

Immunoglobulin Superfamily Proteins in *Caenorhabditis elegans*

Sarah A. Teichmann and Cyrus Chothia

MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, UK

The predicted proteins of the genome of *Caenorhabditis elegans* were analysed by various sequence comparison methods to identify the repertoire of proteins that are members of the immunoglobulin superfamily (IgSF). The IgSF is one of the largest families of protein domain in this genome and likely to be one of the major families in other multicellular eukaryotes too. This is because members of the superfamily are involved in a variety of functions including cell-cell recognition, cell-surface receptors, muscle structure and, in higher organisms, the immune system. Sixty-four proteins with 488 I set IgSF domains were identified largely by using Hidden Markov models. The domain architectures of the protein products of these 64 genes are described. Twenty-one of these had been characterised previously. We show that another 25 are related to proteins of known function. The *C. elegans* IgSF proteins can be classified into five broad categories: muscle proteins, protein kinases and phosphatases, three categories of proteins involved in the development of the nervous system, leucine-rich repeat containing proteins and proteins without homologues of known function, of which there are 18. The 19 proteins involved in nervous system development that are not kinases or phosphatases are homologues of neuroglian, axonin, NCAM, wrapper, klin-gon, ICCR and nephrin or belong to the recently identified *zig* gene family. Out of the set of 64 genes, 22 are on the X chromosome. This study should be seen as an initial description of the IgSF repertoire in *C. elegans*, because the current gene definitions may contain a number of errors, especially in the case of long sequences, and there may be IgSF genes that have not yet been detected. However, the proteins described here do provide an overview of the bulk of the repertoire of immunoglobulin superfamily members in *C. elegans*, a framework for refinement and extension of the repertoire as gene and protein definitions improve, and the basis for investigations of their function and for comparisons with the repertoires of other organisms.

© 2000 Academic Press

Keywords: I set; Hidden Markov models; cell adhesion molecules; cell surface receptors; muscle proteins

Abbreviations used: BPTI, bovine pancreatic trypsin inhibitor; EGF, epidermal growth factor; FGFR, fibroblast growth factor receptor; FnIII, fibronectin type III; HMM, Hidden Markov model; IgSF, immunoglobulin superfamily; LDL, low density lipoprotein receptor; LRR, leucine-rich repeat; NCAM, neural cell adhesion molecule; PH, pleckstrin homology domain; PK, protein kinase; RhoGEF, Rho guanine nucleotide exchange factor; SH3, Src homology 3 domain; tsp_1, thrombospondin type I; unc, uncoordinated.

E-mail address of the authors: sat and chc1@mrc-lmb.cam.ac.uk

Introduction

With completion of the genome sequence of the nematode *Caenorhabditis elegans* (The *C. elegans* Sequencing Consortium, 1998), we can begin to define the repertoire of protein families that form different metazoa. The descriptions of protein family repertoires will help in understanding the biology of the particular organism in which they occur. In addition, comparisons of the repertoires found in different organisms should allow us to understand how gene duplications and recombinations have contributed to the evolution of development. Here, we present an initial view of the

immunoglobulin superfamily proteins that can be found in the genome sequence of the nematode *C. elegans*. Domains from the immunoglobulin superfamily (IgSF) are widespread in multicellular eukaryotes, where they play a central role in important aspects of physiology: in cell-cell recognition, as cell-surface receptors, in the structure of muscle and, in higher organisms, in the immune system. In all, we find 64 genes in *C. elegans* which encode IgSF proteins. Of these, only 21 had been characterised previously. We describe in some detail their domain architectures and compare them to the IgSF proteins currently known for other organisms.

IgSF proteins

The basic evolutionary unit of IgSF proteins is a domain of approximately 100 residues. These domains are mainly formed by two β -sheets packed face to face. Although the members of the IgSF are diverse, inspection of the known structures shows that they can be grouped into four structural "sets": V, C1, C2 and I (Harpaz & Chothia, 1994; Williams & Barclay, 1988). Members of a given set are more similar to each other than they are to any member of any other set. Within a set, members usually have the same conformations over 70% or more of their structures even when their sequence identities are low. IgSF domains in lower organisms largely belong to the I set (Bateman, 1997; Harpaz & Chothia, 1994; and see below).

A few of the proteins in the IgSF family contain just one domain; most have several and the largest, titin, has 112 IgSF domains (Labeit & Kolmerer, 1995). In addition, many members of the superfamily, particularly muscle proteins and cell-adhesion molecules, have additional domains from other superfamilies, the most common of which are from the fibronectin type III family.

