **C** are the mean of the sequences translated by each method. The high scores shown by ESTScan at 5,000 and 10,000 non-redundant cosing nucleotides is due to the method returning at most one polypeptide out of the 2,316 nucleotides provided.

The translations produced by prot4EST_ed were the most robust across all training data sets, for both coverage and bit score (Figures 2.5b&c). As the size of training set decreased, both measures show slight reductions.

The *ab initio* Longest_ORF method does not require any training, instead using rules that describe mRNA structure as its guide. It provided translations for 80% (1,859) of the consensuses (E value cut off e-3). The relative coverage of these translations was 35%, and the normalised bot score was 0.33. Many of the predictions were much longer than the region that matched their cognate WormPep entry; using the Longest_ORF failed to identify frame-shifts and so identify the correct stop codon. This method also fails to correct for any incorrect base-calls, which result in the wrong nucleotide or ambiguous base 'N' being used.

2.7.2 Performance of similarity search

Seven sequences out of 2,316 were identified as rRNA in step 1. Steps 2 and 3 of prot4EST exploit any significant sequence similarity between the query sequence and known proteins for coding determination. This approach identified coding regions for just under half of the consensus, 1,131. Nineteen were identified as derived from the mitochondrial genome. Three mitochondrial genes (ATPase 6, cytochrome oxidase subunits 2 and B) genes each had a representative in the test set and two query sequences matched cytochrome oxidase subunit 1. To compare the similarity approach against the other component methods, thus test its position at the top of the hierarchy, the 1,131 consensuses were passed through DECODER, ESTScan and Longest_ORF and their accuracy measured as before. Translations predicted from the BLAST module were more accurate than those from the other methods (Figure 2.6).

2.7.3 Performance of alternative prot4EST architectures

prot4EST_ed produced more robust translations for larger training sets. However when smaller totals of training nucleotides were used the translations produced by the alternative architecture prot4EST_de were slightly better (see Table 2.3). The differences in performance between the two set-ups were examined by following the fate of individual test sequences through the prot4EST pipeline. Once the 1,131 nucleotides with significant sequence similarity are removed from the process, the remaining sequences are passed to DECODER and ESTScan. Of these 67% can be translated by DECODER to satisfy the quality controls (Figure 2.7). The accuracy of these translations does not change, regardless of the size of the sampling set used to build the codon usage table (discussed above). ESTScan's poor performance when trained with small data sets has also been discussed (see section 2.7.1). For architecture comparison, no polypeptides are predicted by ESTScan with these small training sets, so its relative placement in the hierarchy is irrelevant.

Size of dataset (coding nucleotides)	mean bit score (averaged over all four replicates)	
	prot4EST_ed	prot4EST_de
5000	0.6509	0.6510
10000	0.6507	0.6509
20000	0.5037	0.5617
30000	0.5556	0.5880
40000	0.5873	0.6021
50000	0.6003	0.5994

Table 2.3

Performance of two architectures of prot4EST with small training sets.

_ed : ESTScan was used first first then DECODER; _de : DECODER first then ESTScan.

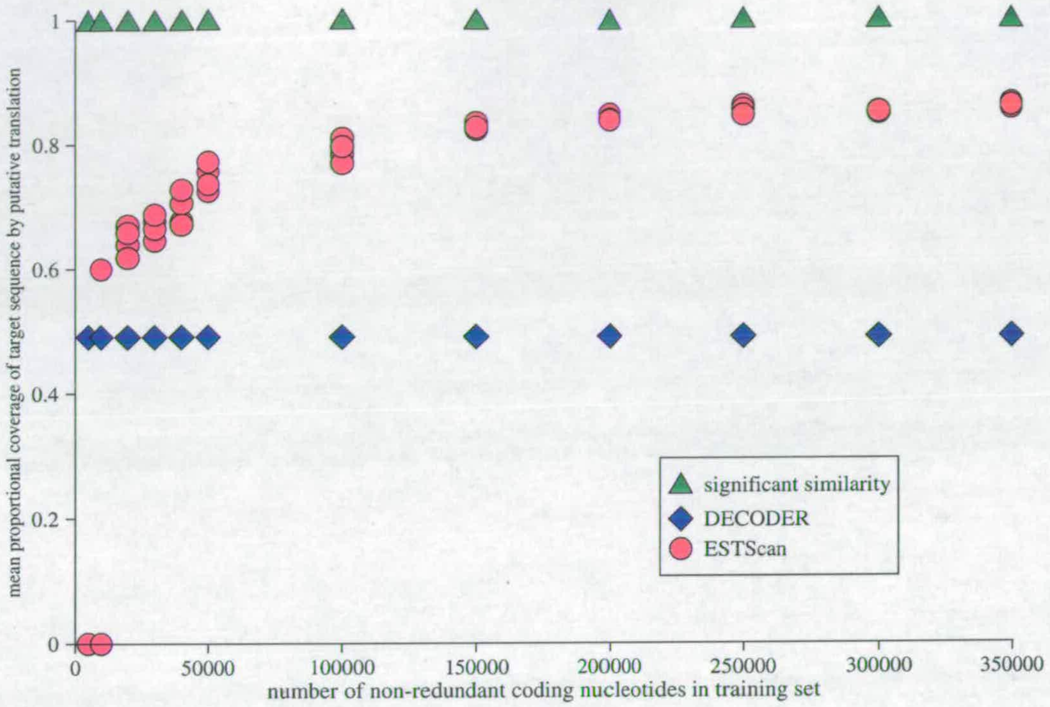


Figure 2.6

Comparison of HSP tiling, ESTScan and DECODER performance.

The accuracy of translation was compared for the 1,131 consensuses that prot4EST translated using similarity criteria.

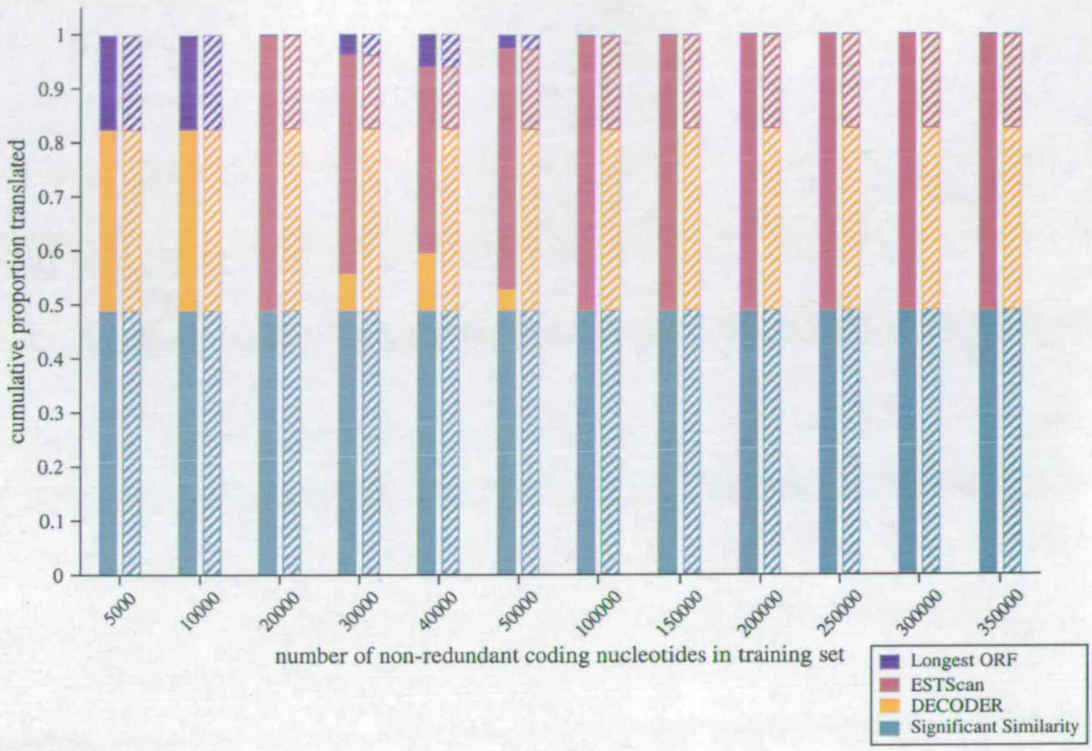


Figure 2.7

The relative efficiency of different organisations of DECODER and ESTScan in the prot4EST pipeline.

The proportion of consensus sequences produced by each part of the pipeline for each level of training is shown.

Bold bars: prot4EST_ed – ESTScan translations were considered before those from DECODER.

Hashed bars: prot4EST_de – robust DECODER translations were used in preference to those from ESTScan.

2.7.4 Effect of training set and target set sequence composition

As a significant proportion of any EST set will not share similarity with known sequences, *de novo* translation methods need to be trained to as accurate a level as possible. The question is how this should be done given the paucity of prior sequence data available for individual species. Should CDS from species considered phylogenetically related be combined or should a large set from a model organism be used? A recent study of gene finding in novel genomes has shown a significant effect of sequence composition upon gene structure prediction, with more closely related model genomes providing poor training models if the codon usage differs significantly from the genome of interest [96]. This behaviour is expected in ESTScan; the six tuple signatures are affected by codon usage, implying a consequence to the parameters used in the HMMs. In the absence of robust methods for global codon usage comparisons, I examined the effect of AT content on the accuracy of translation. The complete CDS complements of 156 prokaryotes were assembled as described in the Methods. This gave a range of AT contents from 28% (*Streptomyces coelicolor*) to 78% (*Wigglesworthia glossinindia*). The rationale behind using prokaryote genomes is the removal of any bias due to the organisms' relatedness to *C. elegans*. The lowest number of non-redundant nucleotides was 461,299, in excess of the minimum number suggest for robust training. To introduce a phylogenetic signal training datasets from more closely related organisms were built. All available CDS entries of the nematode order Spirurida, last common ancestor with *C. elegans* was 475-500 MYA [111], and the model Viridiplantae *Arabidopsis thaliana* were collected as described in the Methods (2.6.1).

This selection of training data was used to estimate emission probabilities for the HMM employed by ESTScan. These models were then used to predict the open reading frames from the 2,309 *C. elegans* consensuses used earlier (the putative rRNA genes were removed).

There was significant correlation between AT content of the training set and the coverage by the putative polypeptides of their reference *C. elegans* proteins (r = 0.49; P > 0.001) (Figure 2.8). The most robust predictions were produced by HMMs trained on datasets with an AT content similar to that of *C. elegans*. For the prokaryote training sets the number of nucleotides available for training had no significant effect upon the performance. I accept that the some prokaryote training sets with AT contents close to *C. elegans* performed poorly; homogeneity of AT content is thus not a panacea. The best performance was obtained using the *A. thaliana* training set, with significantly better coverage than achieved with the more closely related Spirurida. As the plant dataset was two orders of magnitude larger than the Spirurida, four *A. thaliana* training sets of comparable size to the Spirurida were randomly built. These smaller training sets still performed better than the Spirurida training set but are now indistinguishable from the prokaryote. Whether this reduction is performance is due to a bias in the six-tuple signatures detected in the new training sets is unclear. A decrease in the amount of untranslated region is likely to have an effect on the models which characterise the start and stop points of the coding region, therefore affecting the length of the prediction.

The size of these eukaryote training sets, ~230,000 nucleotides, is at least half that of the smallest prokaryote proteome. This level is in excess of the minimum requirement shown previously.

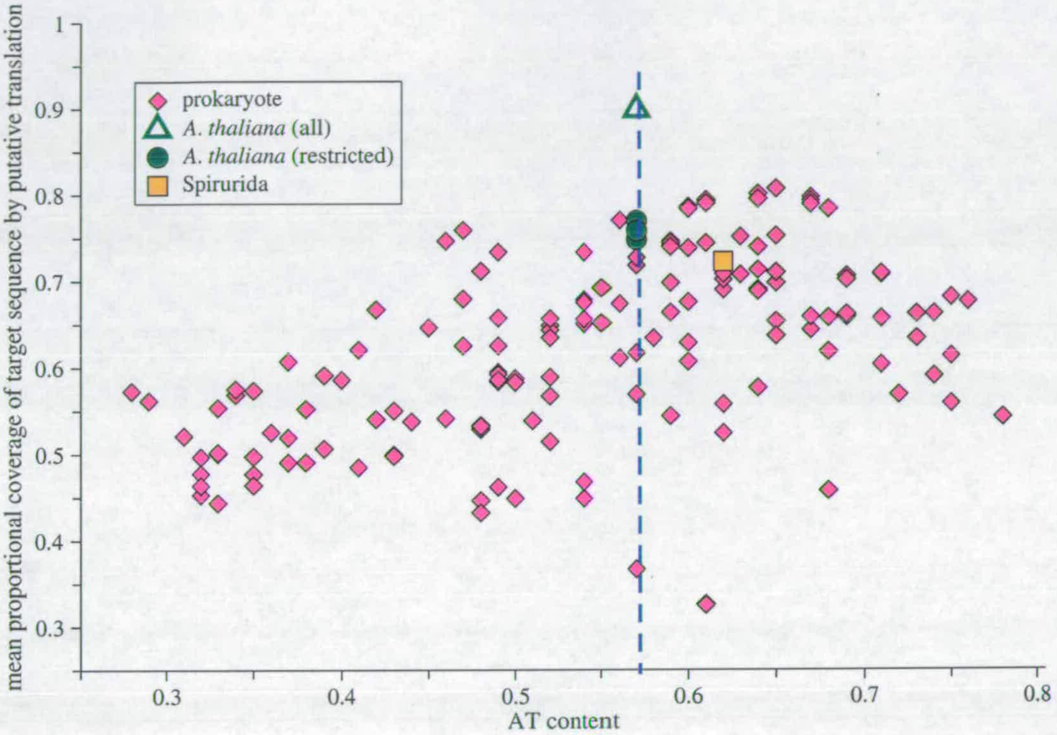


Figure 2.8

Effect of AT content of training set upon translation accuracy.

Each purple diamond represents a complete CDS set from a prokaryote genome. The orange square represents all CDS available from the nematode order Spirurida (~230,000 non-redundant coding nucleotides). The green triangle symbolises the complete *Arabidopsis thaliana* RefSeq collection (~30,000,000 non-redundant coding nucleotides). The green circles are training sets of *A. thaliana* CDS Ref Seq entires randomly selected to total ~230,000 non-redundant coding nucleotides. The AT content of *C. elegans* is shown by the vertical dashed line.

2.8 Conclusion

prot4EST is the most accurate coding region prediction tool available. It outperforms all other standalone methods by combining methods and ensuring they are used to their fullest potential, while recognising each method's shortcomings. This is particularly important when the nucleotide sequences are form a species for which there are few or no full-length cDNA information available. Such situations are common for EST projects.

2.8.1 Recent developments

Since the benchmarking of prot4EST a number of other coding region prediction methods have been published. Nadershahi and colleagues compared a number of approaches for the correct identification of the translation initiation sites [112]. It was found that ATGpr [113], a program that considers up to six discriminative features of the EST sequence, could identify 76% of true start codons while keeping the number of false positives to a minimum. Features considered by ATGpr include, hexanucleotide frequencies and a position scoring matrix for the region immediately surrounding putative start codons. The approach is similar to that used by ESTScan2 to locate the correct start of the coding region. Unfortunately ESTScan2 was not available at the time of this study, forcing ESTScan1 to be used as a comparison with the expected poor performance. With the new architecture of ESTScan2 I would expect a much improved attainment for finding the true initiation methionine.

While the study highlighted many of the problems with finding the correct start codon, most of the methods considered are of limited value for coding region prediction. With the exception of ESTScan1, none attempt to correct for frame-shift errors or try and find the end of the coding region.

The motivation behind predicting the coding regions from error-prone sequences is to provide a platform for high quality annotation. The AutoFact system endeavours to deliver functional annotation from ESTs is a hierarchical manner [20]. BLAST is used to identify significant sequence similarities with annotated protein sequences. The description lines attributed to the proteins are parsed for common, informative terms and, if in agreement, they are assigned to the EST. This allows more detailed characterisation with COG functions [69], KEGG Pathways [114] and Gene Ontologies 115]. RPS-BLAST searches [13] against the Pfam [77] and SMART [99] databases are performed against the remaining, unannotated, ESTs in an attempt to assign a protein domain. Remaining ESTs are classified as 'unassigned protein', 'unknown EST' and 'unclassified'. Such a system is effective for the study of those ESTs with meaningful annotation, however there is little benefit for the remaining sequences. As there is no prediction of the location of coding region it is not possible to perform many of the analyses which may elucidate a putative function. As a consequence the group have begun to use prot4EST for the large number of Protist ESTs (Liisa Koski, pers. comm.).

2.8.2 Future Work

The hierarchical pipeline has been shown to perform best in finding the coding regions. Such approaches are now common in the field of bioinformatics. Using a number of methods and considering all the possible results, with the possible inclusion of a ranking system, have been developed for problems such as gene finding [36] and 3D structural prediction [79].

The relative dependence upon the components is vital to the accuracy of such methods and should be explored empirically, rather than rely upon widely held opinions. Once correctly optimised one would expect improved results over an individual component, The modular form also permits upgrade of the system to be relatively simple. If prot4EST is to be taken forward then possible improvements include the incorporation of more accurate prediction algorithms. Currently there is nothing to challenge the current set-up, except for the possible incorporation of ATGpr [113] to identify the most likely start codon. Comparisons between ATGpr and ESTScan2 are needed, as is ATGpr's prediction if there is no true start codon present, not uncommon given ESTs' error-prone infamy.

2.9 Acknowledgements

I would like to thank Ann Hedley for finding many bugs in prot4EST's earlier versions, particularly with the user interface. Without Al Anthony and Martin Jones to bounce ideas off the code would not be refined (in places). A chat with Ian Korf led to the investigation of the effect of the training set's sequence composition on ESTScan performance. Finally, as always, Mark Blaxter came up with insightful suggestions and improvements.

Chapter Three - Construction and preliminary analysis of a pan-nematode protein database, NemPep

3.1 Abstract

Protein database resources are under continual development for those metazoan species with completely sequenced genomes. They present a superb opportunity to study the proteome of each organism through both bioinformatic and more traditional experimental technologies. Thus the majority of comparative genomic studies involving *C. elegans* use the collection of protein predictions held in the genome database WormPep.

In this chapter I describe how protein sequences were predicted for 37 nematode species using the translation software prot4EST. The use of simulated full-length mRNA for training set construction is explored and justified. The result is NemPep3, a collection of approximately 122,000 polypeptide sequences robustly translated from EST datasets from the phylum Nematoda, excluding *Caenorhabditis* species. I present an initial survey of compositional features for each proteome as well as a comparative review of their content. The problem of usefulness of unclustered expressed sequence tags and identification of sequences with features of experimental error is also investigated.

3.2 Introduction

3.2.1 Available protein resources

The completely sequenced genome of an organism potentially provides the amino acid sequence of all protein gene products. The field of proteomics aims to elucidate a function for each protein through large-scale systematic analysis. A database of gene products can be be used to coordinate the findings from any number of functional genomic techniques, including 2-D electrophoresis, SDS-PAGE, microarray analysis and detection of single nucleotide polymorphisms. Such database resources have been developed for many completed genomes, especially the model eukaryotes. Perhaps the most complete is that for *Caenorhabditis* species. WormBase is the major repository for *C. elegans* information: sequence, cell and gene expression, anatomy and literature [37]. Much of this information is linked allowing the user to move from a genomic region to the genes contained therein and subsequently to details of gene expression and the anatomy of RNAi phenotypes. Of particular interest to my analyses is WormPep, the nonredundant set of predicted proteins from the *C. elegans* genome. Updated at regular intervals, the current version of WormPep (140 – March 2005) contains 22,240 entries. The data contains identifiers which links each entry to WormBase and UniProt, brief annotation, and the amino acid sequence.

3.2.2 Studying proteomes

The majority of comparative metazoan genomic studies focus upon species with complete genomes [3,65,116]. One reason for this is the availability of predicted protein collections. However an increasing number of projects have produced expressed sequence tags (ESTs), focusing on species for which little or no genomic sequence is available. To date the majority of analyses have focused on using ESTs in expression studies, and there has been little direct use of polypeptide sequences derived from the ESTs. Once the coding regions of an EST dataset are identified, it is possible to perform the sorts of comparative analyses previously confined to complete genomes. The polypeptide sequences present a better template for almost all annotation, including domain determination with Interpro [76] and Pfam [77], as well as construction of more accurate multiple sequence alignments, the creation of protein-mass fingerprint libraries for proteomic studies, structural threading and modelling to provide secondary and tertiary structures. A partial proteome also allows analysis of metabolic and other characterised pathways. This work is of particular relevance to parasitic nematodes, as it offers a promising identification screen for new anthelmintic drug targets.

It was this need for robust identification of coding regions from error-prone sequence that led to the development of prot4EST (see Chapter Two and Wasmuth & Blaxter [8]). It was designed for neglected genomes, those with few previously identified full-length mRNA for model training. prot4EST derives partial proteomes from EST datasets, facilitating proteomic research both within and between species datasets.

3.2.3 Analyses of the nematode (partial) proteome

All the expressed sequence tags (ESTs) studied in the analyses described here have been passed through the Edinburgh EST-pipeline, PartiGene [21]. The sequences are first processed by trace2dbEST. Any sequence that resembles a vector, bacterial contamination or poly(A) is trimmed. These sequences are then clustered into putative gene objects using CLOBB [15]. Consensus sequence(s) are assembled from each cluster by phrap [17,18]. It is these consensus sequences that form the basis for subsequent studies [51,47,117,49,48,16]. The primary focus of recent studies has been the identification of genes restricted to specific taxonomic groups. As there are over 200 cDNA libraries for the 37 species, genes whose expression appears to be limited to a particular life-cycle stage or tissue type are also attracting a lot of attention. The interest in these clusters is motivated by the search for new anthelmintics, as well as insights to the evolution of the Nematoda.

prot4EST (version 0.9) was used to derive a collection of translated coding regions, NemPep, for the EST datasets from the 30 species of nematodes. The properties of NemPep (version one) were subject to preliminary analysis in the first pan-phylum transcriptomic study of the Nematoda [51]. As part of the analysis, putatively nematode-specific proteins were clustered into families using Tribe-MCL [118]. The families were then mapped to the phylogeny to identify when they arose. In this analysis the size of each species' datasets affected the number of gene families, though no statistical analysis was performed. Most events of origin of nematode-unique genes were mapped to early in the phylogeny, with approximately 6,500 genes unique to the three clades of the Rhabditida and almost 2,000 genes whose phylogenetic distribution suggests that they were present in the ancestral nematode. The inclusion of additional species datasets and improvements to prot4EST led to the update of NemPep (version two).

However, there are a number of reasons for caution in accepting this high estimate of genic novelty found in the phylum Nematoda.

Limitations of prot4EST

Close examination of NemPep versions one and two revealed some limitations in the implementation of several concepts in prot4EST version 0.9. These mostly stemmed from high-complexity repetitive nucleotide sequence, which were problematic for the BLAST parser and extension processes. Other problems were caused by limitations of the ESTScan code, leading to incorrectly translated coding regions. prot4EST has since been updated (version two) to manage these problem sequences, and corrects the ESTScan code.

Training sets

Version one of NemPep was created without fully optimising the training of ESTScan. Thus, the estimation of emission parameters for the hidden Markov models may not have accurately reflected the hexamer frequency of each species. Techniques have now been devised to overcome the problem and are described in this chapter.

Confidence in coding regions

Ideally, functional analysis should only be performed upon those sequences with a coding region. Spurious polypeptides that have been included in an analysis may, through weak similarity, be assigned some functional remark. Such errors can be propagated through the database. prot4EST produces a coding region prediction for every sequence presented. However, the hierarchical ranking of incorporated methods allows one to make an informed decision about the quality of each translation. I have shown, in the previous chapter, that when sufficiently optimised, BLAST-based similarity and ESTScan predictions are superior to those using DECODER or the longest open reading frame (longest_ORF). In analyses such as that performed by Parkinson and colleagues, the method of translation should be considered. This is particularly important when searching for novelty.

Singletons

From the 341,000 ESTs produced for nematodes other than *Caenorhabditis*, approximately 80,000 (24%) were not assigned to a cluster containing another EST. These singletons share no significant sequence similarity within the parameters of CLOBB [15] and represent almost two-thirds of distinct gene objects. A singleton may represent a gene with a very low level of expression or a sequence segment with no coding potential. In the original pan-Nematoda study the bulk of the genes considered unique to the nematodes were singletons. However it has yet to be determined whether these singletons are indeed real transcribed sequence or not. Without a qualitative exploration testing the biological meaning of singletons, any study performed where singletons are included in the proportion of novel genes [48,47,117,49,7] may be considered with a degree of skepticism.

Concerns regarding valid training sets and confidence in the coding region have been overcome by the development of prot4EST and investigating bootstrapping methods, respectively. This has informed subsequent releases of NemPep, with versions one and two being retired. Version three (July 2005), incorporates 119,668 clusters from the 37 species; the original phylum wide analysis considered 93,645 clusters from the 30 species. I have used NemPep3 to perform a series of analyses. These include examining the validity of a consensus given its coding region prediction with particular reference to singletons and method of translation used. These analyses are described in this chapter, with commentary on their usefulness in downstream analyses. Later chapters describe evolutionary studies performed upon NemPep3.

3.2.4 Foreseen problems with generating NemPep

The methods used by prot4EST that provide the most accurate coding region predictions are BLAST-similarity and ESTScan. The performance of both approaches is dependent upon the information used to characterise the coding regions. The BLAST component uses a database of protein sequences to compare against all possible translations of the contig. Any regions of significant sequence similarity are used to identify the coding region of the contig. It is vital, therefore, that the database only contains accurately reported protein sequences. The UniProt database is a comprehensive catalogue of protein information [86].

Preliminary analysis of the nematode EST datasets show that between 40-60% of contigs share significant sequence similarity to a protein in UniProt. For the remaining contigs there is a reliance upon *de novo* programs to identify the coding regions. The ESTScan algorithm combines hidden Markov models (HMMs) to characterise the structure of a representative mRNA molecule for a particular species. The principal feature used is the *n*-tuple nucleotide frequency for the translated region of the mRNA. The optimal size for *n* is six (a dicodon) and it is the distribution of all possible hexamers in the coding region of full-length mRNA that is used to estimate the HMM probabilities. This reliance upon sequence composition means that the hexamer frequency of the mRNA training set should correspond closely to that of the species providing the EST contigs. Full-length coding sequences (CDS) are not readily available for many species that are the subject of EST sequencing efforts. The nematode species studied in this work are no exception (Table 3.1).

In Chapter Two, I showed that the accuracy of coding region prediction by ESTScan is extremely sensitive to sequence composition (section 2.3.2). The measure of composition was the AT content of the training sequences, which dictates, and/or is a consequence of, codon usage [119]. Differences in AT content between two species will lead to variation in their observed hexamer frequency. This suggests that the use of a large number of full-length CDS from another, better-studied species, even if closely related, as a training set may generate HMMs that do not identify genuine coding regions. Reliance upon a small, yet authentic, training set leads to poorly parameterised HMMs. Both, I have shown, have drastic and deleterious effects upon the predictive power of ESTScan.

In summary, to ensure that the estimation of HMM probabilities is as robust as possible, data must be available that accurately reflects the hexamer frequency of each species. To achieve this I have used the almost unique advantage the phylum Nematoda has over other groups of neglected genomes, the presence of a well-studied model organism. The free-living rhabditine nematode, *Caenorhabditis elegans* was the first multicellular organism for which a complete genome sequence was available [35,37]. The non-redundant collection of predicted proteins for *C. elegans*, WormPep, forms the basis for the generation of large training sets for each nematode species by back-translating using specific codon usage tables. It is expected that these artificial training sets will improve the performance of ESTScan to identify coding regions in genes which do not share similarity with characterised proteins. This portion of the transcriptome is likely to contains many nematode-specific genes, so the robust identification of their encoded polypeptide sequence is an important issue.

Species	Number of CDS
Ancylostoma caninum	52
Ascaris suum	167
Brugia malayi	200
Caenorhabditis briggsae[1]	13,258
Caenorhabditis elegans	22,992
Globodera rostochienesis	40
Haemonchus contortus	184
Heterodera glycines	176
Heterodera schachtii	9
Meloidogyne hapla	13
Meloidogyne javanica	21
Necator americanus	43
Nippostrongylus brasiliensis	11
Parastrongyloides trichosuri	3
Pratylenchus penetrans	10
Pristionchus pacificus	24
Trichinella spiralis	74
Trichuris muris	1
Toxocara canis	33
Wuchereria bancrofti	31
Zeldia punctata	2

Table 3.1

Number of CDS available for nematode species.

With only a small number of CDS available for the parasitic nematodes, the need for developing prot4EST is evident. Only the caenorhabditid species would provide enough sequence data for adequate training sets for the ESTScan algorithm.

(1) *C. briggsae* actually has ~19,500 CDS, but many of these have yet to be incorporated into the EMBL database.

3.3 Methods

3.3.1 Building synthetic training sets

All the consensuses from each species' dataset were searched using BLASTX against the UniProt protein database. Potential coding regions were identified with a slightly modified version of the tile_path algorithm used in prot4EST (E value cut off: e-8). The codon usage for these regions was then calculated with a Perl script.

The proteome of *C. elegans* (WormPep140) was used as the template for the simulated transcriptomes. To reverse translate the proteome the following pseudo-code was implemented in a Perl script:

```
foreach (protein) {
    split_into_amino_acids
    foreach (amino_acid)    {
        pick codon from corresponding distribution
        return codon
    }
}
```

This gives a transcriptome based on the *C. elegans* proteome with the codon usage of the particular species. Any CDS available for a species were downloaded from EMBL using the following search terms:

Organism – as required

Molecule – `'RNA|mRNA'`

Description – `'complete CDS'`

These were added to the corresponding simulated-transcriptome. This training set was used as the input for the build_model script (available with ESTScan installation) which estimates the emission probabilities populating ESTScan's HMM. Default parameters were used. The resultant matrix file (.smat) was applied to prot4EST for the translation of that species' EST contigs.

3.3.2 prot4EST

Version 2.2 of prot4EST was used to generate polypeptide predictions of EST contigs of 37 species of nematode.

All the BLAST searches were performed separate from prot4EST.

rRNA database

The sequences were obtained from the European rRNA database [107]. The E value cut off for the BLASTN search was e-65.

Mitochondrial database

All available proteins of mitochondrial genomes from metozoan lineages were extracted from GenBank using a script written by Martin Jones. This set of sequences was reduced to so that no two sequences shared more than 70% identity. This was done to considerably speed up search time. The E value cut off for the BLASTX search was e-8.

Protein Database

The UniRef100 database (v4 Feb 2005) available through UniProt knowledgebase was used for the BLASTX searches. The E value cut off was e-8.

ESTScan and DECODER requirements

Codon usage tables for each species were generated as described in 3.3.1. The ESTScan matrix files were generated as described in 3.3.1.

3.3.3 AT content of coding regions

Only the coding regions predicted by BLAST-similarity were used for calculating the coding AT content using a Perl script. As the translations predicted by ESTScan used a matrix that was generated with the codon usage from BLAST-similarity matches, they would provide no additional information and were not used here.

3.3.4 Spliced leader library study

Information about cDNA libraries is stored in the NEMBASE database. ESTs generated from SL libraries were identified, and the proportion of SL-library derived ESTs within each cluster was calculated. Singletons were not analysed. Comparisons between the lengths of coding regions in different subsets was performed using a t-test carried out in the R-statistics package [120]. Homogeneity of variance was tested, again with R. Due to the large population sizes, the variances of some pair-wise comparisons were heterogeneous. There is a debate as to the validity of this assumption for t-tests; I chose to use p-values as a guide to biological relevance. Due to the large number of test performed, the Bonferroni adjustment was used to correct for multiple testing [121].

3.3.5 Non-coding singletons v library comparison

Identifying singletons and mapping them to specific cDNA libraries was carried out using Perl scripts that performed SQL queries against NEMBASE. For singletons, only coding regions detected by either BLAST-similarity or ESTScan components were considered.

The proportion of non-coding singletons in the cDNA libraries for a single species were compared for *Brugia malayi* and *Onchocerca volvulus*. A G-test was used to test whether the distribution of non-coding singletons was random between the cDNA libraries.

3.3.6 Distance measure for amino acid usage

A discussion regarding the use of distance measures and the chi-squared statistic is given later in this chapter and considered in Echols *et al.* 2002. In Equation 3.1 ΔFi(A,B) is the difference in the frequency of feature *i* observed between species *A* and *B*. This represents the contribution to the distance between *A* and *B* made by feature *i*.

In Equation 3.2, $D_{(A,B)}$ is a measure of the total distance between *A* and *B*, for a set of features of size *N*. For this analysis the features are amino acids, therefore *N*=20. The 39 species were compared in a pairwise manner, producing a matrix of distances.

$$\Delta F_i(A,B)=|F_i(A)-F_i(B)| \qquad \textbf{Equation 3.1}$$

$$\mathrm{D}_{(A,B)}=\sqrt{\sum_{i=1}^{N}|\Delta F_i(A,B)|^2} \qquad \textbf{Equation 3.2}$$

These formulas were implemented in a Perl script.

3.4 Results and Discussion

prot4EST (version 2.2) was used to predict the coding regions for EST contigs from 37 species of the phylum Nematoda as described above. To improve the performance of ESTScan, simulated training sets were used to estimate emission probabilities for the hidden Markov model (HMM). For each species a synthetic-transcriptome was assembled by the reverse-translation of the WormPep protein set using the distribution of codon usage for the particular species (see Methods 3.2.1). The synthetic-transcriptome was complemented by full-length CDS available for each species.

A total of 121,694 polypeptide sequences were produced. Combined with the complete proteomes of *C. elegans* and *C. briggsae* and previously characterised proteins from other nematode species, NemPep3 includes 167,126 polypeptides available for nematode species.

3.4.1 AT content

The mean AT content of coding regions within a species' datasets ranged from 0.674 (*Strongyloides ratti)* to 0.466 (*Rhadopholus similis*) (Table 3.2). The average across all species' means was 0.552. There is little divergence of AT content within ordinal level taxonomic groups. The outlying position of *Nippostrongylus brasiliensis* (μ=0.491) within the Strongyloidea (μ=0.520) may be due to the small number of clusters available. However, given its basal position in the SSU rRNA phylogeny, it could be a true reflection of divergence. The two clade III orders, the Spiruromorpha (μ=0.582) and Ascasidomorpha (μ=0.516), show significantly different AT contents. Within the Tylenchomorpha, the AT content of coding regions shows a wide variation, but is less variable in more terminal taxonomic groups. The *Meloidogyne* species have a higher AT content (μ=0.624) than members of the subfamily Heteroderinae (*Heterodera glycines, H. schachtii, Globodera* *pallida* and *G. rostochiensis;* μ=0.486). Species belonging to the family Pratylenchidae (*Pratylenchus vulnus, P. penetrans* and *Rhadopholus similis*), display surprisingly large variation in coding AT content. These differences may be the consequence of low sequence sampling for this group (only 418 – 856 contigs are available depending on the species).

3.4.2 rRNA and mitochondrial genes

A total of 482 ribosomal RNA genes were identified, representing between 0 and 1.4% of any species collection (Table 3.3). Proteins encoded by genes found on the mitochondrial genome contributed an average of 0.4% towards each dataset (Table 3.3). The species for which the proportion of clusters representing mitochondrially-encoded proteins was very high (>2%), such as *N. brasiliensis* (2.7%) and *P. vulnus* (5.0%), were generally those with a small number of clusters. This skew was most likely down to the stochastic sampling of clones chosen for sequencing, and the high representation of mitochondrial transcripts in cDNA libraries. In absolute terms there were a high number of putative mitochondrial genes in the datasets of *Brugia malayi* and *Ancylostoma caninum*, with 159 and 111 contigs respectively matching mitochondrial gene products. The nematode mitochondrial genome has between 15 and 16 genes [122,123]. Manual examination of the BLAST reports revealed only three to be spurious matches. However, the number of clusters represented by these contigs is low. The *B. malayi* contigs come from only 48 clusters, a redundancy of 3.75. Similarly, the 111 contigs from *A. caninum* represent only 43 clusters, which is slightly under 3 contigs per cluster. Across the other species' mitochondrially-encoded proteins the redundancy is typically less than two. When all clusters are considered the mean number of contigs varies from 1.01 to 1.14 depending upon the species. The high redundancy in the mitochondrial complement of certain species is probably due to the genetic heterogeneity in the starting populations of nematodes sampled. A recent study on an EST dataset from *Fundulus heteroclitus*, showed that there were 10 clusters which represented the cytochrome oxidase I gene [22], showing that this problem is one indicative of EST analyses.

3.4.3 Polypeptide length

The mean length of translations produced by prot4EST was 136.52 amino acids (s.d.=64.35), or 410 base pairs (Table 3.3). The proteins in WormPep140 have a mean length of 443.78 amino acids, or 1,300bp. The difference is certainly due to the length of the EST contig, the mean of which is 496bp. There was a substantial range of mean lengths per species from 105.3 amino acids (*Brugia malayi*) to 185.1 (*Globodera pallida*). There were no significant correlations between average length of translation and size of the dataset or AT content. Any influence by phylogenetic relatedness is difficult to test statistically, but seems absent.

Of the four methods used by prot4EST, prediction of coding regions through similarity with a known protein provided the longest polypeptide sequences for most species; notable exceptions where ESTScan predictions are longer include *G. pallida*, *Litomosoides sigmodontis* and *Pratylenchus penetrans*. Translations by DECODER and longest ORF are substantially shorter than those from BLAST-similarity and ESTScan. Across the entire collection of EST contigs, slightly more than half have significant sequence similarity to a protein in UniProt. Even ignoring the very small datasets, the use of BLAST-similarity ranges from 37.2% (*B. malayi*) to 65.3% (*Haemonchus contortus*). The overwhelming majority of remaining contigs are translated using the ESTScan component (86%).

Table 3.2
AT proportion of coding regions for 39 species of nematodes.

Species	Code	Clade	Taxonomic Order[1]	AT prop.
Strongyloides ratti	SRC	IV	Panagrolaimomorpha	0.674
Strongyloides stercoralis	SSC	IV	Panagrolaimomorpha	0.661
Meloidogyne chitwoodi	MCC	IV	Tylenchomorpha	0.637
Meloidogyne hapla	MHC	IV	Tylenchomorpha	0.629
Meloidogyne arenaria	MAC	IV	Tylenchomorpha	0.625
Meloidogyne javanica	MJC	IV	Tylenchomorpha	0.620
Meloidogyne incognita	MIC	IV	Tylenchomorpha	0.618
Meloidogyne paranaensis	MPC	IV	Tylenchomorpha	0.617
Dirofilaria immitis	DIC	III	Spiruromorpha	0.601
Brugia malayi	BMC	III	Spiruromorpha	0.591
Onchocerca volvulus	OVC	III	Spiruromorpha	0.583
Parastrongyloides trichosuri	PTC	IV	Panagrolaimomorpha	0.580
Zeldia punctata	ZPC	IV	Cephalobomorpha	0.580
Trichinella spiralis	TSC	I	Trichinellida	0.574
Wuchereria bancrofti	WBC	III	Spiruromorpha	0.572
Caenorhabditis elegans		V	Rhabditoidea	0.571
Litosomosides sigmondontis	LSC	III	Spiruromorpha	0.564
Caenorhabditis briggsae		V	Rhabditoidea	0.558
Pratylenchus penetrans	PEC	IV	Tylenchomorpha	0.540
Haemonchus contortus	HCC	V	Strongyloidea	0.534
Necator americanus	NAC	V	Strongyloidea	0.530
Xiphinema index	XIC	I	Dorylaimida	0.526
Pratylenchus vulnus	PVC	IV	Tylenchomorpha	0.524
Ascaris suum	ASC	III	Ascaridomorpha	0.523
Ancylostoma caninum	ACC	V	Strongyloidea	0.522
Teladorsagia circumcincta	TDC	V	Strongyloidea	0.520
Ostertagia ostertagi	OOC	V	Strongyloidea	0.520
Toxocara canis	TCC	III	Ascasidomorpha	0.516
Trichuris vulpis	TVC	I	Trichinellida	0.516
Ascaris lumbricoides	ALC	III	Ascasidomorpha	0.508
Heterodera schachtii	HSC	IV	Tylenchomorpha	0.505
Trichuris muris	TMC	I	Trichinellida	0.504
Ancylostoma ceylanicum	AYC	V	Strongyloidea	0.500
Heterodera glycines	HGC	IV	Tylenchomorpha	0.490

Continued overleaf...

Species	Code	Clade	Taxonomic Order[1]	AT prop.
Nippostrongylus brasiliensis	NBC	V	Strongyloidea	0.491
Globodera pallida	GPC	IV	Tylenchomorpha	0.486
Globodera rostochiensis	GRC	IV	Tylenchomorpha	0.485
Pristionchus pacificus	PPC	V	Diplogasteromorpha	0.484
Rhadopholus similis	RSC	IV	Tylenchomorpha	0.466

Table 3.2

AT proportion of coding regions for 39 species of nematodes.

The coding regions used are those identified with the BLAST component of prot4EST, expect for *C. elegans* and *C. briggsae* where WormPep140 and BrigPep2 are used, respectively. The colours reflect those used in Figure 1.4.

(1) The taxonomic ranks are those used Parkinson et al. 2004. See Introduction 1.3.1 for an explanation for this choice.

Table 3.3

Summary of NemPep3.

Species[1]	Number of Contigs	Mean length[2]	Mitochondrial[3]	BLAST-similarity		ESTScan		proportion produced by BLAST or ESTScan	DECODER		Longest_ORF	
				number	*mean len[4]*	*number*	*mean len[4]*		*number*	*mean len[4]*	*number*	*mean len[4]*
XIC	4616	170.7	59	2804	190.6	1551	154.8	0.943	126	58.0	175	70.2
TSC	3765	133.4	40	1760	140.5	1845	132.2	0.958	96	64.8	64	78.4
TVC	1265	112.8	17	499	128.9	621	113.7	0.885	44	47.3	101	56.3
TMC	1591	134.1	15	478	147.6	664	136.8	0.718	84	59.8	95	74.9
ALC	875	116.8	26	401	106.7	445	129.0	0.967	10	58.8	19	73.1
ASC	8761	135.9	89	4195	139.3	4397	135.9	0.981	52	53.9	117	48.7
TCC	1562	135.2	45	817	147.4	651	131.2	0.940	33	73.3	61	47.9
OVC	5109	123.3	84	2249	153.0	1720	135.8	0.777	238	60.7	902	42.1
LSC	1651	149.7	34	888	141.8	751	160.3	0.993	4	50.2	8	76.1
BMC	9845	105.3	159	3659	143.0	3312	109.4	0.708	806	52.3	2068	52.5
WBC	2252	123.1	36	1008	166.3	534	146.1	0.685	122	67.2	588	39.6
DIC	1796	116.2	14	761	118.7	1014	116.1	0.988	4	37.5	17	28.8
ZPC	210	154.1	4	179	160.3	27	123.8	0.981	2	52.5	2	109.0
SSC	3721	142.2	46	2419	151.5	1244	128.0	0.984	12	47.2	46	63.5
SRC	3988	136.4	34	2238	149.8	1700	121.6	0.987	10	84.7	40	29.8
PTC	3162	116.6	30	1735	136.7	1085	105.4	0.892	90	51.1	252	49.9
PVC	856	136.8	43	401	162.9	391	121.8	0.925	0	N/A	54	65.4
PEC	418	158.3	2	274	153.8	143	165.9	0.998	0	N/A	1	285.0

Continued overleaf...

Species[1]	**Number of Contigs**	**Mean length**[2]	**Mitochondrial**[3]	**BLAST-similarity**		**ESTScan**		**proportion produced by BLAST or ESTScan**	**DECODER**		**Longest_ORF**	
				number	*mean len*[4]	*number*	*mean len*[4]		*number*	*mean len*[4]	*number*	*mean len*[4]
HSC	1370	149.6	12	772	161.4	557	139.1	0.970	11	86.5	30	63.5
HGC	9256	155.3	51	4884	166.8	4291	144.2	0.991	19	61.2	62	37.8
GPC	2569	185.1	21	1350	175.8	1091	206.8	0.950	47	79.0	81	109.3
GRC	2885	154.2	30	1607	154.6	1265	154.8	0.995	2	94.0	11	39.9
RSC	597	106.5	6	276	125.7	256	98.9	0.891	25	52.8	40	56.4
MCC	3811	117.1	35	1744	136.3	1787	109.4	0.927	54	48.6	226	45.6
MJC	3496	123.3	16	1422	131.5	2019	119.6	0.984	16	59.1	39	42.1
MPC	1644	138.8	11	843	163.4	634	128.1	0.898	50	57.6	117	54.5
MHC	6984	126.2	33	3266	141.7	3293	121.1	0.939	95	51.3	330	46.7
MIC	6062	146	47	3120	153.8	2869	140.3	0.988	18	53.8	55	36.1
MAC	2659	123.2	15	1256	141.5	1165	119.1	0.910	39	43.0	199	46.8
PPC	4306	146.5	55	2580	154.8	1649	137.1	0.982	36	71.4	41	68.2
HCC	5265	155.4	68	3440	165.6	1754	139.0	0.987	27	95.0	44	45.9
AYC	3720	149.8	35	2402	158.6	1293	135.2	0.993	10	56.2	15	56.1
ACC	4368	125.8	111	2144	131.4	2210	120.9	0.997	3	43.0	11	50.3
NBC	773	137.9	21	465	135.5	308	141.6	1.000	0	N/A	0	N/A
NAC	2327	152.5	24	1157	135.6	1168	169.3	0.999	1	34.0	1	91.0
OOC	2566	127	49	1556	133.7	970	119.3	0.984	18	56.4	22	51.2
TDC	1874	147.2	31	1075	152.0	789	140.8	0.995	0	N/A	1	48.0

Table 3.3
Summary of NemPep3.
legend overleaf...

Table 3.3

Summary of NemPep3.

Each species dataset is broken down to show the components used in the prot4EST pipeline. The percentages are the number of sequences translated by that method relative to the available contigs for that species. For more details on prot4EST components see the accompanying text and Chapter Two.

1: species codes - see Table 3.2

2: mean length of all coding regions predicted by prot4EST.

3: contigs that are predicted to represent genes that are encoded for by the mitochondrial genome.

4: the mean length of the coding regions predicted for a particular component of prot4EST.

3.4.4 Spliced leader libraries

Nematodes modify pre-mRNA precursors through *trans*-splicing to generate mature mRNA [124,125]. The most common *trans*-spliced exon is a non-coding 22 nucleotide sequence known as the splice-leader 1 or SL1. It is highly conserved across many species and is spliced to the 5' end of pre-mRNAs. Use of the SL1 transcript is estimated to be 80% of mRNAs in *C. elegans* [126], more than 80% in *Ascaris suum* [127] and approximately 60% in *Globodera rostochiensis* [49]. A second spliced leader exon family has also been identified [128]. The SL2-like spliced leader family is the predominant spliced leader found *trans*-spliced to downstream genes in operons of many nematodes (Blaxter *in prep.*). The widespread occurrence of these spliced leaders on nematode mature mRNAs has been used in the construction of several cDNA libraries, where only full-length cDNAs should be found. A total of 67 cDNA libraries from 23 species were constructed using PCR primers designed against either SL1 (62) or SL2 (five) sequences.

I examined the relationship between SL1/SL2 PCR-derived cDNAs. Clusters were split into four groups: those which contained only ESTs from SL-libraries; those where 50% of ESTs were from SL-libraries; remaining clusters with ESTs from SL-libraries; and those clusters without any SL-library ESTs (Table 3.4). Three species with ESTs from SL-libraries (*Meloidogyne arenaria*, *P. penetrans* and *Zeldia punctata*) did not provide clusters for all the categories, so were excluded. The translations from clusters that contained only ESTs from SL-libraries were significantly shorter than those clusters with no SL-primer ESTs for 14 species (Table 3.4). It is expected that ESTs in exclusively SL-library clusters derive exclusively from the 5' end of the putative mRNA thus reducing the total coverage. There was no detectable difference in the methods used to translate these clusters, so there is no variation in the accuracy of the translations. When SL-primed ESTs comprised the majority of a cluster's membership, the coding regions were shorter in seven species but longer in six

(Table 3.4). Strikingly, where ESTs from SL-libraries make up less than half a cluster's total, the average length of coding region was greater for 13 species, compared to clusters with no SL-library ESTs. These clusters are further advantaged with an increased proportion of coding regions predicted by the BLAST-similarity component of prot4EST. Extensive benchmarking of prot4EST has shown this method to be the most accurate for predicting coding regions.

Clusters containing a mix of ESTs from both SL-primer and more conventional primer-ligation based libraries, provide longer and more robust coding regions. There was a clear benefit in being able to anchor the cluster at the 5' end of cognate mRNA with SL-primed ESTs and improve the overall coverage with the other ESTs.

Species[5]	All SL[1]	>50% SL[2]	≤ SL[3]	None[4]
ACC	93.38	127.07	148.46	153.82
ASC	127.19	166.74	219.19	164.18
AYC	119.26	146.31	190.16	209.64
BMC	79.86	112.87	157.32	144.18
DIC	112.64	139.78	169.59	138.36
HCC	144.05	160.22	192.86	192.17
HGC	127.21	144.28	162.29	193.30
HSC	119.90	143.80	205.57	173.32
MAC	149.00	n/a	n/a	150.95
MCC	124.51	167.00	163.46	136.55
MHC	143.69	173.17	180.63	143.39
MIC	157.60	174.08	183.81	163.87
MJC	134.13	153.07	178.26	146.82
MPC	149.76	155.44	213.55	143.97
OOC	123.30	168.69	198.41	144.92
PEC	175.80	n/a	n/a	n/a
PPC	157.41	166.41	194.89	182.51
PTC	139.46	167.53	199.53	137.70
RSC	124.06	152.00	128.00	51.12
SRC	114.50	148.48	250.88	169.65
TCC	142.46	151.30	223.40	156.78
TSC	155.87	185.50	255.43	157.41
ZPC	185.62	n/a	n/a	n/a

Table 3.4

Effect of SL-primer libraries on coding region length.

A total of 67 cDNA libraries were generated using PCR primers against the splice leader exons. These libraries are from 23 species. All the clusters from these species were split into four categories: 1) all the ESTs in the cluster are from SL-primer libraries, 2) over 50% are from SL-primer libraries, 3) there are ESTs from SL-primer libraries, but less than 50% and 4) the cluster contains no ESTs from SL-primer libraries.

The average length of each group of clusters was calculated. Groups 1-3 were, in turn, compared to the more orthodox group 4 using a t-test to identify whether differences in length of coding regions were significant ($p < 0.05$; coloured red).

3.4.5 The effects of cluster size and singletons

Three quarters of the largest clusters (> 21 ESTs) were translated using the BLAST component of prot4EST (Table 3.5), a proportion similar to that in the *C. elegans* proteome. For clusters containing two or more ESTs, 97-99% of coding regions were identified with BLAST-similarity or ESTScan. There was a noticeable skew in this pattern when the cluster contains only one EST. Singletons may be genes with very low levels of expression, or an artefactual string of nucleotides, due to technical errors. Seven percent of singletons cannot be robustly translated. This may be due to a number of reasons:

1. The coding region contained within the singleton has a sequence composition significantly different from that detected in approximately 50% of contigs that used to generate the simulated transcriptome.
2. The singleton does not contain a coding region, and is entirely made up from either 5' or 3' untranslated regions.
3. The singleton does not contain a coding region, and is a contaminant or artefact of the sequencing process or subsequent quality controls.