Methods of Identifying IgSF Proteins in *C. elegans*

The IgSF domains in *C. elegans* were identified by two methods: Hidden Markov models and key residue analysis. These two methods will be described first, and then the programs and procedures used to annotate the non-Ig sequence regions of the IgSF proteins will be addressed.

Hidden Markov models (HMMs)

Hidden Markov models (Eddy, 1996; Krogh *et al.*, 1994) are probably the most sensitive sequence comparison method currently available (Park *et al.*, 1998). The particular program we used was the iterative procedure SAM-T98 (Karplus *et al.*, 1998). The starting point for the model-building was a set of sequences of 11 I set domains aligned on the basis of their three-dimensional structures. The

PDB identifiers of the protein structures from which these sequences were taken are 1bih (which has four domains), 1tlk, 1koa, 1nct, 2ncm, 1vca, 1zxq and 1iam.

In SAM-T98, it is possible to constrain regions to remain aligned during the model-building process (Mark Diekhans, unpublished results). The constraints used were on the strands that form the central regions of the two β -sheets of the Ig domain. The HMM built by SAM-T98 was used to search the set of predicted *C. elegans* proteins available in November 1998 (Steven Jones, personal communication) and Wormpep18 (The *C. elegans* Sequencing Consortium, 1998). According to an assessment of SAM-T98 (Park *et al.*, 1998), a threshold of -15 bits should yield an error rate of approximately 1%. Hence this score was used as the cut-off in the search. Using this method, 79 proteins with 470 Ig domains were identified. (18 more Ig domains were found using other methods.) These 79 proteins come from 70 genes, with six proteins having two known splice variants and one protein having four known splice variants. In seven cases, two or more genes which were described as being separate in the *C. elegans* database should actually be merged to form a single gene (see below). Therefore, the current set is made up of 73 proteins from 64 genes.

Key residue inspection

Sequences that were expected to be Ig domains by homology to non-*C. elegans* sequences were examined for the pattern of residues characteristic of I set structures (Bateman *et al.*, 1996). As described below, 18 Ig domains and one FnIII domain were found in this way. Some of these domains appear to be "incomplete", as described below.

Non-Ig sequence regions

To identify any FnIII domains that occur in these IgSF proteins, an HMM was built in the same way as for Ig domains, starting from an alignment of ten FnIII domains based on their three-dimensional structure. The resulting HMM was then matched to the *C. elegans* proteins with Ig domains, and 136 FnIII domains were identified.

The domains in the IgSF proteins that are not members of the IgSF or FnIII families were characterised as far as possible by matching the 73 *C. elegans* proteins to the HMM in two databases:

(1) Pfam database (Bateman *et al.*, 1999), the 73 proteins were used as targets for all the HMMs attached to this database.

(2) SMART HMM server (Schultz *et al.*, 1998), some of the 73 proteins were submitted to the HMM server which is part of this database.

All the domains that are not Igs or FnIIIs were found using the HMMs in the Pfam and SMART databases.

(1) SignalP server (Nielsen *et al.*, 1997), regions in the 73 proteins that are signal sequences were detected using this server.

(2) SEG (Wootton & Federhen, 1993), low complexity domains were found using the program SEG with the options 25 3.0 3.3 and 45 3.4 3.75. The transmembrane regions were identified using the GES hydrophobicity method described by Gerstein (1997), and the TopPred program (von Heijne, 1992).

The regions that were not assigned a domain using the methods described above, as well as the complete proteins in the set of 73, were searched against the NRDB90 (Holm & Sander, 1998), SwissProt (Bairoch & Apweiler, 1999) and GenBank (Benson *et al.*, 1999) protein sequence databases using FASTA (Pearson & Lipman, 1988), with an expectation value threshold of 0.001. Proteins identified in this way are described as *C. elegans* protein homologues; see below.

Improving gene predictions

For gene predictions that were suspect, we used the following methods to try to improve their definition.