In the absence of complete genomes, it is difficult to determine the precise reasons why ESTs cannot be translated. However, with the forthcoming release of the genome for the filarial nematode *Brugia malayi* [129], further investigation is possible. The genome for *B. malayi* is still in a draft stage and so there are as yet unsequenced regions. There have also been issues with assembling the AT-rich genome. Gene identification is an ongoing process, benefited by coding region identification in ESTs. The 26,000 ESTs generated for *B. malayi* were an integral part of gene prediction in the raw genome sequence. To date 11,894 coding genes have been identified from the genome. It is unlikely that this number is exhaustive, but it is close to the estimated gene count [130]. All the singleton contigs were compared to the first release of the *B. malayi* proteome using BLASTX. A little over half of these sequences

(3,596 out of 7,085) had a significant match to a *B. malayi* protein (E value cut off e-3). The majority of these contigs (84%) contain coding regions which are identified by the BLAST or ESTScan component of prot4EST. Surprisingly only half of the singleton contigs which were translated by ESTScan had a significant hit to the *B. malayi* proteome. There were no detectable differences in length between those contigs that do match *B. malayi* CDS and those without hits. As these contigs have stretches of nucleotides which share hexamer frequencies with *B. malayi* mRNA, it can be argued that many of these represent genes that have yet to be identified from the raw genomic sequence, or that were from genomic regions yet to be assembled to the current draft.

Over 2,500 singletons did not have coding regions that could be detected by BLAST searches or ESTScan. Of these only 300 (12%) had a significant match to the *B. malayi* proteome. The remaining 2,200 singletons could derive from genes that have yet to be identified in the draft genomic sequence. However, as they seem to lack the *B. malayi* coding signal characterised by ESTScan, their validity is doubtful. If this skepticism was applied across the nematode datasets, it would remove approximately 7,000 contigs, 6% of the total. This level should be kept in mind when the level of novelty in a dataset is announced. Additional peculiarities of the EST dataset from *B. malayi* are investigated below.

Cluster size	BLAST-similarity	ESTScan	DECODER	Longest_ORF
1	0.45	0.46	0.02	0.07
2	0.56	0.41	0.01	0.02
3	0.64	0.34	0.01	0.01
4	0.68	0.31	0.01	0.01
5	0.68	0.31	0.00	0.01
6-10	0.72	0.27	0.00	0.01
11-20	0.74	0.25	0.00	0.01
21+	0.75	0.23	0.01	0.01

Table 3.5

The distribution of method used depending on the size of the cluster.

The number of ESTs which constitute a cluster (its size) is a measure of how valid the cluster is. The larger clusters are more likely to share significant BLAST-similarity with a characterised protein. Singletons make up, relatively, the largest proportion of clusters for which a coding region cannot be accurately predicted.

3.4.6 Focus on the Spiruromorpha

For most nematode species over 90% of contigs were translated by BLAST-similarity or ESTScan. Noticeable exceptions were three species from the Spiruromorpha, *B. malayi* (71%), *Onchocerca volvulus* (78%) and *Wuchereria bancrofti* (68%). Between a fifth and a quarter of coding regions were predicted using the longest ORF method (Figure 3.1). Little confidence can be attached to polypeptide sequences produced by this method due to its rather basic assumptions on coding region identification. These three datasets are not the only ones from members of the Spiruromorpha and thus to distinguish them from *Litomosoides sigmondontis* and *Dirofilaria immitis*, I shall hereafter use the abbreviation 'BmOvWb'.

The proportion of BmOvWb contigs that shared significant BLAST-detected similarity with a protein from the UniProt database was below average. In fact *B. malayi* had the lowest proportion of the 37 collections studied (Figure 3.1). The number of contigs translated with ESTScan was also below average. Datasets of other species, for example *Haemonchus contortus* and *Strongyloides stercoralis,* had a comparable proportion of contigs translated by ESTScan, but these usually complemented a large proportion of coding regions predicted by BLAST-similarity. The most striking feature was the proportion of contigs in which ESTScan could identify a coding region regardless of its accuracy. Across all datasets 95% of contigs without BLAST-similarity have some sort of predicted coding region, although approximately 10% of these are later removed by prot4EST's quality filters. For BmOvWb datasets approximately 60% of contigs have an ESTScan-determined coding region, before the use of filters. The poor performance of ESTScan can be attributed to two potential sources of error. The first is the composition of the training data, the second, the quality of the sequences under analysis.

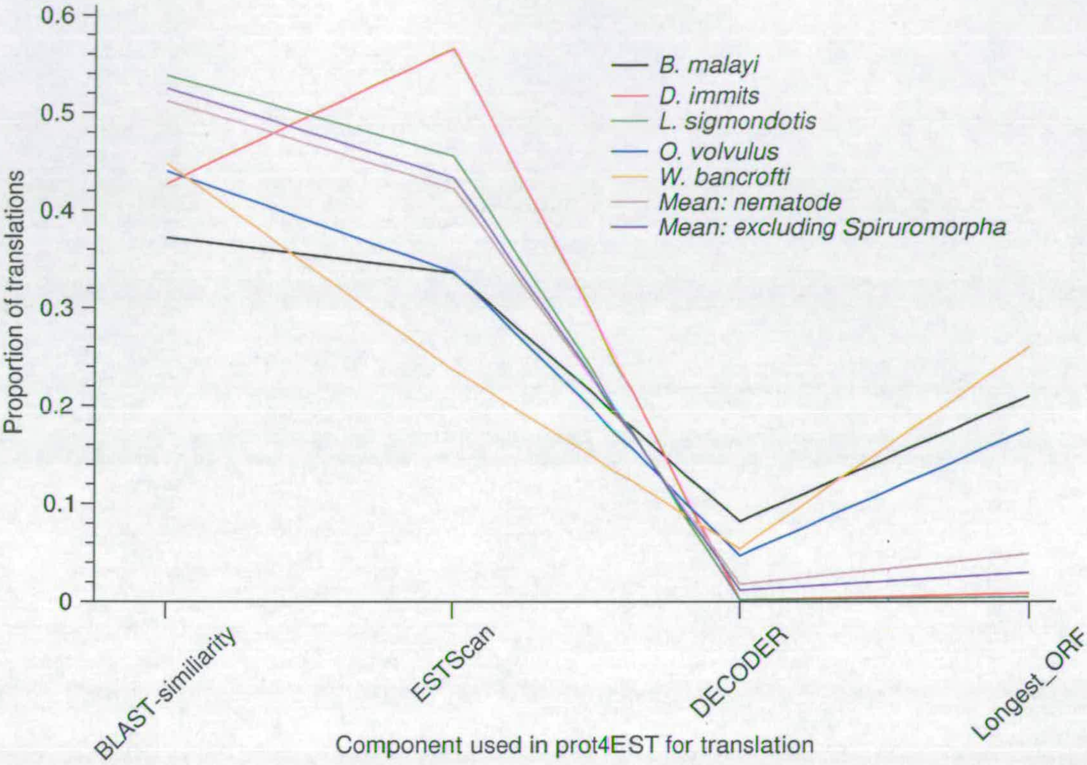


Figure 3.1 Problems in detecting coding regions in the Spiruromorpha.

There is a striking increase in the over-reliance on DECODER and Longest_ORF components for three species of Spiruromorpha. This suggests that a large number of EST contigs from these species contain no coding region, or one that cannot be characterised.

Training set fidelity

The training set for each species was made by reverse translating WormPep using a species-specific codon usage table. The codon usage was determined from those coding regions determined from BLAST-similarity. One of the primary assumptions with this method was that these coding regions would provide an accurate distribution of the hexamer frequency for each species. Translation efficacy from other species datasets suggested that this assumption was legitimate. It was possible that a low proportion of coding regions detected by BLAST-similarity could have had a deleterious effect upon the hexamer frequency. However this was not seen in datasets with a similar proportion of BLAST-translated contigs, such as *Trichuris vuplis* and *Meloidogyne javanica.*

The nucleotide hexamer frequency is heterogeneous across the transcriptome of a single species [131,132]. This is typically true in highly expressed genes, and is one reason why small training sets have previously produced poor results for ESTScan. The variation could be such that certain genes could not be recognised by an algorithm trained with supposedly representative information. It is unlikely, however, that such a large number of unrelated genes would display this behaviour. To explore this issue further, I used mRNAs predicted from the draft *B. malayi* genome. These sequences formed the training set from which the HMM emission probabilities were estimated. Despite this bootstrapping procedure the number of *B. malayi* contigs without any ESTScan prediction was unchanged. This confirms the lack of a coding signal in these contigs and supports the use of a pseudo-transcriptome as the training set for HMM parameter estimation.

Sequence quality

The majority of sequences without a detectable coding signal are singletons. The BmOvWb datasets did not have a greater proportion of singletons relative to other nematode taxa.

However they were less likely to contain a detectable coding region. One possible source of this disparity is the quality of the cDNA library from which the ESTs were selected, and the technology used for sequencing. The library information for each EST is stored within NEMBASE [26]. The proportion of ESTs without a detectable coding region were compared between libraries to identify libraries that contained an elevated level of such ESTs. Of 25 libraries for *B. malayi* five had more low quality ESTs than would be expected by random sampling (G-statistic = 682; p<<0.001) (Table 3.6). The proportion of low quality ESTs found in libraries for other nematode species varies between 1-5% and only once was found to be above 10%. Two libraries from the eight available for *O. volvulus* were shown to contain an excessive number non-coding ESTs. A statistical comparison for *W. bancrofti* libraries was not possible due to the small number of libraries. However it is obvious that some of these libraries contain a large number of ESTs without a detectable coding region (Table 3.6). The information provided in the library reports offers plausible explanations for the low quality of sequences. The libraries highlighted were made in the mid 1990s, and were constructed and sequenced using methods and technologies that have improved greatly in the last few years. The *W. bancrofti* libraries were made from a small number of nematodes, which increases the relative concentration of artefacts in the library. Finally there may have been lab-specific technical issues for many of the highlighted libraries.

Species	Library Title[1]	lib_code[2]	Poor clusters[3]
Brugia malayi	*Brugia malayi* infective L3 JHU93SL-BmL3	279	14%
Brugia malayi	*Brugia malayi* adult male cDNA (SAW94NL-BmAM)	281	13%
Brugia malayi	*Brugia malayi* infective larva cDNA (SAW94WL-BmL3)	282	13%
Brugia malayi	*Brugia malayi* microfilaria cDNA (SAW94LS-BmMf)	283	13%
Brugia malayi	*Brugia malayi* L4 larva (JHU93SL-BmL4)	389	40%
Onchocerca volvulus	*Onchocerca volvulus* adult male cDNA (SAW98MLW-OvAM)	1375	11%
Onchocerca volvulus	*O. volvulus* adult female cDNA 26 h following ivermectin (PF99PF-OvAF26)	3756	18%
Wuchereria bancrofti	*Wuchereria bancrofti* microfilaria cDNA (SAW95SjL-WbMf)	518	18%
Wuchereria bancrofti	*Wuchereria bancrofti* L3 cDNA (SAW96MLW-WbL3)	1630	10%

Table 3.6

cDNA libraries for the Spiruromorpha that contain a disproportionate number of clusters for which no coding region can be detected. These libraries were identified using a G-test to compare the fraction of "non-coding" clusters across all the cDNA libraries of each species. The fraction of "non-coding" clusters for libraries across all nematode datasets is 1-5%

1: taken from NEMBASE – http://www.nematodes.org

2: the unique identifier for the library in question

3: those clusters which did not have a coding region identified by either the BLAST-similarity or ESTScan components of prot4EST.

3.4.8 Number of robustly translated ESTs?

The data so far presented authenticates nearly all coding regions are predicted by either the BLAST-similarity or ESTScan components of prot4EST. However there are serious doubts over the validity of those EST contigs whose coding region must be identified by other less accurate methods. Validation of these contigs is not possible without the complete or robust draft genome sequence. As these do not exist for most of the species under study, in the subsequent analyses I have used only those polypeptides predicted by the preferred first two components of prot4EST. The dataset thus comprises 158,349 polypeptide sequences from the 39 species of the phylum Nematoda (Table 3.7). NemPep3 now allows a number of comparative studies to be undertaken on a scale that have to date only been performed across bacterial species and yeast proteomes. Subsequent chapters will describe analysis of protein relationships and an in depth protein domain comparison.

3.4.9 Effect of AT content and amino acid usage - global

The amino acid composition for each proteome has been calculated (Table 3.8). To investigate whether there is any correlation between these amino acid frequencies and the AT content of each species, a series of statistical analyses were performed. A global comparison of amino acid frequency between nematode species was conducted. However, 94% of the variation observed was between the amino acid frequency within each species. No significant difference was found for the variation of each amino acid between the said species (two-way ANOVA, p=1.0). To detect any trend it was necessary to conduct a series of pairwise comparisons. These comparisons at first appeared to demand a chi-squared test. However the scale of the analysis renders the chi-squared statistic ineffective [133]. With the large sample sizes used here, approximately 1.5 million amino acids for *Heterodera glycines* to 32,000 for *Zeldia punctata*, the chi-squared value would be extremely high even with only a slight compositional difference. This would have resulted in a highly significant statistic even if frequencies were indistinguishable on a plot. This problem confronted Echols and co-workers, when comparing features of genes and pseudogenes [133]. To quantitatively evaluate differences, the datasets for comparison were treated as an N-dimensional vector and the distance between the two calculated. See section (Methods 3.3.6) for a detailed description of the calculation. Pairwise distance calculations were performed for all combinations, including the proteomes of *C. elegans* and *C. briggsae* (Figure 3.2). There is no statistical correlation between the number of contigs that comprise a dataset and its average distances (Spearman's rank; rho=-0.24; p>0.5). The scale of this analysis is sufficiently large to ensure that any bias in composition is real rather than random. Unsurprisingly the pairwise distances are lowest between closely related organisms, such as the *Meloidogyne* species. As the AT content difference between species increases, the trend is for the distance to increase. It is difficult to separate AT content from phylogenetic relatedness, because, as mentioned earlier, closely related species have a similar AT content. A method called comparative analysis of independent contrast (CAIC) finds and calculates phylogenetically independent contrasts in one or more variables enabling hypotheses of correlated evolution to be tested [134]. Unfortunately, the method should not be applied to the relative data that is under scrutiny here. It is striking that the different families comprising the Tylenchomorpha have contrasting AT contents. For example, there is a 13% difference in AT content between the Heteroderidae and the Meloidogynidae but their average amino acid composition distance is only 2.4% (expected distance would be approximately 3.5%). A similar pattern is observed when comparing the Pratylenchidae with other tylenchomorph groups.

Taxon	Total	Translated EST contigs[1]	UniProt/SP[2]	UniProt/TrEMBL[2]	Others
Dorylaimia (clade I)	10,745	10,583	6	156	0
Enoplia (clade II)	0	0	0	0	0
Spirurina (clade III)	28,226	27,289	165	772	0[4]
Tylenchina (clade IV)	52,706	52,039	15	652	0
Rhabditina (clade V)	67,956	25,363	2,749	20,510	19,334[3]

Table 3.7

Taxonomic distribution of NemPep3.

NemPep3 (21/08/05) combines the predictions offered by prot4EST[1] and the complete *Caenorhabditis* genome projects, with all other nematode proteins available from the UniProt database.

[1]: only using translations from BLAST-similarity and ESTScan components of prot4EST.

[2]: as described by the NEWT database (http://www.ebi.ac.uk/newt/display)

[3]: not all of the predicted proteins from *C. briggsae* are available in UniProt. I have included BrigPep2 from WormBase. The *C. elegans* proteome (WormPep140) is included in UniProt.

[4]: the predicted proteome for the *Brugia malayi* draft genome is not yet publicly available, so is excluded from NemPep3. I anticipate its inclusion at a later date.

Table 3.8

Amino acid usage across nematode species.

spID	AT_prop[1]	Ala	Cys	Asp	Glu	Phe	Gly	His	Ile	Lys	Leu	Met	Asn	Pro	Gln	Arg	Ser	Thr	Val	Trp	Tyr
SRC	0.674	0.058	0.020	0.055	0.065	0.047	0.062	0.021	0.072	0.079	0.086	0.024	0.056	0.042	0.034	0.045	0.070	0.054	0.060	0.011	0.037
SSC	0.661	0.060	0.020	0.056	0.069	0.042	0.065	0.022	0.069	0.078	0.084	0.023	0.054	0.043	0.036	0.049	0.069	0.054	0.060	0.011	0.035
MCC	0.637	0.060	0.021	0.051	0.067	0.049	0.061	0.023	0.069	0.074	0.093	0.023	0.052	0.043	0.041	0.054	0.068	0.049	0.057	0.012	0.032
MHC	0.629	0.061	0.021	0.050	0.067	0.054	0.064	0.022	0.066	0.070	0.092	0.023	0.051	0.046	0.041	0.053	0.066	0.048	0.058	0.013	0.032
MAC	0.625	0.062	0.020	0.051	0.068	0.049	0.064	0.022	0.067	0.070	0.092	0.022	0.051	0.044	0.041	0.053	0.068	0.052	0.058	0.013	0.033
MJC	0.620	0.062	0.020	0.051	0.067	0.049	0.062	0.023	0.065	0.073	0.092	0.024	0.049	0.045	0.041	0.055	0.066	0.050	0.058	0.012	0.033
MIC	0.618	0.063	0.020	0.051	0.067	0.048	0.062	0.022	0.063	0.071	0.091	0.023	0.051	0.046	0.043	0.055	0.069	0.051	0.058	0.012	0.031
MPC	0.617	0.062	0.021	0.051	0.067	0.047	0.065	0.022	0.064	0.075	0.089	0.023	0.051	0.046	0.043	0.054	0.067	0.051	0.059	0.011	0.031
DIC	0.601	0.064	0.023	0.052	0.065	0.044	0.056	0.027	0.066	0.069	0.092	0.025	0.047	0.042	0.038	0.061	0.068	0.051	0.061	0.012	0.036
BMC	0.591	0.064	0.022	0.050	0.061	0.046	0.062	0.024	0.063	0.065	0.091	0.027	0.044	0.045	0.038	0.056	0.068	0.052	0.062	0.013	0.034
OVC	0.583	0.066	0.020	0.052	0.061	0.045	0.064	0.023	0.063	0.066	0.088	0.026	0.044	0.046	0.038	0.058	0.066	0.051	0.061	0.014	0.033
PTC	0.580	0.069	0.020	0.053	0.063	0.044	0.068	0.024	0.062	0.078	0.083	0.023	0.048	0.044	0.035	0.053	0.066	0.054	0.064	0.011	0.036
ZPC	0.576	0.075	0.015	0.051	0.058	0.043	0.069	0.025	0.061	0.080	0.086	0.019	0.048	0.046	0.037	0.055	0.056	0.056	0.066	0.011	0.038
TSC	0.574	0.065	0.023	0.052	0.063	0.048	0.059	0.024	0.059	0.065	0.093	0.027	0.046	0.044	0.037	0.056	0.070	0.051	0.067	0.013	0.034
WBC	0.572	0.067	0.021	0.050	0.063	0.043	0.066	0.025	0.058	0.069	0.088	0.025	0.041	0.047	0.037	0.061	0.066	0.050	0.062	0.014	0.032
CAEEL[2]	0.571	0.064	0.020	0.054	0.066	0.047	0.054	0.023	0.061	0.064	0.086	0.026	0.049	0.049	0.042	0.052	0.081	0.059	0.062	0.011	0.031
LSC	0.564	0.068	0.020	0.052	0.063	0.043	0.064	0.023	0.060	0.067	0.087	0.027	0.043	0.048	0.038	0.062	0.069	0.053	0.068	0.013	0.032
CAEBR[3]	0.558	0.063	0.020	0.053	0.068	0.047	0.054	0.023	0.060	0.064	0.085	0.026	0.048	0.050	0.041	0.054	0.080	0.058	0.061	0.011	0.031
PEC	0.540	0.071	0.018	0.054	0.069	0.040	0.068	0.023	0.056	0.080	0.083	0.025	0.045	0.044	0.044	0.063	0.062	0.050	0.060	0.011	0.030
HCC	0.534	0.074	0.022	0.055	0.065	0.043	0.066	0.024	0.054	0.066	0.085	0.026	0.041	0.049	0.036	0.059	0.066	0.052	0.068	0.013	0.033

Continued overleaf...

spID	AT_prop[1]	Ala	Cys	Asp	Glu	Phe	Gly	His	Ile	Lys	Leu	Met	Asn	Pro	Gln	Arg	Ser	Thr	Val	Trp	Tyr
NAC	0.533	0.072	0.022	0.057	0.062	0.044	0.066	0.023	0.054	0.066	0.084	0.025	0.042	0.049	0.038	0.063	0.068	0.053	0.067	0.012	0.032
XIC	0.526	0.072	0.021	0.055	0.059	0.043	0.069	0.024	0.053	0.062	0.089	0.026	0.042	0.050	0.042	0.056	0.068	0.055	0.068	0.013	0.033
PVC	0.524	0.065	0.019	0.048	0.060	0.058	0.065	0.025	0.058	0.060	0.101	0.028	0.044	0.048	0.042	0.051	0.069	0.048	0.062	0.014	0.032
ASC	0.523	0.075	0.025	0.049	0.065	0.042	0.064	0.025	0.053	0.060	0.084	0.022	0.043	0.052	0.036	0.066	0.071	0.051	0.065	0.014	0.033
ACC	0.522	0.073	0.020	0.054	0.063	0.044	0.065	0.025	0.056	0.065	0.087	0.027	0.042	0.046	0.036	0.057	0.067	0.053	0.068	0.013	0.034
TDC	0.520	0.077	0.025	0.053	0.062	0.043	0.066	0.024	0.052	0.068	0.086	0.027	0.042	0.049	0.036	0.057	0.066	0.052	0.067	0.013	0.033
OOC	0.520	0.075	0.022	0.055	0.064	0.043	0.069	0.026	0.053	0.066	0.084	0.026	0.041	0.051	0.038	0.060	0.064	0.051	0.068	0.012	0.032
TCC	0.516	0.072	0.027	0.050	0.061	0.045	0.069	0.025	0.051	0.065	0.083	0.025	0.041	0.051	0.036	0.061	0.067	0.055	0.066	0.013	0.032
TVC	0.516	0.068	0.026	0.052	0.061	0.044	0.064	0.024	0.053	0.064	0.092	0.029	0.042	0.047	0.037	0.057	0.069	0.053	0.068	0.014	0.035
ALC	0.508	0.078	0.021	0.044	0.059	0.043	0.070	0.021	0.049	0.062	0.081	0.021	0.045	0.060	0.035	0.067	0.075	0.047	0.067	0.017	0.034
HSC	0.505	0.071	0.020	0.052	0.071	0.048	0.063	0.025	0.055	0.067	0.092	0.025	0.044	0.044	0.044	0.061	0.065	0.050	0.062	0.012	0.027
TMC	0.504	0.071	0.028	0.052	0.063	0.044	0.067	0.023	0.052	0.065	0.088	0.027	0.040	0.049	0.037	0.060	0.066	0.050	0.066	0.012	0.033
AYC	0.500	0.080	0.021	0.055	0.068	0.041	0.069	0.024	0.052	0.066	0.083	0.025	0.040	0.050	0.038	0.061	0.066	0.052	0.067	0.012	0.030
HGC	0.494	0.074	0.020	0.053	0.064	0.048	0.064	0.026	0.053	0.061	0.093	0.025	0.043	0.049	0.043	0.061	0.069	0.051	0.063	0.012	0.027
NBC	0.491	0.080	0.019	0.055	0.064	0.042	0.073	0.025	0.051	0.070	0.081	0.026	0.039	0.047	0.035	0.062	0.066	0.051	0.071	0.012	0.029
GPC	0.486	0.073	0.022	0.052	0.062	0.046	0.071	0.026	0.052	0.062	0.091	0.024	0.043	0.050	0.043	0.059	0.066	0.050	0.065	0.012	0.030
GRC	0.485	0.074	0.020	0.052	0.062	0.048	0.068	0.025	0.055	0.061	0.093	0.025	0.041	0.047	0.041	0.061	0.067	0.051	0.065	0.012	0.029
PPC	0.484	0.076	0.019	0.054	0.069	0.041	0.069	0.024	0.054	0.067	0.084	0.025	0.039	0.048	0.034	0.062	0.070	0.053	0.065	0.013	0.029
RSC	0.466	0.076	0.021	0.048	0.064	0.039	0.069	0.026	0.053	0.081	0.083	0.024	0.041	0.044	0.045	0.073	0.054	0.047	0.066	0.012	0.030

Table 3.8

Amino acid usage across nematode species.

legend overleaf...

Table 3.8 (previous page)

Amino acid usage across nematode species.

The frequency of each amino acid was calculated for all nematode proteomes. These metrics were used to calculate pairwise distances between species and identify significant correlations between the AT content of a species coding regions and the usage of certain amino acids. Both significant positive (red) and negative (blue) correlations are highlighted.

1: The proportion of adenine and thymine found in the species' coding regions

2: Coding regions from WormPep140

3: Coding regions from BrigPep2

Figure 3.2 (next page)

Pairwise comparisons of amino acid distances.

A distance measure of amino acid usage was calculated for all possible pairs of species (see Methods 3.3.6). The species are ranked according to AT content (see Table 3.8).

The species codes are available from Table 3.2.

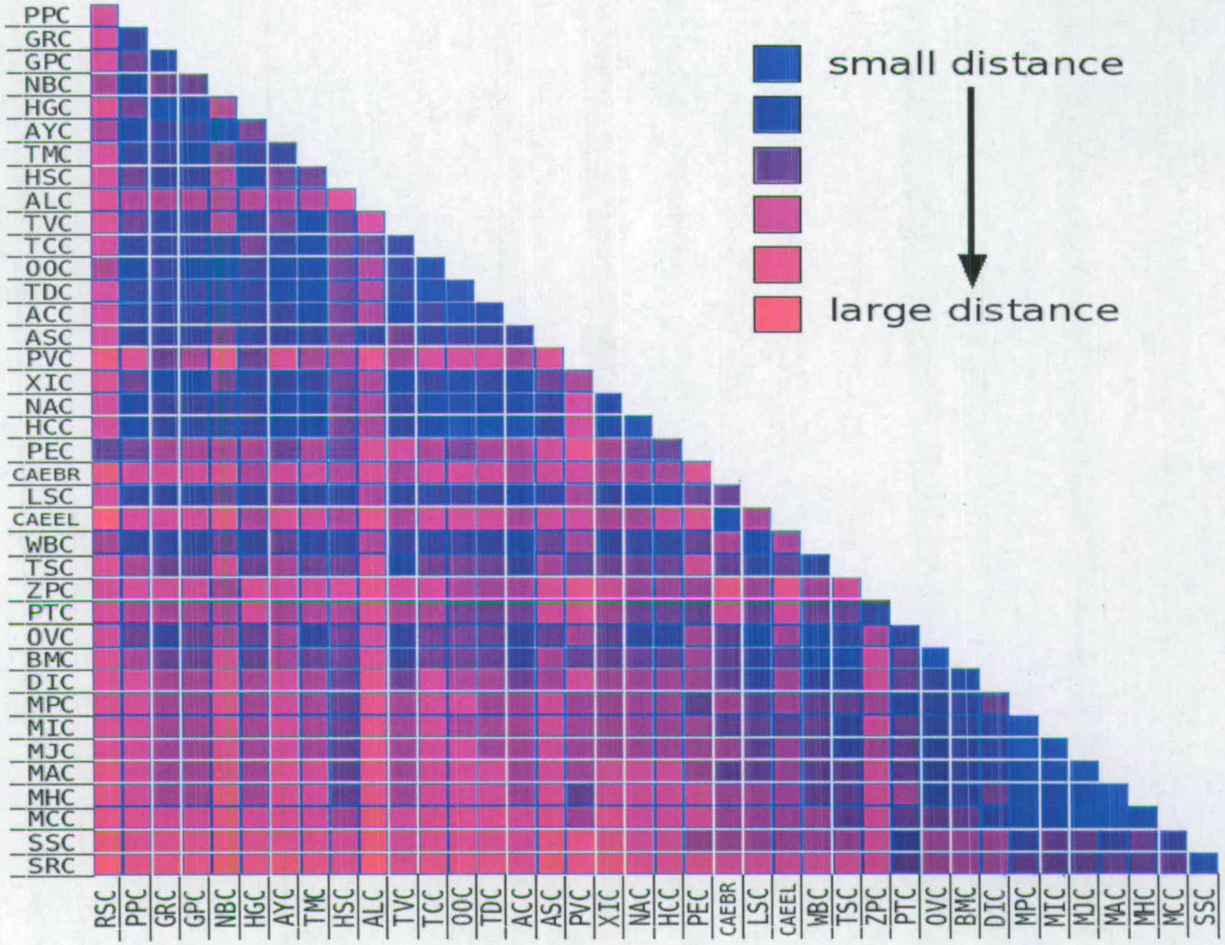


Figure 3.2
Pairwise comparisons of amino acid distances.
legend on previous page

3.4.10 Effect of AT content and amino acid usage - local

It is clear that there are obstacles to performing pairwise species comparisons of amino acid frequencies. A simpler comparison tests the association between AT content and frequency of individual amino acids. The usage of nine amino acids show very strong correlations (Spearman's rank; $p<0.001$) with the coding AT content of species (Figure 3.3). The directions of the correlation (3 positive and 6 negative) are as expected given the AT proportion of the codons for the corresponding amino acids. With the exception of tryptophan (Trp), all amino acids with a clear bias in their codons' AT content showed a significant correlation ($p<0.05$). Tryptophan is the rarest amino acid, which may explain why there is no fluctuation in its frequency between species. The strong correlation shown in the inter-species frequency of glycine (Gly) was a surprise, despite the very low AT content of its codons. Glycine is the smallest amino acid and tends to terminate rigid secondary structure elements because of the entropy cost of restraining its flexibility. Glycine also allows backbone angles, seen on a Ramachandran plot, which cannot be taken on by other amino acids. Any amino acid replacing glycine must be small and avoid physio-chemical properties that conflict with the local environment. The most curious usage profiles are those for valine (Val) and histidine (His), which both showed a significant negative correlation. These two amino acids have no bias in AT content in their codon sets. The increased frequency in species with low coding AT content may be compensation for other amino acids which have an increased abundance in high AT species. Valine may replace isoleucine (Ile) for internal packing in the protein. Histidine can substitute for a number of amino acids depending on their role: indeed it can substitute for tyrosine (Tyr) in its function as a nucleophile and lysine (Lys) for protonation. However there are no correlations in frequency between amino acid pairs. To investigate these trends further, comparison within homologous, preferably orthologous, protein families is required. On a small scale this is eminently possible and the focus of future work. However a larger analysis is dependent upon the accurate automated classification of these orthologous relations, a task that is described in the next chapter.

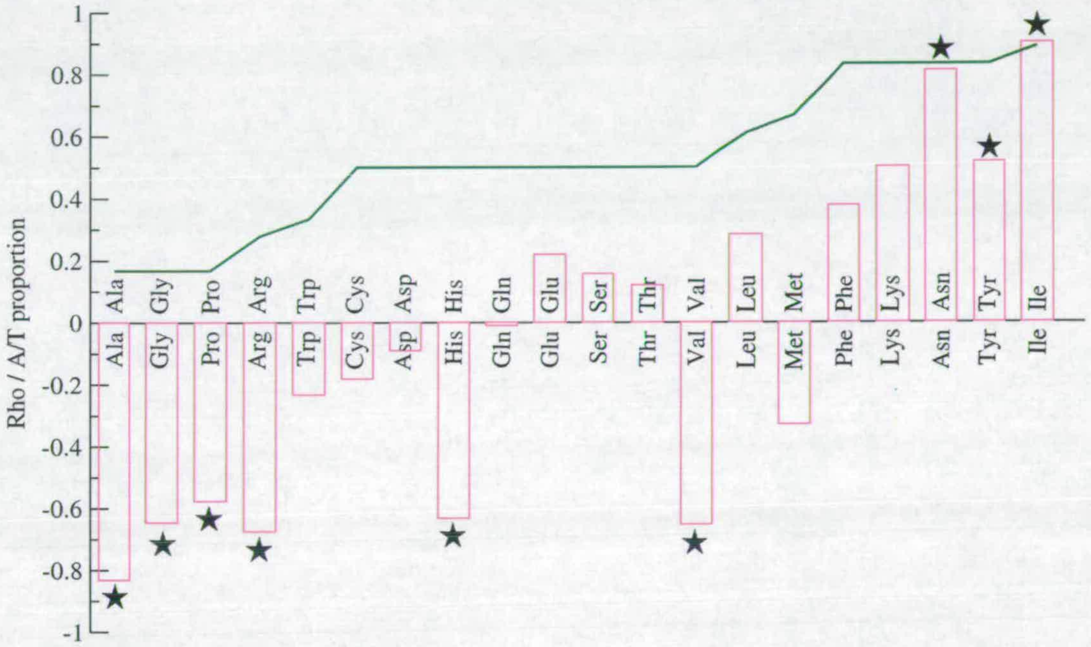


Figure 3.3

Amino acids: their AT content and usage

For each amino acid two metrics were calculated: the mean AT content of its codons, and the correlation between the usage of the amino acid (green) and the coding AT content across nematode transcriptomes (magenta). Significant correlations are denoted by a star. The trend is for amino acids with a high AT content to be more frequently used in species with high proportion of A and T nucleotides in their coding regions. Similarly the usage of amino acids with low AT contents across their codons is higher in species with a low proportion of coding AT.

Standard three letter amino acid codes are used.

3.5 Conclusions

I have presented NemPep3, a robust collection of proteins predicted from EST contigs, predominantly from parasitic nematodes. These partial proteomes have been combined with the complete proteomes from *C. briggsae and C. elegans*. The nematode datasets were the motivation behind the creation of the translation pipeline prot4EST, as they offered problems that had not been considered with other EST translation solutions, but problems that were pertinent to the overwhelming majority of EST projects. The most important was creating a training set for ESTScan so those EST contigs that did not share similarity with a known protein could be translated accurately. Training sets of sufficient size were created using a reverse-translation approach to generate a simulated transcriptome. I took advantage of availability of the *C. elegans* proteome and while, I accept that assumptions were made, they are valid and the method enabled better coding region prediction. Using simulated training sets, while powerful should be done with care, especially in the choice of proteome used as a template.

ESTs that pass through the PartiGene system are subjected to a number of routines that remove sequences of low 'quality'. These include nucleotides at the 5' or 3' termini of a sequence with low phred scores, and BLAST comparisons with bacterial sequences to highlight probable contaminants. In addition, prot4EST identified EST contigs which putatively lacked coding regions. It is possible that some of these are simply untranslated regions (UTRs), probably from the 3', although for the datasets from *B. malayi*, *O. volvulus* and *W. bancrofti* the number of untranslated sequences was too high to be simply a consequence of UTR sequence. Regardless of the reason for not identifying a coding region, such sequences should not be included in subsequent analyses, as they would inflate the number of species-specific proteins in certain proteomes. Therefore the search for coding regions presents another opportunity to 'clean' the datasets and ensure only high quality sequence is analysed.

3.7 Acknowledgements

I wish to thank Asher Cutter for putting the idea of reverse-translating the *C. elegans* proteome in my brain. Without it I shudder at the thought of having to use some of the earlier translations in NemPep1 and NemPep2. Also to Constance Finney, who reeled in all my ideas about statistics and pointed me in the right direction.

Chapter Four - Exploring nematode proteinspace

4.1 Abstract

Diversity between organisms can be explored through any number of avenues. On way is to compare the proteomes of each species, whose individual proteins can be though of as occupying proteinspace. The creation of NemPep3 permits such comparisons between nematode species. Of importance is the level of novelty in nematode proteinspace; what is the proportion of proteins restricted to the phylum, or to parasitic lineages? In this chapter I report that approximately 70,000 proteins (46% of NemPep3) share no detectable (BLAST) similarity to a sequence outside the Nematoda, and that there is no slowing down in new protein discovery as more nematode species are added. The extent of gene loss in *Caenorhabditis elegans* is also investigated, as NemPep3 provides a more robust picture of nematode evolution. In addition NemPep3 was used to update a popular form of generating protein families. The consequence of which is to alter the previously reported dynamic of metazoan relationships in the dataset.

4.2 Introduction

4.2.1 Proteinspace

The different proteins that comprise a species' proteome can be thought of as occupying a slice of proteinspace, defined here as the *composite of properties for all proteins*. The relative position of two proteins, either within or between organisms' proteomes, is defined by their possession of features, such as domain architecture, sequence divergence and protein biochemistry. The proteomes of species with more complex genomes usually occupy a larger region of proteinspace, with a phylum containing the proteinspace that is the union of the constituent species' proteinspace.

As additional sequence becomes available, the area in proteinspace occupied by an organism's proteome grows, until its genome has been completely sequenced. Sequence data from a related species will show significant overlap, but it is likely that there will be some region of proteinspace that the two will not share. As related species adapt to distinct ecological niches the selective pressures upon their proteomes will alter, changing their position in proteinspace. This concept can be applied to the nematodes, with species evolving a range of diverse life-cycles and alternative feeding strategies. How disparate is nematode proteinspace? Do species which have independently evolved similar feeding strategies [43] share regions of proteinspace to the exclusion of more closely-related taxa. An analysis of proteomes from complete prokaryote genomes showed that sequencing additional species has produced diminishing returns in terms of novelty in proteinspace [135]. If the proteinspace of nematodes is similarly restricted then, given the completely sequenced genomes of two *Caenorhabditis* species, further sampling will result in a reduced rate of protein discovery.

The determination of a protein's relative position is reliant on annotation being assigned. In this age of high-throughput sequencing, annotations of protein function are derived from a small set of proteins, perhaps less than 5% of the total number of known protein sequences, whose function has been determined experimentally [64,136]. In the absence of experimental data, protein function is inferred by sequence similarity to a protein of known function. This is certainly the situation that confronts the analysis of EST datasets.

4.2.2 Annotating gene products

There are common features in publications describing the completion of a eukaryote genome sequence [35,137,138,139]:

- genome sequencing – techniques and type of clones used,

- assembly – software, parameters, number of scaffolds and contigs,
- gene finding – both coding and non-coding,
- gene annotation – functional assignment.

Each step has its own problems and controversies, but the work presented in this and the next chapter is concerned only with annotation of putative gene sequences, in particular, their evolutionary origin (this chapter) and their function (Chapter Five). Each month sees more publications describing the annotation of individual species' EST datasets, and occasionally comparative analysis across several related species (Table 4.1).

All these studies describe high-throughput, automated annotation processes which attempt the assignment of various functional classifications. These include:

- protein domains or motifs identified through Pfam [77] or Interpro [76],
- Gene Ontology (GO) terms [115],
- comparison to the KEGG metabolic pathways database with KEGG [114],
- assignment of functional description from similarity to another sequence.

Species / taxonomic group	Additional information	Reference
Mesocricetus auratus	Hamster testis	140
ambystomatid salamanders	Two related species	141
Apis mellifera	Honey bee	142
Fundulus heteroclitus	Mummichog fish	22
Solanaceae	Six related species	143
Xenopus laevis	Detail HOX gene study	144
Daphnia	Water flea (crustacean)	145
Gallus gallus	With genome paper	24
Ancylostoma	Two related species	146
Strongyloides ratti	nematode	48
Meloidogyne incognita	nematode	49

Table 4.1

Recent examples of analyses on species EST datasets.

4.2.3 Detection of patterns of evolution using BLAST

The assignment of functional description is often carried out using BLAST search algorithms to detect significant sequence similarity. Often in the analyses the entire complement of gene objects (clustered or individual ESTs) is compared to two or more model species or taxonomic groupings. For example, BLAST searches were used to search for similarity between the EST dataset for the Cnidarian *Acropora millepora* and model metazoan species (*Caenorhabditis elegans, Drosophila melanogaster* and *Homo sapiens*) [74]. This type of global analysis is often used to summarise how closely related genes of one species are to those of other species with the data displayed through Venn diagrams, or more elegantly with an interactive program [44]. The general trend of the analysis is frequently predictable depending upon the taxa compared. Occasionally; however, unexpected patterns are reported. Comparison of EST contigs from the honey bee, *Apis mellifera*, to complete genomes uncovered 23 contigs that were more similar to human proteins than the fellow neopterans *D. melanogaster* and *Anopheles gambiae* [142]. The authors propose that these putative genes have diverged less in honey bee and mammals, possibly due to selective pressures. The more likely explanation is that the orthologues for these proteins have been lost in the dipteran lineage of the two other insects in the study, and hence the comparison was between paralogues. Gene loss specific in the *D. melanogaster* lineage is well reported [147,148,149]. The effect of gene loss is further highlighted by analysis of ESTs from the cnidarian *Acropora millepora*, which revealed that 12% of EST clusters had a putative homologue in human but not in *D. melanogaster* and *C. elegans* [74]. These studies show how a series of BLAST searches can, when carefully considered, uncover interesting evolutionary patterns, but can also generate ill-considered hypotheses for protein evolution.

A further feature often investigated in EST analyses is the identification of genes that are thought to be novel to a species or particular monophyletic taxonomic group. These are the so-called orphan proteins [36]. There are two slightly different definitions of an orphan protein:

1. A protein that shares no significant similarity to sequence in other taxa (putative homologue) or other proteins from the studied taxon, i.e. a gene duplication event restricted to that species, an inparalogue [75].
2. A protein that shares no significant similarity with a sequence in another taxon (putative homologue).

I favour the second definition as it includes species-specific expansions of protein families that are an important mechanism for the evolution of new or modified protein functions. Comparing the level of novelty across different organisms' transcriptomes is difficult because each analysis varies BLAST parameters and other criteria, most notably E value cut off. However, it is clear from a number of studies on different species that searches have identified a noteworthy proportion of lineage-restricted genes (Table 4.2).

Taxonomic Group	**Number of orphan genes**	**Reference**
Ambystomatid salamanders	3,273 contigs (~17% of EST contigs generated).	150
Solanaceae	10-15% of contigs were specific to each species.	151
Legume	5.5% EST contigs restricted to legumes.	152
Nematodes	7-46% of contigs were specific to each species. Across all datasets 23% of nematode proteins were species-specific.	153
Helicosporidium	299 (43%) of contigs were orphans.	154

Table 4.2

Studies into lineage-restricted genes.

Of course the proportion of orphan genes in a genome is dependent upon a number of factors aside from search parameters. The depth of sequence survey from each species and the presence of datasets from closely-related species will have an effect. The proportions of orphan genes from the solanaceous species were the lowest reported from a comparative analysis of EST datasets [143]. The six species analysed belong to the same family and, while phylogenetic rank assignment is not equivalent throughout taxonomy, the rank of family usually designates a closely related group of species. Therefore, the expectation is for lower sequence divergence and hence fewer orphan genes. The number of EST contigs assembled for each species is also relatively high (4,466-38,239 contigs). For the larger solanaceous datasets, the majority of that species' transcriptome is represented. This increases the likelihood that similar sequences from closely-related species are present in the EST dataset and so detection of the homologue is more probable. The Parkinson *et al.* pan-nematode study presented data that pointed to a greater diversity in the combined transcriptome, with up to 46% of the EST contigs in an individual species' dataset restricted to that species [51]. Drawing general conclusions based on the relative diversity of the solanacean and nematode transcriptomes is not entirely appropriate, as the latter study represents a wider spectrum of species, and few nematode transcriptomes have been surveyed to the same depth as those of the Solanaceae.

Not only is there a large number of nematode species for which EST datasets are available, but they have been selected to span the phylogenetic diversity of the phylum [155,51]. This coverage has been made possible by the resolution of nematode phylogeny based upon small subunit ribosomal RNA gene alignments [28,40]. The presence of a robust molecular evolutionary framework has allowed the data generated to be mapped onto the nematode tree and evolutionary-informed hypotheses proposed and tested. The original pan-nematode study was able to identify not only genes that were specific to a particular species but also genes whose presence was restricted to lineages with a deeper last common ancestor. The work presented in this chapter updates the initial study with an additional seven taxa, which increases the number of nematode orders examined. The rigorous examination of NemPep3, described in Chapter Three, led to the removal of many EST contigs because a coding region could not be robustly assigned, the effect of which is discussed here. Proteins lacking any detectable homologue outwith the Nematoda are identified. Taken together with the phylogeny, these BLAST comparisons allow relationships between proteins to be considered and both the shared and unique diversity to be examined.

4.2.4 Orthology and Paralogy

The concepts of orthology and paralogy have become two of the most essential and controversial in post-genomic molecular biology [156,157,158,159,160]. The history, definition, classification, and exploitation of orthologues and paralogues is described at length in a recent review by Koonin [161]. However the principles are important and their application to the work in this thesis is described here.

A popular way to determine the function of a new protein is to identify orthologues in other genomes. The assumption is that orthologous proteins will have equivalent functions. Proteins whose relationship is due to a gene duplication event (paralogues) can evolve to perform distinct, if related, functions. One copy of the gene is free to evolve, altering its function to one that was not performed by the ancestral gene, this is termed neofunctionalisation. It is obvious that the use of similarity to an annotated protein as a way of assigning function must be based on a correct phylogenetic reconstruction. Herein lies the problem: all, except two, of the nematode proteomes that contribute to NemPep3 are incomplete. To illustrate these potential problems consider a gene that has been duplicated in *Haemonchus contortus*. The stochastic sampling of cDNA clones may only generate an EST for one of the genes. The inferred polypeptide is compared to the *C. elegans* proteome and similarity detected. A simplistic analysis would assign the function from the *C. elegans* protein to that of *H. contortus*. Across multiple datasets, even restricted to a closely-related subset, there could be a complex relationship of proteins that cannot be determined from the partial sequence. Throughout the rest of this thesis I will presume that orthology cannot easily be determined for a given pair of proteins. It is certainly safer to assume that seemingly related proteins are in- and outparologues and that functional assignments should be considered tentative.

4.2.5 Creation of protein families with BLAST

A powerful method to ascribe functional annotation to gene/protein sequences is to assemble them into related groups, often called families. Family classifications provide a higher level of function designation that can be more informative than simple pairwise annotation transfer. A further advantage is the reduction in sequence redundancy offered by families; a large number of proteins are contained within a small number of families [162]. There have been several attempts to assemble proteins into related families, with an emphasis on large-scale automation [163,164,118,135,165]. Currently the most widely cited, and accessible, is the COG system (Clusters of Orthologous Genes) [135,165], which was updated to include euKaryotes in the KOG database [69]. The procedure for K/COG construction relies upon symmetrical best BLAST hits and consists of the following steps [161]:

i. All-against-all comparison of protein sequences encoded in multiple genomes using the BLAST algorithm.

ii. Detection and clustering of obvious inparalogues.

iii. Identification of triangles of mutually consistent, genome-specific best hits. The previously detected inparalogues are treated as a single entity.

iv. Merging triangles with a common side to form K/COGs.

Further manual steps are used in the construction of KOGs:

v. Detection of distant homology using PSI-BLAST.

vi. Splitting of KOGs containing proteins linked through different subunits of multi-domain proteins.

vii. Assignment of proteins to KOGs based upon 'common domain architecture.' [69].

A simple examination of metazoan contributions to each KOG showed that *Homo sapiens*, *D. melanogaster* and *C. elegans* shared 3951 KOGs [65]. The number shared between only two of these species differed depending upon the combination: *H. sapiens* and *D. melanogaster* but not *C. elegans* shared membership of 311 KOGs, while 261 KOGs contained proteins from only *C. elegans* and *H. sapiens* (206) or with *D. melanogaster* (55). The protein families derived by this method have been used for many studies, including the controversial question of deep metazoan phylogeny [65,66,67,68].

One traditional view of metazoan phylogeny, based upon morphology, divides the Metazoa into three groups; Coelomata, Acoelomata, and Pseudocoelomata. Under this hypothesis humans and *D. melanogaster* are more closely related to each each other than either is to *C. elegans*. Small subunit ribosomal RNA phylogeny suggested a new arrangement: a clade Protostomia, linking *D. melanogaster* and *C. elegans* in Ecdysozoa and the Deuterostomia (including humans) [166,167]. This view has been adopted by the developmental biology community as it divides those species for which the mouth develops first (protostomes) from those in which it develops second (deuterostomes). The last three years have seen a series of publications which support one or other arrangement of the Metazoa (see 168 for a review of the current state of play). Proponents of both hypotheses agree that increased taxonomic sampling is vital if a robust topology is to be recovered. The additional sampling must cover species in already surveyed phyla, such as the nematodes, as well as including sufficient sequence data from poorly represented taxonomic groups. Including protein sequences from a diverse selection of nematodes will go some way to overcoming the phylogenetic bias caused by accelerated evolution within the phylum that has led to gene loss and rapid divergence. In this chapter I present an initial assignment of proteins from NemPep3 to the KOG system and highlight changes to the phylogenetic distribution of the generated clusters.

4.3 Methods

4.3.1 BLAST analyses

All-against-all BLASTP searches were performed on the combined nematode proteome, NemPep3. This dataset contains only those polypeptide sequences whose coding regions were identified by the BLAST or ESTScan components on prot4EST (Chapters Two and Three). Unless otherwise stated later, the E value cut off was e-5 and bit score threshold was 50. The number of reported alignments was set to 250 which was sufficient for all queries. The default values were used for all other BLAST parameters.

NemPep3 was compared to the protein database UniProt_minusNema, in which nematode sequences were removed from the UniProt database using in house Perl scripts. Parameters for these searches are described above.

4.3.2 Collectors curve

The proteome of *Caenorhabditis elegans* was screened for redundancy. To maintain consistency across the study any two sequences that shared significant sequence similarity (E value < e-5 and bit score > 50) were clustered.

The other proteomes were then sequentially added based upon the species' phylogenetic distance from *C. elegans* at order level (see Table 2.1 for species codes).

Rhabditoidea:	CBP
Strongyloidea:	ACP→ AYP → HCP → NAP → NBP → OOP → TDP
Diplogasteromorpha:	PPP
Panagrolaimomorpha:	PTP → SRP → SSP
Tylenchomorpha:	GPP → GRP → HGP → HSP → MAP → MCP → MHP → MIP → MJP → MPP → PEP → PVP → RSP
Cephalobomorpha:	ZPP
Ascaridomorpha:	ALP → ASP → TCP
Spiruromorpha:	BMP → DIP → LSP → OVP → WBP
Trichinellida:	TMP → TVP → TSP
Dorylaimida:	XIP

For example when the proteome of *Necator americanus* (NAP) was added, a protein with a significant hit to any of the previous proteomes (*C. elegans, C. briggsae, A. caninum* and *H. contortus*) was not counted. Any novel protein was added to the curve. Similarity to proteins outwith the phylum was detected through similarity to a member of the UniProt_minusNema database (see above).

4.3.3 KOG analysis

The KOG clusters (organisation and sequences) used were obtained from the NCBI ftp server: ftp://ftp.ncbi.nih.gov/pub/COG/KOG/

KOGs with the required phylogenetic distribution were identified and their sequences collected.

NemPep3 sequences were assigned to a KOG by symmetrical best BLAST hits with sequences from at least two species in a particular KOG.

4.3.4 Mapping orphan lineages on the phylogeny

Each node in the rooted phylogeny can be defined by the orders that are descended from it. All other species (including the UniProt_minusNema database) are considered as outgroups. The phylogenetic distribution of proteins with significant BLAST similarities for each nematode protein was examined. If a protein shared similarity with a sequence from a species from a different order, but descended from the node in question, without matching one from an outgroup species it was considered to be restricted to that node. For example, to be considered restricted to the Strongyloidea / Rhabditoidea node, proteins from the Strongyloidea must have similarity outside the order exclusively to a rhabditoidid species, and the converse is true.

4.4 Results and Discussion

4.4.1 Genes and proteins

It is important to define whether the analysis described below considers gene or protein sequences. Expressed sequence tags can be from any region of mRNA that were incorporated into the cDNA library. The mRNA structure comprises a 5' cap, 5' untranslated region (UTR), coding region, 3' UTR and poly-adenine tail, all of which could be found in an EST. In this sense, analysis on an EST (or clustered EST contig) considers the evolution of the gene it tags. One problem with such analysis is that it is unclear exactly what part of the gene is being studied. These components of mRNA are subjected to different evolutionary constraints [169]. The development of prot4EST to identify the coding region of an EST contig and produce the polypeptide sequence, has permitted these issues to be overcome. There has been no attempt to robustly extract ESTs' coding regions in many of the published studies I have mentioned. In this chapter, I refer to sequences derived from the coding region identification as 'proteins', and use the term 'gene' to describe those ESTs for which no attempt at coding region has been made.