(1) GeneWise (Birney & Durbin, 1997): *C. elegans* proteins with homologues in one of the above protein sequence databases over the entire length of the *C. elegans* protein, where the *C. elegans* protein was shorter than its homologue, were checked for their gene definitions. The homologous protein was compared to the DNA sequence containing the *C. elegans* gene and the surrounding DNA using GeneWise. GeneWise tries to find the exons in the DNA which correspond to the protein. Because most *C. elegans* proteins had homologues with less than 30% sequence identity, no useful results were obtained for most proteins. However, comparing *Drosophila* neuroglian, NRG_DROME (which is 30% identical with C18F3.2 and C18F3.3), to the C18F3 DNA sequence did show that C18F3.2 and C18F3.3 form a single gene.

(2) TBLASTN on the *C. elegans* EST database, sequences for which the N-terminal or C-terminal regions were for some reason questionable were compared to the *C. elegans* EST database (Y. Kohara, unpublished) using the BLAST servers at the Sanger Centre (http://www.sanger.ac.uk/Projects/C_elegans/blast_server.shtml) and at the DNA Databank of Japan (http://www.ddbj.nig.ac.jp/c-elegans/html/CE_BLAST.html). In the case of C36F7.3 and C36F7.4, a pair of bridging ESTs exists (O. Hobert, personal communication) such that the correct protein spans these two predictions.

(3) *C. elegans* database. All altered gene definitions were submitted to the *C. elegans* database at the Sanger Centre. Possible but uncertain cases were discussed with Daniel Lawson (Sanger Centre) and revisions are proposed on the basis of additional evidence in ACeDB, such as splice sites predictions and so forth. In three cases, we propose

that proteins predicted as separate are in fact a single protein; C09D8.1 with C09D8.2, C33F10.5 with C33F10.6, R05D8.a with F54E2.3 and F54E2.4.

Comparisons of all 73 proteins to each other

(1) The chromosomal locations of the 64 genes on the six *C. elegans* chromosomes were parsed from the gff files available at http://www.sanger.ac.uk/Projects/C_elegans/Science98/.

(2) GEANFAMMER (Park & Teichmann, 1998). The sequences of all the *C. elegans* Ig superfamily members were compared to each other and clustered by single linkage using the GEANFAMMER programs. Apart from the obvious matches, this revealed that the two proteins of unknown function, C25G4.10 and T04A11.3 are 98.5% identical with each other over the entire length of T04A11.3. These two proteins are adjacent on chromosome IV, indicating that they are recent duplicates. It also showed that the protein of unknown function M02D8.1 is similar to C18A11.7 (DIM-1), and the protein kinase C24G7.5 is similar to the titin-like kinase F12F3.2.

The details of all the matches can be found at the website http://www.mrc-lmb.cam.ac.uk/genomes/CE_igs.html

The Immunoglobulin Superfamily Members in *C. elegans* and their Homologues

Muscle proteins

Eight IgSF *C. elegans* proteins from four genes have been identified previously as being involved in muscle structure and function. These are the four splice variants of UNC-52 and two splice variants of UNC-22 (also known as twitchin) that are known at present, UNC-89 and DIM-1. Sequences with the prefix UNC in their names are proteins in which mutations were found to produce an uncoordinated phenotype in *C. elegans* (Waterston *et al.*, 1980). This can be due to a deficiency in the muscle proteins or in proteins involved in the nervous system or its development.

The domain structure of these UNC proteins and DIM-1 is illustrated in Figure 1(a). The four known splice variants of UNC-52 have 2, 13, 16 or 17 Ig domains interrupted by other domains. They may be involved in myofilament lattice assembly or attachment of the myofilament lattice to the cell membrane. Almost the whole of UNC-52 (residues 132-2440) is homologous to PBGM_HUMAN (residues 270-2800), a basement membrane proteoglycan (Rogalski *et al.*, 1993).

The two splice variants of twitchin (UNC-22), a protein involved in myosin regulation, contain 28 or 29 Ig domains interrupted by FnIII domains and a PK domain towards the C terminus (Benian *et al.*, 1993). The UNC-89 protein (Benian *et al.*, 1996), required for muscle M-line assembly, is a relative of twitchin. (For a study of the evolution of the