4.4.2 Nematode proteinspace

To reveal the extent of nematode proteinspace, NemPep3 was compared to itself and to the UniProt database [86], edited by the removal of nematode proteins, through a series of BLAST searches (see Methods). The proportion of species-specific, or orphan, proteins in the partial proteomes, excluding *Zeldia punctata*, varied between 18-45%, and the proportion of a species' proteome with no non-nematode homologue was 35-62% (Table 4.3). The proportion of orphan proteins from the tylenchid *Z. punctata* is considerably lower than that of other partial proteomes. The transcriptome of *Z. punctata* was not sampled to the same depth as other species (only 390 ESTs) and a critical assessment of the quality of the libraries has not been carried out. These numbers were similar to those presented in the first pan-nematode study [51]. The only striking difference was the drop in the number of proteins that were unique to either *Brugia malayi* or *Onchocerca volvulus*. The majority of these 'missing' putative genes (~98%) were removed from the *B. malayi* or *O. volvulus* proteomes during the creation of NemPep3, because they appeared to lack a coding region (Chapter Three). That most of those removed did not share significant similarity with any reported sequence supports the assumption that the EST contigs are devoid of any coding region.

The proportion of orphan proteins from the two caenorhabditids was much smaller than that of other species (both ~9%). These proteomes are effectively complete and, given the relatedness of the two species, the likelihood of a protein from one species having a homologue in the other is considerably greater. The completion of genome sequences from a further three *Caenorhabditis* species [170], will probably result in further decline in the proportion of orphan proteins. Comparing the recently released proteome of *Brugia malayi* to both *C. elegans* and *C. briggsae* revealed 84 and 57 putative homologues, respectively, that were present in the the spirurid but not in the sister caenorhabditid.

By way of a contrast, for other closely related groups, such as the *Meloidogyne* species, a large fraction of the proteomes were species-specific. One would expect a relatively low proportion of proteins restricted to each species at this taxonomic level. This is seen in the study on the solanaceae family, where between ten and fifteen per cent of the partial transcriptomes (coding regions were not identified in the study) were species-specific [143]. However, for the *Meloidogyne*, up to 28% of the proteome is made from orphan proteins. Scholl and Bird have used ESTs from *Meloidogyne* species to reconstruct the phylogeny of the Tylenchomorpha [50]. Putative gene families were constructed (with the *C.* elegans proteome) using the COG system (symmetrical best BLAST hits), which resulted in 47 groups. This number of shared genes is relatively small given the number of sequences available, although there is nothing currently available with which to compare this finding. The increase in proposed novelty may be a reflection of the diversity within the genus; the *Meloidogyne*, as a group, has a broad plant host range, but individual species are more limited in host preference [171]. Despite this, the number of orphan proteins is likely to be an over-estimate; the stochastic sampling of the cDNA library and analysis of only part of the coding region inflates the level of novelty. This hypothesis will soon be testable; in addition to *B. malayi*, low coverage or complete genome sequences are to be made available for a number of parasitic nematode species: *Haemonchus contortus, Trichinella spiralis* and *Meloidogyne hapla.* Once completed, these genomes will act as models for their representative taxonomic orders, and I expect the proportion of orphan proteins to decline for the taxa in the orders that contain one of these models.

Table 4.3

Contribution by each species to the novelty in the nematode proteome.

Species	Clade	Total number of EST contigs	Polypeptides unique to the species		Polypeptides unique to phylum Nematoda	
			Number	Proportion	Number	Proportion
TMP	I	1,410	471	0.33	649	0.46
TSP	I	3,586	1,647	0.46	1,814	0.51
TVP	I	1,112	453	0.41	586	0.53
XIP	I	4,286	1,328	0.31	1,502	0.35
ALP	III	839	238	0.28	468	0.56
ASP	III	8,472	3,041	0.36	4,789	0.57
TCP	III	1,426	416	0.29	739	0.52
BMP	III	6,812	2,396	0.35	3,674	0.54
DIP	III	1,754	699	0.40	1,065	0.61
LSP	III	1,616	455	0.28	850	0.53
OVP	III	3,915	1,274	0.33	1,936	0.49
WBP	III	1,494	364	0.24	615	0.41
GPP	IV	2,392	738	0.31	1,137	0.48
GRP	IV	2,861	814	0.28	1,417	0.50
HGP	IV	9,064	3,070	0.34	4,608	0.51
HSP	IV	1,305	323	0.25	616	0.47
MAP	IV	2,358	446	0.19	1,194	0.51
MCP	IV	3,429	907	0.26	1,826	0.53
MHP	IV	6,419	1,792	0.28	3,372	0.53
MIP	IV	5,914	1,325	0.22	3,022	0.51
MJP	IV	3,359	952	0.28	2,020	0.60
MPP	IV	1,437	262	0.18	666	0.46
PEP	IV	415	81	0.20	166	0.40
PVP	IV	776	304	0.39	416	0.54
RSP	IV	525	185	0.35	265	0.50
ZPP	IV	205	14	0.07	36	0.18
SRP	IV	3,901	1,174	0.30	1,825	0.47
SSP	IV	3,621	811	0.22	1,409	0.39
PTP	IV	2,758	693	0.25	1,217	0.44
CAEBR	V	19,220	1,808	0.09	8,754	0.46
CAEEL	V	22,296	1,866	0.08	9,899	0.44

Continued overleaf...

Species	Clade	Total number of EST contigs	Polypeptides unique to the species		Polypeptides unique to phylum Nematoda	
			Number	Proportion	Number	Proportion
AYP	V	3,660	887	0.24	1,698	0.46
ACP	V	4,294	1,704	0.40	2,621	0.61
HCP	V	5,157	1,263	0.24	2,348	0.46
NAP	V	2,318	902	0.39	1,427	0.62
NBP	V	770	206	0.27	385	0.50
OOP	V	2,509	642	0.26	1,339	0.53
TDP	V	1,868	516	0.28	1,052	0.56
PPP	V	4,197	1,356	0.32	1,877	0.45

Table 4.3

Contribution by each species to the novelty in the nematode proteome.

NemPep3 was subjected to all-against-all BLASTP searches and against the UniProt_minusNema database (see Methods). The species are arranged into major taxonomic groups. The number of contigs of each species are those for whom a coding region is predicted by the BLAST or ESTScan components of prot4EST (see Chapter Three).

A polypeptide is considered unique to a the phylum if it shared no significant similarity to any proteins from the database UniProt_minusNema (see Methods). The proportion of novelty is with respect to the total number of contigs from that particular species.

4.4.4 Rate of Gene Discovery

The cumulative number of different protein groups, those with no significant sequence similarity to any other protein, was compared to the number of proteins added by the inclusion of each new nematode proteome (Figure 4.1). The starting point was *C. elegans,* with further species added according to their phylogenetic distance. The addition of new proteomes increased the number of protein groups. The linear increase seen (Figure 4.1) shows that the rate of discovery of novel proteins in the phylum Nematoda has not yet started to decline with the inclusion of new proteomes. It is not clear whether this is an accurate indication of nematode proteinspace, or a consequence of stochastic incomplete sampling. If the former were true I would expect to see local plateaus within certain taxonomic levels (for example order), but these are absent suggesting that sampling may have an affect. The number of nematode-specific protein groups is 44,000. As the number of EST generation projects increase, I expect homologues for many of these groups to be found in other species, particularly protostomes. However, a large proportion are likely to be genuinely restricted to the Nematoda, and these should form the basis for more in depth protein family analysis.

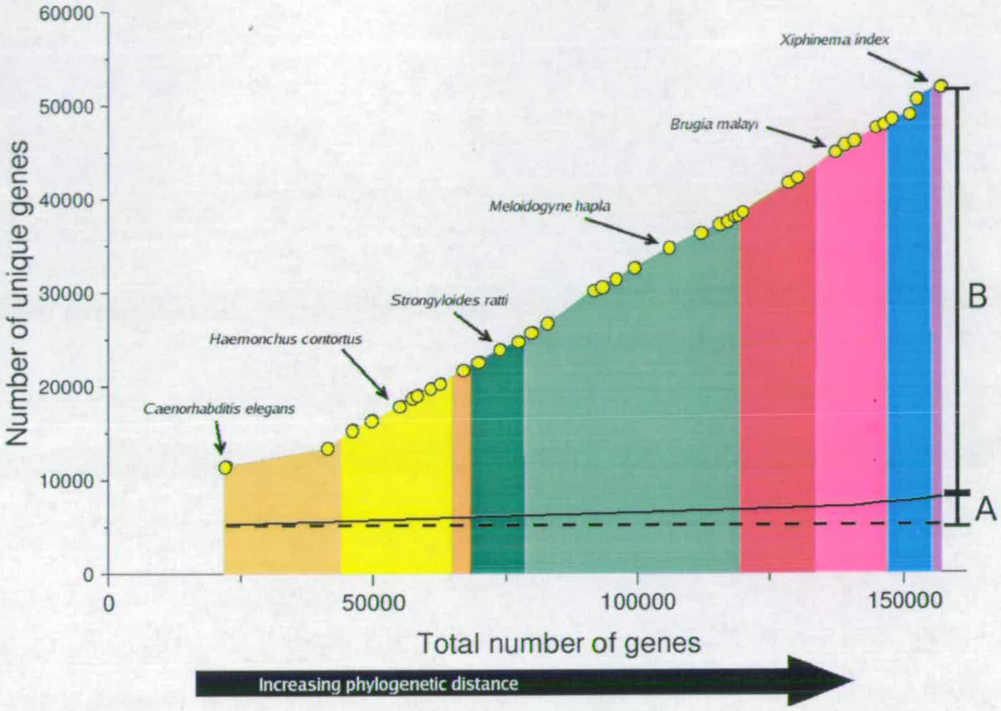


Figure 4.1

Gene discovery and proteinspace in the phylum Nematoda.

The cumulative number of different proteins (those with no significant similarity to any other gene) is compared to the total number of proteins. Starting with *Caenorhabditis elegans* subsequent nematode proteomes (yellow circles) are added in order of the species phylogenetic distance from *C. elegans*. The proteome of *C. elegans* was screened for redundancy, yielding a starting figure of 12,000 proteins

The black dashed line shows the number of distinct proteins from *C. elegans* that have matches in proteomes of non-nematode taxa. The black solid line indicates the cumulative number of nematode protein groups from each species that share similarity with sequences from non-nematode taxa.

Area 'A' represents the proportion of nematode protein types not found in the *C. elegans* proteome but are found in other non-nematodes. This represents possible lineage-specific gene loss in *C. elegans*. Area 'B' indicates those proteins which are specific to the Nematoda.

4.4.5 Origin of novelty in the nematode proteomes

A total of 41,564 proteins (~27%) shared similarity to a protein from species covering all four available clades. The majority of the these (39,740 proteins; 96%) also had significant similarity to a protein outside the Nematoda. Of the 71,016 proteins for which no putative homologue exists outside the phylum, 32,911 had significant similarity with a protein from another nematode species. Mapping these onto the robust nematode phylogeny showed the evolutionary origins of nematode-specific proteins (Figure 4.2). A similar analysis was performed in the earlier pan-nematode study; the data presented here includes 7 additional species and fully incorporates the *C. elegans* and *C. briggsae* proteomes. Within each order, the majority of proteins were orphans to individual species (Figure 4.2 and Table 4.3). As the number of species datasets within an order increased, so did the proportion of proteins shared between these species. Some of these orphans may be novel proteins that have been derived in a specific (terminal) lineage. However, I suggest that many of the putative orphans are not species-specific, but are either the product of pseudogenes that are still transcribed or are the product of genes that have diverged so quickly that the search algorithms cannot detect significant similarity.

The reduced proportion of species-specific proteins from the caenorhabditids is shown again. There were 7,306 proteins in their shared proteome that do not share similarity with another sequence outwith the Rhabditoidea, even from within the phylum. This number is likely to decline as additional nematode genomes are completed. More striking is the abundance of nematode proteins that do not share similarity with a sequence from either *Caenorhabditis* species. Even relatively closely related taxa have many new proteins, for example *Pristionchus pacificus* has 1,356 species-specific proteins.

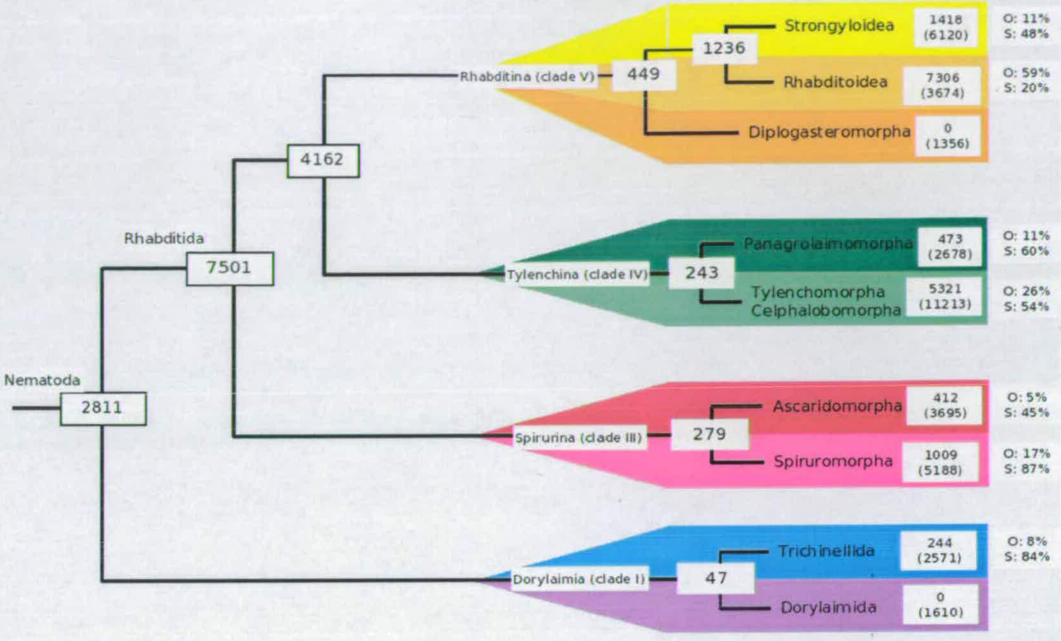


Figure 4.2

Evolutionary origins of lineage-specific proteins in the phylum Nematoda.

The positions inferred by all-against-all comparisons were mapped across the robust SSU rRNA phylogeny. Each node provides the number of proteins whose putative homologues are restricted at that node. The most terminal nodes provide additional information. The upper number is the number of proteins restricted to that level (order) and shared by at least two species. 'O' shows the percentage of all NemPep3 proteins that are restricted to that order. The lower number, in parentheses, shows the number of proteins shown to be species-specific. Similarly 'S' highlights the percentage of proteins in in that order that are species-specific.

Most of the events of gene origin appear to have occurred early in nematode evolution, with for example, 7,501 mapping to an origin at the base of the Rhabditida and 2,811 proteins with an inferred origin in the last common ancestor of all Nematoda. Of those mapping to the base of the Nematoda, 1,824 (65%) are members of related groups represented in each of the four available nematode clades. These proteins may have roles specific to the nematode body plan and life cycle, including feeding strategy and, for parasitic species, immune response evasion. However, given that certain feeding strategies have seemingly evolved multiple times [43], *de novo* generation or accelerated evolution of proteins underpinning these mechanisms are more likely to be restricted to more recent lineages.

The addition of seven species has, as expected, pulled the point of origin for many proteins towards the base of the tree. The principal cause is the inclusion of the dorylaim *Xiphinema index*, which, with a relatively large dataset, contributes as much to the pan-phylum novelty as do the three trichinellids, its sister taxon. The extra tylenchomorphs (*Meloidogyne paranesis, Radophulus similis*, *Heterodera schachtii* and *Pratylenchus vulnus*) are all closely related to a species already present in the analysis. However there is no reduction in the proportion of novel proteins in the tylenchids with ~1,100 proteins per species. Despite the inclusion of two additional spirurids (*Litomosoides sigmondontis* and *Wuchereria bancrofti*), the number of unique proteins has decreased, a consequence of removing those EST contigs without detectable coding regions (see 3.4.6).

Many of these points of origin may not reflect the creation of completely novel proteins. Instead they may signify an event that has forced a protein along a different evolutionary trajectory from that of its orthologous partners. These events could result from, or be consequent to, gene duplication or speciation events. Changes in functional constraint can result, over time, in related sequences not sharing detectable similarity. Alternatively, the novel protein may be the result of domain rearrangements between two loci in the genome. Domain rearrangements are thought to be important for the evolution and adaptation of new functions [172,173]. Domain rearrangements comprise domain duplication (A → AA), swaps (AB → BA), domain deletion (ABC →AC), circular permutations (ABCD → CDAB) and addition (AB → ABC) [174]. Any of these events, especially after a gene duplication event could produce a lineage of seemingly novel proteins.

4.4.6 Discussion on the expansion of Nematode proteinspace

The collectors curve presented in Figure 4.1 shows that the rate of gene discovery in nematode species has not begun to slow, and as new species are added a relatively constant proportion of previously unseen proteins are discovered. This rate is influenced by two, potentially contrasting, factors. As more nematode species are surveyed through EST projects, the number of species-specific proteins is likely to increase further. Projects under way for previously unsampled nematode species include: *Dityocaulus* (Strongylomorpha), *Steinernema* and *Rhabditophanes* (Panagrolaimomorpha), and the chromadorid *Stilbonema* (Mark Blaxter pers. comm.). Although these species, with the exception of *Stilbonema*, are closely related to taxonomic groups which already have representative sequencing projects, the evidence presented here does not suggest this will slow the rate of new gene discovery.

The second factor is deeper sampling of the transcriptomes of species with sequence already available. This is possible in two ways. The first is generation of additional ESTs, which is happening for *Ascaris suum* (Ascaridomorpha) and the hookworms *Ancylostoma* (*caninum* and *ceylanicum*) and *Necator americanus*. Secondly, there will soon be genome sequences available for at least eleven species of nematode (Table 4.4). The effect this increase in sequence on new gene discovery will depend upon the status of closely-related species. The additional caenorhabditid proteomes are likely to reveal a relatively low number of orphan proteins, although they may further highlight lineage-specific gene loss or protein family expansions. Performing the analyses presented here with the inclusion of the complete proteomes of parasitic species will reduce the numbers of orphan proteins found in EST datasets. However, the forthcoming proteomes are from a diverse range of species, so I expect a bounty of newly discovered proteins with no detectable similarity.

Species	Trophic ecology	Status
Caenorhabditis elegans	Bacteriovore	Full sequence complete
Caenorhabditis briggsae	Bacteriovore	Draft sequence complete
Caenorhabditis remanei	Bacteriovore	Draft sequence complete
Caenorhabditis japonica	Bacteriovore	Draft planned
Caenorhabditis sp.	Bacteriovore	Draft planned
Pristionchus pacificus	Bacteriovore	Draft in progress
Brugia malayi	Vertebrate parasite	Draft assembly in progress
Haemonchus contortus	Vertebrate parasite	Draft in progress
Meloidogyne hapla	Plant parasite	Pooled BAC sequence planned
Trichinella spiralis	Vertebrate parasite	Draft in progress
Heterorhabditis bacteriophora	Insect pathogen	Draft planned

Table 4.4

Status of nematode genome projects (August 2005).

4.4.7 Protein loss in *C. elegans*

A total of 7,953 protein groups were shared between the nematodes and other phyla. The *C. elegans* proteome contributes 5,048 of these, suggesting that the remaining 2,905 have been lost in the *C. elegans* lineage (Figure 4.1). Gene loss in *C. elegans* is well reported [175,176,177,178,179,180] and may represent a simplification of the *Caenorhabditis* genome. There is evidence that some of this "streamlining" occurred after the divergence of *Caenorhabditis* from other nematodes; for example the orthologues of Antennapedia and Hox3 are absent in *C. elegans* but present in *B. malayi* [176].

The *C. elegans-C. briggsae* comparison [36] showed that there were a large number of proteins in each species that could not be matched to a gene in the other taxon. Updating this analysis with the current caenorhabditid proteomes revealed the number of orphan proteins to be 2,041 for *C. elegans* and 2,117 for *C. briggsae*. The inclusion of nematode partial proteomes and an updated UniProt database reduces the number of orphan proteins in the caenorhabditids to 1,846 and 1,961, respectively (Table 4.5). Of proteins that were no longer specific to a caenorhabditid (195 and 156), 40% share similarity only with sequences from outside the phylum. There were putative homologues from metazoan proteomes for all these proteins, suggesting that their absence from other nematodes is a consequence of the incomplete nature of the sequencing survey, rather than the less parsimonious explanation of protein loss in multiple independent lineages. There is a steady increase in the number of nematode proteins with putative homologues in UniProt, with the inclusion of each nematode proteome. There are two equally valid explanations for this observation:

1. There has been protein family loss in each nematode lineage.
2. A protein has undergone accelerated evolution; as a result sequence similarity to its orthologues is undetectable.

To confirm and subsequently characterise protein loss events in any detail requires complete proteomes; this would make the identification of orthologous relationships easier.

Orphan after comparison with:	***Caenorhabditis elegans***	***Caenorhabditis briggsae***
Sister caenorhabditid	2,041	2,117
Strongylids + diplogasteromorpha	1,963	2,055
Other nematodes	1,924	2,023
Metozoan proteomes	1,846	1,961

Table 4.5

Number of orphan proteins in two caenorhabditids.

Comparisons of the caenorhabditid proteomes against the partial proteomes of other nematodes resulted in a reduction in the number of proteins previously shown to be unique to each species. Serial BLAST searches were performed. Reported is the number of proteins without a significant hit to the collection of proteomes under consideration or the previous searches. The comparisons were with the proteomes of : (1) the sister caenorhabditid; (2) closely-related Rhabditoidea *(A. caninum, A. ceylanicum, H. contortus, N. americanus, N. brasiliensis. O. ostertagi, P. pacificus, T. circumcincta)*; (3) remaining nematode species; (4) the Metazoa – collected from UniProt.

4.4.9 KOG analysis of NemPep3

Protein families have been used to explore metazoan phylogeny, in particular the relationships between human, *Drosophila melanogaster* and *C. elegans*. There are several studies that consider molecular data that provide conflicting conclusions [166,181,65]. Wolf and colleagues selected candidate proteins from families derived from the KOG classification [69] to explore this question. KOGs (clusters of euKaryote Orthologous Genes) are built from seven eukaryote proteomes, using symmetrical BLAST hits to derive orthologous relationships. Patterns of presence and absence of proteins from KOGs formed part of the evidence for the Coelomata (wherein *Homo sapiens* and *D. melanogaster* are more closely related than either is to *C. elegans*). A character matrix for all KOGs was generated and the Dollo parsimony method applied. The method assumes irreversibility of character loss, so once a KOG is lost from a species' repertoire it cannot be regained. Examining the species contributions to each KOG showed that *H. sapiens*, *D. melanogaster* and *C. elegans* shared membership of 3,951 KOGs. The number shared between only two of these species differed according to the combination (table 4.6). The number of protein families shared by *D. melanogaster* and *H. sapiens* to the exclusion of *C. elegans* was the larger than other combinations. A major concern with the original analysis was that *C. elegans* proteins may be unassigned to a KOG due to a higher rate of evolution, a trait that has been observed in *C. elegans* [70,71,72,73], or that extensive gene loss in the *C. elegans* lineage would misrepresent the true phylogenetic relationship [74]. The availability of NemPep3 allows a larger protein complement from the Nematoda to be considered. Through the use of symmetrical BLAST searches, non-caenorhabditid proteins were assigned to 279 KOGs that did not contain a protein from *C. elegans* (Table 4.6). The most striking change was the reduction in KOGs that contained *D. melanogaster* and *H. sapiens* to the exclusion of *C. elegans*. By adding nematode proteins the phylogenetic distribution of KOGs with only two Metazoa shifted, leaving the most frequent pairing as *C. elegans* and *H. sapiens*.

Shortcomings with the symmetrical BLAST detection

It would be foolhardy to repeat the Dollo parsimony analysis upon these new clusters. Both *H. sapiens* and *D. melanogaster* will show lineage-specific gene loss [74,182], therefore an honest study should include proteomes from additional vertebrates and arthropods. As the newly incorporated nematode datasets (and any subsequent vertebrate or arthropod) are incomplete, there will be some doubt over orthologous relationships determined by symmetrical BLAST analysis. It is assumed that symmetrical best BLAST hits are most likely to be found between orthologues [161], but in the partial dataset the true orthologue may not have been sequenced. These problems need to be overcome in any detailed study of phylogenetic relatedness within protein families. Careful consideration of the data allows certain lineage-specific gene losses to be uncovered. For example, as sequence data exists for "all" *C. elegans* proteins, if a protein from a metazoan species shares similarity with sequence from a parasitic species but not *C. elegans*, the simplest – most parsimonious – explanation is that the protein has been lost at some point in the lineage between the last common ancestor of the two nematodes and *C. elegans*. The data presented here highlights how gene loss in one nematode lineage (*C. elegans*) can affect an analysis considering the phylogenetic position of the entire phylum.

This view is supported by an analysis of ESTs from the cnidarian *Acropora millepora* [74]. The gene clusters were compared to the transcriptomes of human, *D. melanogaster* and *C. elegans*. The results showed that 12% of clusters shared similarity with human but not the model invertebrates. Contrast this with only 1% of clusters matching *D. melanogaster* or *C. elegans* and not human. All three model species provided putative homologues for 87% of *A. millepora* gene clusters. However for 41% of these the coral gene had significantly higher levels of similarity to human, while only 8% were more similar to a model invertebrate.

To prevent the reconstruction of erroneous relationships, more species must be considered. Without consideration of proteomes, even partial, from other species, the trees provide no information regarding missing taxonomic groups [168]. Thus, current genome-scale analyses do not include any members of the Lophotrochozoa, which includes Mollusca and Annelida. More importantly, if a species' proteome or genome is rapidly evolving, algorithms for tree reconstruction can be misled by long branch attraction. This phenomenon describes the problem that arises when the probability that closely-related taxa share character states due to common ancestry is surpassed by the probability that more distant relatives share those states due to convergent changes (homoplasies) [183,184,38]. Studies that only consider complete genomes must currently use a distantly-related organism, a yeast, as the outgroup. The rapid evolution of *C. elegans* means that the number of protein families that have been lost or changed beyond detection is increased when compared to more slowly evolving organisms. If a protein family is also missing in yeast then the change is misconstrued as a shared derived change (synapomorphy) pulling *C. elegans* to a basal position in the metazoan clade. To break these long branches more sequence from additional taxa are required [185,186]. By using manually assembled gene families from 35 species, representing 12 animal phyla, Philippe and coworkers found convincingly for a clade containing Ecdysozoa and Lophotrocozoa to the exclusion of the Deutrostomia and thus rejecting Coelomata [187]. There were, however, a number of unresolved problems with the analysis (reviewed in Jones and Blaxter [168]], which must be overcome. Despite these problems, the study has shown that greater sampling of species and molecular sequence diversity is required to answer robustly some of the major phylogenetic questions.

	Dme+Hsa+ Nem+	Dme+ Hsa+ Nem-	Dme+ Nem+ Hsa-	Nem+ Hsa+ Dme-
NemPep3 –	3,951	331	55	206
NemPep3 +	4,171	111	59	261

Table 4.6

Reconstructing KOGs with NemPep3.

The inclusion of NemPep3 (+) increases the number of protein families (KOGs) with a nematode sequence (Nem+). Symmetrical BLAST searches were performed with the original KOG datasets and NemPep3.

4.5 Further Work

The identification of proteins restricted to certain, particularly parasitic, lineages is an exciting development. However this work represents only the first step in identifying the biological relevance of the data. A primary focus must be to determine possible functions of these protein groups. Discussed in the next chapter is the use of protein family and domain databases to provide possible functional clues. Unfortunately the phylogenetic bias present in primary sequence databases that results in a large number of proteins without putative homologues in other phyla is also present in the domain databases. Therefore it is unlikely that many of the nematode-specific proteins will be in receipt of such annotation. One avenue is to identify distant homology undetected by the BLAST algorithm. The current trend is to move away from sequence-sequence comparisons and use more sensitive profile or profile-based methods [188,189]. These methods are more powerful because they implicitly characterise both the pattern of conserved residues and then distribution of variation with a proposed protein family. Such an approach requires organisation of proteins into related families which is discussed in the next chapter.

Detecting homology in the 'twilight zone' of weak sequence similarity [189], will only provide a hint at possible function. Experimental data are required to assign more information about a protein's function. There are a number of useful procedures including localisation of gene expression (e.g., promoter- GFP report transgenesis), co-expression studies with microarrays, protein-protein interactions (e.g., yeast two-hybrid) and gene knock down studies using RNA interference (RNAi). The advantage with the nematode datasets are that there is a vast amount of experimental data available for *C. elegans*. If similarity were shared between a parasitic protein of interest and a protein from *C. elegans* a link could be drawn. However this would conceal any adaption in the protein's biology. Its important to consider experimental data from other, more closely related species.

Experimental models for parasitic nematodes include *Brugia malayi, Litomosoides sigmondontis, Heligmosomoides polygyrus, Strongyloides ratti, Haemonchus contortus* and *Globodera rostochiensis*. Unfortunately a central repository for parasitic nematode data does not currently exist.

4.6 Conclusions

The work performed in this chapter highlights the diversity of the combined nematode proteome. Use of the robust nematode phylogeny has permitted the identification of lineage-specific novelty with the possibility of protein family expansion and subsequent neofunctionalisation. There exist a large number of proteins that have no putative homologue outside the Nematoda. These proteins have probably arisen from domain shuffling events or gene duplication followed by divergence. Both of these events can produce proteins whose function is biologically distinct, whether mechanistically related or not, from the ancestral gene. These proteins offer promising targets for anthelmintic drugs and are worthy of further study [59]. From a evolutionary perspective the coverage of sampled species across the phylum offers an excellent opportunity to study a number of features in restricted lineages, including structural similarity, relative gene expression, rates of substitution and codon usage.

As the EST projects continue to generate more survey sequence and the complete genomes are assembled and annotated for representative species, the number of lineage-specific proteins will increase, and their taxonomic distribution will reveal evolutionary markers for the species adaptation to new trophic niches.

Chapter Five – Nematode protein domains, NemDom

5.1 Abstract

The decoration of polypeptide sequences with protein domains is one of the most popular forms of annotation. There are a number of databases, or libraries, of protein domains which harbour a wealth of information, from proposed function to species distribution, about each domain. Assigning domains to the proteins of NemPep3 is an excellent way to identify sequences that are fundamental for nematode survival. Finding domains on EST derived proteins presents a major problem; the incomplete nature of ESTs means that only a small section of the domain may be present, which would normally go undetected using standard domain models. To ensure maximum, and robust coverage of domain annotation I have explored the affect of using both global and local alignments between the domain model and protein sequences, as well as different scoring threshold. A combinatorial approach was adopted to assign Pfam-A domains to NemPep3, creating the NemDom3 collection. Species distribution of previously characterised metazoan-wide and nematode-restricted domains was investigated, identifying domains that have been lost in the caenorhabditid-lineage but found through the rest of the phylum. There were also domains that were still only found in *Caenorhabditis elegans*, suggesting that they have been acquired, by some mechanism, in that lineage or that the domain models are too restrictive in their predictive power. I also search for domains that may be present in the domain repertoire of certain nematodes as a consequence of convergent evolution or horizontal gene transfer (HGT).

5.2 Introduction

5.2.1 Protein domains

The term "protein domain" is used to describe the structural, functional or evolutionary units of proteins. These definitions are, in essence, separate but overlap for many characterised domains. The assignment of a protein's function is often dependent upon the division of its structure or sequence into domains. Whatever the variation of definition, it is notable that a relatively small number of domains are used in a large number of proteins [162]. The analyses performed and described in this chapter are predominantly based upon sequence comparison, and thus the domains used are considered from that perspective and described as regions of proteins sharing sequence similarity that are often present in different molecular contexts [190].

Delineating the proteins from the major sequence repositories into their constituent domains has led to the creation of domain libraries – collections of protein domains with functional annotation and meta-data attached. These domain libraries have become important tools for sequence analysis; in fact it has become standard procedure when reporting the completion of a genome to infer general annotation based on high-throughput domain prediction. Most domain libraries are constructed in a similar fashion; for each characterised domain the protein sequences are aligned, and the alignment is described by a number of methods, including position-specific scoring matrices (PSSM) [13,191], hidden Markov models (HMM) [94,93] and profiles [192]. These techniques are more sensitive than single-sequence comparisons, as they identify both more and less highly-conserved positions within the protein, thus summarising the evolutionary history of the domain family [188,193].

5.2.2 Protein families

Here it is appropriate to clarify the differences between domain and protein families, and to justify my decision to concentrate exclusively on domain families. It is well known that members of the same protein family share similar, if not identical biochemical functions [194]. A protein family can be defined as a group of polypeptides that are demonstrably related to each other [195]. The metric most widely used to cluster these families has been sequence similarity [196,197,135,165,69,198,118,199]. A protein family differs from a domain family in that it contains the full-length polypeptide sequences rather than conserved fragments from within the sequences. Similarity is usually detected using the BLAST algorithm [13], primarily because the heuristic search strategy employed is computationally very fast. There are two approaches that have become commonplace in molecular biology research. The COG system uses symmetrical BLAST hits to delineate relationships and is available through the NCBI [135]. TRIBE-MCL uses Markov flow clustering to group similar sequences (Box 5.1) [118,200] and is the algorithm of choice for a number of genome projects, including *Caenorhabditis elegans* (Daniel Lawson pers. comm] and *Plasmodium falciparum* [201].

One problem that clustering methods face is that many proteins consist of multiple independently evolving domains [172,194]. Using BLAST,. which detects local regions of similarity, can result in links forged between unrelated proteins [202]. This applies not only to formally classified protein domains, but to any shared motif of sufficient similarity and size to be considered significant. The COG system, the authors state, overcomes this problem by manual inspection of multi-domain proteins. However such an approach is labour-intensive and not transferable to the majority of research groups. The TRIBE-MCL program 'does not require any explicit knowledge of protein domains to detect protein families', but clusters on the observed relationships through the entire similarity graph [118]. However, the performance of the Markov flow clustering algorithm is dependent upon the inflation parameter, whose value should vary to assemble different protein families correctly

[203].

A recent investigation has shown that both these methods fail to correctly assemble the eukaryote hemoglobin protein family [Wasmuth, Elliot, Schmid and Blaxter *in prep.*). In the eukaryote COG database (KOG), the initial symmetrical BLAST searches failed to assemble the family, but over 30 *C. elegans* proteins were subsequently added based on manual assessment of PSI-BLAST searches. Many of these proteins do not contain the necessary number of α-helices or invariant residues characteristic of globins. The TRIBES database [204] separated related globins into many families, some containing a single member. The similarity statistic used to decorate the edges of the graph to be clustered is the E value, and is transformed ($-\log_{10}$(Evalue)) for the MCL algorithm. While this is acceptable for very large databases such as UniProt, it is probably inefficient when clustering smaller datasets. This was observed in an attempt to divide a collection of chelicerate mitochondrial proteins; the unrelated proteins were aggregated into two large groups containing non-homologous proteins, and many single-member families (Jones and Wasmuth unpublished). It is likely that using a similarity statistic independent of the size of the database would yield more faithful families, but this has yet to be assessed.

Given the uncertainty over the robust clustering of full-length protein sequences I considered it more expedient to focus on an investigation of the protein domain complement of the nematode proteomes.

5.2.3 Domain Databases and Pfam

A number of similar domain databases exist, each with their own specialties (Table 5.1). The TIGRFAMs [164] and SMART [99] databases are libraries of HMMs, which focus on prokaryote and eukaryote domain families respectively. Domains delineated on structural similarity are contained within SCOP [205], which focuses on evolutionary classification, and CATH [206], constructed through manual and automatic methods. Searching SCOP and CATH is usually done with BLAST, although a library of HMMs is available for SCOP entires [207]. The limitation of these two databases is that they only contain known three dimensional structures, and therefore reflect the taxonomic bias of currently available structures; that is mammalian and bacterial. There are also Internet accessible resources which combine some of these libraries to permit simultaneous searching. These include InterPro [76] and CDD [208]. However at time of writing, neither of these meta-servers provided stable stand-alone analysis tools, for annotation of a dataset as large as NemPep.

Of the resources. available, the Pfam database [77] is probably the most comprehensive domain database currently available. There are two divisions of Pfam, Pfam-A, which includes some manual curation and Pfam-B, an automatic classification. Each domain family for Pfam-A is constructed by the manual creation of a seed alignment of UniProt sequences considered representative for a domain. Care is taken in this step to identify potential sequence or alignment errors. The alignment is then converted into a profile-HMM using the HMMer software package [209] which is used to search the sequence database for additional members. Pfam-B is derived from the ProDom database [210] and includes those sequences not assigned to a Pfam-A family. The current release of Pfam-A (version 17 – June 2005) contains 6190 domain families, which match 75% of sequences in UniProt [77]. The addition of Pfam-B increases this coverage to 82%. The online version (http://www.sanger.ac.uk/Pfam) provides a number of analysis tools, which include searching by viewing taxonomic distribution; examining the evolution of domain combinations with NIFAS [202] and cross-referenced structural information provided by the SCOP [211] and CATH [206] databases. All Pfam-A and Pfam-B models are available for download, and can be used for sequence searches with the HMMer software. This is gives Pfam a substantial advantage over other methods, and was one of the reasons it was chosen for this analysis.

An assumption in annotating protein sequences with Pfam models is that domains do not overlap. However nested domains are permitted – one domain that is interrupted by the insertion of another. An example is the IMPDH domain (PF00478), which in many instances is continuous, but in a few cases is broken by the insertion of one CBS domain (PF00571). All Pfam-A families contain curated functional information. This includes two classes of functionally uncharacterised domains, known as Domains of Unknown Function (DUFs) and Uncharacterised Protein Families (UPFs). DUFs are the families created within the Pfam project while UPFs are those generated by UniProt and subsequently incorporated into Pfam. Release 10.0 of Pfam-A contains 1,004 DUF and UPF families, representing 16% of Pfam-A. Bateman *et al.* observed a tendency for completely undescribed families to be small and taxonomically restricted [77]. The wealth of information present, and availability of integrated analysis tools, makes Pfam the best resource for domain annotation. I have used the Pfam-A domain library to annotate the sequences of NemPep3, and thus created the NemDom3.0 resource that is available through the NEMBASE server (http://www.nematodes.org).

Database	URL	Description	Reference
Pfam	www.sanger.ac.uk/Software/Pfam	Hidden Markov models covering many common protein domains and families – manual curation	77
SCOP	scop.mrc-lmb.cam.ac.uk/scop/	Structural classification of protein domains. Includes assumptions of evolutionary relationships	205
CATH	cathwww.biochem.ucl.ac.uk	Hierarchical domain classification of protein structures	206
SMART	smart.embl-heidelberg.de/	Sequence-based classification: eukaryote focus	99
TIGRFam	www.tigr.org/TIGRFAMs	Sequence-based classification : prokaryote focus	164
ProDom	protein.toulouse.inra.fr/prodom/current/html/home.php	Automatically generated from the SWISS-PROT and TrEMBL sequence databases.	210
SuperFamily	http://supfam.mrc-lmb.cam.ac.uk/SUPERFAMILY/	HMMs of all protein sequences with a PDB entry – classified at SCOP superfamily level	207
Interpro	www.ebi.ac.uk/interpro/	Meta-site integrating several methods	76
CDD	http://web.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml	Meta-site integrating several methods	208

Table 5.1

Protein domain resources available.

5.2.4 Domain models

Domain v model

When discussing Pfam annotation in this chapter it is important that I distinguish between the terms 'domain' and 'model'. A domain I have defined already as a functional or evolutionary unit shared by a number of proteins; it is the actual amino acid sequences. The model is the probabilistic characterisation of a domain, so in Pfam it is the profile-HMM built for a given protein alignment.

Local v global models and their scoring thresholds

Every domain in Pfam-A has two models built from the same multiple sequence alignment, the global model (ls) and local model (fs). When searches are performed with global models a potential match must start with the first match state of the model and finish with the last. For the profile-HMM architecture implemented in by HMMer (Plan 7), this is from the first column of the alignment to the final column. Matches to a local model can, in principle at least, start at any point in the model and terminate anywhere downstream. To score the matches (E value) the HMMer program uses Extreme Value Distribution (EVD), to which the scores from local models fit well. However global models are known not to produce such well fitted scores [212]. HMMer implements an approximation of the EVD fit, which empirically at least tends to be accurate at the critical region.

When a Pfam model is built three bit-score thresholds are assigned which describe the parameters used in its creation and hence how to optimise the search for that particular domain (taken from Eddy 2003):

Gathering cut-offs (GA) : these scores are the primary ones in construction of Pfam models. Matches to the model that satisfy these cut offs are included in the full domain alignment.

Trusted cut-offs (TC) : the scores of the lowest-scoring hit that were curated as a member of the particular Pfam family.

Noise cut-offs (NC) : the scores of the high-scoring hits that were considered not part of the Pfam family.

In addition an E value cut-off can be used and so the bit-score is not directly queried, although the E value is intrinsically linked to the bit score calculated for the model aligned to the query sequence.

5.2.5 Difficulties with assigning domains to EST sequences

Decorating predicted gene products with Pfam-A domains is now standard procedure on the completion of an organism's genome. The choice of scoring cut-offs is often not reported, but some simple analyses suggests that the gathering cut-off or E value threshold are most popular. It is almost certain that the annotation is performed with global models from the Pfam-A collection as they are assumed to be complete functional units. The evolutionary constraints on the amino acids across the entire domain are often responsible for a domain's identification in several proteins, therefore protein domain searches should use global models. This approach is the one implemented in nearly all annotation efforts of newly sequenced genomes. However the annotation of EST-derived proteins presents a problem. ESTs are between 200-900 nucleotides in length, therefore cover only part of the messenger RNA (mRNA) for a given gene. While sequence clustering may increase the coverage, a large proportion of ESTs remain singletons and there are no guarantees that a cluster will completely cover the mRNA. It is therefore a possibility that a protein domain that is present on the mature mRNA will be incomplete or missing from the EST-derived protein sequence. If the EST could be extended then the domain would be easily found with current technologies (Figure 5.1). There is nothing that can be done about absent domains; however, partial domains could be identified if 'local' models were carefully used. A further problem is the quality of the sequence being annotated. Despite EST clustering and robust coding region prediction, it is likely that some of the nucleotides have been mis-assigned or are ambiguous (N). If such changes are non-synonymous (at the amino acid level) it is possible that the correct model may not score sufficiently to be assigned to that sequence, which is particularly likely if the altered amino acid is invariant in the model's alignment.

The problem of partial domains due to a prematurely terminated EST could be overcome using local domain models. As only part of the local model needs to align to the query sequence to be considered as a possible match (given a sufficiently high bit score), the model introduces the possibility of finding partial domains. However, it is also possible that by accepting partial matches the number of incorrectly assigned partial domains will increase. This is more of a concern in protein regions which do not have a significant match to a global Pfam-A model.

To ensure that NemPep3 is robustly decorated with Pfam-A annotation, I have benchmarked the use of global and local models for domain identification using the *C. elegans* EST and WormPep resources. The different classes of score thresholds are examined to ensure that as many true domains are returned while reducing the number of false positive identifications which may lead to erroneous theories of parasitic nematode biology.

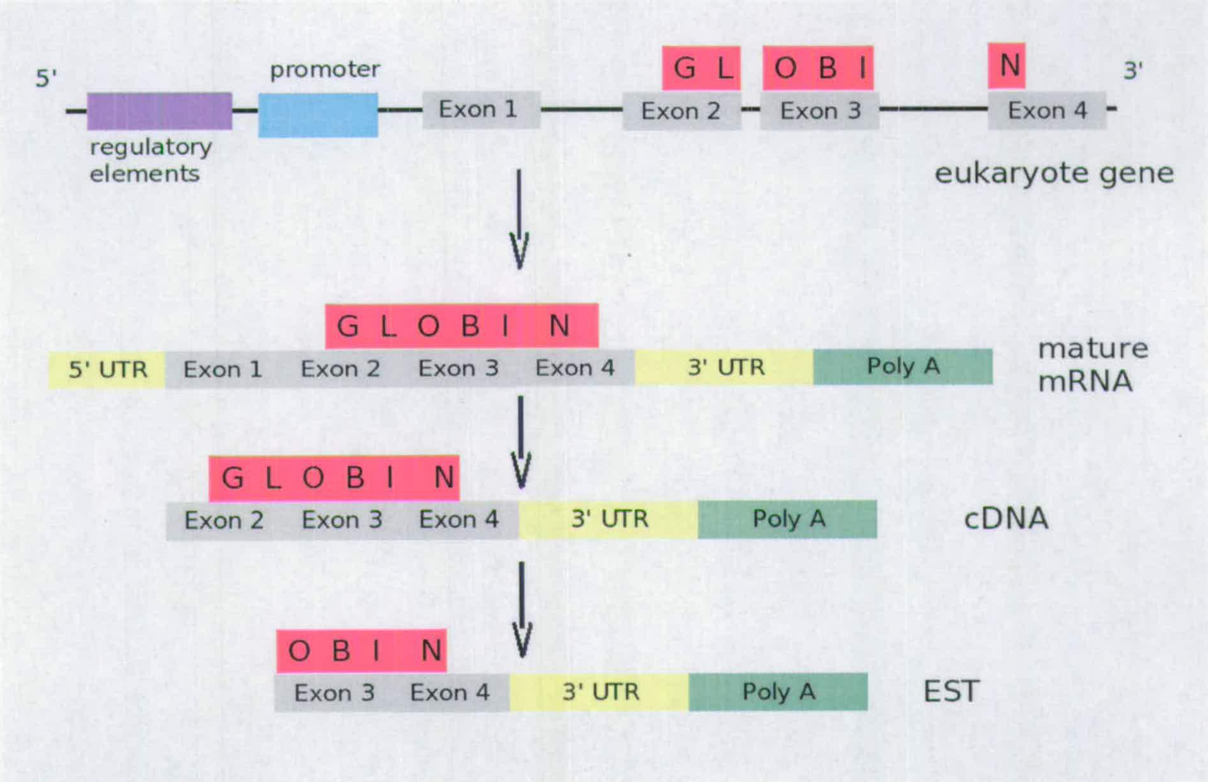


Figure 5.1

The generation of partial domains in an EST dataset.

The partial nature of ESTs, an inherent feature of the experimental design, implies that a protein domain may be incomplete. In this cartoon, the GLOBIN domain (red), spans exons two, three and four. Only the last two exons are present in the sequenced EST, therefore excluding the 5' region of the GLOBIN domain. Standard annotation strategy uses domain models that must align the entire model to the query sequence (global). This approach would probably fail to identify the GLOBIN domain.

5.2.6 Nematode protein domains

There are 2,439 distinct Pfam-A protein domains found in proteins from a nematode species, of which 2,428 are present in the proteome of *Caenorhabditis elegans*. The eleven domains not found in *C. elegans* are, for the most part, well characterised and are the results of specific, targeted studies. Of these eleven Pfam-A domains not found in *C. elegans* half are restricted to the plant parasitic Tylenchomorpha. These domains are either responsible for either the breakdown of the cell wall components cellulose and pectin [213], or interfere with the plant's Shikimate pathway which produces aromatic amino acids and derivative compounds that are involved in growth and defence mechanisms [214,215]. The origin of some of these domains is unclear. There is some preliminary evidence that the chorismate mutase protein (PF01817), a cellulase, found in *Meloidogyne javanica* is the result of a horizontal gene transfer (HGT) from bacteria [214]. However similar assumptions have been made for the glycosyl hydrolases Family 9 found in *Globodera rostochiensis* and *Heterodera glycines* [216], and have been recently refuted by Davison and Blaxter, who showed that this family of cellulases has an ancient origin in the metazoan lineage [217]. The 'Myogenic Basic' domain (PF01586) found in *Trichinella spiralis* is contained within a myogenic transcription factor that may involved in the invasion of the skeletal muscle [218].

The absence of these eleven domain families from *C. elegans* underlines the frequency of lineage-specific sequence loss, of both gene and protein domains. This is further highlighted by the existence of 84 Pfam-A domains which are found only in nematodes, 83 of which are only found in *C. elegans*. With genome sequencing projects producing an exuberance of data, methods were being developed to analyse all the proteins encoded in a genome. One such method was Pfam, whose breakthrough coincided with the completion of the *C. elegans* genome [35]. This enabled Sonnhammer and Durbin to take a closer look at the nematode's proteome with systematic functional classification, clustering of protein domains and comparison to other organisms [75]. The regions of sequence that did not match a Pfam-A domain were clustered, generating 1,516 clusters of between two and 58 members, along with 8,602 unclustered segments. Using consensus sequences to search proteins from other (non-nematode) species, ten domain families were shown to be nematode-restricted. The completion of the *C. elegans* and other metazoan genomes [35,219,139,220] has led to some of those domains once classified as nematode-specific being reclassified as widespread throughout the Metazoa, for example Copine (PF07002), a Ca(2+)-dependent phospholipid-binding protein involved in membrane trafficking. There has also been an increase in the number of protein domains found (so far) only in nematodes. Of the 84 Pfam-A domains found exclusively in nematodes only seven are found in non-caenorhabditid species, with their expanded membership usually a consequence of studies targeted at specific genes (table 5.2). For example, the 'Chromadorea ALT proteins' (PF05535) *alt-1* and *alt-2* are restricted to the Spiruromorpha, and undergo elevated expression while the organism resides in the mosquito vector and are thus possible candidates for vaccine targets [221].

Pfam Description	Pfam Accession	Clade Distribution	Species Distribution
Nematode cuticle collagen N-terminal domain	PF01484	III, VI, V	BMC, *BP*, GPC, MIC, MJC, ASC, TDC, HCC, CAEEL, CBP
DUF148	PF02520	III, IV, V	MIC, ASC, TDC, OOC, NBC, CAEEL
Transthyretin-like family (DUF290)	PF01060	IV, V	HGC, AYC, CAEEL
Nematode fatty acid retinoid binding protein (Gp-FAR-1)	PF05823	III, IV, V	WBC, LSC, BMC, *BP*, OVC, *OG*, *OD*, *AV*, *LL*, GPC, OOC, *HP*, AYC, CAEEL
Pepsin inhibitor-3-like repeated domain	PF06394	III, IV, V	OVC, DIC, *AV*, ASC, *PL*, OOC, *TF*, CAEEL
Tas retrotransposon peptidase A16	PF05585	III, V	ALC, CAEEL
Chromadorea ALT proteins	PF05535	III	OVC, WBC, BMC, DIC, *AV*

Table 5.2

Nematode-restricted Pfam-A domains found in non-caenorhabditid nematodes

The Pfam database was searched with taxonomic queries:

(1) "Caenorhabditis elegans AND (Tylenchida OR Enoplea OR Strongylida OR Panagrolaimoidea ORspirurida OR Ascaridida) AND NOT (Bacteria OR Archaea OR Viruses OR Vertebrata OR Arthropoda OR Fungi OR Viridiplantae)"

(2) "Nematoda AND NOT (Caenorhabditis elegans OR Bacteria OR Archaea OR Viruses OR Vertebrata OR Arthropoda OR Fungi OR Viridiplantae OR Dictyostelium discoideum)"

The species identifiers are described in table 3.1, except for: BP – *Brugia pahangi* (III); OG – *Onchocerca gutturosa* (III); OD – *Onchocerca dukei* (III); OO – *Onchocerca ochengi* (III); AV – *Acanthocheilonema viteae* (III); LL – *Loa loa* (III); HP – *Heligmosomoides polygyrus* (V); PL – *Parelaphostrongylus tenuis* (V); TF – *Trichostrongylus colubriformis* (V);

Of the 77 Pfam-A domains that are found only in *C. elegans* (and/or *C. briggsae*), perhaps the most well-known nematode-restricted protein domain family is the chemoreceptors. These seven-transmembrane G-protein-coupled receptors (7TM GPCRs) comprise approximately 1,280 intact genes, 6% of the total gene count for *C. elegans*, as well as 420 pseudogenes. The eighteen classes of chemoreceptor genes can be split into three superfamilies [222], each comprising a number of related families, although the domain annotation assigned to these families is not entirely congruent with Roberston and Thomas' divisions. The annotation available through WormBase assigns seventeen classes of genes to one of nine Pfam domains (Table 5.3), all of which are found only in *C. elegans*. The original prediction of their chemosensory function was founded upon transgene expression patterns in one or more known pairs of chemosensory neurons [223]. The global RNAi screening initiative failed to assign a phenotype to nearly all these genes [224]; however, targeted RNAi studies and cellular expression studies have begun to shed some light, with many genes expressed only in chemosensory neurons [225,222]. Taken together with multiple alignments of families, the information suggests chemosensory function for the majority, if not all of the genes. Whether these genes are found in other nematodes is unclear. An attempt to identify orthologue pairs with *C. briggsae* were confounded by the highly dynamic gene number with frequent duplications and gene loss, permitting only 7 pairs of orthologues to be robustly resolved [226,222]. The use of symmetrical BLAST searches identified more putative pairs, but their divergent protein sequences suggested that they were in fact paralogues produced prior to the *elegans-briggsae* speciation followed by gene loss in both species [227]. Early analysis of the *Brugia malayi* genome has revealed a number of genes containing chemoreceptor domains, but not in similar numbers to the *Caenorhabditis* species. The availability of EST-derived polypeptide sequences from such a wide distribution of nematodes may lead to the identification of chemoreceptors, although it is important to note that these proteins are expressed at low levels – often from a small (2-5) number of cells.

There are other examples of caenorhabditid-restricted Pfam-A domains which are present in large copy number. Most of these 77 Pfam-A domains were identified by automated sweeps through the *C. elegans* proteome, and a similar approach is necessary if these domains are to have a more species-rich membership. The availability of NemPep3 permits this kind of systematic search. However, caution must temper visions of wholesale expansion of previously caenorhabditid domain families; returning to the chemoreceptors, previous EST data has been of little assistance, with only 81 from ~160,000 *C. elegans* ESTs from chemoreceptor genes [222]. That said, when NemPep3 is decorated with Pfam-A domains, I would still expect a large number of domain families to increase their taxonomic membership.

Superfamily	Family(ies)	Pfam-A domain[1]
Str	*str, srd, srh, sri, srm*	PF01461
	srn	IPR000168
Sra	*sra*	PF02117
	srb	PF02175
	sre	PF03125
Srg	*srg*	PF02118
	srt	PF01748
	sru	PF02688
	srv	PF03375
	srxa	IPR000276
srw	*srw*	IPR000276

Table 5.3

Domain classification of chemosensory receptors.

The Pfam-A accessions were taken from WormBase (WS140) annotation.

(1) Where no Pfam-A domain was assigned the Interpro accession was recorded.

5.2.7 Domain loss and gain

It is pertinent at this point to define one of the evolutionary considerations than underpins my analysis, that of domain loss or gain (adapted from [207]). As with most protein domain studies a domain found in one proteome and not in another is considered to have been gained by one or lost by the other. This should not be confused with gene loss [147] or acquisition. The modification of a gene may change the protein domain affecting the score calculated when the sequence is passed through a domain model. Thus, a small number of amino acid changes in a protein sequence can lead to loss of a domain unit. That is to say, that the domain's function may be altered or removed. Such changes in the amino acid sequence may prevent the identification of the domain with current technologies, especially profile searches.

One of the more noteworthy analyses performed on NemDom3.0 is the identification of protein domains with restricted taxonomic distributions that previously did not include the Nematoda. These domains would be absent from the proteome of the free-living *C. elegans*, and as such may represent an adaptation to parasitism. The domains may be a modification of a pre-existing protein which results in a new, if only slightly varying, function such as regulating a metabolic or immune-response pathway. It is plausible that a duplicated domain, free of the original functional constraints, could adopt a similar local structure to an important protein from the host species. Such cases of convergent evolution are rare, but using tertiary structure has detected instances of convergent evolution. The best-known example of this is the Ser/His/Asp catalytic triad [228]. There is precedent for nematode proteins converging their structure and / or primary sequence to affect the host's metabolic pathways [229]. A dorsal gland polypeptide (HgCLE) taken from *Heterodera glycines* shares regions of similarity to a plant ligand involved in intercellular signalling (CLV3) [230], suggesting that the nematode ligand is used for parasitic modification of plant cells. The amino acid features shared by the plant and nematode sequences appear targeted to functionally important positions, suggesting a convergent evolutionary origin rather than HGT. There is also evidence for localised convergent evolution in two proteins from filarial nematodes. A gene duplication of the *B. malayi* homologue of the Human cytokine macrophage migration inhibitory factor (MIF) (PF01187) has resulted in nematode proteins that have been shown to be hemotactic for human monocytes and activate them to produce the cytokines IL-8 and TNF-α [231,232], thus influencing the host's immune system. Additionally, a cysteine protease from *B. malayi* (Bm-CPI-2) was shown to block the activity of an asparaginyl endopeptidase (AEP) necessary for maturation of the MHC class II receptor [233]. A multiple sequence alignment of the human cysteine protease and nematode homologues highlighted a single amino acid in the region responsible for AEP inhibition, shared between human and *B. malayi* proteins and not the other nematode homologues. Site-directed mutagenesis showed that this residue was necessary for the *B. malayi* protein to inhibit AEP. The mode of convergent evolution described for *H. gylcines*' HgCLEs differs significantly from the example of the filarial nematodes. The former involves a protein from a lineage separate to the sequence it mimics, while the latter proteins share relatively recent ancestry.

Another mechanism for adaptation of a proteome for parasitism is the acquisition of genes from other species, HGT, and is a means for rapid diversification in the proteomes of both prokarya and unicellular eukarya [234,235]. The extent of HGT in the Metazoa is intensely disputed, with many cases of HGT refuted by later studies [236,237,238,239]. The focus of HGT in nematodes has been on the plant parasites, with possible bacterial origin for genes identified in *Meloidogyne javanica*, *Heterodera glycines* and *Globodera rostochiensis* [214,213,216,240]. These findings were the result of single gene studies. The availability of EST data for plant parasitic nematodes enabled a sweep of the transcriptomes of three

Meloidogyne species [62]. Serial BLAST searches identified six candidate genes whose presence in the nematode genome was confirmed through cloning. Phylogenetic analysis advocated rhizobial bacteria, with which the nematodes share a trophic niche, as donors of the genes.

Confirmation of the candidate genes as present in the nematodes' genomes was an important step. There are many sources of potential contamination – both bacterial and human. This poses a problem when identifying protein domains previously absent from nematode proteomes. Bacterial contamination, in particular *Escherichia coli* and *Pseudomonas* species [241], can be prevalent in cDNA library construction and has been highlighted in Chapter Three. An additional source of contamination is from obligate bacteria; symbionts such as the *Wolbachia* found in filarial nematodes [242,243,244], and *Candidatus* Xiphinematobacter, which is found in the gut epithelium and ovaries of the dorylaim *Xiphinema americanum* group [245]. In the construction of the cDNA library it is practically unavoidable that bacterial mRNA will be included [246], therefore it is important to consider this problem when drawing conclusions from data analysis.

To explore domain dynamics in the Nematoda I have extracted those domains that were previously absent from the phylum, in particular those that were candidates for HGT or putatively involved in regulating some part of the host species metabolism or immune / defense response. Each candidate was examined throughly to determine whether it was a consequence of contamination or worthy of further investigation.

5.2.8 Domain family expansion

Coupled to domain loss and gain is lineage-specific expansion of a domain's frequency. If a particular domain is found to represent a greater proportion of one species proteome than that of another, it is likely to be a consequence of positive selection on one lineage. If such expansion was seen in a parasitic nematode compared to the free-living *C. elegans*, it may represent a promising drug target. However such a comparison is dependent upon a proteome derived from a complete genome sequence. EST data is influenced by the level of expression for mRNA. Even the creation of clusters does not guarantee non-redundant gene sets, which has been observed both in the nematode dataset (Parkinson, Blaxter and Wasmuth unpublished) and published for the *Fundulus heteroclitus* EST set, with 15 clusters for the apolipoprotein and 10 for cytochrome oxidase [22]. With EST data it is possible to highlight those domains that come from highly expressed genes in particular species datasets. This approach led to the identification and characterisation of two novel proteins (NIM-1 and NIM-2) in the nematode *Haemonchus contortus* [16]. Comparisons between the proportion of ESTs in a dataset and expression measured through SAGE (Serial Analysis of Gene Expression) has shown high positive correlation [58,57], supporting the use of ESTs in this way.

Box 5.1 The MCL algorithm and TRIBE-MCL

The graph clustering paradigm postulates that natural groups in graphs have the following property (adapted from van Dongen [200]): a random walk in the graph that visits a dense cluster is unlikely to leave that cluster until many of its vertices have been visited. The Markov Cluster (MCL) algorithm simulates flow within a graph, and promotes flow where the 'current is strong', and downgrades flow where the 'current is weak'. According to the above paradigm, if natural groups are present in the graph, then current between different groups will wither and reveal the cluster structure of the graph. The transformation of a graph into a Markov graph, where all the edges sum to one, and hence a Markov matrix is necessary for this flow process. The MCL algorithm simulates flow within the graph by alternating two operators called **expansion** and **inflation**. Expansion calculates the power of the Markov matrix (matrix squaring) with the inflation step taking the Hadamard power of the matrix. This alters the relative probabilities of within-cluster and between-cluster random walks, and is the parameter by which the users can alter the tightness (granularity) of clusters. Higher inflation is more conservative, yielding more clusters.

The concept of graph clustering is one readily adapted in the search for protein families. The MCL algorithm has been incorporated into the TRIBE-MCL program, where the origin graph represents similarity between proteins. The vertex connecting the proteins (nodes) is weighted by a similarity measure, here a log transformed BLAST E value (Figure 5.2).

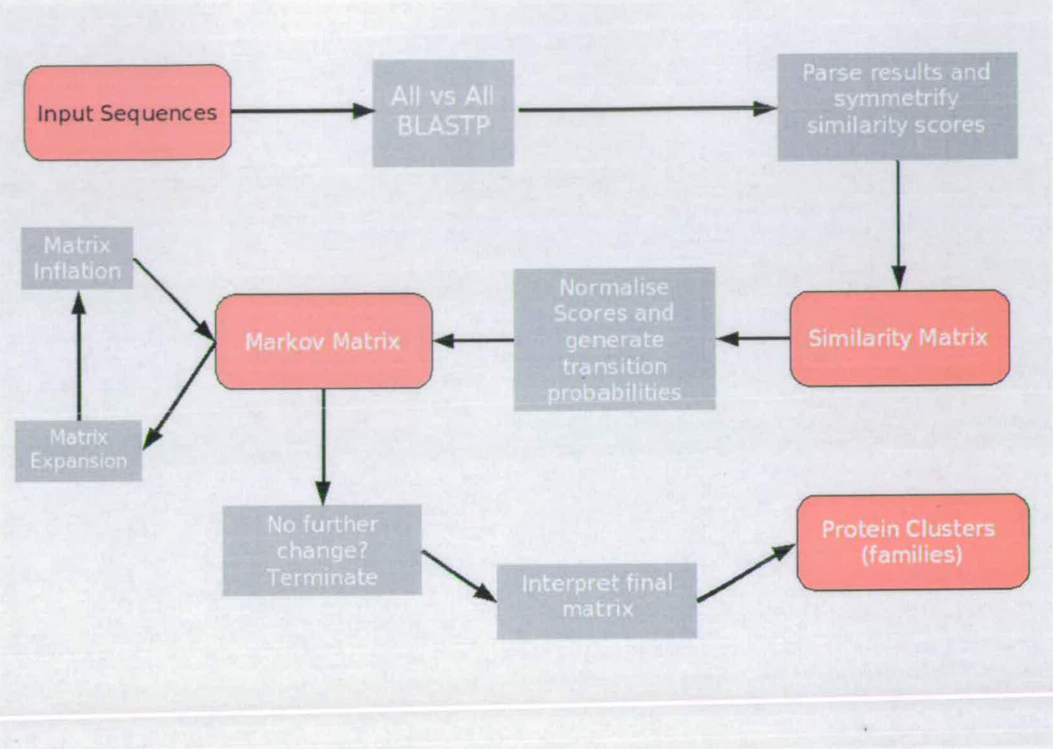

Input Sequences
All vs All BLASTP
Parse results and symmetrify similarity scores
Matrix Inflation
Normalise Scores and generate transition probabilities
Markov Matrix
Similarity Matrix
Matrix Expansion
No further change? Terminate
Interpret final matrix
Protein Clusters (families)

Figure 5.2
Schematic of the TRIBE-MCL process.

5.3 Methods

5.3.1 EST datasets for domain assignment

Four independent sets of 4,000 *C. elegans* ESTs were randomly selected from dbEST. These sets were clustered using CLOBB, and phrap used to derive a consensus contig for each cluster. To ensure that each EST contig corresponded to a known *C. elegans* coding region, I carried out BLASTN searches against the WormPep cDNA collection (version 140). A match was considered significant if the HSP covered ~75% of the EST contig. The significant matches were then compared to the WormPep protein set (BLASTX E > e-8), thus associating each EST contig with a WormPep sequence. The size of the test sets ranged from 2,316 to 2,346 sequences. The coding regions for these EST contigs were predicted using prot4EST version 2.2. To simulate the situation facing the majority to EST projects the codon usage table and HMM matrix were those assembled from 50,000 coding nucleotides, as described in Chapter Two.

5.3.2 Assigning protein domains

All the domain assignments from WormPep and the EST-derived proteins were stored in a custom postgreSQL database, to facilitate searching.

The Pfam-A domain models (version 17 – June 2005) were downloaded from the Sanger Institute's ftp server (ftp.sanger.ac.uk). These models are formatted to be used with the hmmpfam program, part of the HMMer software suite [209]. There are in fact two search programs in HMMer – hmmpfam was used because the E value calculation uses the size (length) of the domain library opposed to hmmsearch where the database is considered the query sequence file. Using hmmpfam ensures that E values are comparable across a range of different sized replicates.

Pfam-A domains in WormPep

Pfam-A domain assignments are available from both the Pfam database and WormBase. To ensure data provenance the WormMart facility at WormBase was used. Surprisingly, there were a small number of conflicts between WormBase and Pfam, where in one database a protein was assigned a particular domain but was absent from the other. Where a disagreement occurred, I accepted the positive domain annotation.

Measuring accuracy

As the alignment between the EST contig and its cognate WormPep sequence was known, the location of the domain assignments could be transferred between sequences. The performance of domain annotation was evaluated with regard to specificity (Equation 5.1) and sensitivity (Equation 5.2). Figure 5.3 shows how classifications were assigned.

$$specificity = \frac{TN}{(TN + FP)}$$ **Equation 5.1**

$$sensitivity = \frac{TP}{(TP + FN)}$$ **Equation 5.2**

5.3.3 Taxonomically-restricted Pfam-A collections

Unless otherwise stated the identifiers of taxonomically-restricted Pfam-A identifiers were obtained from the Pfam website using the taxonomy query facility.

1. *C. elegans* and at least one other nematode to the exclusion of all non-nematodes

```
"Caenorhabditis elegans AND (Tylenchida OR Enoplea OR Strongylida OR Panagrolaimoidea OR spirurida OR Ascaridida) AND NOT (Bacteria OR Archaea OR Viruses OR Vertebrata OR Arthropoda OR Fungi OR Viridiplantae)"
```

2. Only non-caenorhabditid nematodes

"Nematoda AND NOT (Caenorhabditis elegans OR Bacteria OR Archaea OR Viruses OR Vertebrata OR Arthropoda OR Fungi OR Viridiplantae)"

3. Only Metazoa

"Metazoa AND NOT (Bacteria OR Archaea OR Viruses OR Fungi OR Viridiplantae)"

4. Only Prokaryote

"Bacteria AND NOT (Eukaryota OR Archea)"

5. Found in Vertebrates but not nematodes or arthropods

"Vertebrates AND NOT (Nematoda OR Arthropoda)"

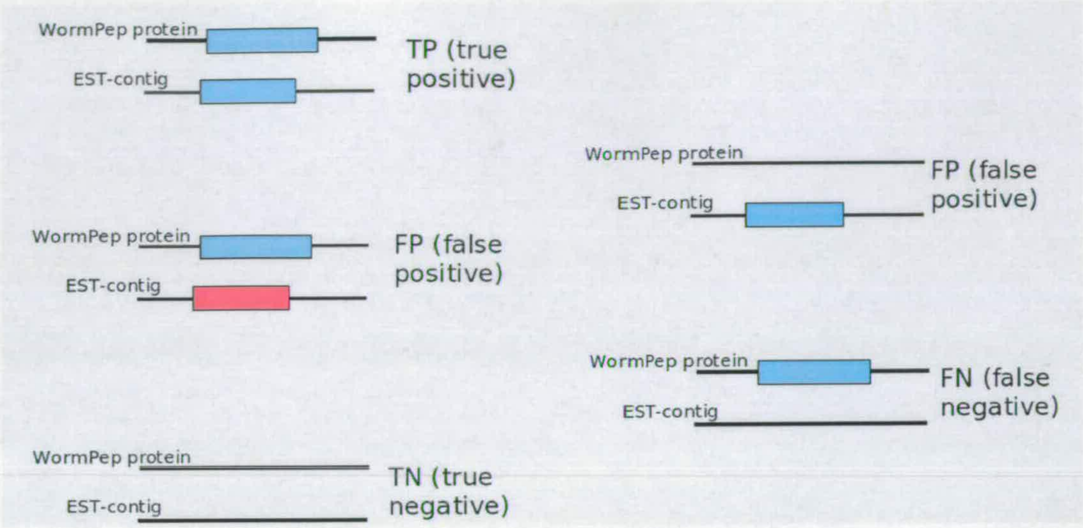


Figure 5.3

Categories for testing accuracy of domain model assignment

The blue and red boxes define different domains. Thus, if an EST contig is annotated with a red domain when a blue one was expected, it is considered a false positive.

5.4 Results and Discussion

5.4.1 Accuracy of domain identification in translated EST contigs

Before the proteins of NemPep3 could be decorated with Pfam annotation two potentially confounding effects had to be explored. One was how scoring thresholds affected the assignment of protein domains to polypeptides derived from EST contigs. The second was whether correct Pfam annotation is possible for partial protein sequences that cover only part of a protein domain (Figure 5.1). The test set comprised four independent replicates of 4,000 randomly selected ESTs from *C. elegans*. Each set was clustered, and polypeptides derived by prot4EST, as described in the Methods (5.3.1). Each polypeptide from a *C. elegans* EST contig, (known as a 'CXP'), was mapped to its cognate WormPep sequence. The CXPs were searched with the Pfam-A library using four scoring thresholds, three bit score cut-offs (trusted, gathering and noise) and one E value cut-off (<0.1) recommended by Eddy [209]. The accession and location of each assigned protein domain in a CXP was compared to the annotation of the cognate complete protein in WormPep (see Method 5.3.2). This was performed for both the global and local models.

Definitions of coverage and measuring accuracy

The *coverage* of a protein domain is the proportion of a domain's length that is covered by the CXP when aligned to its cognate protein in WormPep.

To evaluate how accurate the assignment of domain annotation on CXPs I used a binary classification; each domain was considered to be present in a certain location on the protein or not. The performance was tested using *specificity* and *sensitivity* (see the Methods (5.3.2) for how each is assigned). The higher the specificity the less often a domain was incorrectly assigned to a region of the protein. The higher the sensitivity the fewer true domains were not identified.

Performance of global and local Models

The differences in model scoring parameters between global and local HMM suggests that their performances would differ. As local models permit matches that cover only part of the domain, it is likely that a region with no domain could be erroneously assigned Pfam-A annotation, increasing the number of false positives. However if the EST-derived polypeptide contains only part of a protein domain, it is unlikely that a global model will score sufficiently to cause a domain assignment, thus inflating the number of false negatives and decreasing the sensitivity of the search.

As expected, whether a model was local or global affected the accuracy of domain identification (Figures 5.4a&b). The specificity of domain assignment was consistently higher using the global models (μ=0.97±0.009) than local models (μ=0.93 ± 0.011). The proportion of domain coverage made little difference to the specificity because the regions that were of lower coverage usually did not return any domain. An increase in domain coverage led to a reduction in incorrectly assigned domains (false positives). Relatively few domains (<0.5%) were assigned to regions of the CXP proteins where no domain was expected, this number decreased as the domain coverage increased. The difference in the sensitivity of the searches was more pronounced. Searching with local models saw the sensitivity reach a plateau when the match between the model and a CXP covered 40-50% of the domain (Figure 5.4b). Similar performance was achieved with the global models only when almost all of the domain was covered by the model-CXP alignment (Figure 5.4a). Use of the local models identified the majority of the domains predicted from the WormPep annotation which overlapped with the EST-derived polypeptides. Many of these domains overlapped only part of the CXP proteins; from a total of 21,153 Pfam domains assigned to CXP proteins, 3,876 domains were found at the 5' end of the CXP protein, 3,799 on the 3' end and 3,412 domains covered the entire CXP protein without either boundary found in the polypeptide (Figure 5.5).

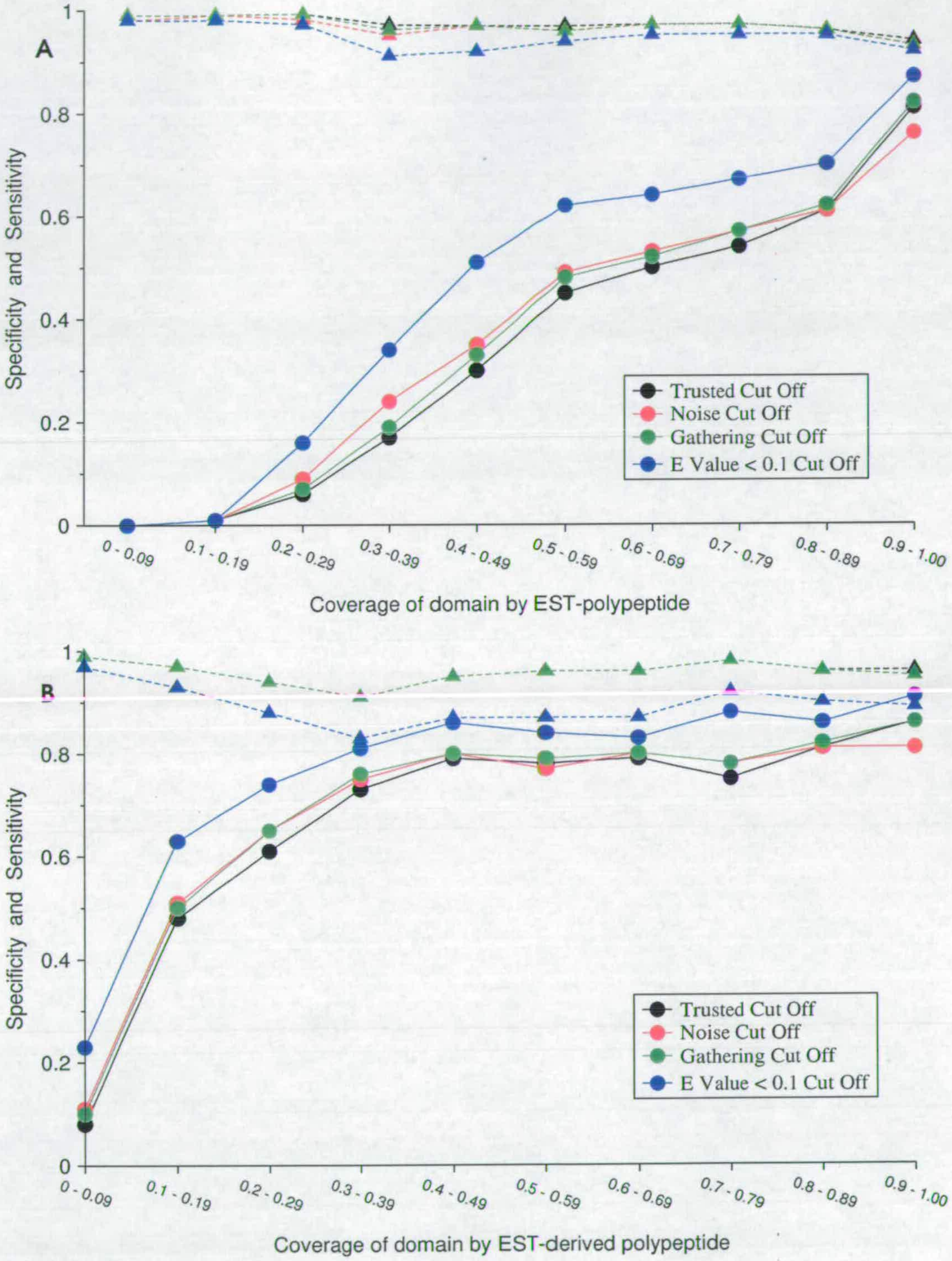


Figure 5.4

Specificity (dashed line) and Sensitivity (continuous line) of assigning Pfam-A domains to EST-derived proteins.

Legend overleaf

Figure 5.4 (previous page)

Specificity (dashed line) and Sensitivity (continuous line) of assigning Pfam-A domains to EST-derived proteins.

ESTs from *C. elegans* were clustered, their coding regions predicted and a cognate full-length *C. elegans* protein assigned. The hmmpfam program was used to assign Pfam-A domain models to each sequence. The domain annotation was compared to the reference set from WormPep.

Two model architectures were used, (A) global Pfam models, where an alignment must be global with respect to the model but local for the query sequence, and (B) local Pfam models, in which the alignment may be local for both the model and sequence. See Methods and Figure 5.3 for categories measuring accuracy of assignment. Some of the data points for the specificity are hidden behind the Gathering cut off .

Figure 5.5 (overleaf)

Assigning partial domains to EST-derived proteins.

The numbers correspond with the number of times this type of assignment was made. Total of 21,153 local models were assigned. The black bar refers to the model; the continuous section is the part of the model that aligns with the query sequence and the hashed region is the remainder of the domain model that does not match the query sequence

(A) the Pfam-A domain overlapped with the 5' region of the EST contig.

(B) there was a significant match with the 3' region of the query sequence.

(C) the model covers the entire EST contig. Usually seen for short EST-derived proteins.

(D) the local model matches an internal region of the query sequence. These were not included in NemDom.

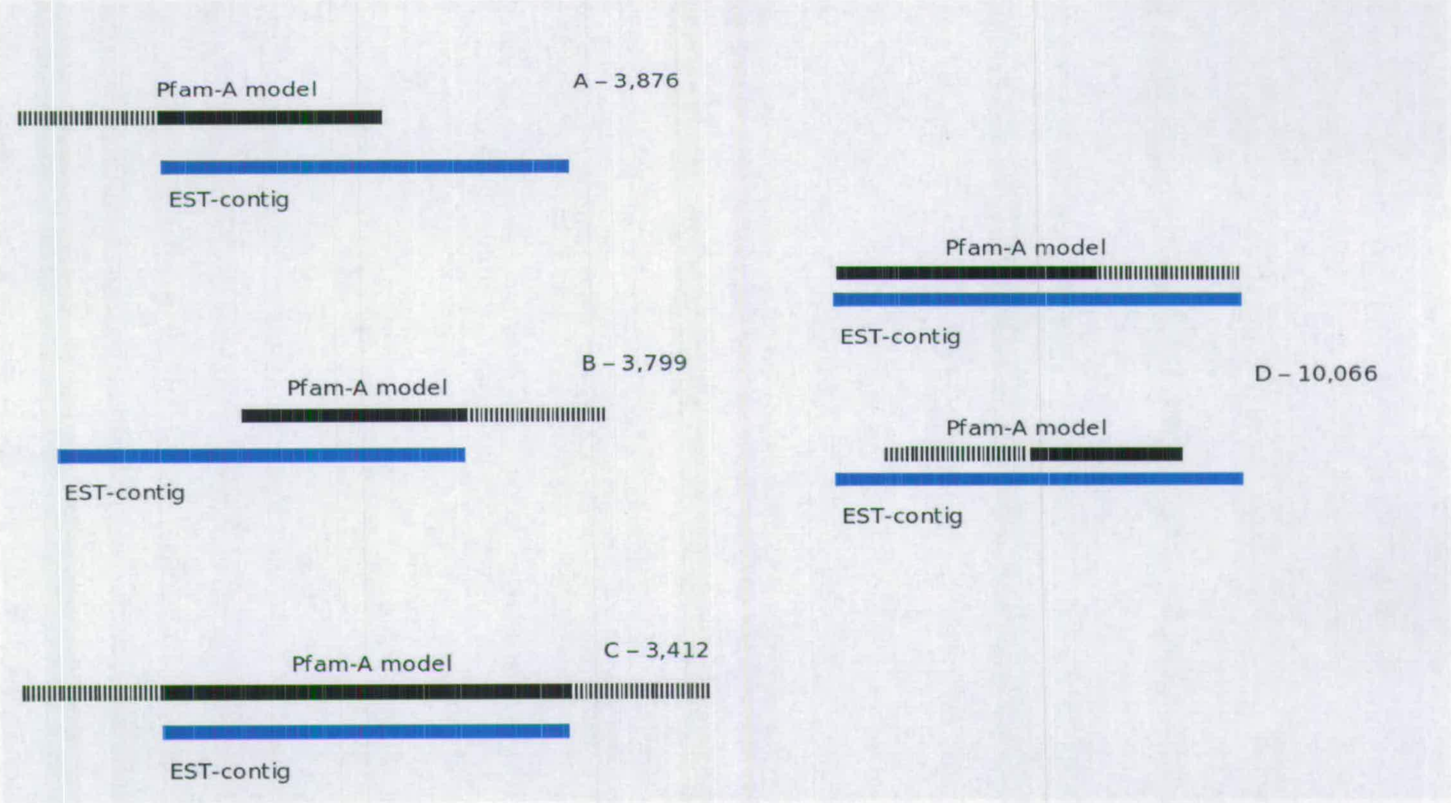


Figure 5.5

Assigning partial domains to EST-derived proteins.

Legend on previous page

Scoring thresholds

The choice of cut-off for bit scores (TC, GA or NC) made a nugatory difference to the performance of searches. The trusted cut-off (most conservative), missed several domains when the coverage was less than 90%, as the insertions introduced into the model-CXP alignment caused a drop in the bit score. When the domain coverage was maximal, use of the noise cut-off resulted in a reduction in the sensitivity of the search. The poorer performance was predominantly caused by the incorrect identification of the 'IMP dehydrogenase / GMP reductase' domain (PF00478) when another domain was anticipated. The expected domain varied in these circumstances, and for many there was no feature (e.g. sequence similarity or transmembrane regions) explaining the erroneous assignment. PF00478 belongs to the 'common phosphate binding-site TIM barrel superfamily' (CL0036) that contains 27 Pfam-A domains, but none of the related domains were an expected domain. An explainable miss-annotation was PF00153, where 14 of the expected 31 'mitochondrial carrier protein' were attributed to PF00478. Both domains contain transmembrane regions, which are predominantly hydrophobic and so explain the small regions of similarity. It is likely that the CXP domains are sufficiently different to their cognate protein and the domain model, prohibiting correct assignment of the model. The bit score cut-offs for the PF00478 domain are extremely low (-186 to -195), therefore encouraging miss-assignment in the absence of another domain match. It should be noted that the E values awarded for these spurious matches were many orders of magnitude greater than those for the correctly assigned PF00478 domains (cf. e-17 and e+3). Altering the bit score threshold did little to resolve the annotation as the differences between thresholds is often relatively small, for any given model the score rarely varies by more than 5 bits. This fine tuning is crucial for separating domains from a closely related group, or 'clan', for example the '*C. elegans* chemosensory receptor' clan (CL0138) [223,247,222], but probably less consequential when distinguishing between unrelated domains.

There was a much more marked effect upon classification of domains when an E value threshold (<0.1) was used. The specificity was reduced, with an increase in the number of domains incorrectly assigned to the CXP proteins. The E value approximates how significant a match is given the size of database searched, and, when used as the cut-off, the search program (HMMer) ignores the bit score threshold assigned to each Pfam-A model. This was most striking for the local models (Figure 5.4a), where domains were erroneously assigned due to partial hits to the domain's model which satisfied the E value cut-off. The bit score thresholds, curated by Pfam, overcome this error as they are frequently higher than the corresponding global model thresholds, making the identification of a partial hit, at the level of individual amino acids at least, more conservative. In contrast, the sensitivity of the searches were improved with the E value cut-off. This was most striking at low levels of domain coverage. The data suggests that the use of E value as a score threshold is less conservative than the Noise bit score cut-off curated for each model. Hence, the increase in the proportion of correctly assigned protein domains was a consequence of this relaxation.

Finding domains in NemPep3

Taken together, these results supported a combinatorial approach to identifying protein domains in the partial proteins of NemPep3. Global models should be used to identify full-length protein domains in the EST-derived polypeptides, as the number of false positive predictions would be lower. Next, partial domains identified by local models would only be accepted into NemDom3.0 if they resided at either terminus of the polypeptide.

The choice of scoring threshold was less clear. An E value cut-off of <0.1 correctly identified a greater proportion of domains when the coverage was reduced (sensitivity). However, such searches also returned a greater proportion of domains when, according to WormPep annotation, none existed. For global models the number of false positives had to be kept to a minimum, and therefore I selected the gathering (GA) bit score cut-off. The local Pfam models were to be used to find domains at the termini of the polypeptides, and could involve low domain coverage. Once again, reducing the number of false positives was important, so the gathering bit score threshold was used. For local models, this decision would mean the failure to identify true domains at the protein's termini. However, subsequent *de novo* protein domain identification would hopefully cluster these regions that could then be manually annotated.

Number of Domains per CXP

To provide an expectation for the number of Pfam-A domains assigned to EST-derived proteins, the number of domains assigned to each CXP was compared to the annotation for the full-length WormPep proteins (Figure 5.6). The most striking contrast was that, regardless of the type of models used, more than half of CXP polypeptides had no Pfam domain assigned, compared with 38% for the full-length WormPep proteins. The number of domains per protein then followed a similar distribution to the WormPep annotation. The CXP polypeptides which were annotated with at least one domain were significantly longer (m=164; sd=56.1) than CXPs without a domain (m=105; s.d.=51.8) (t = 46.7039, df = 4803.55, p-value < 2.2e-16). There is also a significant difference between the lengths of CXPs with one domain and two (t = -7.9429, df = 266.055, p-value = 5.591e-14), but there were no significant difference for further increments in domain complement. The CXPs with more than two domains generally contained short, tightly-packed, homogeneous repeats, e.g. the 'tetratricopeptide repeat' (PF00515) or 'WD-40' (PF00400). This is verified by manual inspection of the domain annotation. Searching with the local models assigned at least one Pfam-A domain to a higher proportion of CXP polypeptides.

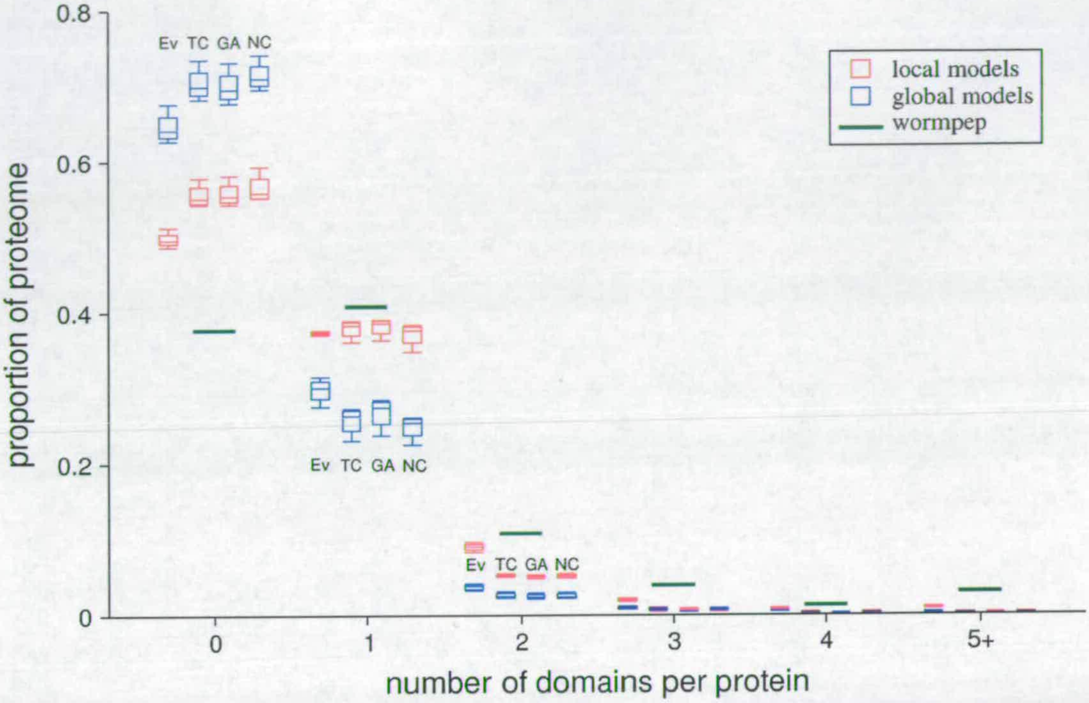


Figure 5.6

Number of domains found per protein in the EST-derived collections and WormPep140.

The effect on the number of domains assigned by using two model architectures (global and local) and varying the score cut off was investigated.

Use of the local models annotated more EST-derived contigs with Pfam-A domains than global models. This was expected as partial matches are permitted with this architecture. A domain was more likely to be assigned if an E value cut off (<0.1) was used. However it should be noted that there was no assumption that the assignments were correct.

The cut offs used were: Ev – E; TC – Trusted cut off; GA – Gathering cut off; NC – Noise cut off. For clarity these are absent in some positions, but the order remains consistent.

In the box and whisker plot the following is presented: mid-line is the median; the box boundaries are the upper and lower quartiles; and the whiskers are the outlying (extreme) values.

5.4.2 Annotating NemPep3

The HMMs from the Pfam-A database were used to search NemPep3 to identify known protein domains, creating the NemDom3.0 collection. Between 36.5% (*Ascaris* suum) and 41.4% (*X. index*) of the partial proteomes (excluding *Zeldia punctata*) were decorated with at least one Pfam-A model (Table 5.4), in agreement with earlier expectations (Figure 5.5). The mean protein length (MLP) varied between the proteomes (Table 3.2), which suggested a positive correlation between MLP and domain content. The correlation was rho=0.51, which was significant (t = 3.416, df = 34, p-value = 0.0017). There was a negative relationship between the size (number of proteins) of the proteome (modified from Table 3.2) and the proportion of unique (non-redundant) Pfam domains (rho=-0.8089218; t = -8.0229, df = 34, p-value = 2.375e-09). This is a logical consequence of the random nature of EST sampling. As anticipated from the work described earlier in this chapter, the polypeptides derived from ESTs were assigned fewer protein domains than the full-length caenorhabditid proteins (Figure 5.6 and Table 5.5). Selective use of the local models in the searches added a further 15,514 protein domain annotations. Almost 900 domains, present in caenorhabditid proteins, were not found in the EST datasets. This absence of many of these was a consequence of the incomplete sampling of the proteomes of the parasitic species.

Table 5.4

Summary of Pfam-A domain assignment by species.

Species[1]	# domains[2]	# unique[3]	Number of proteins with X domains					Mean[3]	Standard deviation[4]	Mode[5]
			X = 0	X = 1	X = 2	X = 3	X ≥ 4			
ACP	5,492	672	7,166	3,348	902	59	31	0.48	0.71	0
ALP	1,003	124	1,466	698	131	10	3	0.43	0.64	0
ASP	10,871	789	14,496	6,668	1,587	185	100	0.47	0.72	0
AYP	5,240	749	5,490	2,463	1,021	126	71	0.57	0.85	0
BMP	8,927	896	11,552	5,253	1,527	122	56	0.48	0.71	0
CAEEL	36,559	2,297	9,216	17,180	3,085	989	1,166	1.16	3.00	1
CBP	28,173	2,741	8,928	15,975	1,768	729	862	1.00	1.87	1
DIP	2,175	340	2,991	1,406	332	18	11	0.46	0.67	0
GPP	3,449	545	3,750	1,619	681	89	43	0.56	0.82	0
GRP	3,896	623	4,438	1,988	770	75	30	0.53	0.77	0
HCP	7,359	928	7,664	3,358	1,581	164	73	0.57	0.82	0
HGP	12,109	1,217	14,642	6,615	2,174	229	94	0.51	0.76	0
HSP	1,744	334	2,019	975	273	46	19	0.52	0.77	0
LSP	2,155	381	2,555	1,161	420	31	14	0.52	0.74	0
MAP	3,078	449	3,902	1,852	503	43	19	0.49	0.71	0
MCP	4,518	662	5,689	2,622	810	57	24	0.49	0.71	0
MHP	8,313	879	10,783	4,968	1,406	111	43	0.48	0.7	0
MIP	7,860	914	9,572	4,323	1,459	138	42	0.51	0.74	0

Continued overleaf...

Species[1]	# domains[2]	# unique[3]	Number of proteins with X domains					Mean[3]	Standard deviation[4]	Mode[5]
			X = 0	X = 1	X = 2	X = 3	X ≥ 4			
MJP	4,293	553	5,753	2,667	699	50	17	0.47	0.68	0
MPP	2,048	417	2,209	983	419	47	20	0.56	0.79	0
NAP	2,997	440	3,870	1,764	488	45	23	0.48	0.74	0
NBP	1,058	241	1,156	514	236	14	7	0.55	0.77	0
OOP	3,446	511	3,850	1,755	672	65	32	0.54	0.79	0
OVP	5,298	742	6,403	2,794	1,034	103	28	0.51	0.74	0
PEP	633	151	594	237	149	22	7	0.63	0.88	0
PPP	5,790	811	6,506	2,862	1,219	107	37	0.54	0.77	0
PTP	3,906	585	4,150	1,888	812	86	28	0.56	0.80	0
PVP	1,060	189	1,261	567	186	23	11	0.52	0.77	0
RSP	763	176	823	330	181	15	5	0.56	0.81	0
SRP	5,305	760	6,168	2,822	936	119	55	0.53	0.78	0
SSP	5,183	789	5,446	2,415	1,071	115	57	0.57	0.83	0
TCP	2,015	320	2,196	988	408	53	12	0.55	0.78	0
TDP	2,556	381	2,879	1,282	524	42	23	0.54	0.77	0
TMP	1,956	365	2,164	982	369	34	27	0.55	0.82	0
TSP	4,618	636	5,836	2,730	763	67	34	0.49	0.72	0
TVP	1,423	235	1,804	852	227	26	9	0.49	0.71	0
WBP	2,197	419	2,265	936	475	68	24	0.58	0.84	0
XIP	6,575	954	6,106	2,562	1,394	243	106	0.63	0.91	0
ZPP	346	131	220	84	107	14	1	0.81	0.96	0

Table 5.4
Summary of Pfam-A domain assignment by species – ***legend overleaf***

Table 5.4 (previous page)

Summary of Pfam-A domain assignment by species.

(1) For species codes see table 3.1.

(2) The total number of non-overlapping instances of Pfam-A domains assigned to each species' proteome. This includes the global and local models.

(3) The number of matches between proteins in a species' proteome to Pfam-A domains that are restricted to the Nematoda.

(4) The mean average of domain instances in a given species' proteome.

(5) The mode average of domain instances in a given species' proteome.

	caenorhabditids	NemPep3 incl. caenorhabditids	NemPep3 excl. caenorhabditids
Total # of proteins	41,754	155,128	113,374
# of proteins with domain >= 1	24,281 (58.1%)	62,725 (40.4%)	38,444 (33.9%)
Total # of domains	46,588	101,500	54,912
Global Models	46,588	85,986 (84%)	39,398 (72%)
Local Models	0	15,514 (16%)	15,514 (28%)
Unique Pfam accessions	2,877	3,241	2,361

Table 5.5

Differences in domain assignment between complete and partial proteomes.

The incomplete nature of EST-derived proteins means that the proteomes of parasitic nematodes have a smaller proportion of Pfam-A assignments, compared to the full-length proteins from *C. elegans* and *C. briggsae*. However the use of local models has increased the number of domain matches, and the parasitic species' domain complement includes 364 domains not found in currently surveyed caenorhabditid proteomes.

5.4.3 Phylogenetic distribution of Pfam domains

There are 3,730 domain models in the Pfam database that are derived from proteins from at least one metazoan species, 2,543 of which are found in NemPep3. The presence of these metazoan domains were mapped onto the Nematode phylogenetic tree (Figure 5.7). Half (1,222) of the metazoan domains found in the Nematoda appear to be widespread throughout the phylum, although only 883 were identified in each of the four clades. As expected from previous analyses presented in this thesis (Chapter Four), the complete proteomes for *C. elegans* and *C. briggsae* accounted for the majority of diversity in the datasets. The most striking feature of the distribution was that 228 domains were not found in either of the *Caenorhabditis* sp. proteomes (see project web-site: www.nematodes.org/thesis/james/supp). Previous analyses on the datasets suggest that experimental contamination is a very real possibility, and is examined later in this chapter. However, by considering only those domains that are found in proteomes from two distinct orders, the likelihood of being misled by a contaminant is much reduced. A total of 31 domains matched this criteria. Many of these domains are from ancient lineages, whose last common ancestor are deep taxonomic divisions: Eukaryota plus Prokaryota (11), Eukaryota plus Archea (2), Eukaryota (11), Metazoa (5). The most parsimonious explanation is that these domains were present in the last common ancestor of all nematodes, and have since been lost in the *C. elegans* lineage. I note that "loss" could be through rapid divergence of the domain sequence so it is no longer recognised by the Pfam-A model or through deletion of the gene.

Two of the domains are worthy of a brief description. The Phage minor tail protein L (PF05100) domain had previously only been found in proteins from proteobacteria and viruses. It was identified in three proteins, occurring once in *Haemonchus contortus*, *Nippostrongylus brasiliensis* and *Wuchereria bancrofti.* The three proteins are all derived from singleton ESTs (unclustered) and, given that there is no obvious role for the protein in the nematodes, it is likely that they represent contaminants. The BESS motif (PF02944), a domain of unknown function and named after the proteins in which it is found (BEAF [248], Suvar [249] and Stonewall [250]), was thought to be specific to Drosophila, although Pfam annotation identified matches in the proteomes of *Anopheles gambiae* and *Xenopus tropicalis*. Identification of the BESS domains in two distantly related nematodes (*Pratylenchus vulnus* and *H. contortus*) suggests a more ancient origin for the domain. It must be noted that while both nematode proteins (PVP00645 and HCP07484) match the model with relatively poor E values (0.12 and 0.046), the bit scores awarded to the protein-model alignment exceed the conservative trusted cut-off. The function of this domain remains unknown [251], but inclusion of representatives from additional species may yield testable predictions. If the BESS domain was lost in the ancestral mammal then given the taxonomic bias in sequencing towards mammals and the reported high level of gene loss in *C. elegans*, it easy to understand why the domain's lineage was miss-classified.

Four hundred and sixty-four domains that are restricted to a single nematode order but also found in animals outside the phylum were identified. Many of these were found in only one species and are therefore possibly the result of contamination. Each domain must be considered in turn, a project that is currently underway. Despite this, it is note worthy that the dorylaim, *Xiphinema index*, was the only nematode in which 25 metazoan domains were found. It is likely that a proportion of these are present in other nematodes but have yet to be sequenced. Some may derive from plant host material or from *Xiphinema*'s symbiont – *Xiphinematobacter* (see Introduction). However, some may reflect the life-cycle and feeding strategy of this plant parasite. An example is the "dihydrodipicolinate synthetase family" (PF00701), a domain that forms a TIM barrel structure and is a key enzyme in the diaminopimelate pathway (lysine biosynthesis) of higher plants and prokaryotes [252,253]. If this protein is from the *Xiphinema* genome, its evolutionary origin is of obvious interest;

this issue is discussed in more detail later in this chapter.

Figure 5.7 (overleaf)

Distribution of metazoan protein domains in the phylum Nematoda.

The tree represents the phylogenetic relationships of the Nematoda as presented by Mark Blaxter [40]. There are a total of 3,730 Pfam-A domains that are found in a metazoan species, of which 2,543 are present in NemDom3. It can be inferred that these are ancient domain families. The distribution of these domains have been mapped onto the tree, with each node detailing the number of domains that are distributed in the daughter taxa. That is, each protein domain has been assigned to the most recent common ancestor of the taxa in which is found. The numbers are mutually exclusive. This allows one to surmise in which lineage a domain has been lost or modified beyond detection. For example, two domains were found in both a trichinellid and a dorylaimid; 1,222 domains had representation in the Dorylaimida and at least one order of the Rhabditida.

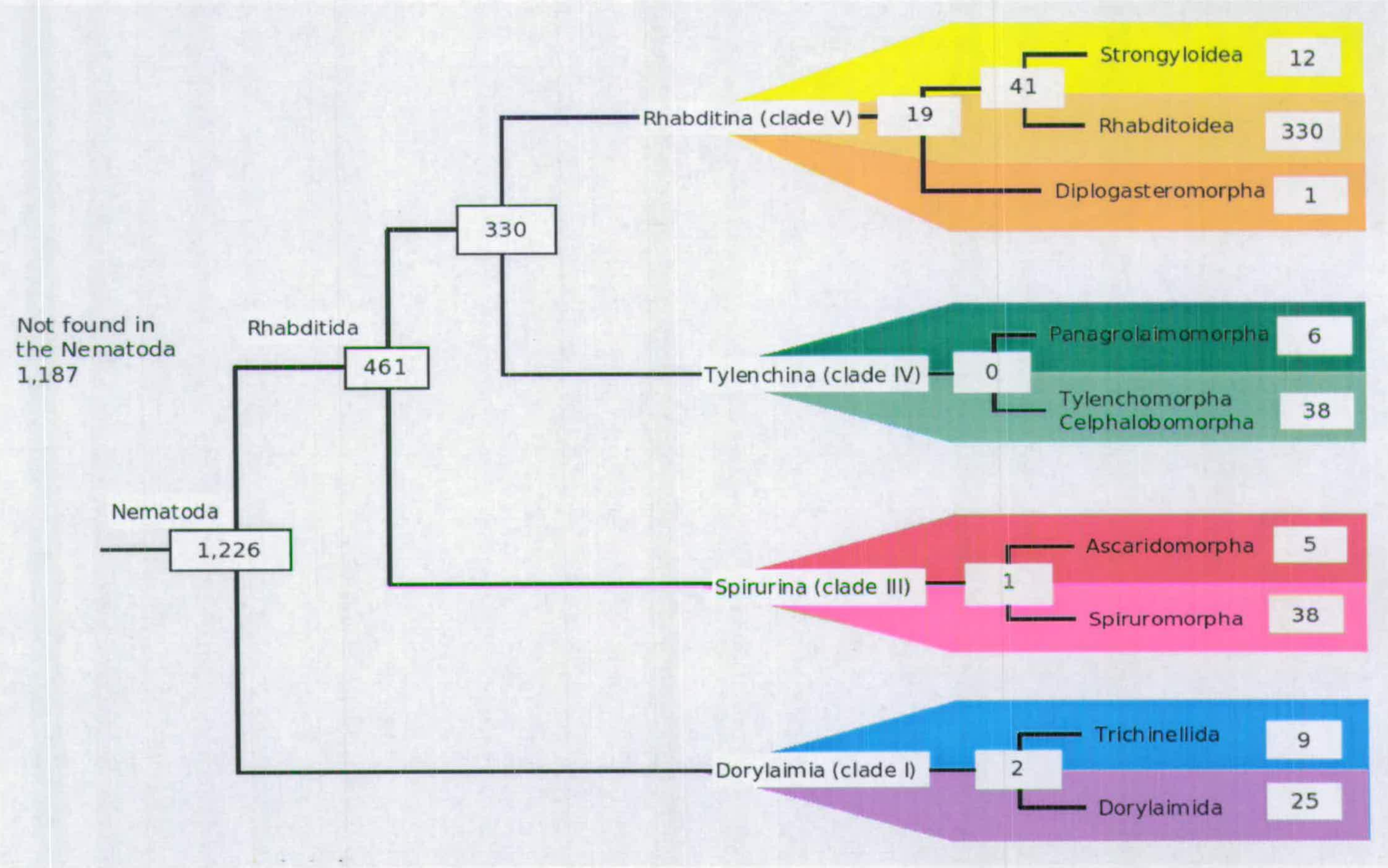


Figure 5.7

Distribution of metazoan protein domains in the phylum Nematoda.