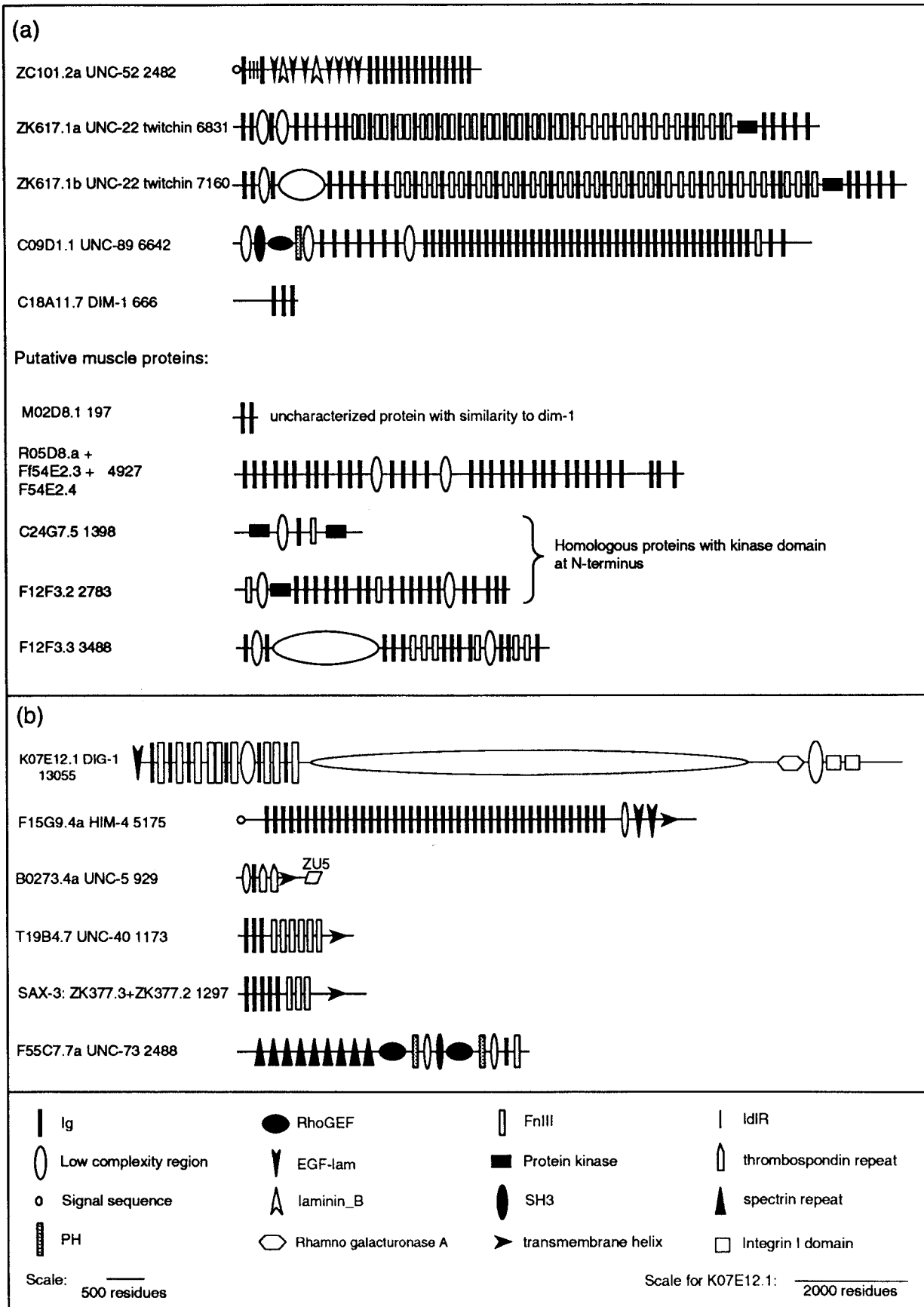


Figure 1 (legend opposite)

pattern of Ig and FnIII domains in these *C. elegans* muscle proteins, see Kenny *et al.*, 1999.)

DIM-1 is required for myofilament lattice stability in the body wall muscle (Rogalski *et al.*, 1998). It has an N-terminal region of unknown character and, at its C terminus, three adjacent Ig domains. The M02D8.1 protein, with only two Ig domains, has 30% sequence identity with two Ig domains of DIM-1 (Table 1A), so it is likely that it is also a muscle protein rather than involved in the nervous system.

Five *C. elegans* proteins, R05D8.a, F54E2.3, F54E2.4, F12F3.2 and F12F3.3 are homologous to different regions of human titin, a very large elastic protein involved in the organisation of thick filaments in muscle (Labeit & Kolmerer, 1995), to myosin light chain kinase, and to other muscle members of the IgSF (see Table 1A), though their actual function is unknown. Three genes, R05D8.a, F54E2.3 and F54E2.4, are adjacent to each other on chromosome V. Inspection of regions between the three genes, as displayed in ACeDB, shows that there are features that indicate that fusion of these three genes into one gene is a reasonable alternative to the present predictions (D. Lawson, personal communication). The resulting long gene with 34 Ig domains is likely to be a structural muscle protein, although the domain architecture does not exactly correspond to titin or any other known muscle protein. F12F3.2 and F12F3.3 are adjacent genes which are also homologues to IgSF muscle proteins, but there was no evidence to support fusion of the two predicted proteins. (The domain structure of the *C. elegans* proteins is illustrated in Figure 1(a).)

Neural UNC proteins and other previously characterised *C. elegans* IgSF proteins

The remaining UNC proteins are involved in neural development. Their domain structure is shown in Figure 1(b). The two UNC-5 splice variants are cell surface receptors required by motile cells and motor neurons to orient movements away from cells that have UNC-6 (netrin) on their surface (Leung-Hagesteijn *et al.*, 1992). UNC-40 is a receptor for UNC-6. It is expressed on motile cells and pioneer neurons (Chan *et al.*, 1996).

The axon guidance protein SAX-3 also mediates cell interactions during axon guidance (Zallen *et al.*, 1998). The gene had been split into two parts in the *C. elegans* database: the known sequence is coded by two adjacent gene predictions: ZK377.3 and ZK377.2. UNC-73 is involved in cell migration and axon guidance and is thought to regulate actin dynamics during cell and growth cone migrations

(Steven *et al.*, 1998). The two splice variants differ greatly in length, but each contain only one Ig domain.

A protein involved in cell adhesion, but not in the nervous system, is HIM-4, a uterine cell adhesion molecule (Hodgkin *et al.*, 1979; B. E. Vogel, personal communication). It contains 45 Ig domains in both splice variants.

The longest non-muscle protein in the Ig superfamily in *C. elegans* is K07E12.1, 13055 residues long, with intercalated Ig and FnIII domains. This protein has many regions of low complexity as well as other types of domains, such as integrin I domains and a rhamnogalacturonase A domain. It has been called DIG-1 and sensory processes of *dig-1* mutants often follow aberrant paths to the worm's nose, sometimes branching abnormally, which appears to be due to an adhesive defect (Ryder (1999) <http://www.wpi.edu/Academics/Depts/Bio/People/ryder.html> and personal communication).

Tyrosine kinases

Protein kinases, mentioned above in connection with muscle proteins, can also be part of proteins involved in development. The main type of development in which Ig superfamily members are involved in *C. elegans* is development of the nervous system. One of the two characterised non-muscle IgSF kinases in *C. elegans* is EGL-15, which is the receptor for an extracellular signal required for sex myoblast migration during gonad development (Devore *et al.*, 1995). It belongs to the fibroblast growth factor receptor (FGFRs) superfamily. Like most tyrosine protein kinases (Tyr-PK) involved in development, it is a type I membrane protein, with a signal sequence followed by an extracellular domain, a transmembrane domain and C terminus, which contains the Tyr-PK domain (see Figure 2). There is no other Tyr-PK in *C. elegans* that is homologous to EGL-15 or other FGFRs. The other characterised Tyr-PK kinase in *C. elegans* involved in the development of the nervous system is CAM-1/KIN-8 (Forrester *et al.*, 1999; Koga *et al.*, 1999), which belongs to the Ror kinase family. This protein is a fusion of the predicted proteins D2013.4 and C01G6.8. Mutants in CAM-1/KIN-8 display defects in axon outgrowth and cell migration as well as asymmetric cell division. CAM-1/KIN-8 contains one Ig domain, a cysteine-rich region and a kringle domain in its extracellular region. There appears to be no other protein with this domain architecture in *C. elegans*.

F59F3.1 and F59F3.5 are homologous to the vascular growth factor receptors (VGRs) of humans

Figure 1. (a) Muscle proteins. Four experimentally characterised muscle proteins are depicted. The remaining proteins are uncharacterised, but similar to muscle proteins. (b) Experimentally characterised proteins in the nervous system. These six proteins are previously characterised nervous system molecules and are the only proteins with these domain architectures in *C. elegans*.