Legend previous page

5.4.4 Nematode Specific Pfam Domains

There are 84 protein domains in the Pfam-A database that are exclusive to the Nematoda. All are found in *C. elegans* (see later regarding the ALT protein), while only seven are found in non-caenorhabditid species (see Table 5.2). Considering these seven domains first, six were identified in the domain complements of non-caenorhabditid nematodes, expanding their taxonomic distribution (Table 5.6). The Nematode cuticle collagen domain (PF01484) and transthyretin-like family (PF01060) were found to be distributed throughout the phylum Nematoda, including the Dorylaimia (clade I). The first of these was expected to be present in all nematodes, given its role in the core structure of the cuticle, one of the defining morphological features of the Nematoda [254,255]. The second domain has weak similarity to hormone transport protein, and although the function of this domain in nematodes is unknown, its ubiquity may aid annotation. Three of the domains (PF02520, PF0583 and PF06394) are seemingly restricted to a broad cross-section of species from the Chromadorea (Clades III, IV, V). Two possibilities for the origins of these domains are that they arose in the ancestral Chromadorea, or that they were present in an earlier nematode and subsequently lost after the separation of Dorylaimia and Chromadorea. Ideally, EST data generated from enoplid nematodes (clade II) would go a long way towards confirming one of these hypotheses.

Pfam Description	Pfam Accession	Clade Distribution[1]	Species Distribution[2]		
			Previously known	Acquired	Missing
Nematode cuticle collagen N-terminal domain	PF01484	I, III, VI, V	ASP, BMP, *BP*, CAEEL, CBP, GPP, HCP, MIP, MJP, TDP	ALP, AYC, DIP, HGP, HSP, LSP, MAP, MCP, MHP, MPP, NAP, NBP, OOP, OVP, PEP, PPP, PTP, RSP, SRP, SSP, TCP, TMP, WBP, XIP, ZPP	
DUF148	PF02520	III, IV, V	ASP, CAEEL, OOP, TDP	ACP, AYP, CBP, GRP, HCP, PEP, PPP, PTP, SSP, TCP	MIP, NBP
Transthyretin-like family (DUF290)	PF01060	I, II, IV, V	AYP, CAEEL, HGP	ACP, ALP, ASP, BMP, CBP, DIP, GPP, GRP, HCP, LSP, MAP, MCP, MHP, MIP, MJP, MPP, NAP, NBP, OOP, OVP, PEP, PPP, PTP, PVP, RSP, SRP, SSP, TCP, TDP, TMP, TSP, WBP, XIP, ZPP	
Nematode fatty acid retinoid binding protein (Gp-FAR-1)	PF05823	III, IV, V	*AV*, AYP, MBP, *BP*, CAEEL, GPP, *HP*, *LL*, *OD*, *OG*, OOP, OVP, WBP	ACP, ALP, ASP, CBP, DIP, GRP, HCP, HGP, MAP, MCP, MHP, MIP, MJP, MPP, NAP, NBP, PPP, PTP, PVP, SRP, SSP, TCP, TDP, ZPP	LSP
Pepsin inhibitor-3-like repeated domain	PF06394	III, IV, V	ASP, *AV*, CAEEL, DIP, OOP, OVP, *PL*, *TF*	ACP, ALP, AYP, BMP, CBP, HCP, HGP, LSP, MAP, MCP, MHP, MIP, MJP, MPP, PEP, PPP, PTP, SSP, TCP, TDP	
Tas retrotransposon peptidase A16	PF05585	III, V	CAEEL	CBP	ALP
Chromadorea ALT proteins	PF05535	III	*AV*, BMP, DIP, OVP, WBP	ALP, LSP	

Table 5.6.

NemDom3.0 update: Nematode-restricted Pfam-A domains found in non-caenorhabditid nematodes.

Legend overleaf

Table 5.6 (previous page)

NemDom3.0 update: Nematode-restricted Pfam-A domains found in non-caenorhabditid nematodes.

The taxonomic distribution of the seven domain-subset of nematode-restricted Pfam-A domains was queried using the NemDom3.0 annotations.

(1) Nematode clades not previously known to contain the domain but now found in NemDom3 are in blue. If a domain is characterised in a clade but not found in NemDom3, the clade is in red.

(2) *Previously known* are those species associated with the domain in the Pfam database, and the domain is found in NemDom3.

Acquired are species for which the domain is assigned in NemDom3, but had not been previously characterised

Missing are those species associated with the domain in the Pfam database, but which is not found in NemDom3.

The species identifiers and Pfam descriptions are available from Table 3.1 and Table 5.3, respectively.

The DUF148 (PF02520) domain is found in 32 *C. elegans* and 22 *C. briggsae* proteins, but relatively fewer EST-derived proteins. This may be a consequence of an expansion in the caenorhabditid lineage or a reflection of low expression level of this protein. The *C. elegans* proteins that contain the DUF148 domain are represented by a mean of seven ESTs per gene, and 0.09% of available *C. elegans* ESTs. The expression of EST contigs in non-caenorhabditid nematodes is variable (Table 5.7), but, strikingly, is two orders of magnitude greater in *A. suum* and *Toxocara canis* (both ascaridomorphs) compared to *C. elegans*. For these two species the ESTs were predominantly expressed in the adult stage, although this may reflect a bias in the type of cDNA library. No DUF148 domains were identified in the five spiruromorph species, from the sister order to the Ascaridomorpha. Comparison of the DUF148 containing proteins to the draft *B. malayi* proteome revealed a number of proteins which shared sequence similarity, especially for the DUF148 region, and the domain could be identified 15 times in the *B. malayi* proteome with a Pfam-A search. It is probable, therefore, that the gene encoding this protein in the spiruromorphs is expressed at levels too low to be detected through an EST survey. The absence from the Dorylaimia of the nematode fatty acid retinoid binding protein (PF05823) and Pepsin inhibitor-3-like repeated domain (PR06394) also warrants inspection and demands more robust functional annotation and explanation of their role in the evolution of parasitism.

Species	Number of EST contigs	Number of ESTs	ESTs per protein	Proportion of total EST collection
ACP	6	15	2.50	0.0016
ASP	20	737	36.85	0.0192
AYP	5	28	5.60	0.0026
CAEEL	32	228	7.13	0.0009
CBP	22	N/A	N/A	N/A
GRP	1	2	2.00	0.0003
HCP	10	35	3.50	0.0016
HGP	1	8	8.00	0.0003
OOP	8	26	3.25	0.0039
PEP	1	9	9.00	0.0047
PPP	3	4	1.67	0.0047
PTP	3	5	1.67	0.0007
SSP	2	6	3.00	0.0005
TCP	1	64	64.00	0.0138
TDP	2	9	4.50	0.0021

Table 5.7

Expression profiles of proteins containing the domain DUF148.

The EST to protein mappings for *C. elegans* were extracted from WormBase, and those for the parasitic nematodes from NEMBASE.

There is one domain that, according to the Pfam database is restricted to the filarial (spiruromorph) nematodes. The ALT domain (PF05535) has an unknown function but has been shown to be a possible vaccine candidate. Its expression is primarily in the L3 larval stage while developing in the mosquito vector. The domain was first characterised through an abundant mRNA in *B. malayi* [256], and homologues were found in *Dirofilaria immitis* [257], *O. volvulus* [258], *Acnthocheilonema viteae* [259] and *W. bancrofti*. Use of the EST datasets has identified a large number of *alt* genes in *B. malayi*, of which seven have been confirmed through cDNA cloning. A BLAST search against the *C. elegans* genome provided a match in which the *C. elegans* sequence shared the 'core' region in the third and fourth exons [221]. This region did not correspond to a characterised gene in the WormBase, but has been confirmed as a cDNA and the missing 5' end has been sequenced using 5' RACE PCR (William Gregory pers. comm) and is awaiting inclusion in the *C. elegans* proteome. The lack of a WormPep sequence meant that the Pfam curators could not include *C. elegans* in the family, thus restricting it to the Spiruromorpha. Mining NemDom3 revealed the ALT domain in *B. malayi* (27 clusters), *O. volvulus* (19 clusters), *D. immitis* (one cluster) and *W. bancrofti* (6 clusters), as well as *Litomosoides sigmodontis* (two clusters) which was expected. Four ALT containing clusters had previously been identified through BLAST searches in *Ascaris lumbricoides* (one cluster) and *A. suum* (three clusters). Of these only ALP00802 has been assigned the ALT domain. A detailed search of the clusters from *A. suum* showed that they share similarity with the spiruromorph *alt* genes and, more critically, when the bit score threshold for a HMM search was relaxed, the Pfam-A model was found in all three, with highly significant e-values (e-7 to e-14). The matches were all the 'core' region, with upstream region of the domain absent from the ascarid proteins. Inspection of the alignment between the protein and EST contig showed that the translations for ASP12236 and ASP00978 cover the 5' end of the contig, starting at nucleotide positions 94 and 62 respectively. These proteins were not assigned an ALT domain because the PF005535 model is built from an alignment of only five spiruromorph proteins and the 'core' region is invariant in many positions (Figure 5.8). The problem of restrictive models is discussed below (5.4.5). It is hoped that once an alignment using the recently characterised *alt* genes is ready, a more representative domain model can be built and additional ALTs entered in a future a release of NemDom.

Figure 5.8 (overleaf)

Pfam-A ALT domain.

The ALT protein domain was first characterised in filarial nematodes but was recently been found as a mis-identified in *C. elegans*.

(A) The alignment used by Pfam curators to build the Pfam-A model for the ALT domain. The domain regions are mapped according to Gregory *et al.* [221].

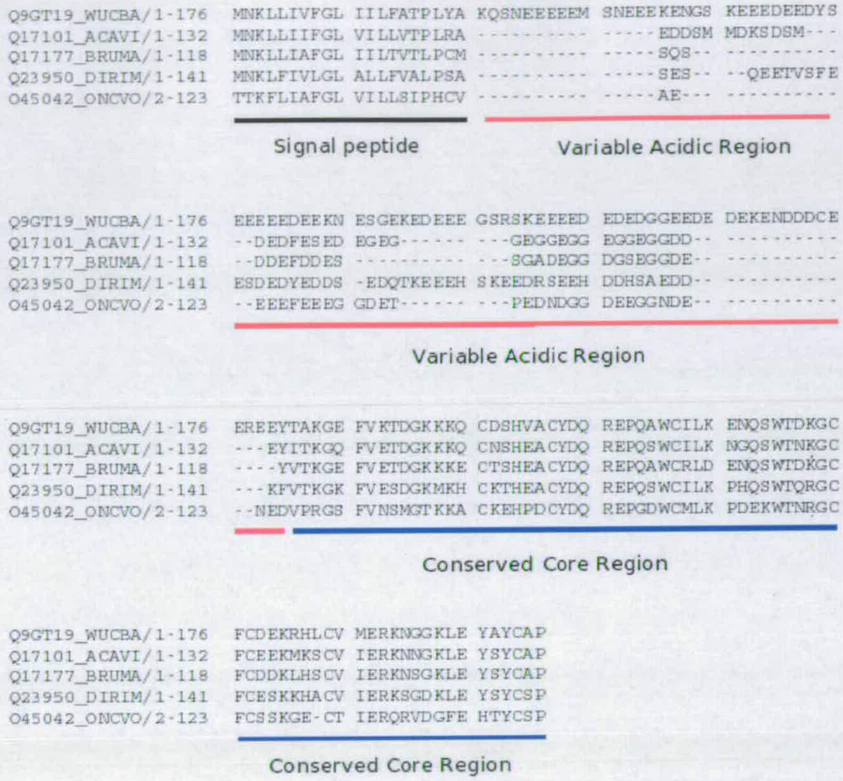
(B) A HMM logo easing visualisation of the variation in the model alignment (position: 105aa to 176aa). The graphic shows only the conserved core region of the domain, and is intended to highlight why this model fails to identify ALT domain characterised in non-filarial nematodes. The tall letters show that these positions are, in practice, invariant in the model. Therefore if they differ in the query sequence a punitive score is calculated, and the domain is not assigned.

Species key: WUCBA – *Wuchereria bancrofti*; ACAVI – *Acnathocheilonema viteae*; BRUMA – *Brugia malayi*; DIRIM – *Dirofilaria immitis*; ONCVO – *Onchocerca volvulus*

Figure 5.8

Pfam-A ALT domain.

legend on previous page

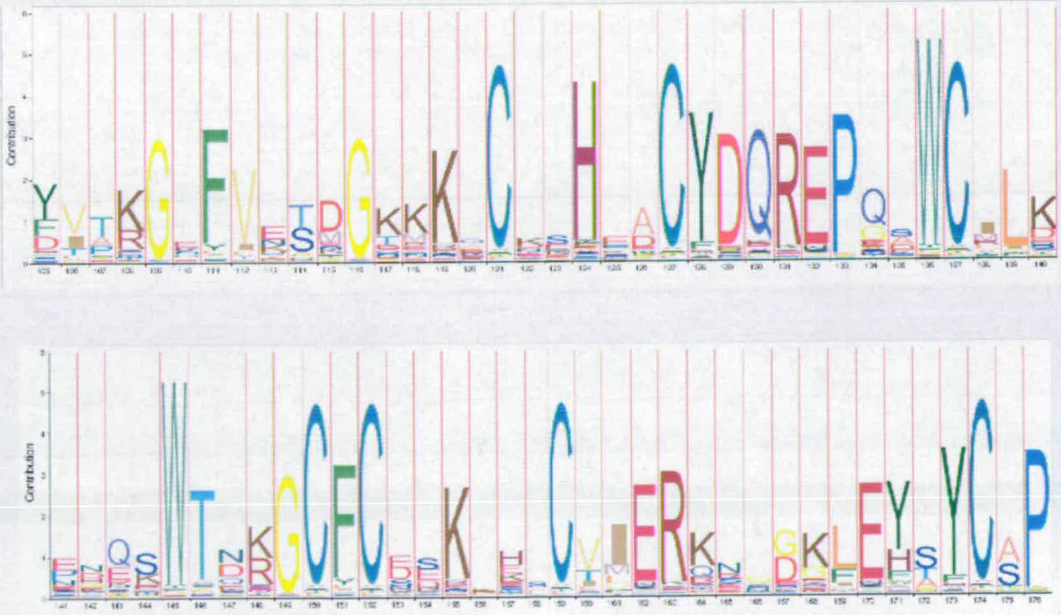


```
Q9GT19_WUCBA/1-176   MNKLLIVFGL IILFATPLYA KQSNEEEEEM SNEEEKENGS KEEEDEEDYS
Q17101_ACAVI/1-132   MNKLLIIFGL VILLVTPLRA ---------- -----EDDSM MDKSDSM---
Q17177_BRUMA/1-118   MNKLLIAFGL IILTVTLPCM ---------- -----SQS-- ----------
Q23950_DIRIM/1-141   MNKLFIVLGL ALLFVALPSA ---------- -----SES-- --QEETVSFE
O45042_ONCVO/2-123   TTKFLIAFGL VILLSIPHCV ---------- -----AE--- ----------

Q9GT19_WUCBA/1-176   EEEEEDEEKN ESGEKEDEEE GSRSKEEEED EDEDGGEEDE DEKENDDDCE
Q17101_ACAVI/1-132   --DEDFESED EGEG------ ---GEGGEGG EGGEGGDD-- ----------
Q17177_BRUMA/1-118   --DDEFDDES ---------- ---SGADEGG DGSEGGDE-- ----------
Q23950_DIRIM/1-141   ESDEDYEDDS -EDQTKEEEH SKEEDRSEEH DDHSAEDD-- ----------
O45042_ONCVO/2-123   --EEEFEEEG GDET------ ---PEDNDGG DEEGGNDE-- ----------

Q9GT19_WUCBA/1-176   EREEYTAKGE FVKTDGKKKQ CDSHVACYDQ REPQAWCILK ENQSWTDKGC
Q17101_ACAVI/1-132   ---EYITKGQ FVETDGKKKQ CNSHEACYDQ REPQSWCILK NGQSWTNKGC
Q17177_BRUMA/1-118   ----YVTKGE FVETDGKKKE CTSHEACYDQ REPQAWCRLD ENQSWTDKGC
Q23950_DIRIM/1-141   ---KFVTKGK FVESDGKMKH CKTHEACYDQ REPQSWCILK PHQSWTQRGC
O45042_ONCVO/2-123   --NEDVPRGS FVNSMGTKKA CKEHPDCYDQ REPGDWCMLK PDEKWTNRGC

Q9GT19_WUCBA/1-176   FCDEKRHLCV MERKNGGKLE YAYCAP
Q17101_ACAVI/1-132   FCEEKMKSCV IERKNNGKLE YSYCAP
Q17177_BRUMA/1-118   FCDDKLHSCV IERKNSGKLE YSYCAP
Q23950_DIRIM/1-141   FCESKKHACV IERKSGDKLE YSYCSP
O45042_ONCVO/2-123   FCSSKGE-CT IERQRVDGFE HTYCSP
```



5.4.5 Caenorhabditid-restricted domains

A large number of protein domains were originally found exclusively in *C. elegans*, and latterly *C. briggsae*. Given the expansion in taxonomic membership described above for a subset of domains it was likely that these caenorhabditid-restricted domains would be found in other nematodes. It was therefore a surprise when only 24 domains with this profile were found in non-caenorhabditid proteomes (Figure 5.9). There are three possible explanations for the presence of so many domains specific to one genus of the Nematoda:

1) It is a true reflection of taxonomic distribution. The domains 'arose' after the Rhabditoidea – Strongyloidea split. An alternative, though less parsimonious explanation, is that some of the domains were present in an older lineage but subsequently lost multiple times.
2) Incomplete sampling of non-caenorhabditid nematodes has missed those sequences which contain the domains. This is especially likely to be the case if the protein domains are in genes that are expressed at low levels.
3) The models which characterise the domain alignments are too specific. Caenorhabditid-restricted domains were identified by comparing proteins within the *C. elegans* proteome [75]. The alignments in these models may contain a large proportion of invariant positions, and thus not reflect any deep evolutionary variation in the domain. This is especially likely to be true if the domain family contains a large number of inparalogues.

The most probable scenario is that all three possibilities contributed to this observation. However, only the third could be tested readily. To investigate whether proteins from non-caenorhabditid proteomes could be assigned to these domains, the caenorhabditid proteins from each domain family were used as a query sequence in BLAST searches of the partial proteomes. The sequence coordinates of any significant matches were then compared to the location of the domains on the query proteins.

A domain of unknown function (DUF225), PF02795, is found in seven caenorhabditid proteins, repeated up to five times. The domain contains 4 conserved cysteine residues which presumably form disulphide bridges, suggesting an extracellular role for proteins with this domain. The domain model was constructed from an alignment that contained both *C. briggsae* and *C. elegans* proteins. Only one protein from the partial proteomes met the criteria to be considered as a new member to this domain. NAP00385_1 was from the human hookworm and sister taxa to the Rhabditoidea, *Necator americanus*. It had strong BLAST similarity to two PF02795-containing proteins, one from each caenorhabditid, with the conserved cysteines present. The *N. americanus* protein could be aligned to the model for PF02795 but the bit score for the alignment was -16.9, below the Gathering threshold (GA) applied to searches to construct NemDom3. The Noise cutoff (NC) for the model was -27.8, suggesting that this domain was in fact present on the protein NAP00385_1.

It is important to retain a level of skepticism when assigning function in an automatic fashion. The criteria described above (using an E value cut off) that putatively added NAP00385_1 to the family PF02795, thus widening its taxonomic distribution, identified six proteins from parasitic species that may contain the DUF1280 domain (PF06918). However, for all six the region identified as DUF1280 overlapped with another predicted domain, PF03357. The 'ESCRT-III complex subunit' is implicated in protein sorting and transport from the endosome to the vacuole or lysosome in eukaryotic cells. The BLAST alignments between the caenorhabditid-restricted domain and the proteins from other nematodes had a high proportion of identical and positive positions (41-60% and 66-76%). I, therefore, suggest that DUF1280 is related to, and possibly a divergent form of ESCRTIII.

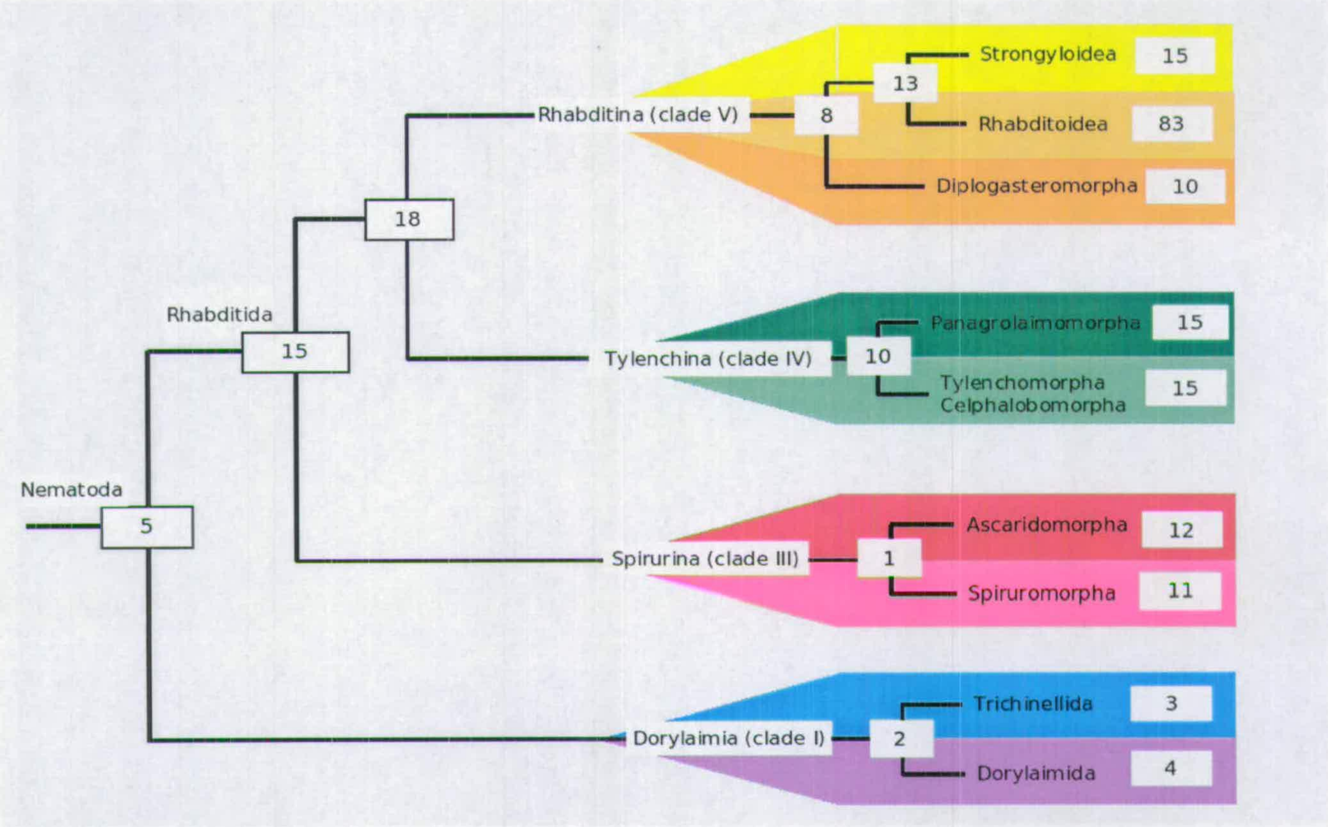


Figure 5.9

Distribution of nematode-restricted Pfam-A domains.

legend overleaf

Figure 5.9 (previous page)

Distribution of nematode-restricted Pfam-A domains.

The tree represents the phylogenetic relationships of the Nematoda as presented by Mark Blaxter [40]. There were 84 Pfam-A domains that, at the time of writing, were found only in the nematodes, mostly in *C. elegans*. The creation of NemDom3 allowed me to ask how widespread these domains were in the phylum. At each node is described the number of domains shared by daughter taxa. A domain may be described in more than one node, which differs from previous uses of this visualisation approach (e.g. Figure 5.6). For example, the 10 domains shared by the Tylenchina are included in each order's haul of 15 domains. The ALT domain has been identified in *C. elegans* as a PCR clone and therefore the Rhabditoidea should contain all 84 domains. However the *C. elegans* ALT protein is absent from the current release of gene models, and is therefore excluded from the *C. elegans* proteome (WormPep). It is because of this, and a desire for continuity of data that the ALT domain is considered missing from the Rhabditoidea in this figure.

5.4.6 Domains not previously found in nematodes

The complete proteomes of *C. elegans* and *C. briggsae* are the source of 95% of nematode protein sequences in the major databases. Therefore, the majority of inferences of protein evolution in the Nematoda are based solely on this lineage. The analysis of NemPep3 in Chapter Four revealed that almost 3,000 nematode protein groups (putative families) did not have a match from the *C. elegans* proteome but shared similarity with a sequence from a non-nematode species. I argued that some of these represented gene loss events in the caenorhabditid lineage. Turning to protein domains, it is clear that there has been domain loss or modification in *C. elegans*; before this study eleven domains, found in multiple phyla, were already known to be present in nematodes excluding *C. elegans*. This number was increased to 31 domains, and possibly 200 plus, with the first sweep of NemDom3.0 annotation (see 5.4.3). There were, however, some domains whose taxonomic distribution did not suggest that they were present in an older lineage and subsequently lost in certain nematodes. These domains present the possibility that other genetic mechanisms were responsible for their presence in the organism. In this section I describe an analysis on the presence of two groups of Pfam-A domains in the nematode proteomes – those thought to be specific to bacteria and those previously found in vertebrate species.

Prokaryote-restricted domains

The are 2,090 Pfam-A models found exclusively in bacterial proteomes. Their presence in the proteome of a nematode would be a consequence of either: (1) bacterial contamination in cDNA library construction, (2) cloning of sequences from a symbiont of the nematode (e.g. *Wolbachia* in spiruromorphs or *Xiphinematobacter* in *Xiphinema*), (3) convergent evolution of a nematode gene or (4) horizontal gene transfer (HGT). These have been introduced earlier in the chapter (5.2.6). Horizontal transfer of a gene from a bacterium to a metazoan is extremely rare [260], so a confirmed instance would be an exciting find, as the fixation of the gene into the genome suggests that it would provide a necessary function. If the function was linked to the organisms parasitic life-cycle it would be a putative target for anti-nematode drugs. Therefore, it is important to distinguish between those NemPep3 proteins that habour a prokaryote domain due to HGT and those which are a result of contamination, either as an experimental artefact or from an obligate symbiont.

Fifty-one models of the 2,090 bacterial-specific domain models were found in 56 proteins from non-caenorhabditid species. Forty (78%) of the models were identified in just two species, *B. malayi* (7 models) and *O. volvulus* (35 models). It was striking that the remaining three spiruromorphs, *Litomosoides sigmodontis*, *Dirofilaria immitis* and *W. bancrofti*, contained only three prokaryote restricted domains (LSC - 1, WBC - 2), which are described below. To identify the possible sources of these domains, serial BLAST searches were performed against different protein databases. The nematode proteins were compared to sequences from species of *Wolbachia*, *Escherichia* and *Pseudomonas*. All 56 sequences had a significant match to *Escherichia coli* proteins, with 34 producing significant alignments with *Pseudomonas*. However, only seven NemPep3 sequences shared significant similarity with a protein from a *Wolbachia* species, of which six were more similar to *Pseudomonas* and *E. coli* proteins. This suggests that 55 proteins are derived from contamination during cDNA library construction. There was one protein from *B. malayi* that was most similar to a sequence from the *Wolbachia* symbiont of *B. malayi*. A search against the draft *B. malayi* nuclear proteome failed to record a match, affirming that the ESTs in that cluster are derived directly from the symbiont.

Level of bacterial contamination

The difficulties compromising the search for horizontally transferred genes highlighted the high level of bacterial contamination in the *O. volvulus* dataset. To examine whether there is widespread contamination, BLAST comparisons between NemPep3 and the UniProt database were examined closely in an attempt to identify how many EST contigs are of bacterial origin. For most species, the proportion of proteins with an apparent bacterial origin was one to six percent (table 5.8). For *C. elegans*, 2%, and *C. briggsae* 3% of proteins yielded a putatively bacterial source identification. However the level of bacterial proteins in both *A. suum* (10%) and *O. volvulus* (15%) datasets were significantly greater (t-test; df = 38, p-value < 2.2e-16). *E. coli* was a major contaminant in the datasets of both species, although the *A. suum* proteome also contained a large number of sequences very similar to *Mycoplasma penetrans* and *Campylobacter jejuni*. The majority of contaminated *A. suum* sequences were derived from seven of the 24 libraries available for *A. suum*, of which six are from a common source. The contaminant protein sequences found in *O. volvulus* were derived from the same two cDNA libraries that were responsible for EST contigs with no detectable coding region (see table 3.5).

Table 5.8

Level of prokaryote contamination of NemPep3.

legend overleaf

Species Dataset	EST contigs	
	Number	Proportion
ACP	43	0.026
ALP	3	0.008
ASP	380	0.103
AYP	36	0.018
BMP	136	0.043
CAEEL	240	0.019
CBP	305	0.029
DIP	30	0.044
GPP	35	0.028
GRP	26	0.018
HCP	38	0.013
HGP	162	0.036
HSP	22	0.032
LSP	53	0.069
MAP	30	0.026
MCP	27	0.017
MHP	80	0.026
MIP	77	0.026
MJP	32	0.024
MPP	11	0.014
NAP	12	0.013
NBP	0	N/A
OOP	34	0.029
OVP	302	0.152
PEP	8	0.032
PPP	54	0.023
PTP	31	0.020
PVP	4	0.011
RSP	3	0.012
SRP	35	0.017
SSP	38	0.017
TCP	23	0.033

continued overleaf...

Species Dataset	EST contigs	
	Number	Proportion
TDP	9	0.011
TMP	9	0.012
TSP	42	0.024
TVP	5	0.010
WBP	51	0.055
XIP	51	0.018
ZPP	5	0.030

Table 5.8

Level of prokaryote contamination of NemPep3.

A BLAST search of NemPep against UniProt was performed. Those nematode proteins that matched only prokaryote proteins (E < e-10) were considered contaminants.

Vertebrate domains

Nematode parasites of mammals survive in extracellular locations such as the lymphatic system, gastrointestinal tract and blood stream. In these exposed locations the nematode must divert, suppress or subvert the host's immune response through regulation [261]. Research into how the parasite down modulates the immune system has usually been performed at the individual gene level [262,263], but EST data has presented a large number of potential immunomodulators [51]. So far all parasite encoded immunomodulators discovered have been homologues of mammalian genes, and are thus regarded as ancient families. There are cases where a gene duplication in the nematode lineage has permitted one copy to undergo convergent evolution to mimic the function of the host gene.

Many of the protein domains specific to vertebrates identified in the nematode ESTs are involved in the immune system, such as the class II histocompatibility antigen (PF00969), cytokines like Interleukin-1 (PF00340) and certain receptors like the one for Interferon gamma (PF07140). It is the function of these molecules that the nematode must copy if it is to modulate the immune response of its host. One mechanism to achieve this is the convergent evolution of a nematode protein to a mammalian one. To detect such instances I generated a sub-set of Pfam-A domain models that were found in vertebrate species to the exclusion of the Arthropoda and Nematoda. It is not sufficient to restrict the collection to vertebrate-only models because certain domains integral to immune regulation (e.g. Class I Histocompatibility antigen (PF00129), are found in viruses which have been acquired from the vertebrate host [264]

NemDom3.0 contained 34 domains that are otherwise vertebrate restricted from 57 separate proteins. Excitingly, the matched domains included a number involved in the immune system, such as histocompatibility antigens, intercellular adhesion molecules and immunoglobulin constant chain domains. It was also striking that 23 (40%) of the proteins were from the filarial nematodes, for whom immunomodulation is the subject of intense study. The analysis of HGT candidates (above) showed that contamination accounted for all the putatively interesting findings. Therefore, it was imperative that each instance in this data set be considered in detail. The model-sequence alignments were inspected, in conjunction with BLAST similarity searches of the UniProt database. Finally, the sequences were compared to both the *C. elegans* and (draft) *B. malayi* proteomes. Only four domain families could be confidently confirmed in the nematode proteomes (Table 5.9), none of which are annotated as involved in the immune response. Three of these domains (PF07967, PF04970, PF07092) had been identified as part of an ancient family lost to *C. elegans*, but finding similarity to the *B. malayi* proteome confirms their validity. For domain PF07967, the parasitic nematode representatives share weak similarity (e-5 to e-6) with a protein from *C. elegans*, C49H3.9, which has no Pfam-A annotation. This *C. elegans* protein shares significant similarity with a human sequence (UniProt ID: Q86WB0) which is annotated as a Zinc finger C3HC-type protein. This domain is found in NIPA proteins (Nuclear interacting partner of anaplastic lymphoma kinase), and the protein from *Schizosaccharomyces pombe* containing this domain has been shown to be involved in mRNA export from the nucleus [265]. The region of similarity between the human and *C. elegans* proteins includes the domain, suggesting that the domain sequence has been modified, either by positive selection to a new function or by neutral drift, as the function is now redundant in *C. elegans*. This would also explain why the protein from *Meloidogyne incognita* does not share significant similarity with the *B. malayi* proteome. The Tat binding domain was found only in *Ancylostoma caninum*, and shares similarity with a protein from *B. malayi* in the region covered by the domain.

Pfam-A accession	Pfam-A description	NemPep ID	*B. malayi* match[1]	*C. elegans* match
PF07106	Tat binding protein 1 -interacting protein (TBPIP)	ACP05087_1	BmaPep - 14992_11292	none
PF07967	C3HC zinc finger-like	MIP05281_1	none	C49H3.9
		WBP02124_1	BmaPep – 14972_7621	C49H3.9
		BMP10466_2	BmaPep – 14972_7621	C49H3.9
		DIP00043_1	BmaPep – 14972_7621	C49H3.9
		HGP11288_1	BmaPep – 14972_7621	C49H3.9
		LSP00284_1	BmaPep – 14972_7621	C49H3.9
PF04970	NC domain	MHP00115_1	BmaPep – 14972_7019	none
		MJP00561_1	BmaPep – 14972_7019	none
		PTP00339_1	BmaPep – 14972_7019	none
		ACP03654_1	BmaPep – 14972_7019	none
		BMP15955_1	BmaPep – 14972_7019	none
		MCP04473_1	BmaPep – 14972_7019	none
		HCP04887_2	BmaPep – 14972_7019	none
		PTP00339_1	BmaPep – 14972_7019	none
		ACP03654_1	BmaPep – 14972_7019	none
PF07092	DUF1356	TSP04007_1	BmaPep – 14990_8051	none
		XIP01189_1	BmaPep – 14990_8051	none
		HCP03094_1	BmaPep – 14990_8051	none

Table 5.9

Are the vertebrate restricted domains really found in nematodes?

The NemPep3 proteins that were decorated with vertebrate domains not previously identified in nematodes were compared to the draft *B. malayi* proteome and WormPep (*C. elegans*). Those proteins that had a significant match are reported in this table and described in the accompanying text.

5.5 Further discussion and conclusion

This chapter describes the robust annotation of NemPep3 with protein domain annotation from the Pfam-A database to create NemDom3.0. This involved exploring the parameter space offered by the profile HMM searches to ensure that domains were assigned correctly to sequences which contain a number of ambiguous or erroneous characters. The problem of domain identification was compounded two-fold. If, as was predicted to be frequently the case, the EST contig did not cover the mature mRNA, the derived protein would be incomplete. This meant that in annotating NemPep3, I had to be mindful of regions at the termini of the sequence that had matches which were local with respect to the domain model. Benchmarking of the process suggested that a combination of Pfam-A global and local models would produce optimal results.

NemDom3.0 is the first attempt at carefully identifying protein domains in EST-derived sequences, and also represents the largest collection of domain annotation for a restricted taxonomic group. Combining NemDom3.0 with the nematode phylogenetic tree identified domains found throughout the Nematoda but absent in the *Caenorhabditis* proteomes.

One branch of research into the nematodes that parasitise mammals looks at the organisms' ability to modulate the response of their hosts immune system. To date, the discovery of such genes has identified homologues between nematode and host where the parasite's protein has undergone convergent evolution. It was hoped that using Pfam-A annotation would identify candidate proteins that perhaps shared no direct ancestry with the host protein and arose due to convergent evolution. Unfortunately, no such domains were found, with many hopeful annotations likely to be due to contamination. Given that convergent evolution usually affects a small number of key, often enzymatically important residues, it was perhaps unsurprising that models built from taxonomically restricted alignments, likely to lack much variation, failed to identify candidate proteins.

The presence of domains from contaminating organisms resulted in a large number of prokaryote-specific domains in the datasets, highlighted in the search for possible horizontally transferred proteins. The issue of identifying true contamination is a difficult one to satisfactorily overcome. In Chapter Three, I identified 8,072 EST contigs, 7% of all non-caenorhabditid ESTs, whose coding region could not be identified through BLAST similarity or hexamer-frequency characteristic of that species. These ESTs were probable contaminants and removed from further analyses. However, given the number of EST contigs that shared similarity only to prokaryote proteins without tell-tale signs of HGT, the original estimate of contamination was too low. Possible improvements to screening methods are discussed in the next chapter.

Between 30-40% of EST-derived amino acids were covered by a Pfam-A domain. It is likely that there are other evolutionary conserved units present in these sequences that have yet to be discovered; in particular those specific to the nematodes. Several future improvements for subsequent releases of NemDom are possible. The SuperFamily domain library [266] stores HMMs for structural inference and, it is hoped, will provide a more powerful way to detect localised convergence of structural (functional) motifs. To identify these domains, polypeptide regions not covered by a Pfam-A domain and longer than 30 amino acids will be passed through the ProDom software [210]. These resulting clusters will be aligned and used to build profile HMMs, which can be used to search the proteomes of other organisms to determine their taxonomic distribution. It will be important to repeat such a search on a regular basis as more sequence becomes available from previously neglected taxa, especially EST data.

Chapter Six - Summary Discussions

The work described in this thesis not only explores the diversity of the nematode proteome, but also provides a framework for other comparative studies involving EST data and provides a clear direction for future work to ensure robust analysis of partial proteomes.

6.1 Getting more from an EST analysis

It is common practice to reduce the redundancy of an EST dataset by clustering the sequences and performing analysis on the consensus-contigs of each cluster. Contemporary analysis of an EST dataset then involves performing BLAST searches against popular databases such as GenBank and UniProt, and the most significant hit providing annotation in the form of a description line and frequently a gene ontology (GO) term. EST contigs are highlighted if they lack similarity to another sequence (species-specific), or are high-expressed (containing many ESTs), or are made up by ESTs exclusively from a single cDNA library. Published reports often conclude proposing the future direction of the project, which is inevitably towards microarray experiments. In my opinion this represents a missed opportunity to ensure robust annotation of the dataset thus promoting better targeted, hypothesis-driven studies. One aspect of the work presented in this thesis was a technical appraisal to improve sequence annotation and reduce the possibility for erroneous interpretation to be propagated through the database. The study of multiple datasets in this thesis made it imperative that I establish robust annotation of the EST contigs, through the recognition of their shortcomings and implementing biologically relevant processes. Here I outline a framework for a comparative phylogenomic approach of EST data.

1. Identification of the coding region

The protein sequences present a better template for almost all approaches of assigning annotation, including domain determination, construction of more accurate multiple sequence alignments, the creation of protein-mass fingerprint libraries for proteomic studies and structural threading and modelling to provide secondary and tertiary structures. It is necessary to identify the coding region of an EST contig, which is complicated by frame-shifts, ambiguous nucleotides and the presence of untranslated regions. Previously published software that corrected for these errors require a level of species-specific training data that is unrealistic for the majority to EST projects. In Chapter Two, I described prot4EST, an EST translation program that implemented several algorithms to achieve the most accurate coding region predictions. Benchmarking prot4EST with ESTs from *Caenorhabditis elegans* highlighted that the performance of the ESTScan algorithm was the most variable, and dependent upon the sequence composition of the training set. I showed that training ESTScan with mRNA from a plant or bacterium with a similar sequence composition to *C. elegans* produced better predictions than a training set from another nematode. This observation prompted the use of simulated training sets (Chapter Three) in which the *C. elegans* proteome, WormPep, was reverse-translated using the codon distribution for the nematode under study. Thus it provided the minimum number of nucleotides in the training set required by ESTScan. This approach should be adopted for other species, though the choice of model proteome is still a matter of debate.

2. Probable contamination

The identification of EST contigs, whose presence in the dataset is a consequence of contamination, is paramount to avoid large scale errors in analysis. There are two, relatively simple approaches to uncover such contigs. First, in depth analysis of NemPep3 showed that those EST contigs which were translated using either BLAST-similarity or ESTScan (hexamer composition) components of prot4EST could confidently be assigned to the transcriptome of the organism. If a coding region could not predicted by either of these stages, it was likely that the EST contig represented contamination in either cDNA library construction or a sequencing artifact. It was noteworthy that nearly 90% of EST contigs in this second category were singletons, compared to 65% across the entire nematode EST dataset. Secondly, inspection of BLAST search reports is likely to reveal bacterial contamination. The criteria for such a search depends predominantly on the species being studied. For example, the nematode EST contigs that shared very high level of sequence identity (<95%) with proteins from bacteria known to contaminate cDNA libraries (*Escherichia coli*, *Shigella* sp. and *Pseudomonas* sp.) without similar or better matches to eukaryote proteins were probably not of nematode origin.

3. Protein domains

The delineation of protein sequences into conserved building blocks, or protein domains, is an exceedingly useful template from which putative function and evolutionary histories can be elucidated. I have explored the effect of using two representations of the domain model, one which forms alignments which are global with respect to the model (and local to the query sequence), and one whose alignments are local for both the model and query. Standard analysis, using full-length protein sequences, favours searches performed with more accurate global models, but EST-derived proteins are likely to be incomplete sequences, and as such contain only part of the domain. Comparing EST contigs from *C. elegans* with WormPep advocated a hierarchical approach, in which the results from a search with global models are acceptable matches and alignments formed with local models are only approved if they occur at either the 5' or 3' end of the proteins sequence. The inclusion of partial domains led to a 18% increase in the number of sequence regions annotated, corresponding to an additional 10% of Pfam identifiers.

4. Confirmation of observations

Comparative studies often point to the absence or acquisition of a protein or domain from one particular lineage over another, or the difference in expression level (EST number). Such observations are usually the most interesting and passionately reported. However, occasionally, such observations in the nematode EST datasets have been a consequence of contamination. It is therefore critical that the comparative study is supplemented by experimental verification, such as isolation and cloning of a particular gene or by expression studies, e.g. SAGE or real time PCR. Once a gene has been identified, elucidating its function becomes the next task. There are a number of methods available such as: RNAi; yeast 2-hybrid to look at interactions, knock out studies, enzyme assays and replacement / rescue of genetic defects in yeast. It may be possible to predict the three dimensional structure of the protein, and thus allow comparisons to other protein structures along with studies of the structure to identify possible catalytic binding sites. Additionally if the encoded protein was possibly involved in a metabolic pathway, then the organism could be fed radiolabelled substrate, and the fate of the products assayed.

6.2 ESTs of the Nematoda

The development of the protocol outlined above has been motivated by the study of multiple EST datasets from predominantly parasitic species of nematodes. Over 340,000 ESTs had been clustered into approximately 120,000 gene objects These had been a source of intense study, predominantly at an individual species level. There were a handful of comparative studies on the taxonomic slices of the dataset [47,117,50], and a comparative investigation across 30 non-caenorhabditid species [51]. The focus of my work has been to concentrate on the proteomes generated from, an expanded set of 37 species, with particular reference to creating a resource that benefits future exploitation of the data. To this end I created prot4EST, a software pipeline to predict the coding regions from error-prone EST contigs. The shortfall in training data experienced by all the taxa was overcome by creating simulated transcriptomes. This was achieved by reverse translating WormPep with the organism's codon usage table. The collection of partial proteomes, combined with those from *C. elegans* and *C. briggsae*, formed NemPep3. This step provided the basis for a number of analyses, enabling comment to be passed on both technical aspects of the dataset and biological rationale.

A total of 67 cDNA libraries exploited the nematode-specific phenomenon of a 5' spliced leader (SL) motif, on which the PCR primers are designed. I have shown that coding regions from EST contigs containing sequences from a mixture of SL-primer and more conventional primer-ligation based libraries were significantly longer and, frequently, better quality, than contigs made exclusively from one type. This should guide future nematode EST sequencing strategies to ensure that the maximum coverage of a tagged genes coding region. The poor quality of many translations from spiruromorph species, detected by a lack of sequence similarity or matching nucleotide hexamer frequency, suggest a high level of sequencing artifacts in a number of cDNA libraries. Comparison between the *Brugia malayi* EST contigs and the organism's draft genome lent support to the dubious nature of these sequences. I therefore considered that there was sufficient doubt to exclude approximately 8,000 contigs (7,100 singletons) that were not translated by BLAST or ESTScan algorithms.

All-against-all BLAST analysis revealed that between 18 and 45% of partial proteome was species-specific (orphan proteins), which was significantly greater than that for the fully-sequenced caenorhabditids (~9%). As sequencing surveys generate more ESTs for currently described datasets or commence for other nematode species more putative homologues will be identified, reducing the proportion of orphan proteins. A small proportion of proteins from partial proteomes shared significant similarity with caenorhabditid sequences previously characterised as orphans, pointing to gene-loss after the *elegans-briggsae* split and that the figure of 9% is likely to be an upper-bound on the level of species-specificity from the currently sampled species. Mapping the similarity searches onto the nematode phylogenetic tree revealed two important features of nematode proteinspace. The first was that 44,000 groups of proteins were seemingly unique to the Nematoda and, secondly, the great extent of gene-loss in certain lineages. The level of novelty in proteinspace will certainly fluctuate in the near future; additional sequencing from nematodes, both ESTs and complete genomes (Table 4.4), will, given the current rate of gene discovery (Figure 4.1), reveal more nematode-specific contigs. This will be offset by the inclusion of non-nematode EST projects in the comparative studies. The assembly of partial genomes for ~300 eukaryotic organisms is currently underway (http://www.partigenedb.org). The next step in the process is to automate creation of simulated training sets for prot4EST to ensure robust coding region predictions. There were almost 3,000 proteins with putative homologues from outside the Nematoda without a match within the *C. elegans* proteome, promoting these as candidates for gene-loss. The collectors curve indicates that such lineage-specific gene-loss is widespread throughout the phylum, but it can only be tested in those proteomes considered complete. This was applicable for the protein domain studies, where 31 domains (possibly as high as 228), were identified in a number of nematode orders but absent in the caenorhabditid proteomes. The mechanisms for gene-loss and gain can differ to those for domain modification, and the work I have presented here should be continued and include the forthcoming *C. remanei* and *B. malayi* proteomes, thus offering insights into *C. elegans* evolution.

6.3 Protein families

It is well known that members of the same protein family share similar, if not identical biochemical functions [194]. A protein family can be defined as a group of polypeptides that are demonstrably related to each other [195]. The criteria most widely used to assemble these families has been sequence similarity, and subsequent clustering [196,197,165,118,199]. A protein family differs from a domain family in that it contains the entire polypeptide sequence rather than inter-sequence conserved fragments. Similarity is usually detected using the BLAST algorithm, primarily because its heuristic search is very fast. There are two approaches that are used throughout molecular biology. The COG system uses symmetrical BLAST hits to delineate relationships [135] and is available through the NCBI. TRIBE-MCL uses Markov flow clustering to group similar sequences [118,200], and is the algorithm of choice for a number of genome projects, including *Caenorhabditis elegans* (Daniel Lawson pers. comm} and *Plasmodium falciparum* [201].

One problem that the clustering methods face are that many proteins consist of multiple independently evolving domains [172,194]. Using BLAST, which detects local regions of similarity, can result in links forged between unrelated proteins. This applies not only to formally classified protein domains, but any shared motif of sufficient size and similarity to be considered significant. The COG system, the authors state, overcomes this problem by manual inspections of multi-domain proteins. However such an approach is labour-intensive and not transferable to the majority of research groups. The TRIBE-MCL program 'does not require any explicit knowledge of protein domains to detect protein families', rather clusters on the observed relationships through the entire similarity graph [118]. However the Markov flow clustering algorithm is dependent upon the inflation parameter, whose value should vary to assemble different protein families correctly [203].

A recent investigation has shown that both these methods fail to correctly resolve the eukaryote hemoglobin protein-family (Wasmuth, Elliot, Schmid and Blaxter *in prep*).

Symmetrical BLAST searches failed to assemble the family, but over 30 *C. elegans* proteins were assigned by the COG curators based on the manual assessment of PSI-BLAST searches. Many of these proteins did not contain the necessary number of α-helices or invariant residues characteristic of globins. The TRIBES database [204] separated related globins into many families, some containing a single member. The similarity statistic used to decorate the edges of the graph is the E value, which is transformed (-log) for the MCL algorithm. While, this is acceptable for very large databases such as UniProt, it is probably inefficient when clustering small datasets, as observed when Chelicerate mitochondrial proteomes were clustered (Jones and Wasmuth unpublished). It is likely that using a similarity statistic independent of the size of the database would yield more faithful families, but this has yet to be assessed.

Given the uncertainty over the robust clustering of full-length protein sequences I considered it more expedient to focus on an investigation of the protein domain complement of the nematode proteomes.

6.4 Future work

At time of writing, a number of analyses have been initiated, which primarily focus around NemDom3:

1. Other domain model libraries, including SUPERFAMILY and SMART, are being searched to identified additional, characterised domains.
2. The Pfam-A models containing whose membership is increased by nematodes, will be rebuilt to include the additional information and used to search NemPep3 for any more diverged members.
3. Regions of NemPep3 proteins which are not decorated with domain annotation and longer than 30 amino acids are passed through the ProDom program to identify

conserved polypeptide regions, akin to nematode-restricted domains. Such an undertaking for the ALT domain will aid the discovery of this domain in other nematode species.

4. NemDom3 will be integrated into NemBase with the inclusion of analysis tools, similar to xdom [267].
5. The identification by rigourous analysis of domains whose copy number or expression levels (EST number) differ from other species.

Completion of these studies will further phylogenomics for the Phylum Nematoda, and identify proteins who should be investigated for their biological relevance, thus providing a promising avenue for the identification for anthelmintic drug targets.

Bibliography

[1] Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W and Lipman DJ (1997) "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs." Nucleic Acids Res 25: 3389--3402.

[2] Bernal A, Ear U and Kyrpides N (2001) "Genomes OnLine Database (GOLD): a monitor of genome projects world-wide." Nucleic Acids Res 29: 126-127.

[3] Müller A, MacCallum RM and Sternberg MJE (2002) "Structural characterization of the human proteome." Genome Res 12: 1625-1641.

[4] Williams KL, Gooley AA and Packer NH (1996) "" Today's Life Science 8: 16-21.

[5] Eddy SR (2001) "Non-coding RNA genes and the modern RNA world." Nat Rev Genet 2: 919-929.

[6] Szymanski M, Barciszewska MZ, Zywicki M and Barciszewski J (2003) "Noncoding RNA transcripts." J Appl Genet 44: 01/01/19.

[7] Blaxter M (2002) "Opinion piece. Genome sequencing: time to widen our horizons." Brief Funct Genomic Proteomic 1: 07/09/06.

[8] Wasmuth JD and Blaxter ML (2004) "prot4EST: translating expressed sequence tags from neglected genomes." BMC Bioinformatics 5: 187.

[9] Boguski MS, Lowe TM and Tolstoshev CM (1993) "dbEST--database for ""expressed sequence tags""." Nat Genet 4: 332-333.

[10] Miller RT, Christoffels AG, Gopalakrishnan C, Burke J, Ptitsyn AA, Broveak TR and Hide WA (1999) "A comprehensive approach to clustering of expressed human gene sequence: the sequence tag alignment and consensus knowledge base." Genome Res 9: 1143-1155.

[11] Burke J, Davison D and Hide W (1999) "d2_cluster: a validated method for clustering EST and full-length cDNAsequences." Genome Res 9: 1135-1142.

[12] Hide W, Burke J and Davison DB (1994) "Biological evaluation of d2, an algorithm for high-performance sequence comparison." J Comput Biol 1: 199-215.

[13] Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W and Lipman DJ (1997) "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs." Nucleic Acids Res 25: 3389-3402.

[14] Quackenbush J, Cho J, Lee D, Liang F, Holt I, Karamycheva S, Parvizi B, Pertea G, Sultana R and White J (2001) "The TIGR Gene Indices: analysis of gene transcript sequences in highly sampled eukaryotic species." Nucleic Acids Res 29: 159-164.

[15] Parkinson J, Guiliano DB and Blaxter M (2002) "Making sense of EST sequences by CLOBBing them." BMC Bioinformatics 3: 31.

[16] Geldhof P, Whitton C, Gregory WF, Blaxter M and Knox DP (2005) "Characterisation of the two most abundant genes in the Haemonchus contortus expressed sequence tag dataset." Int J Parasitol 35: 513-522.

[17] Ewing B and Green P (1998) "Base-calling of automated sequencer traces using phred. II. Error probabilities." Genome Res 8: 186-194.

[18] Ewing B, Hillier L, Wendl MC and Green P (1998) "Base-calling of automated sequencer traces using phred. I. Accuracy assessment." Genome Res 8: 175-185.

[19] Rudd S (2003) "Expressed sequence tags: alternative or complement to whole genome sequences?" Trends Plant Sci 8: 321-329.

[20] Koski LB, Gray MW, Lang BF and Burger G (2005) "AutoFACT: an automatic functional annotation and classification tool." BMC Bioinformatics 6: 151.

[21] Parkinson J, Anthony A, Wasmuth J, Schmid R, Hedley A and Blaxter M (2004) "PartiGene--constructing partial genomes." Bioinformatics 20: 1398-1404.

[22] Paschall JE, Oleksiak MF, VanWye JD, Roach JL, Whitehead JA, Wyckoff GJ, Kolell KJ and Crawford DL (2004) "FunnyBase: a systems level functional annotation of Fundulus ESTs for the analysis of gene expression." BMC Genomics 5: 96.

[23] Boardman PE, Sanz-Ezquerro J, Overton IM, Burt DW, Bosch E, Fong WT, Tickle C, Brown WRA, Wilson SA and Hubbard SJ (2002) "A comprehensive collection of chicken cDNAs." Curr Biol 12: 1965-1969.

[24] Hubbard SJ, Grafham DV, Beattie KJ, Overton IM, McLaren SR, Croning MDR, Boardman PE, Bonfield JK, Burnside J, Davies RM, Farrell ER, Francis MD, Griffiths-Jones S, Humphray SJ, Hyland C, Scott CE, Tang H, Taylor RG, Tickle C, Brown WRA, Birney E, Rogers J and Wilson SA (2005) "Transcriptome analysis for the chicken based on 19,626 finished cDNA sequences and 485,337 expressed sequence tags." Genome Res 15: 174-183.

[25] Rudd S (2005) "openSputnik--a database to ESTablish comparative plant genomics using unsaturated sequence collections." Nucleic Acids Res 33: D622-D627.

[26] Parkinson J, Whitton C, Schmid R, Thomson M and Blaxter M (2004c) "NEMBASE: a resource for parasitic nematode ESTs." Nucleic Acids Res 32: D427-D430.

[27] Wylie T, Martin JC, Dante M, Mitreva MD, Clifton SW, Chinwalla A, Waterston RH, Wilson RK and McCarter JP (2004) "Nematode.net: a tool for navigating sequences from parasitic and free-living nematodes." Nucleic Acids Res 32: D423-D426.

[28] De Ley P and Blaxter M (2002) "Biology of Nematodes" ed. Lee, D.L, Taylor and Francis, London.

[29] Platt HM (1994) "The phylogenect systematics of free-living nematodes" ed. Lorenzen, S., The Ray Society, London.

[30] De Ley P and Blaxter M (2002) "Biology of Nematodes" ed. Lee, D.L, Taylor and Francis, London.

[31] de Meeûs T and Renaud F (2002) "Parasites within the new phylogeny of eukaryotes." Trends Parasitol 18: 247-251.

[32] Lambshead PJD (1993) "Recent developments in marine benthic biodiversity research" Oceanis 19: 01/05/24.

[33] Lambshead PJD, Brown CJ, Ferrero TJ, Hawkins LE, Smith CR and Mitchell NJ (2003) "Biodiversity of nematode assemblages from the region of the Clarion-Clipperton Fracture Zone, an area of commercial mining interest." BMC Ecol 3: 1.

[34] Platonova TA and Gal'tsova VV (1976) "Nematodes and Their Role in the Meiobenthos" Nakua, Leningrad.

[35] C. elegans Sequencing Consortium (1998) "Genome sequence of the nematode C. elegans: a platform for investigating biology." Science 282: 2012-2018.

[36] Stein LD, Bao Z, Blasiar D, Blumenthal T, Brent MR, Chen N, Chinwalla A, Clarke L, Clee C, Coghlan A, Coulson A, D'Eustachio P, Fitch DHA, Fulton LA, Fulton RE, Griffiths-Jones S, Harris TW, Hillier LDW, Kamath R, Kuwabara PE, Mardis ER, Marra MA, Miner TL, Minx P, Mullikin JC, Plumb RW, Rogers J, Schein JE, Sohrmann M, Spieth J, Stajich JE, Wei C, Willey D, Wilson RK, Durbin R and Waterston RH (2003) "The genome sequence of Caenorhabditis briggsae: a platform for comparative genomics." PLoS Biol 1: E45.

[37] Chen N, Harris TW, Antoshechkin I, Bastiani C, Bieri T, Blasiar D, Bradnam K, Canaran P, Chan J, Chen CK, Chen WJ, Cunningham F, Davis P, Kenny E, Kishore R, Lawson D, Lee R, Muller HM, Nakamura C, Pai S, Ozersky P, Petcherski A, Rogers A, Sabo A, Schwarz EM, Auken KV, Wang Q, Durbin R, Spieth J, Sternberg PW and Stein LD (2005) "WormBase: a comprehensive data

resource for Caenorhabditis biology and genomics." Nucleic Acids Res 33: D383-D389.

[38] Anderson FE and Swofford DL (2004) "Should we be worried about long-branch attraction in real data sets? Investigations using metazoan 18S rDNA." Mol Phylogenet Evol 33: 440-451.

[39] Blaxter M and Bird D (1997) "C. elegans II" ed. Riddle, D.L and Blumenthal, T and Barbara, J.M and Priess, J.R, CSHL Press, NY.

[40] Blaxter M (1998) "Caenorhabditis elegans is a nematode." Science 282: 2041-2046.

[41] Chan MS (1997) "The global burden of intestinal nematode infections--fifty years on." Parasitol Today 13: 438-443.

[42] Barker KR, Hussey RS and Krusberg LR (1994) "Plant and soil nematodes: Societal impact and focus on the future" Society of Nematologists, Missouri.

[43] Dorris M, Ley PD and Blaxter ML (1999) "Molecular analysis of nematode diversity and the evolution of parasitism." Parasitol Today 15: 188-193.

[44] Parkinson J and Blaxter M (2003) "SimiTri--visualizing similarity relationships for groups of sequences." Bioinformatics 19: 390-395.

[45] Daub J, Loukas A, Pritchard DI and Blaxter M (2000) "A survey of genes expressed in adults of the human hookworm, Necator americanus." Parasitology : 171-184.

[46] Blaxter M, Daub J, Guiliano D, Parkinson J, Whitton C and Project FG (2002) "The Brugia malayi genome project: expressed sequence tags and gene discovery." Trans R Soc Trop Med Hyg 96: 01/07/17.

[47] Mitreva M, McCarter JP, Martin J, Dante M, Wylie T, Chiapelli B, Pape D, Clifton SW, Nutman TB and Waterston RH (2004) "Comparative genomics of gene expression in the parasitic and free-living nematodes Strongyloides stercoralis and Caenorhabditis elegans." Genome Res 14: 209-220.

[48] Thompson FJ, Mitreva M, Barker GLA, Martin J, Waterston RH, Waterson RH, McCarter JP and Viney ME (2005) "An expressed sequence tag analysis of the life-cycle of the parasitic nematode Strongyloides ratti." Mol Biochem Parasitol 142: 32-46.

[49] McCarter JP, Mitreva MD, Martin J, Dante M, Wylie T, Rao U, Pape D, Bowers Y, Theising B, Murphy CV, Kloek AP, Chiapelli BJ, Clifton SW, Bird DM and Waterston RH (2003) "Analysis and functional classification of transcripts from the nematode Meloidogyne incognita." Genome Biol 4: R26.

[50] Scholl EH and Bird DMK (2005) "Resolving tylenchid evolutionary relationships through multiple gene analysis derived from EST data." Mol Phylogenet Evol 36: 536-545.

[51] Parkinson J, Mitreva M, Whitton C, Thomson M, Daub J, Martin J, Schmid R, Hall N, Barrell B, Waterston RH, McCarter JP and Blaxter ML (2004) "A transcriptomic analysis of the phylum Nematoda." Nat Genet 36: 1259-1267.

[52] Harcus YM, Parkinson J, Fernández C, Daub J, Selkirk ME, Blaxter ML and Maizels RM (2004) "Signal sequence analysis of expressed sequence tags from the nematode Nippostrongylus brasiliensis and the evolution of secreted proteins in parasites." Genome Biol 5: R39.

[53] Zang X and Maizels RM (2001) "Serine proteinase inhibitors from nematodes and the arms race between host and pathogen." Trends Biochem Sci 26: 191-197.

[54] Sommer A, Nimtz M, Conradt HS, Brattig N, Boettcher K, Fischer P, Walter RD and Liebau E (2001) "Structural analysis and antibody response to the extracellular glutathione S-transferases from Onchocerca volvulus." Infect Immun 69: 7718-7728.

[55] Jones SJ, Riddle DL, Pouzyrev AT, Velculescu VE, Hillier L, Eddy SR, Stricklin SL, Baillie DL, Waterston R and Marra MA (2001) "Changes in gene expression associated with developmental arrest and longevity in Caenorhabditis elegans." Genome Res 11: 1346-1352.

[56] Velculescu VE, Zhang L, Vogelstein B and Kinzler KW (1995) "Serial analysis of gene expression." Science 270: 484-487.

[57] Skuce PJ, Yaga R, Lainson FA and Knox DP (2005) "An evaluation of serial analysis of gene

expression (SAGE) in the parasitic nematode, Haemonchus contortus." Parasitology 130: 553-559.

[58] Knox DP and Skuce PJ (2005) "SAGE and the quantitative analysis of gene expression in parasites." Trends Parasitol 21: 322-326.

[59] McCarter JP (2004) "Genomic filtering: an approach to discovering novel antiparasitics." Trends Parasitol 20: 462-468.

[60] Werren JH (1997) "Biology of Wolbachia." Annu Rev Entomol 42: 587-609.

[61] Bandi C, Anderson TJ, Genchi C and Blaxter ML (1998) "Phylogeny of Wolbachia in filarial nematodes." Proc Biol Sci 265: 2407-2413.

[62] Scholl EH, Thorne JL, McCarter JP and Bird DM (2003) "Horizontally transferred genes in plant-parasitic nematodes: a high-throughput genomic approach." Genome Biol 4: R39.

[63] Onwuliri CO (1984) "The effect of anaerobiosis on adenosine triphosphate levels in larval Nippostrongylus brasiliensis and Haemonchus contortus." Z Parasitenkd 70: 667-671.

[64] Valencia A (2005) "Automatic annotation of protein function." Curr Opin Struct Biol 15: 267-274.

[65] Wolf YI, Rogozin IB and Koonin EV (2004) "Coelomata and not Ecdysozoa: evidence from genome-wide phylogenetic analysis." Genome Res 14: 29-36.

[66] Karev GP, Wolf YI and Koonin EV (2003) "Simple stochastic birth and death models of genome evolution: was there enough time for us to evolve?" Bioinformatics 19: 1889-1900.

[67] Makarova KS, Wolf YI, Mekhedov SL, Mirkin BG and Koonin EV (2005) "Ancestral paralogs and pseudoparalogs and their role in the emergence of the eukaryotic cell." Nucleic Acids Res 33: 4626-4638.

[68] Lehnert U, Xia Y, Royce TE, Goh CS, Liu Y, Senes A, Yu H, Zhang ZL, Engelman DM and Gerstein M (2004) "Computational analysis of membrane proteins: genomic occurrence, structure prediction and helix interactions." Q Rev Biophys 37: 121-146.

[69] Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV, Krylov DM, Mazumder R, Mekhedov SL, Nikolskaya AN, Rao BS, Smirnov S, Sverdlov AV, Vasudevan S, Wolf YI, Yin JJ and Natale DA (2003) "The COG database: an updated version includes eukaryotes." BMC Bioinformatics 4: 41.

[70] Mushegian AR, Garey JR, Martin J and Liu LX (1998) "Large-scale taxonomic profiling of eukaryotic model organisms: a comparison of orthologous proteins encoded by the human, fly, nematode, and yeast genomes." Genome Res 8: 590-598.

[71] Hirsh AE and Fraser HB (2001) "Protein dispensability and rate of evolution." Nature 411: 1046-1049.

[72] Coghlan A and Wolfe KH (2002) "Fourfold faster rate of genome rearrangement in nematodes than in Drosophila." Genome Res 12: 857-867.

[73] Lynch M (2002) "Genomics. Gene duplication and evolution." Science 297: 945-947.

[74] Kortschak RD, Samuel G, Saint R and Miller DJ (2003) "EST analysis of the cnidarian Acropora millepora reveals extensive gene loss and rapid sequence divergence in the model invertebrates." Curr Biol 13: 2190-2195.

[75] Sonnhammer EL and Durbin R (1997) "Analysis of protein domain families in Caenorhabditis elegans." Genomics 46: 200-216.

[76] Mulder NJ, Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D, Bradley P, Bork P, Bucher P, Cerutti L, Copley R, Courcelle E, Das U, Durbin R, Fleischmann W, Gough J, Haft D, Harte N, Hulo N, Kahn D, Kanapin A, Krestyaninova M, Lonsdale D, Lopez R, Letunic I, Madera M, Maslen J, McDowall J, Mitchell A, Nikolskaya AN, Orchard S, Pagni M, Ponting CP, Quevillon E, Selengut J, Sigrist CJA, Silventoinen V, Studholme DJ, Vaughan R and Wu CH (2005) "InterPro, progress and status in 2005." Nucleic Acids Res 33: D201-D205.

[77] Bateman A, Coin L, Durbin R, Finn RD, Hollich V, Griffiths-Jones S, Khanna A, Marshall M,

Moxon S, Sonnhammer ELL, Studholme DJ, Yeats C and Eddy SR (2004) "The Pfam protein families database." Nucleic Acids Res 32: D138-D141.

[78] Rost B, Schneider R and Sander C (1997) "Protein fold recognition by prediction-based threading." J Mol Biol 270: 471-480.

[79] Kelley LA, MacCallum RM and Sternberg MJ (2000) "Enhanced genome annotation using structural profiles in the program 3D-PSSM." J Mol Biol 299: 499-520.

[80] Marti-Renom MA, Madhusudhan MS and Sali A (2004) "Alignment of protein sequences by their profiles." Protein Sci 13: 1071-1087.

[81] Li BW, Rush AC, Tan J and Weil GJ (2004) "Quantitative analysis of gender-regulated transcripts in the filarial nematode Brugia malayi by real-time RT-PCR." Mol Biochem Parasitol 137: 329-337.

[82] Gordon D, Abajian C and Green P (1998) "Consed: a graphical tool for sequence finishing." Genome Res 8: 195-202.

[83] Pearson WR, Wood T, Zhang Z and Miller W (1997) "Comparison of DNA sequences with protein sequences." Genomics 46: 24-36.

[84] Zhang Z, Pearson WR and Miller W (1997) "Aligning a DNA sequence with a protein sequence." J Comput Biol 4: 339-349.

[85] Cuff JA, Birney E, Clamp ME and Barton GJ (2000) "ProtEST: protein multiple sequence alignments from expressed sequence tags." Bioinformatics 16: 111-116.

[86] Apweiler R, Bairoch A, Wu CH, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, Martin MJ, Natale DA, O'Donovan C, Redaschi N and Yeh LSL (2004) "UniProt: the Universal Protein knowledgebase." Nucleic Acids Res 32: D115-D119.

[87] Boeckmann B, Bairoch A, Apweiler R, Blatter MC, Estreicher A, Gasteiger E, Martin MJ, Michoud K, O'Donovan C, Phan I, Pilbout S and Schneider M (2003) "The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003." Nucleic Acids Res 31: 365-370.

[88] Wheeler DL, Church DM, Lash AE, Leipe DD, Madden TL, Pontius JU, Schuler GD, Schriml LM, Tatusova TA, Wagner L and Rapp BA (2001) "Database resources of the National Center for Biotechnology Information." Nucleic Acids Res 29: 01/11/16.

[89] Hatzigeorgiou AG, Fiziev P and Reczko M (2001) "DIANA-EST: a statistical analysis." Bioinformatics 17: 913-919.

[90] Lottaz C, Iseli C, Jongeneel CV and Bucher P (2003) "Modeling sequencing errors by combining Hidden Markov models." Bioinformatics : II103-II112.

[91] Fukunishi Y and Hayashizaki Y (2001) "Amino acid translation program for full-length cDNA sequences with frameshift errors." Physiol Genomics 5: 81-87.

[92] Durbin R, Eddy SR, Krogh A and Mitchison GJ (1998) "Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids" Cambridge University Press.

[93] Eddy SR (1998) "Profile hidden Markov models." Bioinformatics 14: 755-763.

[94] Krogh A, Brown M, Mian IS, Sjölander K and Haussler D (1994) "Hidden Markov models in computational biology. Applications to protein modeling." J Mol Biol 235: 1501-1531.

[95] Burge C and Karlin S (1997) "Prediction of complete gene structures in human genomic DNA." J Mol Biol 268: 78-94.

[96] Korf I (2004) "Gene finding in novel genomes." BMC Bioinformatics 5: 59.

[97] Borodovsky M and McIninch J (1993) "Recognition of genes in DNA sequence with ambiguities." Biosystems 30: 161-171.

[98] Solovyev VV, Salamov AA and Lawrence CB (1994) "Predicting internal exons by oligonucleotide composition and discriminant analysis of spliceable open reading frames." Nucleic Acids Res 22: 5156-5163.

[99] Letunic I, Copley RR, Schmidt S, Ciccarelli FD, Doerks T, Schultz J, Ponting CP and Bork P (2004) "SMART 4.0: towards genomic data integration." Nucleic Acids Res 32: D142-D144.

[100] Löytynoja A and Milinkovitch MC (2003) "A hidden Markov model for progressive multiple alignment." Bioinformatics 19: 1505-1513.

[101] Iseli C, Jongeneel CV and Bucher P (1999) "ESTScan: a program for detecting, evaluating, and reconstructing potential coding regions in EST sequences." Proc Int Conf Intell Syst Mol Biol : 138-148.

[102] Pruitt KD, Tatusova T and Maglott DR (2003) "NCBI Reference Sequence project: update and current status." Nucleic Acids Res 31: 34-37.

[103] Pearson WR (1991) "Searching protein sequence libraries: comparison of the sensitivity and selectivity of the Smith-Waterman and FASTA algorithms." Genomics 11: 635-650.

[104] Lottaz C (2002) "Modelling Expressed Sequence Tags using a Hidden Markov Model", Thesis, Swiss Institute of Bioinformatics and Universities of Lausanne and Geneva, Switzerland.

[105] Kozak M (1987) "An analysis of 5'-noncoding sequences from 699 vertebrate messenger RNAs." Nucleic Acids Res 15: 8125-8148.

[106] Min XJ, Butler G, Storms R and Tsang A (2005) "OrfPredictor: predicting protein-coding regions in EST-derived sequences." Nucleic Acids Res 33: W677-W680.

[107] Wuyts J, Perrière G and Peer YVD (2004) "The European ribosomal RNA database." Nucleic Acids Res 32: D101-D103.

[108] Nakamura Y, Gojobori T and Ikemura T (2000) "Codon usage tabulated from international DNA sequence databases: status for the year 2000." Nucleic Acids Res 28: 292.

[109] Phan IQH, Pilbout SF, Fleischmann W and Bairoch A (2003) "NEWT, a new taxonomy portal." Nucleic Acids Res 31: 3822-3823.

[110] Cavener DR (1987) "Comparison of the consensus sequence flanking translational start sites in Drosophila and vertebrates." Nucleic Acids Res 15: 1353-1361.

[111] Vanfleteren JR, de Peer YV, Blaxter ML, Tweedie SA, Trotman C, Lu L, Hauwaert MLV and Moens L (1994) "Molecular genealogy of some nematode taxa as based on cytochrome c and globin amino acid sequences." Mol Phylogenet Evol 3: 92-101.

[112] Nadershahi A, Fahrenkrug SC and Ellis LBM (2004) "Comparison of computational methods for identifying translation initiation sites in EST data." BMC Bioinformatics 5: 14.

[113] Salamov AA, Nishikawa T and Swindells MB (1998) "Assessing protein coding region integrity in cDNA sequencing projects." Bioinformatics 14: 384-390.

[114] Kanehisa M, Goto S, Kawashima S, Okuno Y and Hattori M (2004) "The KEGG resource for deciphering the genome." Nucleic Acids Res 32: D277-D280.

[115] Gene Consortium (2001) "Creating the gene ontology resource: design and implementation." Genome Res 11: 1425-1433.

[116] Copley RR, Aloy P, Russell RB and Telford MJ (2004) "Systematic searches for molecular synapomorphies in model metazoan genomes give some support for Ecdysozoa after accounting for the idiosyncrasies of Caenorhabditis elegans." Evol Dev 6: 164-169.

[117] Mitreva M, Appleton J, McCarter JP and Jasmer DP (2005) "Expressed sequence tags from life cycle stages of Trichinella spiralis: application to biology and parasite control." Vet Parasitol 132: 13-17.

[118] Enright AJ, Dongen SV and Ouzounis CA (2002) "An efficient algorithm for large-scale detection of protein families." Nucleic Acids Res 30: 1575-1584.

[119] Knight RD, Freeland SJ and Landweber LF (2001) "A simple model based on mutation and selection explains trends in codon and amino-acid usage and GC composition within and across genomes." Genome Biol 2: RESEARCH0010.

[120] Team RDC(2005) "R: A language and environment for statistical computing.." url: http://www.R-project.org .

[121] Bonferroni CE (1935) "Il calcolo delle assicurazioni su gruppi di teste" Studi in Onore del Professore Salvatore Ortu Carboni. : 13-60.

[122] Hu M, Chilton NB and Gasser RB (2004) "The mitochondrial genomics of parasitic nematodes of socio-economic importance: recent progress, and implications for population genetics and systematics." Adv Parasitol 56: 133-212.

[123] Blaxter ML (2003) "Nematoda: genes, genomes and the evolution of parasitism." Adv Parasitol 54: 101-195.

[124] Krause M and Hirsh D (1987) "A trans-spliced leader sequence on actin mRNA in C. elegans." Cell 49: 753-761.

[125] Takacs AM, Denker JA, Perrine KG, Maroney PA and Nilsen TW (1988) "A 22-nucleotide spliced leader sequence in the human parasitic nematode Brugia malayi is identical to the trans-spliced leader exon in Caenorhabditis elegans." Proc Natl Acad Sci U S A 85: 7932-7936.

[126] Blumenthal T (1995) "Trans-splicing and polycistronic transcription in Caenorhabditis elegans." Trends Genet 11: 132-136.

[127] Maroney PA, Denker JA, Darzynkiewicz E, Laneve R and Nilsen TW (1995) "Most mRNAs in the nematode Ascaris lumbricoides are trans-spliced: a role for spliced leader addition in translational efficiency." RNA 1: 714-723.

[128] Kuersten S, Lea K, MacMorris M, Spieth J and Blumenthal T (1997) "Relationship between 3' end formation and SL2-specific trans-splicing in polycistronic Caenorhabditis elegans pre-mRNA processing." RNA 3: 269-278.

[129] Ghedin E, Wang S, Foster JM and Slatko BE (2004) "First sequenced genome of a parasitic nematode." Trends Parasitol 20: 151-153.

[130] Whitton C, Daub J, Quail M, Hall N, Foster J, Ware J, Ganatra M, Slatko B, Barrell B and Blaxter M (2004) "A genome sequence survey of the filarial nematode Brugia malayi: repeats, gene discovery, and comparative genomics." Mol Biochem Parasitol 137: 215-227.

[131] Louie E, Ott J and Majewski J (2003) "Nucleotide frequency variation across human genes." Genome Res 13: 2594-2601.

[132] Tabaska JE and Zhang MQ (1999) "Detection of polyadenylation signals in human DNA sequences." Gene 231: 77-86.

[133] Echols N, Harrison P, Balasubramanian S, Luscombe NM, Bertone P, Zhang Z and Gerstein M (2002) ."Comprehensive analysis of amino acid and nucleotide composition in eukaryotic genomes, comparing genes and pseudogenes." Nucleic Acids Res 30: 2515-2523.

[134] Purvis A and Rambaut A (1995) "Comparative analysis by independent contrasts (CAIC): an Apple Macintosh application for analysing comparative data." Comput Appl Biosci 11: 247-251.

[135] Tatusov RL, Koonin EV and Lipman DJ (1997) "A genomic perspective on protein families." Science 278: 631-637.

[136] Watson JD, Laskowski RA and Thornton JM (2005) "Predicting protein function from sequence and structural data." Curr Opin Struct Biol 15: 275-284.

[137] El-Sayed NM, Myler PJ, Bartholomeu DC, Nilsson D, Aggarwal G, Tran AN, Ghedin E, Worthey EA, Delcher AL, Blandin G, Westenberger SJ, Caler E, Cerqueira GC, Branche C, Haas B, Anupama A, Arner E, Aslund L, Attipoe P, Bontempi E, Bringaud F, Burton P, Cadag E, Campbell DA, Carrington M, Crabtree J, Darban H, da Silveira JF, de Jong P, Edwards K, Englund PT, Fazelina G, Feldblyum T, Ferella M, Frasch AC, Gull K, Horn D, Hou L, Huang Y, Kindlund E, Klingbeil M, Kluge S, Koo H, Lacerda D, Levin MJ, Lorenzi H, Louie T, Machado CR, McCulloch R, McKenna A, Mizuno Y, Mottram JC, Nelson S, Ochaya S, Osoegawa K, Pai G, Parsons M, Pentony M, Pettersson U, Pop M, Ramirez JL, Rinta J, Robertson L, Salzberg SL, Sanchez DO, Seyler A, Sharma R, Shetty J, Simpson AJ, Sisk E, Tammi MT, Tarleton R, Teixeira S, Aken SV, Vogt C, Ward PN,

Wickstead B, Wortman J, White O, Fraser CM, Stuart KD and Andersson B (2005) "The genome sequence of Trypanosoma cruzi, etiologic agent of Chagas disease." Science 309: 409-415.

[138] Berriman M, Ghedin E, Hertz-Fowler C, Blandin G, Renauld H, Bartholomeu DC, Lennard NJ, Caler E, Hamlin NE, Haas B, Böhme U, Hannick L, Aslett MA, Shallom J, Marcello L, Hou L, Wickstead B, Alsmark UCM, Arrowsmith C, Atkin RJ, Barron AJ, Bringaud F, Brooks K, Carrington M, Cherevach I, Chillingworth TJ, Churcher C, Clark LN, Corton CH, Cronin A, Davies RM, Doggett J, Djikeng A, Feldblyum T, Field MC, Fraser A, Goodhead I, Hance Z, Harper D, Harris BR, Hauser H, Hostetler J, Ivens A, Jagels K, Johnson D, Johnson J, Jones K, Kerhornou AX, Koo H, Larke N, Landfear S, Larkin C, Leech V, Line A, Lord A, Macleod A, Mooney PJ, Moule S, Martin DMA, Morgan GW, Mungall K, Norbertczak H, Ormond D, Pai G, Peacock CS, Peterson J, Quail MA, Rabbinowitsch E, Rajandream MA, Reitter C, Salzberg SL, Sanders M, Schobel S, Sharp S, Simmonds M, Simpson AJ, Tallon L, Turner CMR, Tait A, Tivey AR, Aken SV, Walker D, Wanless D, Wang S, White B, White O, Whitehead S, Woodward J, Wortman J, Adams MD, Embley TM, Gull K, Ullu E, Barry JD, Fairlamb AH, Opperdoes F, Barrell BG, Donelson JE, Hall N, Fraser CM, Melville SE and El-Sayed NM (2005) "The genome of the African trypanosome Trypanosoma brucei." Science 309: 416-422.

[139] Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P, Agarwala R, Ainscough R, Alexandersson M, An P, Antonarakis SE, Attwood J, Baertsch R, Bailey J, Barlow K, Beck S, Berry E, Birren B, Bloom T, Bork P, Botcherby M, Bray N, Brent MR, Brown DG, Brown SD, Bult C, Burton J, Butler J, Campbell RD, Carninci P, Cawley S, Chiaromonte F, Chinwalla AT, Church DM, Clamp M, Clee C, Collins FS, Cook LL, Copley RR, Coulson A, Couronne O, Cuff J, Curwen V, Cutts T, Daly M, David R, Davies J, Delehaunty KD, Deri J, Dermitzakis ET, Dewey C, Dickens NJ, Diekhans M, Dodge S, Dubchak I, Dunn DM, Eddy SR, Elnitski L, Emes RD, Eswara P, Eyras E, Felsenfeld A, Fewell GA, Flicek P, Foley K, Frankel WN, Fulton LA, Fulton RS, Furey TS, Gage D, Gibbs RA, Glusman G, Gnerre S, Goldman N, Goodstadt L, Grafham D, Graves TA, Green ED, Gregory S, Guigó R, Guyer M, Hardison RC, Haussler D, Hayashizaki Y, Hillier LDW, Hinrichs A, Hlavina W, Holzer T, Hsu F, Hua A, Hubbard T, Hunt A, Jackson I, Jaffe DB, Johnson LS, Jones M, Jones TA, Joy A, Kamal M, Karlsson EK, Karolchik D, Kasprzyk A, Kawai J, Keibler E, Kells C, Kent WJ, Kirby A, Kolbe DL, Korf I, Kucherlapati RS, Kulbokas EJ, Kulp D, Landers T, Leger JP, Leonard S, Letunic I, Levine R, Li J, Li M, Lloyd C, Lucas S, Ma B, Maglott DR, Mardis ER, Matthews L, Mauceli E, Mayer JH, McCarthy M, McCombie WR, McLaren S, McLay K, McPherson JD, Meldrim J, Meredith B, Mesirov JP, Miller W, Miner TL, Mongin E, Montgomery KT, Morgan M, Mott R, Mullikin JC, Muzny DM, Nash WE, Nelson JO, Nhan MN, Nicol R, Ning Z, Nusbaum C, O'Connor MJ, Okazaki Y, Oliver K, Overton-Larty E, Pachter L, Parra G, Pepin KH, Peterson J, Pevzner P, Plumb R, Pohl CS, Poliakov A, Ponce TC, Ponting CP, Potter S, Quail M, Reymond A, Roe BA, Roskin KM, Rubin EM, Rust AG, Santos R, Sapojnikov V, Schultz B, Schultz J, Schwartz MS, Schwartz S, Scott C, Seaman S, Searle S, Sharpe T, Sheridan A, Shownkeen R, Sims S, Singer JB, Slater G, Smit A, Smith DR, Spencer B, Stabenau A, Stange-Thomann N, Sugnet C, Suyama M, Tesler G, Thompson J, Torrents D, Trevaskis E, Tromp J, Ucla C, Ureta-Vidal A, Vinson JP, Niederhausern ACV, Wade CM, Wall M, Weber RJ, Weiss RB, Wendl MC, West AP, Wetterstrand K, Wheeler R, Whelan S, Wierzbowski J, Willey D, Williams S, Wilson RK, Winter E, Worley KC, Wyman D, Yang S, Yang SP, Zdobnov EM, Zody MC, Lander ES and Consortium MGS (2002) "Initial sequencing and comparative analysis of the mouse genome." Nature 420: 520-562.

[140] Oduru S, Campbell JL, Karri ST, Hendry WJ, Khan SA and Williams SC (2003) "Gene discovery in the hamster: a comparative genomics approach for gene annotation by sequencing of hamster testis cDNAs." BMC Genomics 4: 22.

[141] Putta S, Smith JJ, Walker JA, Rondet M, Weisrock DW, Monaghan J, Samuels AK, Kump K, King DC, Maness NJ, Habermann B, Tanaka E, Bryant SV, Gardiner DM, Parichy DM and Voss SR (2004) "From biomedicine to natural history research: EST resources for ambystomatid salamanders." BMC Genomics 5: 54.

[142] Nunes FMF, Valente V, Sousa JF, Cunha MAV, Pinheiro DG, Maia RM, Araujo DD, Costa MCR, Martins WK, Carvalho AF, Monesi N, Nascimento AM, Peixoto PMV, Silva MFR, Ramos RGP, Reis LFL, Dias-Neto E, Souza SJ, Simpson AJG, Zago MA, Soares AEE, Bitondi MMG, Espreafico EM, Espindola FS, Paco-Larson ML, Simoes ZLP, Hartfelder K and Silva WA (2004) "The use of Open Reading frame ESTs (ORESTES) for analysis of the honey bee transcriptome."

BMC Genomics 5: 84.

[143] Rensink WA, Lee Y, Liu J, Iobst S, Ouyang S and Buell CR (2005) "Comparative analyses of six solanaceous transcriptomes reveal a high degree of sequence conservation and species-specific transcripts." BMC Genomics 6: 124.

[144] Sczyrba A, Beckstette M, Brivanlou AH, Giegerich R and Altmann CR (2005) "XenDB: full length cDNA prediction and cross species mapping in Xenopus laevis." BMC Genomics 6: 123.

[145] Colbourne JK, Singan VR and Gilbert DG (2005) "wFleaBase: the Daphnia genome database." BMC Bioinformatics 6: 45.

[146] Mitreva M, McCarter JP, Arasu P, Hawdon J, Martin J, Dante M, Wylie T, Xu J, Stajich JE, Kapulkin W, Clifton SW, Waterston RH and Wilson RK (2005b) "Investigating hookworm genomes by comparative analysis of two Ancylostoma species." BMC Genomics 6: 58.

[147] Krylov DM, Wolf YI, Rogozin IB and Koonin EV (2003) "Gene loss, protein sequence divergence, gene dispensability, expression level, and interactivity are correlated in eukaryotic evolution." Genome Res 13: 2229-2235.

[148] Wagstaff BJ and Begun DJ (2005) "Comparative genomics of accessory gland protein genes in Drosophila melanogaster and D. pseudoobscura." Mol Biol Evol 22: 818-832.

[149] Zdobnov EM, von Mering C, Letunic I, Torrents D, Suyama M, Copley RR, Christophides GK, Thomasova D, Holt RA, Subramanian GM, Mueller HM, Dimopoulos G, Law JH, Wells MA, Birney E, Charlab R, Halpern AL, Kokoza E, Kraft CL, Lai Z, Lewis S, Louis C, Barillas-Mury C, Nusskern D, Rubin GM, Salzberg SL, Sutton GG, Topalis P, Wides R, Wincker P, Yandell M, Collins FH, Ribeiro J, Gelbart WM, Kafatos FC and Bork P (2002) "Comparative genome and proteome analysis of Anopheles gambiae and Drosophila melanogaster." Science 298: 149-159.

[150] Putta S, Smith JJ, Walker JA, Rondet M, Weisrock DW, Monaghan J, Samuels AK, Kump K, King DC, Maness NJ, Habermann B, Tanaka E, Bryant SV, Gardiner DM, Parichy DM and Voss SR (2004) "From biomedicine to natural history research: EST resources for ambystomatid salamanders." BMC Genomics 5: 54.

[151] Rensink WA, Lee Y, Liu J, Iobst S, Ouyang S and Buell CR (2005) "Comparative analyses of six solanaceous transcriptomes reveal a high degree of sequence conservation and species-specific transcripts." BMC Genomics 6: 124.

[152] Graham MA, Silverstein KAT, Cannon SB and VandenBosch KA (2004) "Computational identification and characterization of novel genes from legumes." Plant Physiol 135: 1179--1197.

[153] Parkinson J, Mitreva M, Whitton C, Thomson M, Daub J, Martin J, Schmid R, Hall N, Barrell B, Waterston RH, McCarter JP and Blaxter ML (2004) "A transcriptomic analysis of the phylum Nematoda." Nat Genet 36: 1259--1267.

[154] de Koning AP, Tartar A, Boucias DG and Keeling PJ (2005) "Expressed sequence tag (EST) survey of the highly adapted green algal parasite, Helicosporidium." Protist 156: 181--190.

[155] Parkinson J, Mitreva M, Hall N, Blaxter M and McCarter JP (2003) "400000 nematode ESTs on the Net." Trends Parasitol 19: 283-286.

[156] Petsko GA (2001) "Homologuephobia." Genome Biol 2: COMMENT1002.

[157] Jensen RA (2001) "Orthologs and paralogs - we need to get it right." Genome Biol 2: INTERACTIONS1002.

[158] Koonin EV (2001) "An apology for orthologs - or brave new memes." Genome Biol 2: COMMENT1005.

[159] Sonnhammer ELL and Koonin EV (2002) "Orthology, paralogy and proposed classification for paralog subtypes." Trends Genet 18: 619-620.

[160] Varshavsky A (2004) "Spalog' and 'sequelog': neutral terms for spatial and sequence similarity." Curr Biol 14: R181-R183.

[161] Koonin EV (2005) "Orthologs, paralogs, and evolutionary genomics (1)." Annu Rev Genet 39:

309-338.

[162] Kunin V, Teichmann SA, Huynen MA and Ouzounis CA (2005) "The properties of protein family space depend on experimental design." Bioinformatics 21: 2618-2622.

[163] Linial M, Linial N, Tishby N and Yona G (1997) "Global self-organization of all known protein sequences reveals inherent biological signatures." J Mol Biol 268: 539-556.

[164] Haft DH, Loftus BJ, Richardson DL, Yang F, Eisen JA, Paulsen IT and White O (2001) "TIGRFAMs: a protein family resource for the functional identification of proteins." Nucleic Acids Res 29: 41-43.

[165] Tatusov RL, Natale DA, Garkavtsev IV, Tatusova TA, Shankavaram UT, Rao BS, Kiryutin B, Galperin MY, Fedorova ND and Koonin EV (2001) "The COG database: new developments in phylogenetic classification of proteins from complete genomes." Nucleic Acids Res 29: 22-28.

[166] Aguinaldo AM, Turbeville JM, Linford LS, Rivera MC, Garey JR, Raff RA and Lake JA (1997) "Evidence for a clade of nematodes, arthropods and other moulting animals." Nature 387: 489-493.

[167] Winnepenninckx B, Backeljau T and Wachter RD (1995) "Phylogeny of protostome worms derived from 18S rRNA sequences." Mol Biol Evol 12: 641-649.

[168] Jones M and Blaxter M (2005) "Evolutionary biology: animal roots and shoots." Nature 434: 1076-1077.

[169] Keightley PD, Lercher MJ and Eyre-Walker A (2005) "Evidence for widespread degradation of gene control regions in hominid genomes." PLoS Biol 3: e42.

[170] Sternberg P, Waterston R, Spieth J, Eddy S and Wilson R(2003) "Genome sequence of additional Caenorhabditis species: enhancing the utility of C. elegans as a model organism." url: http://genome.wustl.edu/ancillary/data/whitepapers/Caenorhabditis_WP.pdf.

[171] Sasser JN (1980) "" Plant Disease 64: 36-41.

[172] Doolittle RF and Bork P (1993) "Evolutionarily mobile modules in proteins." Sci Am 269: 50-56.

[173] Vogel C, Bashton M, Kerrison ND, Chothia C and Teichmann SA (2004) "Structure, function and evolution of multidomain proteins." Curr Opin Struct Biol 14: 208-216.

[174] Weiner J, Thomas G and Bornberg-Bauer E (2005) "Rapid motif-based prediction of circular permutations in multi-domain proteins." Bioinformatics 21: 932-937.

[175] Ruvkun G and Hobert O (1998) "The taxonomy of developmental control in Caenorhabditis elegans." Science 282: 2033-2041.

[176] Aboobaker AA and Blaxter ML (2003) "Hox Gene Loss during Dynamic Evolution of the Nematode Cluster." Curr Biol 13: 37-40.

[177] Sheps JA, Ralph S, Zhao Z, Baillie DL and Ling V (2004) "The ABC transporter gene family of Caenorhabditis elegans has implications for the evolutionary dynamics of multidrug resistance in eukaryotes." Genome Biol 5: R15.

[178] Chervitz SA, Aravind L, Sherlock G, Ball CA, Koonin EV, Dwight SS, Harris MA, Dolinski K, Mohr S, Smith T, Weng S, Cherry JM and Botstein D (1998) "Comparison of the complete protein sets of worm and yeast: orthology and divergence." Science 282: 2022-2028.

[179] Remm M, Storm CE and Sonnhammer EL (2001) "Automatic clustering of orthologs and in-paralogs from pairwise species comparisons." J Mol Biol 314: 1041-1052.

[180] Aravind L, Watanabe H, Lipman DJ and Koonin EV (2000) "Lineage-specific loss and divergence of functionally linked genes in eukaryotes." Proc Natl Acad Sci U S A 97: 11319-11324.

[181] Blair JE, Ikeo K, Gojobori T and Hedges SB (2002) "The evolutionary position of nematodes." BMC Evol Biol 2: 7.

[182] Schmid KJ and Tautz D (1997) "A screen for fast evolving genes from Drosophila." Proc Natl

Acad Sci U S A 94: 9746-9750.

[183] Hendy MD and Penny D (1989) "A framework for the quantitative study of evolutionary trees" Systematic Zoology 38: 297-309.

[184] Felsenstein J (1978) "Cases in which parsimony or combatibility methods will be positively misleading" Systematic Zoology 27: 401-410.

[185] Swofford DL and Olsen GJ (1990) "Molecular Systematics" ed. Hillis, D.M and Moritz, C, Sinauer Associates, Sunderland, Massachusetts.

[186] Hillis DM (1996) "Inferring complex phylogenies." Nature 383: 130-131.

[187] Philippe H, Lartillot N and Brinkmann H (2005) "Multigene analyses of bilaterian animals corroborate the monophyly of Ecdysozoa, Lophotrochozoa, and Protostomia." Mol Biol Evol 22: 1246-1253.

[188] Park J, Karplus K, Barrett C, Hughey R, Haussler D, Hubbard T and Chothia C (1998) "Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods." J Mol Biol 284: 1201-1210.

[189] Yona G and Levitt M (2002) "Within the twilight zone: a sensitive profile-profile comparison tool based on information theory." J Mol Biol 315: 1257-1275.

[190] Ponting CP and Russell RR (2002) "The natural history of protein domains." Annu Rev Biophys Biomol Struct 31: 45-71.

[191] Henikoff S and Henikoff JG (1997) "Embedding strategies for effective use of information from multiple sequence alignments." Protein Sci 6: 698-705.

[192] Gribskov M, McLachlan AD and Eisenberg D (1987) "Profile analysis: detection of distantly related proteins." Proc Natl Acad Sci U S A 84: 4355-4358.

[193] Pearson WR and Sierk ML (2005) "The limits of protein sequence comparison?" Curr Opin Struct Biol 15: 254-260.

[194] Hegyi H and Bork P (1997) "On the classification and evolution of protein modules." J Protein Chem 16: 545-551.

[195] Doolittle RF (1981) "Similar amino acid sequences: chance or common ancestry?" Science 214: 149-159.

[196] Zuckerkandl E (1975) "The appearance of new structures and functions in proteins during evolution." J Mol Evol 7: 01/01/57.

[197] Dayhoff MO and Orcutt BC (1979) "Methods for identifying proteins by using partial sequences." Proc Natl Acad Sci U S A 76: 2170-2174.

[198] Heger A and Holm L (2000) "Towards a covering set of protein family profiles." Prog Biophys Mol Biol 73: 321-337.

[199] Li L, Stoeckert CJ and Roos DS (2003) "OrthoMCL: identification of ortholog groups for eukaryotic genomes." Genome Res 13: 2178-2189.

[200] van Dongen S (2000) "Graph Clustering by Flow Simulation", Thesis, University of Utrecht.

[201] Coulson RMR, Hall N and Ouzounis CA (2004) "Comparative genomics of transcriptional control in the human malaria parasite Plasmodium falciparum." Genome Res 14: 1548-1554.

[202] Storm CE and Sonnhammer EL (2001) "NIFAS: visual analysis of domain evolution in proteins." Bioinformatics 17: 343-348.

[203] Harlow TJ, Gogarten JP and Ragan MA (2004) "A hybrid clustering approach to recognition of protein families in 114 microbial genomes." BMC Bioinformatics 5: 45.

[204] Enright AJ, Kunin V and Ouzounis CA (2003) "Protein families and TRIBES in genome sequence space." Nucleic Acids Res 31: 4632-4638.

[205] Andreeva A, Howorth D, Brenner SE, Hubbard TJP, Chothia C and Murzin AG (2004) "SCOP

database in 2004: refinements integrate structure and sequence family data." Nucleic Acids Res 32: D226-D229.

[206] Pearl FMG, Bennett CF, Bray JE, Harrison AP, Martin N, Shepherd A, Sillitoe I, Thornton J and Orengo CA (2003) "The CATH database: an extended protein family resource for structural and functional genomics." Nucleic Acids Res 31: 452-455.

[207] Gough J and Chothia C (2002) "SUPERFAMILY: HMMs representing all proteins of known structure. SCOP sequence searches, alignments and genome assignments." Nucleic Acids Res 30: 268-272.

[208] Marchler-Bauer A, Anderson JB, Cherukuri PF, DeWeese-Scott C, Geer LY, Gwadz M, He S, Hurwitz DI, Jackson JD, Ke Z, Lanczycki CJ, Liebert CA, Liu C, Lu F, Marchler GH, Mullokandov M, Shoemaker BA, Simonyan V, Song JS, Thiessen PA, Yamashita RA, Yin JJ, Zhang D and Bryant SH (2005) "CDD: a Conserved Domain Database for protein classification." Nucleic Acids Res 33: D192-D196.

[209] Eddy S (2003) "HMMer User Guide 2003: http://hmmer.wustl.edu".

[210] Servant F, Bru C, Carrère S, Courcelle E, Gouzy J, Peyruc D and Kahn D (2002) "ProDom: automated clustering of homologous domains." Brief Bioinform 3: 246-251.

[211] Conte LL, Brenner SE, Hubbard TJP, Chothia C and Murzin AG (2002) "SCOP database in 2002: refinements accommodate structural genomics." Nucleic Acids Res 30: 264-267.

[212] Goldstein L and Waterman MS (1994) "Approximations to profile score distributions." J Comput Biol 1: 93-104.

[213] Popeijus H, Overmars H, Jones J, Blok V, Goverse A, Helder J, Schots A, Bakker J and Smant G (2000) "Degradation of plant cell walls by a nematode." Nature 406: 36-37.

[214] Lambert KN, Allen KD and Sussex IM (1999) "Cloning and characterization of an esophageal-gland-specific chorismate mutase from the phytoparasitic nematode Meloidogyne javanica." Mol Plant Microbe Interact 12: 328-336.

[215] Jones JT, Furlanetto C, Bakker E, Banks B, Blok V, Chen Q, Phillips M and Prior A (2003) "Characterization of a chorismate mutase from the potato cyst nematode Globodera pallida" Molecular Plant Pathology 4: 43-50.

[216] Smant G, Stokkermans JP, Yan Y, de Boer JM, Baum TJ, Wang X, Hussey RS, Gommers FJ, Henrissat B, Davis EL, Helder J, Schots A and Bakker J (1998) "Endogenous cellulases in animals: isolation of beta-1, 4-endoglucanase genes from two species of plant-parasitic cyst nematodes." Proc Natl Acad Sci U S A 95: 4906-4911.

[217] Davison A and Blaxter M (2005) "Ancient origin of glycosyl hydrolase family 9 cellulase genes." Mol Biol Evol 22: 1273-1284.

[218] Connolly B, Trenholme K and Smith DF (1996) "Molecular cloning of a myoD-like gene from the parasitic nematode, Trichinella spiralis." Mol Biochem Parasitol 81: 137-149.

[219] International Human Genome Sequencing Consortium (2001) "Initial sequencing and analysis of the human genome." Nature 409: 860-921.

[220] Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD, Amanatides PG, Scherer SE, Li PW, Hoskins RA, Galle RF, George RA, Lewis SE, Richards S, Ashburner M, Henderson SN, Sutton GG, Wortman JR, Yandell MD, Zhang Q, Chen LX, Brandon RC, Rogers YH, Blazej RG, Champe M, Pfeiffer BD, Wan KH, Doyle C, Baxter EG, Helt G, Nelson CR, Gabor GL, Abril JF, Agbayani A, An HJ, Andrews-Pfannkoch C, Baldwin D, Ballew RM, Basu A, Baxendale J, Bayraktaroglu L, Beasley EM, Beeson KY, Benos PV, Berman BP, Bhandari D, Bolshakov S, Borkova D, Botchan MR, Bouck J, Brokstein P, Brottier P, Burtis KC, Busam DA, Butler H, Cadieu E, Center A, Chandra I, Cherry JM, Cawley S, Dahlke C, Davenport LB, Davies P, de Pablos B, Delcher A, Deng Z, Mays AD, Dew I, Dietz SM, Dodson K, Doup LE, Downes M, Dugan-Rocha S, Dunkov BC, Dunn P, Durbin KJ, Evangelista CC, Ferraz C, Ferriera S, Fleischmann W, Fosler C, Gabrielian AE, Garg NS, Gelbart WM, Glasser K, Glodek A, Gong F, Gorrell JH, Gu Z, Guan P, Harris M, Harris NL, Harvey D, Heiman TJ, Hernandez JR, Houck J, Hostin D, Houston KA, Howland TJ, Wei MH, Ibegwam C,

Jalali M, Kalush F, Karpen GH, Ke Z, Kennison JA, Ketchum KA, Kimmel BE, Kodira CD, Kraft C, Kravitz S, Kulp D, Lai Z, Lasko P, Lei Y, Levitsky AA, Li J, Li Z, Liang Y, Lin X, Liu X, Mattei B, McIntosh TC, McLeod MP, McPherson D, Merkulov G, Milshina NV, Mobarry C, Morris J, Moshrefi A, Mount SM, Moy M, Murphy B, Murphy L, Muzny DM, Nelson DL, Nelson DR, Nelson KA, Nixon K, Nusskern DR, Pacleb JM, Palazzolo M, Pittman GS, Pan S, Pollard J, Puri V, Reese MG, Reinert K, Remington K, Saunders RD, Scheeler F, Shen H, Shue BC, Sidén-Kiamos I, Simpson M, Skupski MP, Smith T, Spier E, Spradling AC, Stapleton M, Strong R, Sun E, Svirskas R, Tector C, Turner R, Venter E, Wang AH, Wang X, Wang ZY, Wassarman DA, Weinstock GM, Weissenbach J, Williams SM, , Worley KC, Wu D, Yang S, Yao QA, Ye J, Yeh RF, Zaveri JS, Zhan M, Zhang G, Zhao Q, Zheng L, Zheng XH, Zhong FN, Zhong W, Zhou X, Zhu S, Zhu X, Smith HO, Gibbs RA, Myers EW, Rubin GM and Venter JC (2000) "The genome sequence of Drosophila melanogaster." Science 287: 2185-2195.

[221] Gregory WF, Atmadja AK, Allen JE and Maizels RM (2000) "The abundant larval transcript-1 and -2 genes of Brugia malayi encode stage-specific candidate vaccine antigens for filariasis." Infect Immun 68: 4174-4179.

[222] Robertson H and Thomas J (*in press*) "WormBook (http://www.wormbook.org)" ed. Community, The C. elegans Research, .

[223] Troemel ER, Chou JH, Dwyer ND, Colbert HA and Bargmann CI (1995) "Divergent seven transmembrane receptors are candidate chemosensory receptors in C. elegans." Cell 83: 207-218.

[224] Kamath RS and Ahringer J (2003) "Genome-wide RNAi screening in Caenorhabditis elegans." Methods 30: 313-321.

[225] Ashrafi K, Chang FY, Watts JL, Fraser AG, Kamath RS, Ahringer J and Ruvkun G (2003) "Genome-wide RNAi analysis of Caenorhabditis elegans fat regulatory genes." Nature 421: 268-272.

[226] Chen N, Pai S, Zhao Z, Mah A, Newbury R, Johnsen RC, Altun Z, Moerman DG, Baillie DL and Stein LD (2005) "Identification of a nematode chemosensory gene family." Proc Natl Acad Sci U S A 102: 146-151.

[227] Thomas JH, Kelley JL, Robertson HM, Ly K and Swanson WJ (2005) "Adaptive evolution in the SRZ chemoreceptor families of Caenorhabditis elegans and Caenorhabditis briggsae." Proc Natl Acad Sci U S A 102: 4476-4481.

[228] Dodson G and Wlodawer A (1998) "Catalytic triads and their relatives." Trends Biochem Sci 23: 347-352.

[229] Baldwin JG, Nadler SA and Adams BJ (2004) "Evolution of plant parasitism among nematodes." Annu Rev Phytopathol 42: 83-105.

[230] Olsen AN and Skriver K (2003) "Ligand mimicry? Plant-parasitic nematode polypeptide with similarity to CLAVATA3." Trends Plant Sci 8: 55-57.

[231] Pastrana DV, Raghavan N, FitzGerald P, Eisinger SW, Metz C, Bucala R, Schleimer RP, Bickel C and Scott AL (1998) "Filarial nematode parasites secrete a homologue of the human cytokine macrophage migration inhibitory factor." Infect Immun 66: 5955-5963.

[232] Zang X, Taylor P, Wang JM, Meyer DJ, Scott AL, Walkinshaw MD and Maizels RM (2002) "Homologues of human macrophage migration inhibitory factor from a parasitic nematode. Gene cloning, protein activity, and crystal structure." J Biol Chem 277: 44261-44267.

[233] Murray J, Manoury B, Balic A, Watts C and Maizels RM (2005) "Bm-CPI-2, a cystatin from Brugia malayi nematode parasites, differs from Caenorhabditis elegans cystatins in a specific site mediating inhibition of the antigen-processing enzyme AEP." Mol Biochem Parasitol 139: 197-203.

[234] Jain R, Rivera MC and Lake JA (1999) "Horizontal gene transfer among genomes: the complexity hypothesis." Proc Natl Acad Sci U S A 96: 3801-3806.

[235] Doolittle WF (1999) "Phylogenetic classification and the universal tree." Science 284: 2124-2129.

[236] Stephens RS, Kalman S, Lammel C, Fan J, Marathe R, Aravind L, Mitchell W, Olinger L, Tatusov RL, Zhao Q, Koonin EV and Davis RW (1998) "Genome sequence of an obligate

intracellular pathogen of humans: Chlamydia trachomatis." Science 282: 754-759.

[237] Stanhope MJ, Lupas A, Italia MJ, Koretke KK, Volker C and Brown JR (2001) "Phylogenetic analyses do not support horizontal gene transfers from bacteria to vertebrates." Nature 411: 940-944.

[238] Salzberg SL, White O, Peterson J and Eisen JA (2001) "Microbial genes in the human genome: lateral transfer or gene loss?" Science 292: 1903-1906.

[239] Salzberg SL and Eisen JA (2001) "Lateral gene transfer or viral colonization?" Science 293: 1048.

[240] Yan Y, Smant G, Stokkermans J, Qin L, Helder J, Baum T, Schots A and Davis E (1998) "Genomic organization of four beta-1,4-endoglucanase genes in plant-parasitic cyst nematodes and its evolutionary implications." Gene 220: 61-70.

[241] Lewis AP and Crowther D (2005) "DING proteins are from Pseudomonas." FEMS Microbiol Lett 252: 215-222.

[242] McLaren DJ, Worms MJ, Laurence BR and Simpson MG (1975) "Micro-organisms in filarial larvae (Nematoda)." Trans R Soc Trop Med Hyg 69: 509-514.

[243] Franz M and Büttner DW (1983) "The fine structure of adult Onchocerca volvulus IV. The hypodermal chords of the female worm." Tropenmed Parasitol 34: 122-128.

[244] Sironi M, Bandi C, Sacchi L, Sacco BD, Damiani G and Genchi C (1995) "Molecular evidence for a close relative of the arthropod endosymbiont Wolbachia in a filarial worm." Mol Biochem Parasitol 74: 223-227.

[245] Vandekerckhove TT, Willems A, Gillis M and Coomans A (2000) "Occurrence of novel verrucomicrobial species, endosymbiotic and associated with parthenogenesis in Xiphinema americanum-group species (Nematoda, Longidoridae)." Int J Syst Evol Microbiol : 2197-2205.

[246] Williams SA, Lizotte-Waniewski MR, Foster J, Guiliano D, Daub J, Scott AL, Slatko B and Blaxter ML (2000) "The filarial genome project: analysis of the nuclear, mitochondrial and endosymbiont genomes of Brugia malayi." Int J Parasitol 30: 411-419.

[247] Troemel ER (1999) "Chemosensory signaling in C. elegans." Bioessays 21: 1011-1020.

[248] Reuter G, Giarre M, Farah J, Gausz J, Spierer A and Spierer P (1990) "Dependence of position-effect variegation in Drosophila on dose of a gene encoding an unusual zinc-finger protein." Nature 344: 219-223.

[249] Zhao K, Hart CM and Laemmli UK (1995) "Visualization of chromosomal domains with boundary element-associated factor BEAF-32." Cell 81: 879-889.

[250] Clark KA and McKearin DM (1996) "The Drosophila stonewall gene encodes a putative transcription factor essential for germ cell development." Development 122: 937-950.

[251] Delattre M, Spierer A, Hulo N and Spierer P (2002) "A new gene in Drosophila melanogaster, Ravus, the phantom of the modifier of position-effect variegation Su(var)3-7." Int J Dev Biol 46: 167-171.

[252] Laber B, Gomis-Rüth FX, Romão MJ and Huber R (1992) "Escherichia coli dihydrodipicolinate synthase. Identification of the active site and crystallization." Biochem J : 691-695.

[253] Bhattacharjee JK (1985) "alpha-Aminoadipate pathway for the biosynthesis of lysine in lower eukaryotes." Crit Rev Microbiol 12: 131-151.

[254] Cox GN, Shamansky LM and Boisvenue RJ (1989) "Identification and preliminary characterization of cuticular surface proteins of Haemonchus contortus." Mol Biochem Parasitol 36: 233-241.

[255] der Eycken WV, Engler JA, Montagu MV and Gheysen G (1994) "Identification and analysis of a cuticular collagen-encoding gene from the plant-parasitic nematode Meloidogyne incognita." Gene 151: 237-242.

[256] Gregory WF, Blaxter ML and Maizels RM (1997) "Differentially expressed, abundant trans-

spliced cDNAs from larval Brugia malayi." Mol Biochem Parasitol 87: 85-95.

[257] Frank GR, Tripp CA and Grieve RB (1996) "Molecular cloning of a developmentally regulated protein isolated from excretory-secretory products of larval Dirofilaria immitis." Mol Biochem Parasitol 75: 231-240.

[258] Hussain R, Grögl M and Ottesen EA (1987) "IgG antibody subclasses in human filariasis. Differential subclass recognition of parasite antigens correlates with different clinical manifestations of infection." J Immunol 139: 2794-2798.

[259] Pogonka T, Oberländer U, Marti T and Lucius R (1999) "Acanthocheilonema viteae: characterization of a molt-associated excretory/secretory 18-kDa protein." Exp Parasitol 93: 73-81.

[260] Genereux DP and Logsdon JM (2003) "Much ado about bacteria-to-vertebrate lateral gene transfer." Trends Genet 19: 191-195.

[261] Maizels RM, Balic A, Gomez-Escobar N, Nair M, Taylor MD and Allen JE (2004) "Helminth parasites--masters of regulation." Immunol Rev 201: 89-116.

[262] Maizels RM, Blaxter ML and Scott AL (2001) "Immunological genomics of Brugia malayi: filarial genes implicated in immune evasion and protective immunity." Parasite Immunol 23: 327-344.

[263] Maizels RM, Gomez-Escobar N, Gregory WF, Murray J and Zang X (2001) "Immune evasion genes from filarial nematodes." Int J Parasitol 31: 889-898.

[264] Beck S and Barrell BG (1988) "Human cytomegalovirus encodes a glycoprotein homologous to MHC class-I antigens." Nature 331: 269-272.

[265] Ouyang T, Bai RY, Bassermann F, von Klitzing C, Klumpen S, Miething C, Morris SW, Peschel C and Duyster J (2003) "Identification and characterization of a nuclear interacting partner of anaplastic lymphoma kinase (NIPA)." J Biol Chem 278: 30028-30036.

[266] Gough J, Karplus K, Hughey R and Chothia C (2001) "Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure." J Mol Biol 313: 903-919.

[267] Gouzy J, Eugéne P, Greene EA, Kahn D and Corpet F (1997) "XDOM, a graphical tool to analyse domain arrangements in any set of protein sequences." Comput Appl Biosci 13: 601-608.

Appendix

Papers published during this thesis:

Parkinson J, Anthony A, Wasmuth J, Schmid R, Hedley A and Blaxter M (2004) "PartiGene-constructing partial genomes." *Bioinformatics* **20**: 1398-1404

Wasmuth JD and Blaxter ML (2004) "prot4EST: translating expressed sequence tags from neglected genomes." *BMC Bioinformatics* **5**: 187

Vol. 20 no. 9 2004, pages 1398–1404
DOI: 10.1093/bioinformatics/bth101

PartiGene—constructing partial genomes

John Parkinson*, Alasdair Anthony, James Wasmuth, Ralf Schmid, Ann Hedley and Mark Blaxter

School of Biological Sciences, Ashworth Laboratories, King's Buildings,
West Mains Rd, University of Edinburgh, Edinburgh EH9 3JT, UK

Received on July 31, 2003; revised on December 9, 2003; accepted on December 11, 2003
Advance Access publication February 26, 2004

ABSTRACT
Expressed sequence tags (ESTs) offer a low-cost approach to gene discovery and are being used by an increasing number of laboratories to obtain sequence information for a wide variety of organisms. The challenge lies in processing and organizing this data within a genomic context to facilitate large scale analyses. Here we present PartiGene, an integrated sequence analysis suite that uses freely available public domain software to (1) process raw trace chromatograms into sequence objects suitable for submission to dbEST; (2) place these sequences within a genomic context; (3) perform customizable first-pass annotation of the data; and (4) present the data as HTML tables and an SQL database resource. PartiGene has been used to create a number of non-model organism database resources including NEMBASE (http://www.nematodes.org) and LumbriBase (http://www.earthworms.org/). The packages are readily portable, freely available and can be run on simple Linux-based workstations.

Availability: PartiGene is available from http://www.nematodes.org/PartiGene and also forms part of the EST analysis software, associated with the Natural Environmental Research Council (UK) Bio-Linux project (http://envgen.nox.ac.uk/biolinux.html).

Contact: jparkin@sickkids.ca

INTRODUCTION

The advent of low-cost, high-throughput sequencing has permitted the generation of fully sequenced genomes of a number of model organisms including 122 prokaryotic and 17 eukaryotic species (http://wit.integratedgenomics.com/GOLD/). For these fully sequenced genomes, integrated databases are used to contextualize sequence data within a rich biological information environment. An increasing amount of sequence data is being generated from a range of other, non-model organisms. For eukaryotic species, these sequence data are typically in the form of expressed sequence tags (ESTs). Datasets range from just a few hundred to as many as several hundred thousand sequences. There are over 180 species, with more than 1000 entries, in the database for ESTs (dbEST, http://www.ncbi.nlm.nih.gov/dbEST/index.html) (Boguski *et al.*, 1993). In general, these data are not well organized and are difficult to interpret in a genomic context.

Common problems include significant redundancy in the datasets (some genes may have been sequenced multiple times) and a lack of consistent annotation between projects. An effective way to overcome these problems is to group ESTs into clusters that represent genes and to provide annotations for the clusters. Since ESTs provide only a fraction of the available genes for a particular organism, we refer to these analysed datasets as partial genomes. Informatic solutions to produce partial genomes or 'gene indices' have been developed by several groups (Adams *et al.*, 1995; Boguski and Schuler, 1995; Sutton *et al.*, 1995; White and Kervalage, 1996; Christoffels *et al.*, 2001; Pertea *et al.*, 2003). The analysis of partial genomes has tended to involve complex integrated database solutions and/or a large amount of manual sequence annotation, both of which require a considerable investment in bioinformatic resources and make cross-species and between-lab integration difficult.

Our involvement in a wide range of different EST projects (Allen *et al.*, 2000; Daub *et al.*, 2000; Blaxter *et al.*, 2002; Kenyon *et al.*, 2003; Parkinson *et al.*, 2003) has led us to develop a generic, automated software pipeline, PartiGene, that handles an EST project from raw trace data through to a partial genome database ready for data mining. In this it goes beyond a simple EST-focussed LIMS system and other solutions to EST processing such as that of Paquola *et al.* (2003). PartiGene consists of three integrated scripts, based on the PERL scripting language, which rely on freely available public domain software. PartiGene is readily portable to most UNIX-based operating systems and is freely available from our Web server (http://www.nematodes.org/PartiGene). We have designed PartiGene to be freestanding, permitting installation and operation with a minimum of background expert knowledge. In addition to being portable and customizable, PartiGene offers further advantage over similar pipelines

*To whom correspondence should be addressed at Programs in Genetics and Genomic Biology & Structural Biology and Biochemistry, Hospital for Sick Children, 555 University Avenue, Toronto, Ontario M5G 1X8, Canada.

in that it allows incremental updates to established partial genome datasets.

METHODS

Software and hardware tools

The creation and presentation of partial genomes described here was undertaken on an Intel workstation (Dual Processor Pentium III, 750 MHz) running Red Hat Linux 7.1. PartiGene has also been tested on more recent versions of Red Hat Linux (8.0 and 9.0) and is expected to be fully portable to most UNIX distributions and hardware architectures. PartiGene uses the PERL scripting language, installed as default on most systems: a PERL interpreter of version 5.005 or later is required. In addition to the scripts presented here, PartiGene requires the installation of a number of other publicly available tools (freely available unless otherwise noted): phred, phrap and cross_match (http://www.phrap.org; a license is required for commercial users); DECODER (contact the authors, rgscerg@gsc.riken.go.jp; a license is required for commercial users); ESTscan (http://www.isrec.isb-sib.ch/ftp-server/ESTScan/); postgreSQL (http://www.postgresql.org); NCBI BLAST (http://www.ncbi.nlm.nih.gov/BLAST/); Bioperl (http://www.bioperl.org); and EMBOSS (http://www.hgmp.mrc.ac.uk/Software/EMBOSS/).

Overview

We were concerned with producing a software solution that provided ease of use while maintaining best practice for EST analysis. Therefore we have written a pipeline that takes raw sequence trace (chromatogram) data, performs base calling and vector and low quality sequence removal, preparation of dbEST submission files, clustering into putative genes, consensus sequence prediction, peptide prediction and sequence similarity annotation. The analysed data can be viewed as flat files (in HTML format) or as a standard-format SQL database. Throughout, we have implemented 'best practice' based on our experience with generating and analysing EST sequences. PartiGene is divided into three segments that process the raw sequence traces (trace2dbest), generate the partial genomes (PartiGene) and derive peptide predictions (prot4est). The input may be in the form of raw sequence chromatographic trace data, processed sequence data or more simply the name of the target species for which EST data are available in dbEST. The output can include dbEST submission files, HTML tables describing each putative gene object and/or a set of SQL database tables that may be readily queried using the SQL interpreter.

Process 1: from raw trace data to dbEST submission

The first script, trace2dbest, is an interactive pipeline script that takes raw sequencer trace data and converts them into formatted dbEST submission objects. The script first asks the user for cDNA library-specific information, which may be entered interactively or recalled from a previous session. The program offers two levels of vector elimination stringency via cross_match (P.Green, unpublished data) and also offers the opportunity to add some primary annotation to the dbEST submission in the form of the best similarity match found in a chosen protein database. After specifying a directory containing sequence traces, the process uses phred (Ewing and Green, 1998; Ewing *et al.*, 1998) to perform base calling. Any vector-derived sequence is removed, and user-specified leader/adaptor sequences may also be trimmed. Poly(A) tails are identified and deleted, and sequences that have more than 150 high-quality bases are used to create submission files. At this stage, a BLAST similarity search may be performed against a user-defined database to provide some preliminary annotation ('Similar to xyz...', with appropriate BLAST scores). Finally, the user is given the option to automatically submit the sequences to dbEST.

Process 2: creating partial genome databases

An overview of the construction of a partial genome as implemented in the PartiGene script is given in Figure 1. The script operates as a series of menu-listed steps. Each step may be interrupted at any time and the process can simply be restarted from where it left off. The first step collates the sequences in fasta format. PartiGene is able to download complete species-specific datasets from dbEST. When databases are updated, downloaded sequences are compared against the existing database and only new sequences are extracted. As not all database sequences will necessarily have been processed through trace2dbest (e.g. ESTs submitted by other research groups), these ESTs are first screened for any possible contaminating vector sequence, poly(A) tails, quality (presence of N bases) and size. Non-insert sequences are trimmed, and only those sequences >100 bases in length are used in subsequent processing.

The next step involves clustering the sequences on the basis of sequence similarity into groups that putatively derive from the same gene using our freely available program, CLOBB (Parkinson *et al.*, 2002). CLOBB has an advantage over other clustering solutions in that it readily performs incremental updates of datasets maintaining previous cluster identities. Clusters that contain more than one sequence are then used to derive a consensus sequence (putative gene sequence). This assembly step, based on phrap (P.Green, unpublished data), offers the user the ability to incorporate sequence quality information (produced by the base calling package, phred, in the trace2dbest script). We have used phrap in preference to the alternative cap3 because phrap creates fewer contigs for large clusters and includes the 'single-stranded' regions at the ends of contigs (which are therefore longer). Our

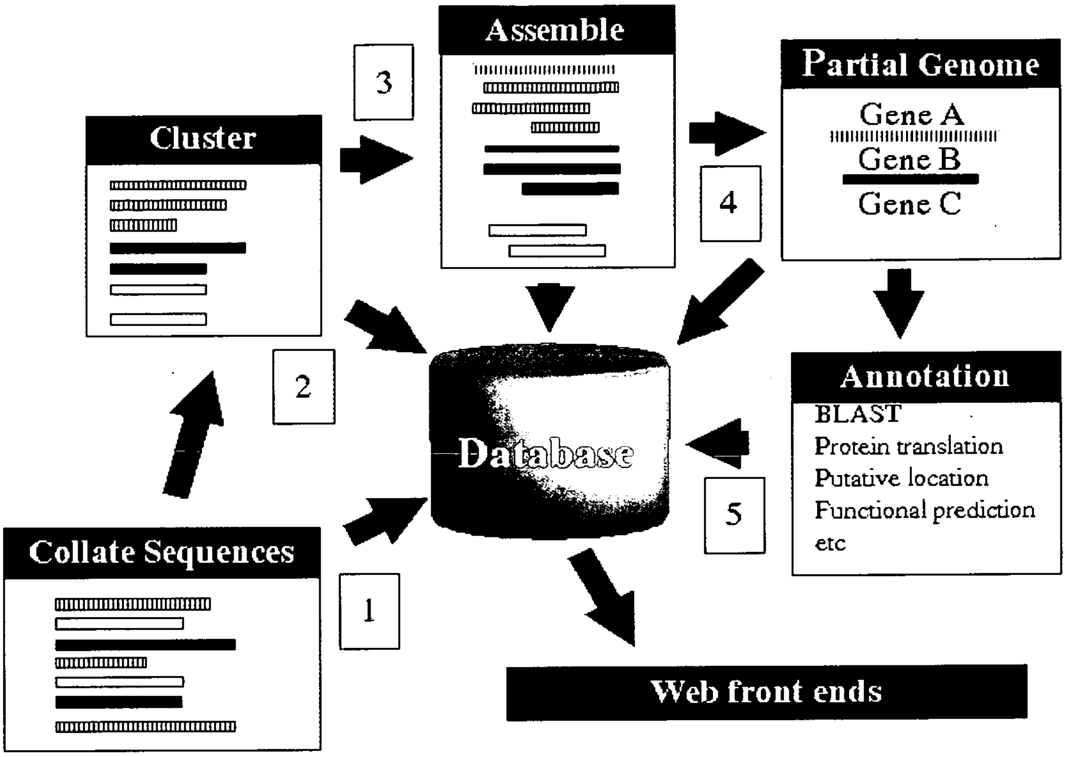


Fig. 1. An outline of the whole PartiGene process. (1) Sequences are collated either from local sources or via automatic download from GenBank dbEST. (2) Sequences are clustered on the basis of sequence similarity using CLOBB into groups that putatively derive from the same gene. (3) Clusters containing more than one sequence are assembled into consensus sequences. (4) The partial genome consists of these consensus sequences along with those clusters that contain only one sequence (termed singletons). (5) Putative genes are annotated by performing custom BLAST searches, peptide predictions, etc. and collated in a central database.

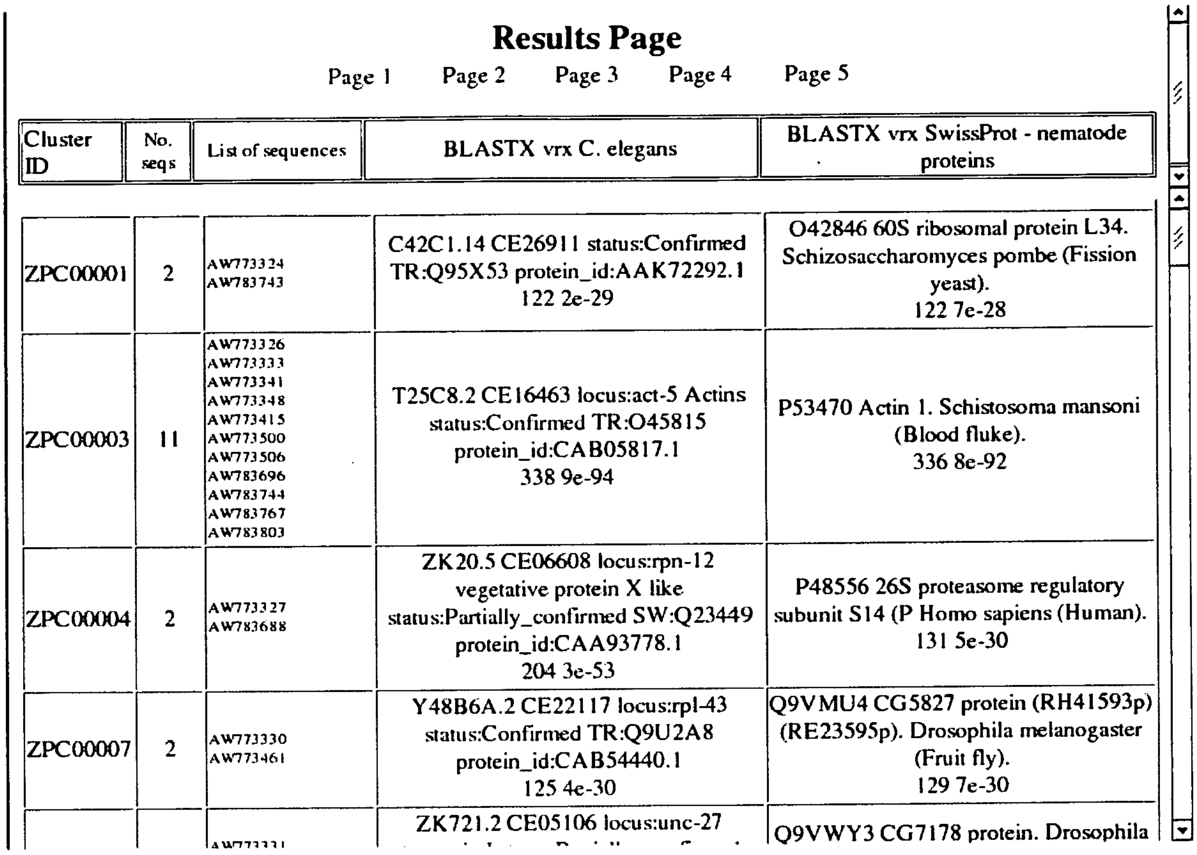
analysis of the recent releases of phrap have not identified the issues of base insertion and non-majority base calls identified in earlier publications. PartiGene offers three options in building consensus sequences: (1) use no trace quality data; (2) use quality data (if available) only for clusters containing two sequences; and (3) use quality data (if available) for all clusters. If no traces are available locally, sequences are given a default, modifiable phred score of 15 for each base position. Our preliminary analyses suggest that option 2 yields the fewest, high-quality contigs per cluster in mixed-source datasets.

The collection of clusters that contain only one sequence (termed singletons) and the sequence consensuses created in the assembly step above form the partial genome of the selected organism. A first-pass annotation of these putative genes is afforded by customizable BLAST similarity analyses. The user may select up to five different searches against locally available databases. As performing a large number of BLAST searches can be time-consuming, the user may halt the PartiGene process and perform such analyses independently: the search results can be imported.

To view and review these analyses, PartiGene offers two levels of access. For smaller datasets, a series of HTML summary tables can be written that provide information on each cluster including constituent members and summary BLAST output (Fig. 2). It is recommended that this be only undertaken for smaller datasets (less than 1000 clusters). The final step of the PartiGene script involves importing the data into a local database. PostgreSQL is implemented for its availability, functionality, development and support. If a database has not yet been created, PartiGene will automatically generate appropriate tables. Sequence, cluster and annotation data are then automatically imported.

Process 3: predicting protein translations

In current EST datasets, a small majority of the putative genes will have a BLAST similarity match to a database protein. However, a significant minority (up to 45%) will remain 'novel'. The majority of ESTs derive from protein-encoding genes, but translation of the putative genes identified by the PartiGene script is not trivial. Sequencing errors are commmon within ESTs and can lead to frameshift errors,



Results Page

Page 1 Page 2 Page 3 Page 4 Page 5

Cluster ID	No. seqs	List of sequences	BLASTX vrx C. elegans	BLASTX vrx SwissProt - nematode proteins
ZPC00001	2	AW773324 AW783743	C42C1.14 CE26911 status:Confirmed TR:Q95X53 protein_id:AAK72292.1 122 2e-29	O42846 60S ribosomal protein L34. Schizosaccharomyces pombe (Fission yeast). 122 7e-28
ZPC00003	11	AW773326 AW773333 AW773341 AW773348 AW773415 AW773500 AW773506 AW783696 AW783744 AW783767 AW783803	T25C8.2 CE16463 locus:act-5 Actins status:Confirmed TR:O45815 protein_id:CAB05817.1 338 9e-94	P53470 Actin 1. Schistosoma mansoni (Blood fluke). 336 8e-92
ZPC00004	2	AW773327 AW783688	ZK20.5 CE06608 locus:rpn-12 vegetative protein X like status:Partially_confirmed SW:Q23449 protein_id:CAA93778.1 204 3e-53	P48556 26S proteasome regulatory subunit S14 (P Homo sapiens (Human). 131 5e-30
ZPC00007	2	AW773330 AW773461	Y48B6A.2 CE22117 locus:rpl-43 status:Confirmed TR:Q9U2A8 protein_id:CAB54440.1 125 4e-30	Q9VMU4 CG5827 protein (RH41593p) (RE23595p). Drosophila melanogaster (Fruit fly). 129 7e-30
		AW773331	ZK721.2 CE05106 locus:unc-27	Q9VWY3 CG7178 protein. Drosophila

Fig. 2. HTML summary table for displaying cluster and associated BLAST annotation for ESTs derived from the nematode *Zeldia punctata*. For each cluster, the number and list of ESTs are provided, along with a brief description of the top hit from a BLAST search to a list of user defined databases (in this instance *Caenorhabdities elegans* proteins and SwissProt, with nematode proteins extracted, were selected). The page features links to individual and cluster consensus sequences and the detailed BLAST output for each cluster.

which may not be corrected by consensus sequence prediction. We have therefore developed prot4est, which combines state of the art programs to produce accurate protein predictions from PartiGene-processed ESTs (Fig. 3). prot4est is a six tier system of prediction. The first three tiers involve the use of BLAST annotation. First, potential RNA genes are identified, tagged and removed from the dataset. Second, all remaining sequences are searched using BLASTX against a protein database of choice [we recommend SwissProt; Boeckmann *et al.* (2003)]. If a sequence is found to share significant sequence similarity (expectation (E)-value $<e^{-20}$) to a database protein, the frame of translation used to resolve the match is assumed to be the correct frame of translation. The frame of translation for local regions within each of these sequences is determined, and using transeq (part of the EMBOSS package) a robust tiling path determined. A series of rules (see supplementary data on Web site) are then invoked to determine which, if any, potential start codons should be used. Potentially incorrect stop codons caused through errors in the sequencing process are identified through comparison of the high-scoring pairs (HSPs) and may be ignored. Third, potential mitochondrial proteins are identified, and for these, further processing implements translations using the relevant mitochondrial genetic codes.

Many sequences will not share significant similarity to a database entry. In such cases, *de novo* prediction software must be employed. prot4est combines two of the more successful programs, DECODER (Fukunishi and Hayashizaki, 2001), and ESTScan (Iseli *et al.*, 1999), to obtain accurate peptide predictions. Both require training sets, in the form of annotated complete coding sequences and codon usage tables, to identify coding regions. prot4est will automatically download this information from the relevant Web-based resource. DECODER uses codon bias tables (http://www.kazusa.or.jp/codon/) and coding sequences, while ESTScan relies on the availability of protein coding sequences (typically available from EMBL/Genbank, e.g. http://srs.ebi.ac.uk/).

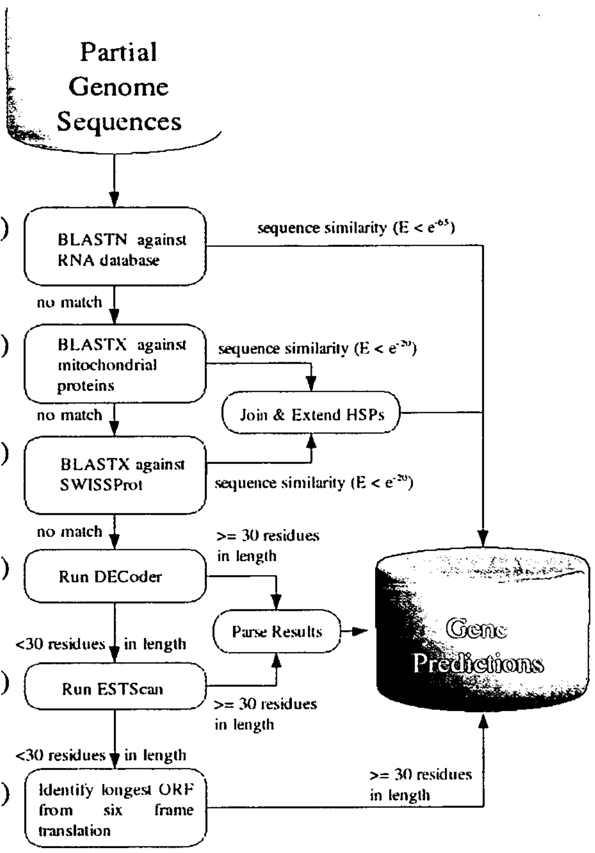


Fig. 3. How prot4est derives peptide sequence from low-quality EST data. Partial genome sequences derived from the PartiGene process are fed through a six tier system. (1) RNA genes are identified by BLASTN matches to a RNA gene database. (2) and (3) Nuclear and mitochondrially encoded protein-coding genes are identified on the basis of BLASTX similarity to known proteins. The BLAST output is analysed to allow extensions beyond the high scoring pairs. Sequences with significant sequence similarity to a known protein use the frame of translation designated by their common alignment to obtain an accurate peptide prediction. Peptides from sequences with no significant sequence similarity to a known protein are determined *de novo* using either DECODER (4) or ESTScan (5). If neither program predicts a peptide above a tunable length cutoff, six-frame translations of the sequence are identified (6) and the longest open reading frame extracted. Results each stage are collated in a central database.

We have determined that, for both these programs, reduced prior information significantly impacts translation quality (data not shown). If the acquired information is likely to be insufficient for accurate translations (less than 50 coding sequences for DECODER and less than 125 sequences for ESTScan), then the user is warned and the alternative of using data derived from a related species is offered.

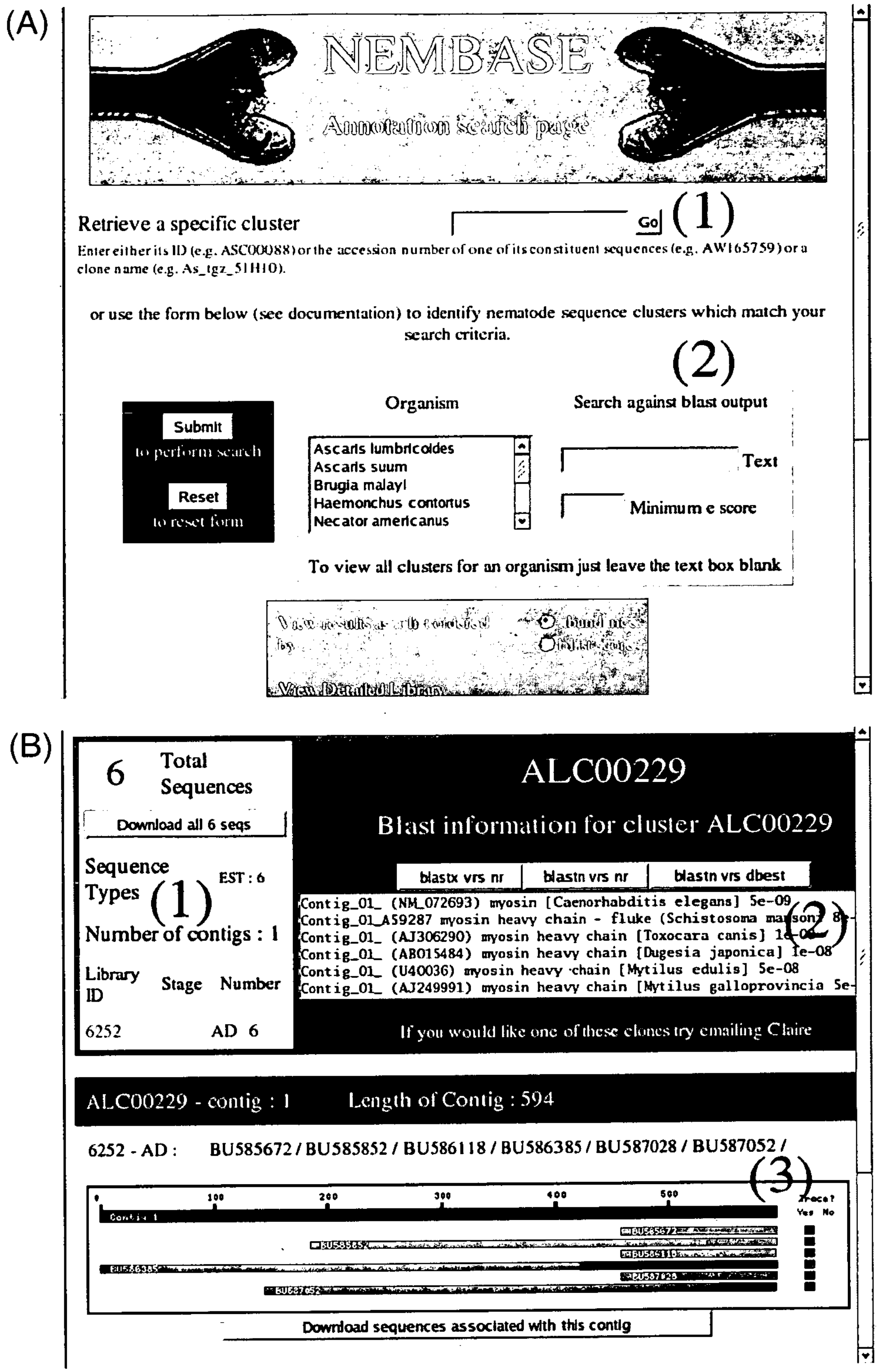
In the fourth step, then, sequences are passed through DECODER, which requires the availability of quality files. These are generated as part of the PartiGene process above. For sequence consensuses, the phrap derived quality file is used. For singletons the original trace quality file is used, or if this is not available, then as above, sequences are given a default, modifiable score of 15 for each base position. As DECODER only makes a prediction in the forward strand, a reverse complement of each singleton and consensus sequence is created to ensure that all frames are considered. DECODER was originally written for use on complete cDNAs, and it expects a start methionine, which may not always be present in incomplete EST sequences. prot4est therefore appends any peptide sequence upstream of the prediction made by DECODER, provided that no stop codons are encountered. If the peptide is less than 30 amino acids in length, the sequence is passed to the fifth step, ESTScan.

ESTScan builds hidden Markov models based on coding sequence nucleotide patterns to derive peptide sequence. prot4est takes these predictions and again adds upstream and downstream in-frame ORF translations. A 30-residue cutoff is again applied. The sixth step takes the remaining sequences, generates a six-frame translation and identifies the longest open reading frame (ORF). If the length of this ORF is less than 30 residues, the sequence is deemed to be non-coding.

prot4est peptide predictions may be imported into the SQL database created by PartiGene. Further annotation of these protein data, including pI, molecular weight and putative location, may then be generated and imported. Comments on the accuracy of translation for certain regions within the sequence are also passed by prot4est to the database.

Presentation of the partial genome

Although PartiGene offers the ability to view results in the form of simple HTML tables (Fig. 2), the creation of a local database provides a powerful resource for querying and presenting the data. PostgreSQL is based on the popular SQL syntax and provides an easily accessible interface to perform complex queries on the data. Alternatively, the user may wish to consider the use of Web-based forms to allow remote users, and those less experienced in computing, access to the data (Fig. 4). We have created a number of Web-accessible sites for presentation of our partial genome data including NEMBASE (http://www.nematodes.org/nematodeESTs/nembase.html), LophDB (http://www.nematodes.org/Lopho/LophDB.php) and LumbriBase (http://www.earthworms.org). Each site utilizes the Apache Web server (http://www.apache.org) to serve pages created using the PHP Web scripting language, which features a database interpreter (http://www.php.net). Examples of these scripts may be obtained from the authors.



Retrieve a specific cluster

Go

Enter either its ID (e.g. ASC00088) or the accession number of one of its constituent sequences (e.g. AW165759) or a clone name (e.g. As_tgz_51H10).

or use the form below (see documentation) to identify nematode sequence clusters which match your search criteria.

Submit to perform search

Reset to reset form

Organism

- Ascaris lumbricoides
- Ascaris suum
- Brugia malayi
- Haemonchus contortus
- Necator americanus

Search against blast output

Text

Minimum e score

To view all clusters for an organism just leave the text box blank

6 Total Sequences

Download all 6 seqs

Sequence Types EST : 6

Number of contigs : 1

Library ID	Stage	Number
6252	AD	6

ALC00229

Blast information for cluster ALC00229

blastx vrs nr | blastn vrs nr | blastn vrs dbest

```
Contig_01_ (NM_072693) myosin [Caenorhabditis elegans] 5e-09
Contig_01_A59287 myosin heavy chain - fluke (Schistosoma mansoni 8e-
Contig_01_ (AJ306290) myosin heavy chain [Toxocara canis] 1e-08
Contig_01_ (AB015484) myosin heavy chain [Dugesia japonica] 1e-08
Contig_01_ (U40036) myosin heavy chain [Mytilus edulis] 5e-08
Contig_01_ (AJ249991) myosin heavy chain [Mytilus galloprovincia 5e-
```

If you would like one of these clones try emailing Claire

ALC00229 - contig : 1 Length of Contig : 594

6252 - AD : BU585672 / BU585852 / BU586118 / BU586385 / BU587028 / BU587052 /

Download sequences associated with this contig

Fig. 4. Screenshots from NEMBASE showing Web pages created using the php scripting language to submit user queries to the underlying postgreSQL database. (**A**) Annotation search page. This form may be used to retrieve individual clusters by their unique ID (1) or groups of clusters by keywords associated with their BLAST annotation (2). (**B**) Detailed cluster page. This page provides information on a single cluster including the number and source of constituent sequences (1), summaries of BLAST annotation (2) and graphical views of the alignment of individual sequences to the cluster consensus (3). For details on the interpretation of this information, please see the NEMBASE help pages.

ACKNOWLEDGEMENTS

The authors would like to thank the Natural Environmental Research Council Environmental Genomics Thematic Data Centre, especially Dr Dan Swan and Dr Bela Tiwari, for their invaluable comments during the creation of the PartiGene system. This work was supported by the Wellcome Trust and the Natural Environmental Research Council Environmental Genomics Thematic Research Programme. J.W. is supported by the Biotechnology and Biological Sciences Research Council and Astra Zeneca, UK.

REFERENCES

Adams,M.D., Kerlavage,A.R., Fleischmann,R.D., Fuldner,R.A., Bult,C.J., Lee,N.H., Kirkness,E.F., Weinstock,K.G., Gocayne,J.D., White,O. *et al.* (1995) Initial assessment of human gene diversity and expression patterns based upon 83 million nucleotides of cDNA sequence. *Nature*, **377** (suppl.), 3–174.

Allen,J.E., Daub,J., Guiliano,D., McDonnell,A., Lizotte-Waniewski,M., Taylor,D.W. and Blaxter,M. (2000) Analysis of genes expressed at the infective larval stage validates utility of *Litomosoides sigmodontis* as a murine model for filarial vaccine development. *Infect. Immun.*, **68**, 5454–5458.

Blaxter,M., Daub,J., Guiliano,D., Parkinson,J., Whitton,C. and The Filarial Genome Project (2002) The *Brugia malayi* genome project: expressed sequence tags and gene discovery. *Trans. R. Soc. Trop. Med. Hyg.*, **96**, 7–17.

Boeckmann,B., Bairoch,A., Apweiler,R., Blatter,M.-C., Estreicher,A., Gasteiger,E., Martin,M.J., Michoud,K., O'Donovan,C., Phan,I., Pilbout,S. and Schneider,M. (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.*, **31**, 365–370.

Boguski,M.S. and Schuler,G.D. (1995) ESTablishing a human transcript map. *Nat. Genet.*, **10**, 369–371.

Boguski,M.S., Lowe,T.M. and Tolstoshev,C.M. (1993) dbEST—database for 'expressed sequence tags'. *Nat. Genet.*, **4**, 332–333.

Christoffels,A., van Gelder,A., Greyling,G., Miller,R., Hide,T. and Hide,W. (2001) STACK: Sequence Tag Alignment and Consensus Knowledgebase. *Nucleic Acids Res.*, **29**, 234–238.

Daub,J., Loukas,A., Pritchard,D.I. and Blaxter,M. (2000) A survey of genes expressed in adults of the human hookworm, *Necator americanus. Parasitology*, **120**, 171–184.

Ewing,B. and Green,P. (1998) Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res.*, **8**, 186–194.

Ewing,B., Hillier,L., Wendl,M.C. and Green,P. (1998) Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res.*, **8**, 175–185.

Fukunishi,Y. and Hayashizaki,Y. (2001) Amino-acid translation for cDNA with frame-shift error. *Physiol. Genomics.*, **5**, 81–87.

Iseli,C., Jongeneel,C.V. and Bucher,P. (1999) ESTScan: a program for detecting, evaluating, and reconstructing potential coding regions in EST sequences. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **7**, 138–158.

Kenyon,F., Welsh,M., Parkinson,J., Whitton,C., Blaxter,M.L. and Knox,D.P. (2003) Expressed sequence tag survey of gene expression in the scab mite *Psoroptes ovis* allergens, proteinases and free radical scavengers. *Parasitology*, **126**, 451–460.

Paquola,A.C., Nishyiama,M.Y.Jr, Reis,E.M., da Silva,A.M. and Verjovski-Almeida,S. (2003) ESTWeb: bioinformatics services for EST sequencing projects. *Bioinformatics*, **19**, 1587–1588.

Parkinson,J., Guiliano,D.B. and Blaxter,M. (2002) Making sense of EST sequences by CLOBBing them. *BMC Bioinformatics*, **3**, 31.

Parkinson,J., Mitreva,M., Hall,N., Blaxter,M. and McCarter,J. (2003) 400,000 nematode ESTs on the net. *Trends Parasitol*, **19**, 283–286.

Pertea,G., Huang,X., Liang,F., Antonescu,V., Sultana,R., Karamycheva,S., Lee,Y., White,J., Cheung,F., Parvizi,B., Tsai,J. and Quackenbush,J. (2003) TIGR gene indices clustering tools (TGICL): a software system for fast clustering of large EST datasets. *Bioinformatics*, **19**, 651–652.

Sutton,G.G., White,O., Adams,M.D. and Kerlavage,A.R. (1995) TIGR assembler: a new tool for assembling large shotgun sequencing projects. *Gen. Sci. Technol.*, **1**, 9–19.

White,O. and Kervalage,A.R. (1996) TDB: new databases for biological discovery. *Methods Enzymol.*, **266**, 27–40.

BMC Bioinformatics

BioMed Central

Software

Open Access

prot4EST: Translating Expressed Sequence Tags from neglected genomes

James D Wasmuth* and Mark L Blaxter

Address: Institute of Evolutionary Biology, School of Biological Sciences, University of Edinburgh, EH9 3JT, UK

Email: James D Wasmuth* - james.wasmuth@ed.ac.uk; Mark L Blaxter - mark.blaxter@ed.ac.uk

* Corresponding author

Published: 30 November 2004

BMC Bioinformatics 2004, 5:187 doi:10.1186/1471-2105-5-187

Received: 23 August 2004
Accepted: 30 November 2004

This article is available from: http://www.biomedcentral.com/1471-2105/5/187

Abstract

Background: The genomes of an increasing number of species are being investigated through generation of expressed sequence tags (ESTs). However, ESTs are prone to sequencing errors and typically define incomplete transcripts, making downstream annotation difficult. Annotation would be greatly improved with robust polypeptide translations. Many current solutions for EST translation require a large number of full-length gene sequences for training purposes, a resource that is not available for the majority of EST projects.

Results: As part of our ongoing EST programs investigating these "neglected" genomes, we have developed a polypeptide prediction pipeline, prot4EST. It incorporates freely available software to produce final translations that are more accurate than those derived from any single method. We show that this integrated approach goes a long way to overcoming the deficit in training data.

Conclusions: prot4EST provides a portable EST translation solution and can be usefully applied to >95% of EST projects to improve downstream annotation. It is freely available from http://www.nematodes.org/PartiGene.

Background

The need for more sequence

Complete genome sequencing is a major investment and is unlikely to be applied to the vast majority of organisms, whatever their importance in terms of evolution, health or ecology. Complete genome sequences are available for only a few eukaryote genomes, most of which are model organisms. The focus of eukaryote genome sequencing has been on a restricted subset of known diversity, with, for example, nearly half of the completed or draft stage genomes being from vertebrates. While Arthropoda and Nematoda have two completed genomes each, with a dozen others in progress, compared to predicted diversity (over a million species each) current genome sequencing illuminates only small parts of even these phyla. The disparity between sequence data and motivation for biological study is significant. Allied to this bias in genome sequence is a bias in functional annotation for the derived proteomes: a vertebrate gene is more likely to have been assigned a function due to the focus of biomedical research on humans and closely related model species such as mouse [1].

Shotgun sample sequencing of additional genomes through expressed sequence tags (EST) or genome survey sequences (GSS) has proved to be a cost-effective and rapid method of identifying a significant proportion of the genes of a target organism. Thus many genome initiatives on non-traditional model organisms have utilised EST and GSS strategies to gain an insight into "wild" biology. An EST strategy does not yield sequence for all of the expressed genes of an organism, because some genes may not be expressed under the conditions sampled, and others may be expressed at very low levels and missed through the random sampling that underlies the strategy. However the creation of EST libraries from a range of conditions, such as different developmental stages or environmental exposures, promotes a closer examination of the biology of these species.

The well documented phylogenetic sequence deficit [2] has led us to coin the term "neglected genomes". Currently many groups are sequencing ESTs from their chosen species to perform studies in a wide-range of disciplines, from comparative ecotoxicology [3] to high-throughput detection of sequence polymorphisms [4,5]. The contribution of EST projects for neglected but biologically relevant organisms is highlighted in Figure 1. As with all sequence data, obtaining high quality annotation requires prior information and is labour intensive. The "partial genome" information that results from EST datasets presents special problems for annotation, and we are developing tools for this task.

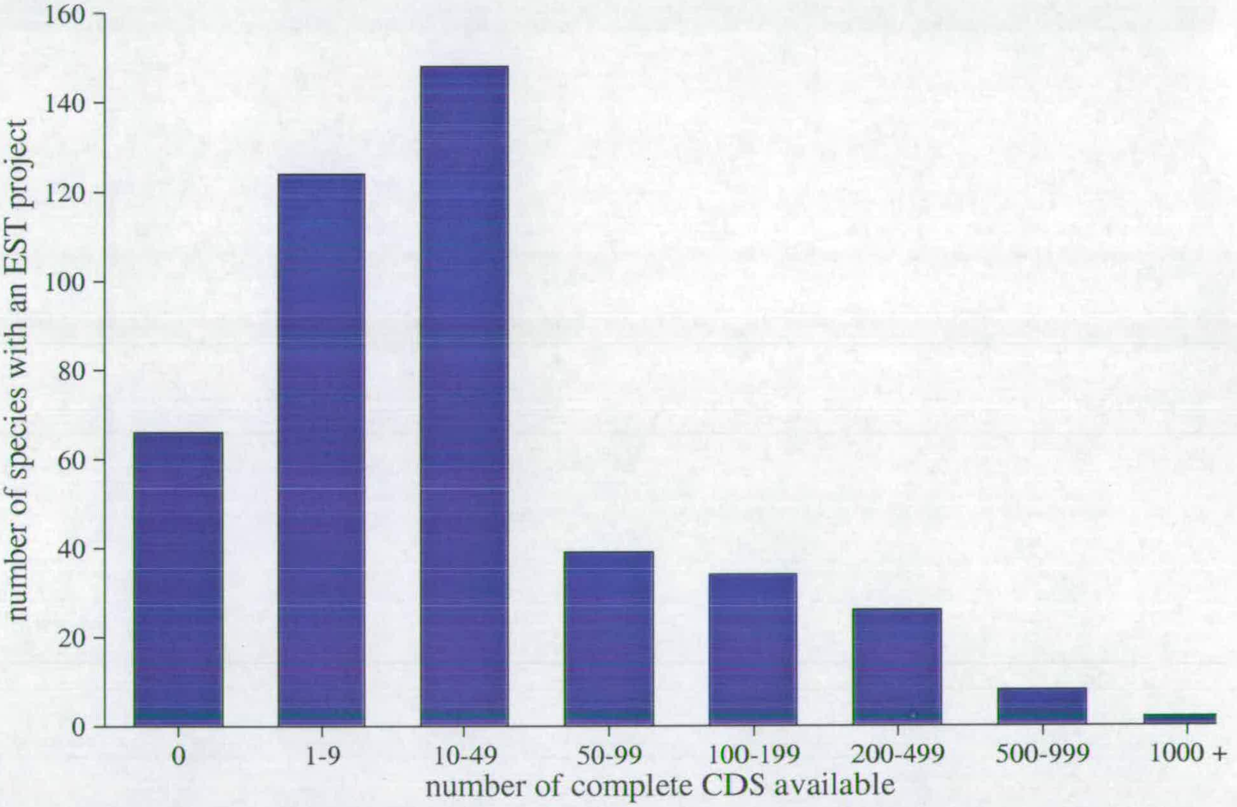


Figure 1
The training set deficit for EST projects. Around 85% of species with representation in dbEST (>100 ESTs) have less than 100 complete CDS entries in the EMBL database. These species comprise ~45% of all ESTs. Sixty-six species, with 246263 dbEST sequences, have no full-length CDS. Source: dbEST and EMBL database (July 2004).

The need for high quality translation

The PartiGene software suite [6] simplifies the analysis of partial genomes. ESTs are clustered into putative genes and consensuses determined. All the data is stored in a relational database, allowing it to be searched easily. While preliminary annotation based on BLAST analysis of nucleotide sequence can be performed, more robust methods are needed to allow high-quality analysis. The error-prone nature of ESTs makes application of most annotation tools difficult. To improve annotation, and facilitate further exploitation, a crucial step is the robust translation of the EST or consensus to yield predicted polypeptides. The polypeptide sequences present a better template for almost all annotation, including InterPro [7] and Pfam [8], as well as the construction of more accurate multiple sequence alignments, and the creation of protein-mass fingerprint libraries for proteomics exploitation. High quality polypeptide predictions can be applied to functional annotation and post-genomic study in a similar way to those available for completed genomes.

Translating Expressed Sequence Tags

Prediction of the correct polypeptide from ESTs is not trivial:

1. The inherent low quality of EST sequences may result in shifts in the reading frame (missing or inserted bases) or ambiguous bases. These errors impede the correct recognition of coding regions. The initiation site may be lost, or an erroneous stop codon introduced to the putative translation.

2. ESTs are often partial segments of a mRNA, and as most cloning technology biases representation to the internal parts of genes, the initiation methionine codon may be missed. This is a problem for some of the *de novo* programs which use the initiation methionine to identify the coding region (described below).

Sequence quality can be improved by clustering the sequences based on identity. For each cluster a consensus can be determined [9]. This approach, however, will not address the whole problem as poor quality EST sequences may not yield high quality consensuses and for smaller volume projects, most genes have a single EST representative. Therefore additional methods must be applied to provide accurate polypeptide predictions.

Similarity-based methods

A robust method to determine the correct encoded polypeptide is to map a nucleotide sequence onto a known protein. This concept is the basis for BLASTX [10], FASTX [11] and ProtEST [12]. BLASTX and FASTX use the six frame translation of a nucleotide sequence to seed a search of a protein database. The alignments generated for each significant hit provide an accurately translated region of the EST. BLASTX is extremely rapid, but the presence of a frameshift terminates each individual local alignment, ending the polypeptide prematurely. FASTX is able to identify possible frameshifts, but its dynamic programming approach is significantly slower than BLASTX. These methods require that the nucleotide sequence shares detectable similarity with a protein in the selected database. Many genes from both well studied and neglected genomes do not share detectable similarity to other known proteins. For example, the latest analysis of the *Caenorhabditis elegans* proteome shows that only ~50% of the 22000 predictions contain Pfam-annotated protein domains [8,13], and 40% share no significant similarity with non-nematode proteins in the SwissProt/trEMBL database [14]. This feature is not unique to the phylum Nematoda, and is likely perhaps to be more extreme for neglected genomes, given the phylogenetic bias of most protein databases.

ProtEST uses a slightly different similarity-based approach [12]. A protein sequence is compared to an EST database. phrap [9] is used to construct a consensus sequence from the ESTs found to have significant similarity. These consensuses are then compared to the original sequence using ESTWISE (E. Birney, unpublished [15]) giving a maximum likelihood position for possible frameshifts. The system is accurate but is not readily adaptable to the high-throughput approach necessary when dealing with very large numbers of ESTs. More crucially, an EST that does not show significant similarity to a known protein is not translated.

'de novo' predictions

To overcome the reliance upon sequence similarity, *de novo* approaches based on recognition of potential coding regions within poor quality sequences, reconstruction of the coding regions in their correct frame, and discrimination between ESTs with coding potential and those derived from non-coding regions have been developed [16-18].

DIANA-EST [16], combines three Artificial Neural Networks (ANN), developed to identify the transcription initiation site and the coding region with potential frameshifts. ESTScan2 [18] combines three hidden Markov models trained to be error tolerant in their representations of mRNA structure (modelling the 5' and 3' untranslated regions, initiation methionine and coding region). DECODER [17] uses an essentially rule-based method for identifying possible insertions and deletions in the nucleotide sequence, as well as the most likely initiation site, and was developed for complete cDNA sequence translation.

Each of these methods has different strengths in their attempt to identify the precise coding region; all require prior data to train their models. Published descriptions of their utility are based on training with human full length coding sequences (mRNAs), and thus tens of thousands of training sequences (many million coding nucleotides) were used to achieve optimum results. As stressed above, this amount of prior data is not available for the vast majority of EST project species (Figure 1).

New solution – prot4EST

Prior to this project, nematode ESTs available through NEMBASE [19] had been translated using DECODER, as a preliminary study had suggested that it outperformed the other available methods (DIANA-EST and ESTScan1 [20]) (Parkinson pers. com.). 7388 out of the 40000 resulting predicted polypeptides were likely to be poorly translated (<30 amino acids), and we suspected many more contained errors. This motivated the creation of a solution using several methods to enhance the quality of the polypeptide predictions, exploiting their strengths while recognising their short-comings. prot4EST is an EST translation pipeline, written in Perl, with a user-friendly interface, that links some of these described methods together. It carries out retrieval and formatting of files from online databases for the user. It has been designed to be used as a stand-alone tool, or as an integral part of the PartiGene process [6].

Implementation

DECODER

The DECODER program [17] was developed to define start codons and open reading frames in full-length cDNA sequences. It exploits the quality scores for the sequence produced from base-calling software, such as phred [21,22], and additional text-based information to identify all possible coding regions. In regions of low sequence quality up to 2 nucleotides are removed or inserted, representing possible frameshifts. A likelihood score is calculated for each possible coding sequence (CDS), and the one with the lowest score is chosen as the correct CDS. The score is computed from the probability of generating a random sequence with a better Kozak consensus (the nucleotide sequence surrounding the initiation codon of a eukaryotic mRNA), ATG position and codon usage. DECODER requires a codon bias table, which is used to determine the putative coding regions optimal codon usage. A penalty term limits the number of insertions/deletions in the corrected CDS.

ESTScan2.0

Hidden Markov models (HMM) can represent known sequence composition in a probabilistic manner [23]. This has been exploited recently in applications to find genes in genomic sequence [24,25], predict domain composition in protein sequences [26], and align multiple sequences [27]. ESTScan [18] exploits the predictive power of Hidden Markov models by combining three models:

1. Modeling mRNA structure: ESTScan separates the probable CDS from the untranslated regions (UTRs). The core of the coding sequence is represented by a 3-periodic inhomogeneous hidden Markov model. Flanking this core model are start and stop profiles for the codons observed at these positions. The profiles for untranslated regions flank the start and stop states.

2. Error tolerance: ESTScan allows insertions and deletions (indels) in the EST sequence. For example, if it is more probable that a particular nucleotide is the result of an insertion event then it is omitted from the 'corrected' sequence. Conversely, if the HMM probability scores suggest that a nucleotide has been deleted then the model inserts an X into the 'corrected' sequence to denote this prediction.

3. EST structure: ESTScan recognises that the EST may be composed of a combination of 5' UTR, CDS and 3' UTR.

ESTScan's hidden Markov models are trained using complete CDS entries from either the EMBL or RefSeq databases. Scripts included with the distribution parse the data files, extracting the necessary sequence information to produce the model files. The major issue considered at this point is redundancy. If the training data is internally redundant then the resultant model will be fully successful only in finding what is known and will have reduced power in detecting novel transcripts. Default parameters were used in ESTScan for building the HMM and in predicting polypeptides.

HSP tiling

The BLASTX program [10] allows a nucleotide sequence to be searched against a protein database. The nucleotide query is translated in all six frames and these are used as the query sequences for a BLASTP search. High scoring segment pairs (HSP) are identified that maximise a bit score derived from an amino acid similarity matrix. If a single indel occurs in the nucleotide sequence, causing a frameshift, the HSP is either terminated at this position or continues out of frame. Downstream of this frameshift the query sequence may be long enough to result in another significant HSP to the same protein sequence, this time in a different frame. Simple extraction of the best BLAST HSP will miss such features. prot4EST implements a rule-based method that considers all the HSPs for a match to a database sequence and considers whether a frameshift can be identified. Where a frameshift is identified the HSPs are joined. Where two HSPs overlap the sequence with the better bit score is used.

The prot4EST pipeline

prot4EST is an integrated pipeline utilising freely available software in a tiered, rule-based system (Figure 2).

Tier 1: Identification of ribosomal RNA (rRNA) genes

The protein databases contain (probably spurious) translations of ribosomal RNA genes and gene fragments, and thus it is important to identify and remove putative rRNA derived sequences before further processing. A BLASTN search is performed against a database of rRNA sequences obtained from the Ribosomal Database II (Table 1; [28]). A BLAST expect value cutoff of e-65 is used to identify matches. The cutoff is a conservative one to reduce the number of false positives. Those nucleotide sequences with significant matches are annotated as rRNA genes and take no further part in the translation process.

Tiers 2 and 3: Similarity search

The second and third stages are similar. First a BLASTX search is performed against proteins encoded by mitochondrial genomes. The mitochondrial protein database is obtained from the NCBI ftp site (Table 1). Any sequences with significant hits (cutoff e-8) are annotated as mitochondrion-encoded genes for the remainder of the process, and the relevant mitochondrial genetic code is used for translation. Sequences that do not have significant similarity to mitochondrial proteins are compared using BLASTX to the SwissProt database [14]. Sequences that yield no significant similarity are moved onto tier 4 of the process.

For those sequences that show significant similarity to a protein sequence from either database a HSP tile path is constructed. prot4EST then considers whether the nascent translation can be extended at either end in the same reading frame.

Tier 4: ESTScan prediction

The hidden Markov models used by ESTScan to identify the coding region are constructed from EMBL format files for complete CDS using scripts supplied with the package. Preprocessing is integrated within prot4EST, including the downloading of the EMBL files. A pair of length threshold criteria are applied to each putative polypeptide before it is accepted. The open reading frame must be at least 30 codons in length, and cover at least 10% of the input sequence. Polypeptides that satisfy these criteria undergo the extension process described above, sequences that fail any of the criteria are passed onto the next tier. The extension process is carried out on those sequences that exceed the thresholds.

Tier 5: DECODER prediction

The DECODER program is used to predict CDS and thus polypeptide translations for the remaining nucleotide sequences. For each sequence a quality file in phrap format is required. When a quality file is unavailable a file with quality values of 15 is generated for each sequence. The codon usage table required by DECODER can be specified by the user or downloaded from CUTG, the codon usage table database [29]. By default DECODER only processes the forward strand of each sequence, and therefore the reverse complement of each sequence is taken and processed through DECODER. Two putative polypeptides are generated for each nucleotide sequence. The longer polypeptide is selected as the more probable translation. The polypeptide predictions are checked using the same length threshold criteria as for ESTScan (above).

Tier 6: Longest ORF

This last attempt to provide a putative polypeptide translation determines the longest string of amino acids uninterrupted by stop codons from a six-frame translation of the sequence. If a methionine is present in this string it is flagged as a potential initiation site.

Output

The primary output from prot4EST consists of the putative polypeptides in FASTA format, complemented with files containing information describing the translated sequences. This information includes:

position of the translation with respect to the nucleotide sequence, the genetic code used for translation,

position and BLAST statistics of HSPs used in the tile path.

All this additional information is stored in two CSV format files, permitting parsing and simple insertion into a database.

Speed

This is highly dependent upon the composition and size of the dataset. As a guide, each prot4EST run carried out in the benchmarking (below), took less than an hour for a 2316-sequence input with an Athlon 1400 Mhz processor. The BLASTX searches were carried out separately and used as input to prot4EST (for details see the userguide, availabile from the program web page).

Benchmarking EST translation methods

We benchmarked five translation methods to test their relative performance. DECODER is designed to consider only the forward strand of the nucleotide sequence, as it was originally designed for full-length CDSs. When applied to ESTs it is imperative that both strands are

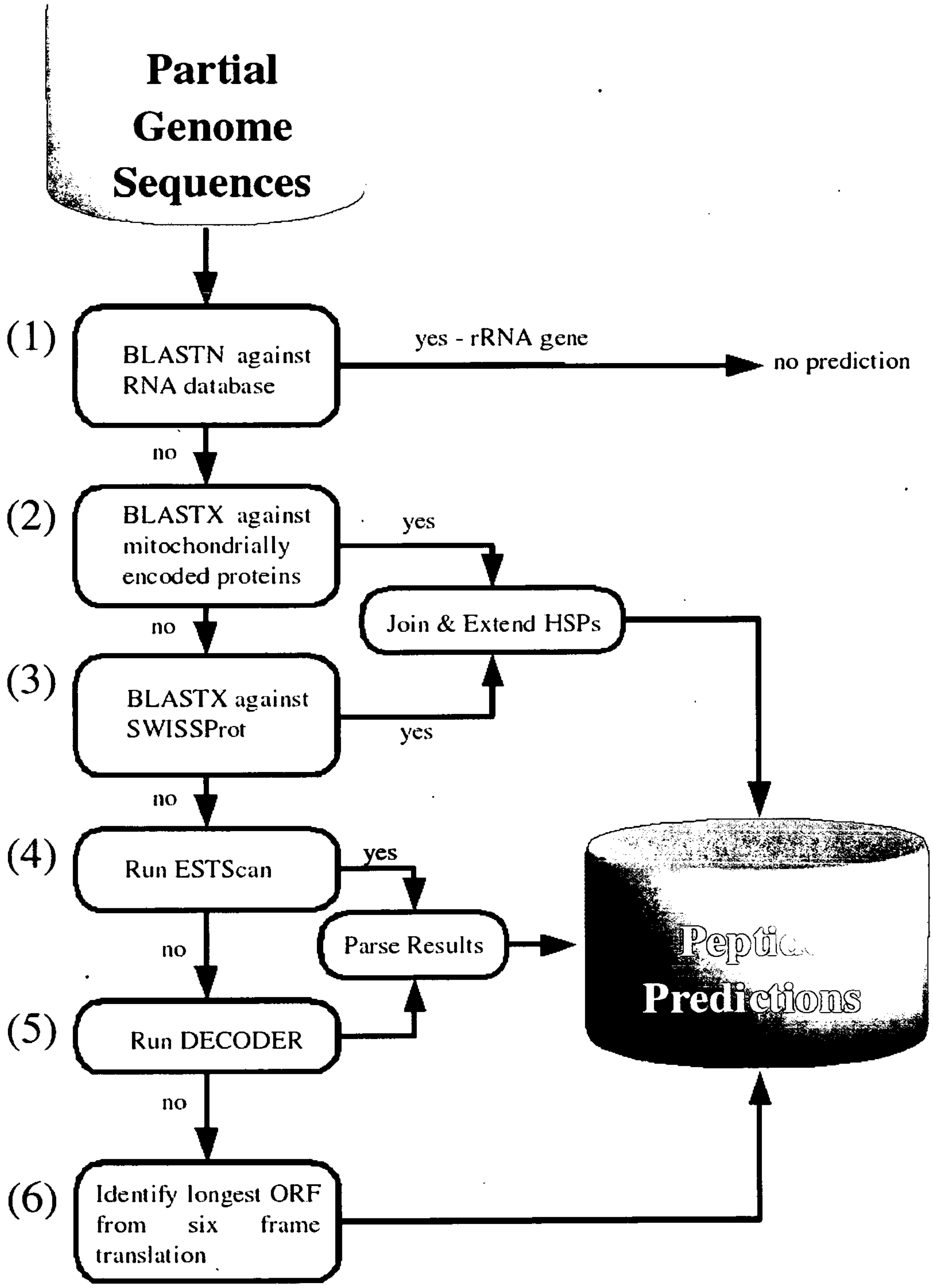


Figure 2
The prot4EST pipeline.

Table 1: Description of databases used for similarity searches.

Source	*Tier**	*Database*	*Link*
ribosomal RNA	1	RBP II	http://rdp.cme.msu.edu
mitochondrial proteins	2	NCBI	ftp://ftp.ncbi.nih.gov/blast/db/
protein sequences	3	SwissProt/TrEMBL	http://ca.expasy.org/sprot/

*the stage in which the database is used in prot4EST pipeline (see Figure 2).

analysed, as both 5' and 3' ESTs are generated. Therefore the reverse complement of each nucleotide consensus was also analysed. DECODER_default (1) considers only the prediction from the forward strand, whilst DECODER_best (2) uses the more accurate prediction. ESTScan (3) considers both strands of the nucleotide sequence, and was run as a stand-alone process with default settings.

Two arrangements of components within prot4EST were tested. prot4EST_ed (4) implements ESTScan before using DECODER on any remaining untranslated sequences. Conversely, prot4EST_de (5) uses DECODER first followed by ESTScan. The DECODER module in prot4EST considers translations on both the foward and reverse strands of the query sequence.

1 Data Sets

Test EST dataset for translation

We randomly selected 4000 *Caenorhabditis elegans* ESTs from dbEST [30]. To reduce redundancy, the ESTs were clustered using CLOBB [31]. phrap [9] was then used to derive a consensus sequence for each cluster. This resulted in 2899 nucleotide sequences. To ensure that the consensuses corresponded to a coding region, we carried out a BLASTN search for each consensus against the complete *C. elegans* cDNA dataset available from Wormbase (version 117) [32]. Significant matches were found for 2372 consensuses. Finally, this set was used to query the *C. elegans* protein dataset (Wormpep version 117), thus associating each nucleotide sequence with a corresponding reference polypeptide. A final test set of 2316 consensus sequences was produced.

Training datasets

1: Caenorhabditis elegans

Both ESTScan and DECODER require prior gene sequence. The *C. elegans* RefSeq collection was obtained, comprising 21033 entries (December 2003; [33]). A Perl script constructed random training sets giving differing totals of coding nucleotides from 10000 to 350000. Four sets were assembled for each level. The build_tables script (part of the ESTScan package) was used to filter out sequences [18].

We used the same training sets to build the codon usage tables required by DECODER. CUSP from EMBOSS [34] was used to build the tables, and a separate Perl script written to convert the output to that required by DECODER. For any given run of prot4EST the ESTScan HMM training set and codon usage table used were derived from the same training set of *C. elegans* cDNAs.

2: Prokaryote genomes

GenBank entries from 167 complete prokaryote genomes were obtained (May 2004). A Perl script was written to extract the CDS entries and construct a RefSeq-style resource for each prokaryote species (available upon request). If a taxon's genome consisted of more than one megaplasmid the sequences were combined. CDS annotation was not available for 11 genomes. We used the CDS collections for the 156 taxa to determine AT content, construct hidden Markov models and codon usage tables.

3: Arabidopsis thaliana

28960 complete CDS entries for *A. thaliana* were obtained from the RefSeq database [35].

4: Spirurida (Nematoda)

We queried GenBank for all complete CDS entries from species in the nematode order Spirurida.

BLAST databases

SwissProt (release 42.7) and TrEMBL (release 25.7) [14] were combined to give a SwissAll database. To recreate the situation facing neglected genome analysis, the accession numbers for all proteins from species in the nematode orderRhabditida were retrieved from the NEWT taxonomic database [36] and these entries (~23000) were removed from SwissAll.

2 Data collection and analysis

Comparison of predicted polypeptides to the 'true' polypeptide

We compared each putative polypeptide predicted from the *C. elegans* test dataset to its cognate reference protein using bl2seq from the NCBI distribution. Default parameters were used except for the theoretical database size (-d), set to 130000, the size of SwissProt. The blast reports were parsed using BioPerl modules [37]. Each *C. elegans* reference protein sequence was also compared to itself using bl2seq with default parameters. The raw and bit scores were recorded.

Calculation of comparison statistics

The raw and bit scores were normalised for length and against their theoretical maximum using equation 1, where:

BITlocal is the bit score of the local alignment between the predicted polypeptide and its cognate reference protein,

BITmax is the bit score for the alignment between the reference protein and its self,

WPlength is the length of the wormpep protein that is the reference of the nucleotide consensus translated,

ESTlength is the length of the nucleotide consensus that has been translated.

$$\text{Normalised Bit Score} = \frac{\text{BITlocal}}{\text{BITmax}} \times (\frac{3 \times \text{WPlength}}{\text{ESTlength}})$$

(equation 1)

Results and discussion

To measure the accuracy of translation two statistics were derived from the comparison of the predicted and reference polypeptides. The coverage is the percentage of the predicted polypeptide that aligns with the reference. The bit score represents the total of the alignment's pair-wise scores, normalised with respect to the substitution matrix used to calculate these scores. In this study the bit score was itself normalised to compensate for EST length and the maximum possible bit score for each comparison (see Methods, equation 1). The number of consensuses translated that had a significant match to their cognate reference *C. elegans* protein was also recorded for each run.

The influence of number of training codons

Both variants of DECODER were unable to produce robust translations for over half the nucleotide sequences no matter how many nucleotides were in the training set (Figure 3). As expected, the inclusion of the reverse complement in the DECODER analysis improved its performance. The inability of DECODER to translate more than 50% of the polypeptides can be traced to its core assumptions. One criterion used is the determination of the most likely initiation methionine. While this is almost always present in full length cDNAs (for which it was designed), the occurrence of any ATG codon in EST consensuses is less certain. We noted that DECODER will try any ATG codon to start its prediction, even if this results in a polypeptide of 2 amino acids in length.

The effect of the number of training nucleotides on ESTScan performance is pronounced. For the majority of the replicates, at each training set size the fraction of predictions that have significant matches to their reference sequence was around 75%, but the number of translations dropped significantly below 250000 training nucleotides. However, for 10000 coding nucleotides or less no robust translations are produced. Additionally, there was variance in the performance of ESTScan when there were between 20000 and 50000 training nucleotides. Examination of these training sets showed no difference in AT content compared to larger training sets, but did suggest that fluctuations in codon usage bias might be involved. The replicates that performed less well comprised sequences with shorter mean length, and had codon biases that were at the extremes of the distribution (not shown). This variation in sequence composition clearly has an effect on the probabilities that populate the HMM used by ESTScan. We suspect that the ability of ESTScan to predict robust translations when trained by datasets of 150000 to 200000 coding nucleotides is inflated as a consequence of the random selection of the training set from the complete *C. elegans* transcriptome. In a genuine situation, when only a small number of full-length CDS exist in the public databases, a significant number .will be from highly expressed genes with atypical codon bias and structure. This bias will be evident in real-world CDS sets with fewer than 200 members (150000–200000 coding nucleotides).

When the training sets contained a large number of non-redundant coding nucleotides (> 150000), prot4EST_ed and ESTScan performed equally well (Figure 3a). When the number of coding nucleotides available for training and codon bias determination were reduced, prot4EST translations still showed significant similarity to the correct protein in at least 80% of instances.

The translations produced by prot4EST_ed were the most robust across all totals of coding nucleotides, for both coverage and bit score (Figures 3b & 3c). As the number of coding nucleotides used in training decreased, both measures showed slight reductions.

Performance of alternative prot4EST architectures

prot4EST_ed produced more robust translations for higher numbers of training sequences. However when smaller totals of training nucleotides were used the translations produced by the alternative architecture, prot4EST_de, were slightly better (Figure 3c), although a smaller proportion of translations were produced with this setup (Figure 3a).

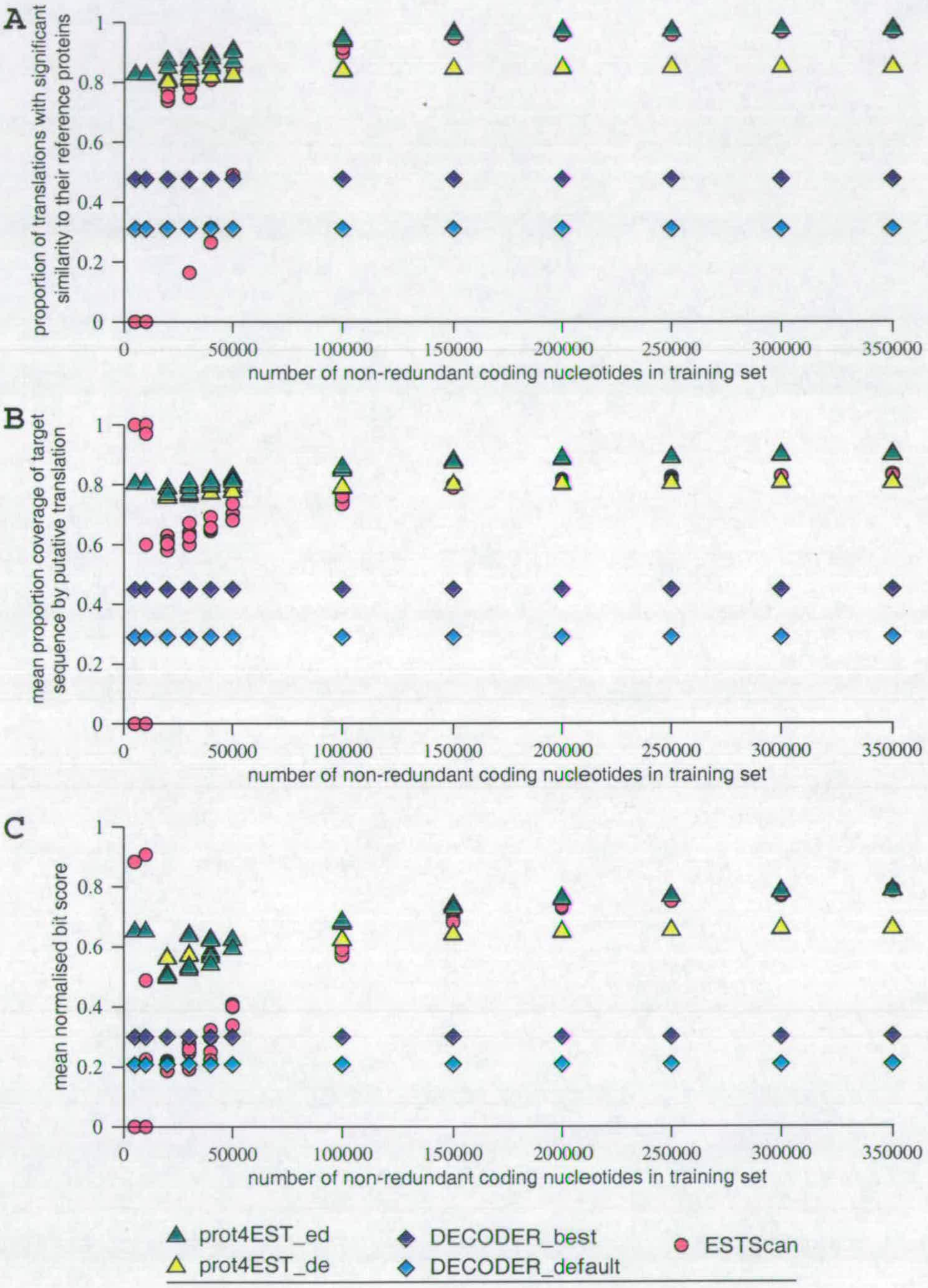


Figure 3

Performance of polypeptide prediction methods under different training regimes. Predicted polypeptides were compared to their reference. Four independent replicates of each training set size were used. a) Proportion of predicted polypeptide peptides having a significant BLASTP match to their reference protein. b) The mean proportion of each sequence covered by a predicted polypeptide. c) The mean relative bit score of each predicted polypeptide compared to its reference protein. The scores in b) and c) are the mean of the sequences translated by each method. The high scores shown by ESTScan at 5000 and 10000 non-redundant coding nucleotides is due to the method returning at most one polypeptide out of the 2316 nucleotides provided.

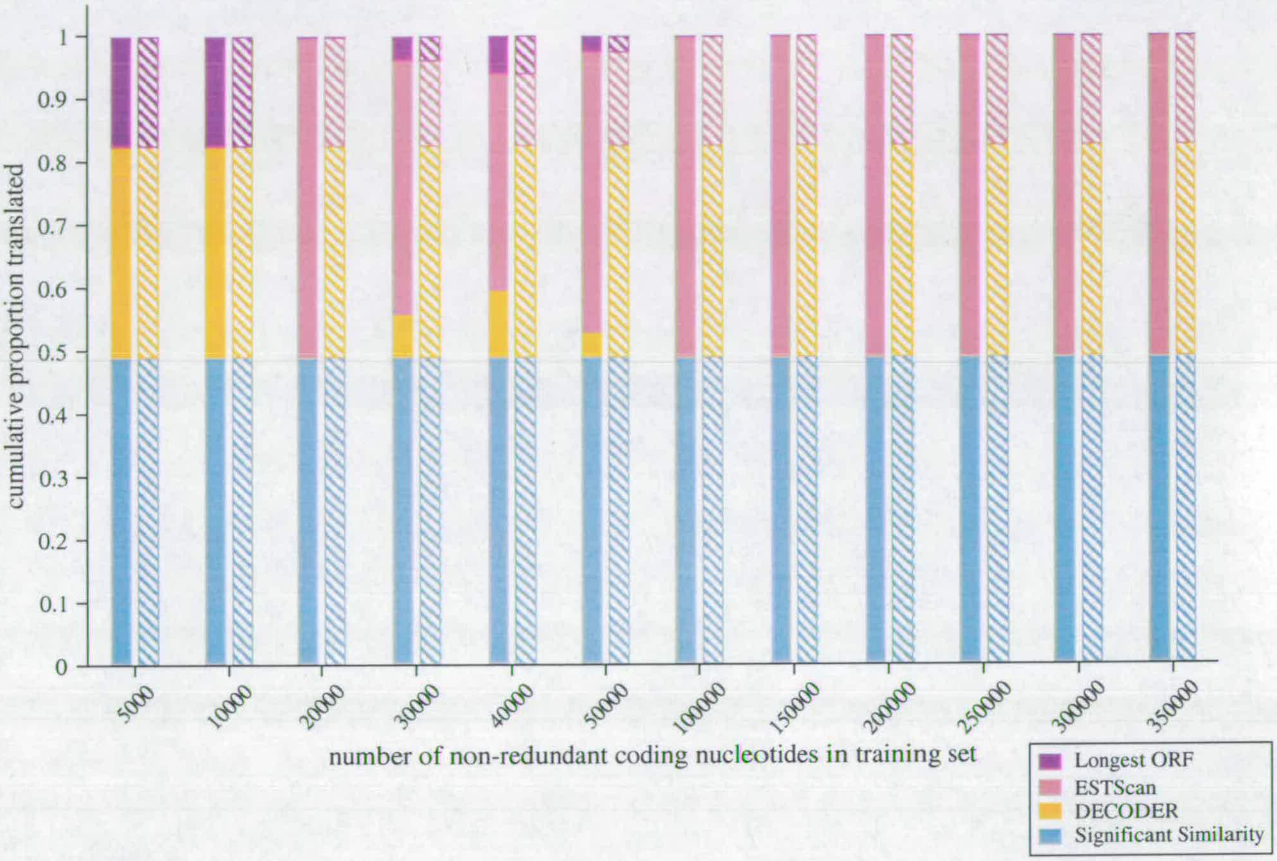


Figure 4
The relative efficiency of different organisations of DECODER and ESTScan in the prot4EST pipeline. The proportion of consensus sequences translated by each part of the pipeline for each level of training is shown. bold bars: prot4EST_ed – ESTScan translations were considered before those from DECODER. hashed bars: prot4EST_de – Robust DECODER translations were used in preference to those from ESTScan.

The better performance of prot4EST_ed was examined by following the fate of individual test sequences through the prot4EST pipeline. By employing ESTScan before DECODER, larger training sets allowed the deployment of well trained HMMs (Figure 4). All predictions satisfied length and quality filters, and so were accepted as robust. The corresponding DECODER predictions, while satisfying length filters, were not as robust. As the training sets decreased in size, the ESTScan predictions failed the filters and so were ignored, and DECODER used instead.

Performance of similarity search

Seven sequences out of 2316 were identified as rRNA in tier 1. Tiers 2 and 3 of the prot4EST pipeline exploit any significant sequence similarity between the query sequence and known proteins for coding region determination. This approach identified coding regions from just under half of the consensuses, 1131. Nineteen were identified as mitochondrial genome derived. To benchmark the similarity approach against the other probabilistic methods, the accuracy of predictions from 1131 consensuses were compared. Translations derived from prot4EST tiers 2 and 3 were more robust than those from ESTScan or DECODER (Figure 5).

Given that an increase in the number of non-redundant coding nucleotides used to train ESTScan produces more robust translations, we attempted to use coding regions determined thus far to create larger training sets, with the expectation of improved translations. The results from the

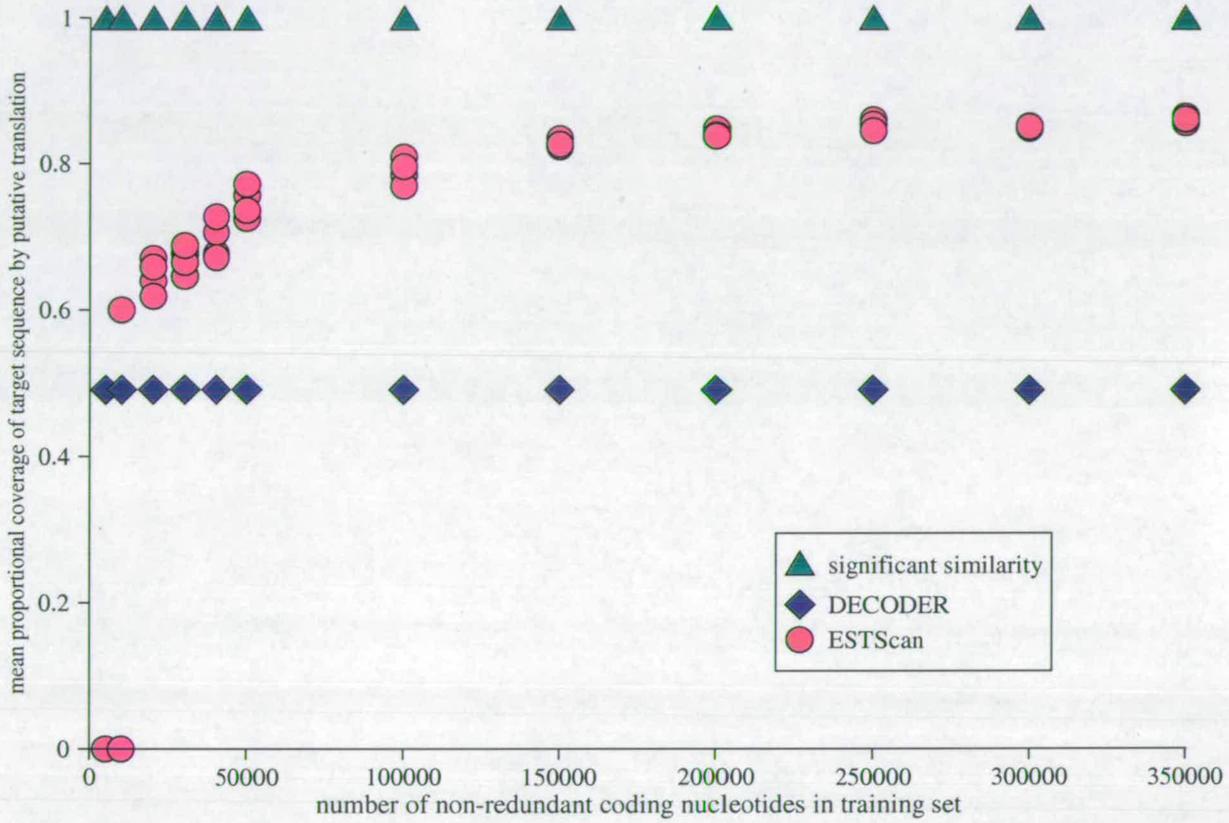


Figure 5
Comparison of HSP tiling, ESTScan and DECODER performance in translating the 1131 consensuses that prot4EST translated using similarity criteria.

BLASTX search against the SwissAll database were checked for matches where the alignment included the start of the protein sequence. These results contained the information required to construct pseudo-CDS entries which can be added to the training set for populating the HMMs of ESTScan. In this study there were only six BLASTX alignments that provided suitable pseudo-CDS, failing to provide any significant increase in the level of non-redundant coding nucleotides. However other species we study have produced higher numbers of pseudo-CDS which prot4EST uses to give improved translations (data not shown).

Effect of training set and target set sequence composition

As a significant proportion of any EST set will not share similarity with known sequences, *de novo* translation methods need to be trained to as high a level as possible. The question is how this should be done, given the paucity of prior sequence data for individual species. Should CDS from species considered phylogenetically related be combined or should a large set from a model organism be used? A recent study of gene finding in novel genomes has shown a significant effect of sequence composition upon gene structure prediction, with more closely related model genomes providing poor training if the codon bias differs significantly from the genome of interest [25]. The performance of ESTScan was affected by even slight fluctuations in sequence composition. We examined the effect of AT content on the accuracy of translation. The complete CDS complements of 156 prokaryotes were assembled as described in the Methods. This gave a range of AT contents from 28% (*Streptomyces coelicolor*) to 78% (*Wigglesworthia glossinidia*), independent of any bias due the organisms' relatedness to *C. elegans*. The lowest number of non-redundant coding nucleotides was 461,299, in excess of

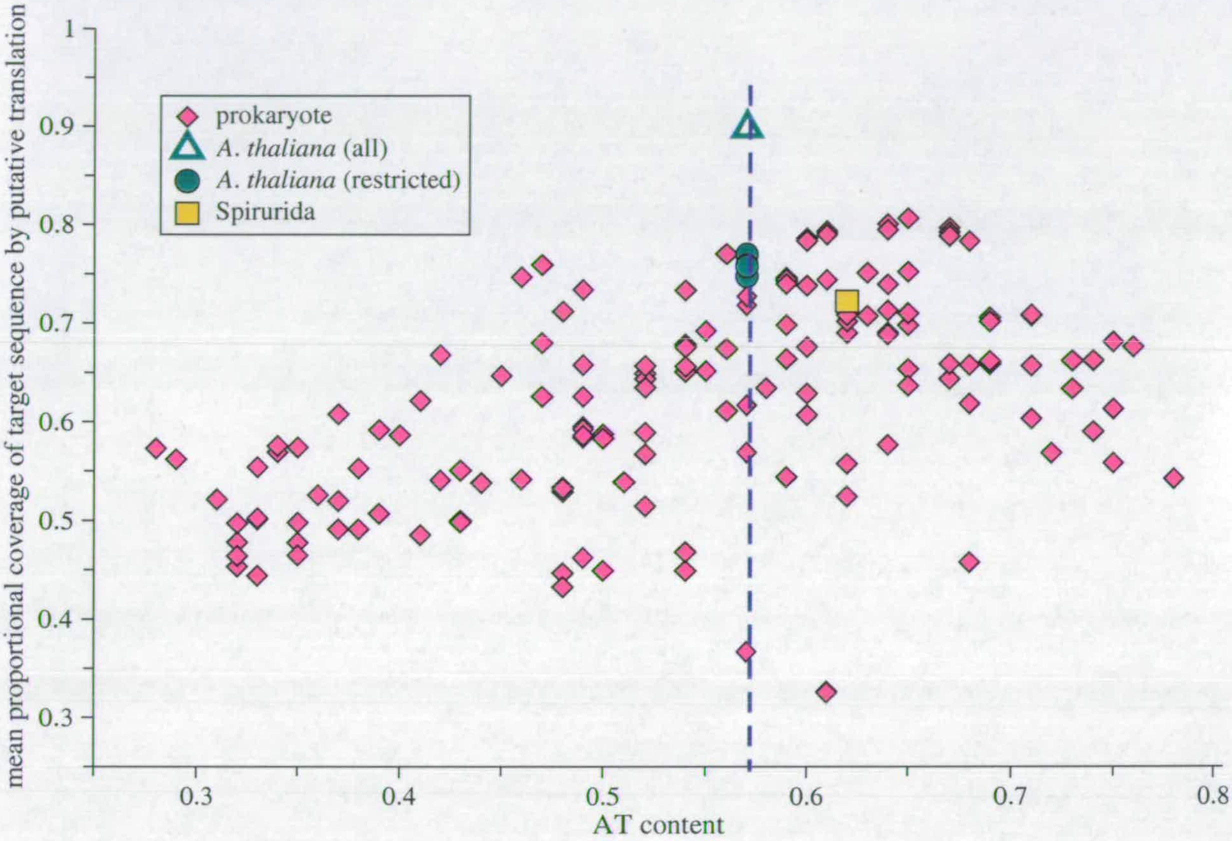


Figure 6

Effect of AT content of training set upon translation accuracy. Each purple diamond represents a complete CDS set from a prokaryote genome. The orange box represents all CDS available from the nematode order Spirurida (~230000 non-redundant coding nucleotides). The green triangle represents the complete *Arabidopsis thaliana* RefSeq collection (~30000000 non-redundant coding nucleotides). The green circles are training sets of *A. thaliana* CDS RefSeq entries randomly selected to total ~230000 non-redundant coding nucleotides. The AT content of *C. elegans* is shown by the vertical dashed line.

the minimum number suggested for robust training. To explore datasets from more closely related sources all available CDS entries for the nematode order Spirurida (last common ancestor with *C. elegans* was 475–500 MYA [38]), and the plant *Arabidopsis thaliana* [39] were obtained.

There was a significant correlation between AT content of the training set and the coverage by the putative polypeptides of their reference *C. elegans* proteins (r = 0.49 P > 0.001) (Figure 6). The most robust predictions were produced by HMMs trained on datasets with an AT content similar to that of *C. elegans*. For the prokaryote training sets, the number of nucleotides used had no significant effect upon performance (data not shown). We note that some prokaryote training sets with AT contents close to *C. elegans* performed poorly: homogeneity of AT content is thus not a panacea. The best performance was obtained using the *A. thaliana* training set, with significantly better coverage than achieved with the more closely related Spirurida. As the plant dataset contained 130 times as many coding nucleotides as did the Spirurida training set, four random *A. thaliana* training sets of comparable size to the Spirurida were built. These smaller training sets still performed better than the Spirurida training set, though not as well as the full CDS collection.

Conclusions

prot4EST is a protein translation pipeline that utilises the advantages of a number of publicly available tools. We have shown that it produces significantly more robust translations than single methods for species with little or no prior sequence data. Around three quarters of current EST projects are associated with training sets of < 50000 coding nucleotides (Figure 1). Thus prot4EST offers significant improvement in this real world situation. Even with substantial numbers of coding nucleotides, the use of similarity searches means prot4EST is able to outperform the best *de novo* methods. Given the increase in protein sequences submitted to SwissProt/TrEMBL, prot4EST's ability and accuracy can only increase over time. These more accurate translations provide the platform for more rigorous down-stream annotation. Currently we are using the prot4EST pipeline to translate ~95000 nematode consensus sequences from 30 species. These translations will then be passed onto other tools we are developing for EST analysis and annotation (see http://www.nematodes.org/PartiGene).

Availability and requirements

Project name: prot4EST

Project home page: http://www.nematodes.org/PartiGene

Operating system(s): Fully tested on Linux - Redhat9.0, Fedora2.0.

Programming language: Perl

Other requirements:

ESTScan2.0 http://www.isrec.isb-sib.ch/ftp-server/ESTScan/

DECODER rgscerg@gsc.riken.go.jp

BioPerl 1.4 http://bioperl.org

Transeq http://www.hgmp.mrc.ac.uk/Software/EMBOSS/

License: GNU GPL

Any restrictions to use by non-academics: None for prot4EST source code. DECODER requires a license. See User Guide.

Authors' contributions

JW performed all the analyses and wrote all the Perl code. MB oversaw the project and suggested additional features. Both authors shared responsibility for writing this manuscript.

Acknowledgements

This work was funded by a BBSRC CASE PhD studentship to JW. We thank Astra Zeneca for supporting the CASE program. Work in MB's laboratory is funded by NERC, BBSRC and the Wellcome Trust. We thank Y. Fukunishi and Y. Hayashizaki of the RIKEN Institute for DECODER, C. Iselli and C. Lottaz for the ESTscan package, and our colleagues Ralf Schmid, John Parkinson, Ann Hedley and Makedonka Mitreva for support and comments on the manuscript.

References

1. Muller A, MacCallum RM, Sternberg MJ: **Structural characterization of the human proteome.** *Genome Res* 2002, **12**:1625-1641.
2. Blaxter ML: **Genome sequencing: time to widen our horizons.** *Briefings in Functional Genomics and Proteomics* 2002, **1**:7-9.
3. Stürzenbaum SR, Parkinson J, Blaxter ML, Morgan AJ, Kille P, Georgiev O: **The earthworm EST sequencing project.** *Pedobiologia* 2003, **47**:447-451.
4. Cheng TC, Xia QY, Qian JF, Liu C, Lin Y, Zha XF, Xiang ZH: **Mining single nucleotide polymorphisms from EST data of silkworm, Bombyx mori, inbred strain Dazao.** *Insect Biochem Mol Biol* 2004, **34**:523-530.
5. Barker G, Batley J, H OS, Edwards KJ, Edwards D: **Redundancy based detection of sequence polymorphisms in expressed sequence tag data using autoSNP.** *Bioinformatics* 2003, **19**:421-422.
6. Parkinson J, Anthony A, Wasmuth J, Schmid R, Hedley A, Blaxter M: **PartiGene - constructing partial genomes.** *Bioinformatics* 2004, **20**:1398-1404.
7. Mulder NJ, Apweiler R, Attwood TK, Bairoch A, Barrell D, Bateman A, Binns D, Biswas M, Bradley P, Bork P, Bucher P, Copley RR, Courcelle E, Das U, Durbin R, Falquet L, Fleischmann W, Griffiths-Jones S, Haft D, Harte N, Hulo N, Kahn D, Kanapin A, Krestyaninova M, Lopez R, Letunic I, Lonsdale D, Silventoinen V, Orchard SE, Pagni M, Peyruc D, Ponting CP, Selengut JD, Servant F, Sigrist CJ, Vaughan R, Zdobnov EM: **The InterPro Database, 2003 brings increased coverage and new features.** *Nucleic Acids Res* 2003, **31**:315-318.
8. Bateman A, Coin L, Durbin R, Finn RD, Hollich V, Griffiths-Jones S, Khanna A, Marshall M, Moxon S, Sonnhammer EL, Studholme DJ, Yeats C, Eddy SR: **The Pfam protein families database.** *Nucleic Acids Res* 2004, **32 Database issue**:D138-41.
9. Gordon D, Abajian C, Green P: **Consed: a graphical tool for sequence finishing.** *Genome Res* 1998, **8**:195-202.
10. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Res* 1997, **25**:3389-3402.
11. Pearson WR, Wood T, Zhang Z, Miller W: **Comparison of DNA sequences with protein sequences.** *Genomics* 1997, **46**:24-36.
12. Cuff JA, Birney E, Clamp ME, Barton GJ: **ProtEST: protein multiple sequence alignments from expressed sequence tags.** *Bioinformatics* 2000, **16**:111-116.
13. Stein LD, Bao Z, Blasiar D, Blumenthal T, Brent MR, Chen N, Chinwalla A, Clarke L, Clee C, Coghlan A, Coulson A, D'Eustachio P, Fitch DH, Fulton LA, Fulton RE, Griffiths-Jones S, Harris TW, Hillier LW, Kamath R, Kuwabara PE, Mardis ER, Marra MA, Miner TL, Minx P, Mullikin JC, Plumb RW, Rogers J, Schein JE, Sohrmann M, Spieth J, Stajich JE, Wei C, Willey D, Wilson RK, Durbin R, Waterston RH: **The Genome Sequence of Caenorhabditis briggsae: A Platform for Comparative Genomics.** *PLoS Biol* 2003, **1**:E45.
14. Boeckmann B, Bairoch A, Apweiler R, Blatter MC, Estreicher A, Gasteiger E, Martin MJ, Michoud K, O'Donovan C, Phan I, Pilbout S, Schneider M: **The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003.** *Nucleic Acids Res* 2003, **31**:365-370.
15. Birney E: **ESTWISE 2 [http://www.ebi.ac.uk/Wise2/]**. .
16. Hatzigeorgiou AG, Fiziev P, Reczko M: **DIANA-EST: a statistical analysis.** *Bioinformatics* 2001, **17**:913-919.
17. Fukunishi Y, Hayashizaki Y: **Amino acid translation program for full-length cDNA sequences with frameshift errors.** *Physiol Genomics* 2001, **5**:81-87.

18. Lottaz C, Iseli C, Jongeneel CV, Bucher P: **Modeling sequencing errors by combining Hidden Markov models.** *Bioinformatics* 2003, **19 Suppl 2:**II103-II112.
19. Parkinson J, Whitton C, Schmid R, Thomson M, Blaxter M: **NEMBASE: a resource for parasitic nematode ESTs.** *Nucleic Acids Res* 2004, **32:**D427-30.
20. Iseli C, Jongeneel CV, Bucher P: **ESTScan: a program for detecting, evaluating, and reconstructing potential coding regions in EST sequences.** *Proc Int Conf Intell Syst Mol Biol* 1999:138-148.
21. Ewing B, Hillier L, Wendl MC, Green P: **Base-calling of automated sequencer traces using phred. I. Accuracy assessment.** *Genome Res* 1998, **8:**175-185.
22. Ewing B, Green P: **Base-calling of automated sequencer traces using phred. II. Error probabilities.** *Genome Res* 1998, **8:**186-194.
23. Durbin R, Eddy S, Krogh A, Mitchison G: **Biological sequence analysis. Probabilistic models of proteins and nucleic acids.**, Cambridge Univerity Press; 1998:356.
24. Burge C, Karlin S: **Prediction of complete gene structures in human genomic DNA.** *J Mol Biol* 1997, **268:**78-94.
25. Korf I: **Gene finding in novel genomes.** *BMC Bioinformatics* 2004, **5:**59.
26. Sonnhammer EL, Eddy SR, Birney E, Bateman A, Durbin R: **Pfam: multiple sequence alignments and HMM-profiles of protein domains.** *Nucleic Acids Res* 1998, **26:**320-322.
27. Loytynoja A, Milinkovitch MC: **A hidden Markov model for progressive multiple alignment.** *Bioinformatics* 2003, **19:**1505-1513.
28. Maidak BL, Cole JR, Lilburn TG, Parker CTJ, Saxman PR, Farris RJ, Garrity GM, Olsen GJ, Schmidt TM, Tiedje JM: **The RDP-II (Ribosomal Database Project).** *Nucleic Acids Res* 2001, **29:**173-174.
29. Nakamura Y, Gojobori T, Ikemura T: **Codon usage tabulated from international DNA sequence databases: status for the year 2000.** *Nucleic Acids Res* 2000, **28:**292.
30. Kohara Y: **[Genome biology of the nematode C. elegans].** *Tanpakushitsu Kakusan Koso* 1999, **44:**2601-2608.
31. Parkinson J, Guiliano D, Blaxter M: **Making sense of EST sequences by CLOBBing them.** *BMC Bioinformatics* 2002, **3:**31.
32. Stein L, Sternberg P, Durbin R, Thierry-Mieg J, Spieth J: **WormBase: network access to the genome and biology of Caenorhabditis elegans.** *Nucleic Acids Res* 2001, **29:**82-86.
33. Stein LD: **Internet access to the C. elegans genome.** *Trends Genet* 1999, **15:**425-427.
34. Rice P, Longden I, Bleasby A: **EMBOSS: the European Molecular Biology Open Software Suite.** *Trends Genet* 2000, **16:**276-277.
35. Pruitt KD, Tatusova T, Maglott DR: **NCBI Reference Sequence project: update and current status.** *Nucleic Acids Res* 2003, **31:**34-37.
36. Phan IQ, Pilbout SF, Fleischmann W, Bairoch A: **NEWT, a new taxonomy portal.** *Nucleic Acids Res* 2003, **31:**3822-3823.
37. Stajich JE, Block D, Boulez K, Brenner SE, Chervitz SA, Dagdigian C, Fuellen G, Gilbert JG, Korf I, Lapp H, Lehvaslaiho H, Matsalla C, Mungall CJ, Osborne BI, Pocock MR, Schattner P, Senger M, Stein LD, Stupka E, Wilkinson MD, Birney E: **The Bioperl toolkit: Perl modules for the life sciences.** *Genome Res* 2002, **12:**1611-1618.
38. Vanfleteren JR, Van de Peer Y, Blaxter ML, Tweedie SA, Trotman C, Lu L, Van Hauwaert ML, Moens L: **Molecular genealogy of some nematode taxa as based on cytochrome c and globin amino acid sequences.** *Mol Phylogenet Evol* 1994, **3:**92-101.
39. The Arabidopsis Sequencing Consortium: **Analysis of the genome sequence of the flowering plant Arabidopsis thaliana.** *Nature* 2000, **408:**796-815.