Table 1. Sequence identities and similarities

A. Sequence matches between CE proteins and those of other organisms

Protein group	Sequence pairs	Length	Match regions	E-value and seq identity	Figure	
Muscle	R05D8.a titin	2284 27710	481-2229 800-2500	10 ⁻²⁸ 23%	1(a)	
	F54E2.3 titin	2269 27710	312-2212 800-2500	10 ⁻¹⁸ 21%	1(a)	
	F54E2.4 titin	473 27710	87-447 900-1300	10 ⁻⁶ 27%	1(a)	
	F12F3.3 titin	3488 27710	1800-3500 18100-19800	10 ⁻²¹ 28%	1(a)	
	F12F3.2 titin	2783 27710	12-2344 24600-27000	0 23%	1(a)	
	ZC101.2a PGBM_HUMAN	2482 4393	132-2440 270-2800	10 ⁻²³ 27%	1(a)	
	F59F3.1 VGR1_HUMAN	1227 1338	95-1160 92-1151	10 ⁻²⁵ 24%	2	
	F59F3.5 VGR1_HUMAN	1199 1338	50-1163 49-1159	10 ⁻¹¹ 24%	2	
Protein kinase	T17A3.8 MATK_MOUSE	649 505	33-405 118-474	10 ⁻¹³ 29%	2	
	Phosphatases	K04D7.4 PTP9_DROME	1156 1301	304-1152 124-1018	0 29%	2
		C09D8.2 LAR_DROME	818 2029	99-792 33-697	0 35%	2
F56D1.4 PTP6_DROME		1422 1462	317-1416 384-1450	10 ⁻²⁵ 29%	2	
Development factors	C09C7.1 IML2_DROME	253 263	60-246 69-257	10 ⁻¹³ 30%	3(a)	
	C14F5.2 IML2_DROME	251 263	28-245 46-257	10 ⁻¹¹ 27%	3(a)	
	F42F12.2 IML2_DROME	238 263	19-236 40-259	10 ⁻¹¹ 29%	3(a)	
	Y48A6A.1 IML2_DROME	264 263	54-259 60-255	10 ⁻⁴ 26%	3(a)	
	NCAM	F02G3.1 NCA1_BOVIN	976 850	40-961 12-777	10 ⁻¹³ 25%	4
Wrapper		F41D9.3 4574736	495 500	7-465 8-466	10 ⁻¹⁵ 23%	4
	Nephrin	C26G2.1 3025699	1270 1241	8-1270 20-1221	10 ⁻³¹ 24%	4
ICCR		K02E10.8 ICCR_DROME	662 764	23-472 28-517	10 ⁻¹⁰ 29%	4
	Neuroglian	C18F3.3 NRG_DROME	304 1239	82-250 28-187	10 ⁻⁷ 29%	4
C18F3.2 NRG_DROME		987 12039	18-965 351-1233	10 ⁻¹² 30%	4	
Y94H6A_148.D NRG_DROME		508 1239	24-465 28-470	0 34%	4	
Axonin	C33F10.5 AXO1_CHICK	964 1036	50-951 154-1015	10 ⁻³³ 25%	4	
	C33F10.5 AXO1_CHICK	309 1036	190-302 51-164	10 ⁻⁵ 30%	4	

Table 1. (Continued)

B. Sequences with high identities within the CE genome					
Protein group	Protein	Length	Match regions	Sequence identity	Figure
Neuroglial and homolog	C18F3.3+C18F3.2	1355	82-534	29%	4
	Y94H6A_148.d	508	24-460	29%	4
Titin homologs	R05D8.a*	2284	13-1556	32%	1(a)
	F54E2.3*	2269	719-2261	32%	1(a)
Unknown	C25G4.10*	1526	498-1526	98.5%	6
	T04A11.3*	1014	1-1014	98.5%	6
dim-1	C18A11.7	666	323-523	30%	1(a)
	M02D8.1	197	4-197	30%	1(a)
Muscle PKs	F12F3.2	2783	30-1051	24%	1(a)
	C24G7.5	1398	280-1377	24%	1(a)
IMP-L2 homologs	C09C7.1*	253	Over whole length to C14F5.2: 45%		3(a)
	C14F5.2*	251			
	F42F12.2	238			
Nervous system PK's	F58A3.2*	1103	307-922	26% to F59F3.5	2
	F59F3.5*	1199	1-1192	45% to F59F3.1	2
	F59F3.1*	1227	1-1197	45% to F59F3.5	2
	T17A3.1+	875	59-875	38% to F59F3.1	2
	T17A3.8+	649	12-440	43% to F59F3.1	2

* and + indicate proteins which are adjacent to each other on the chromosome.

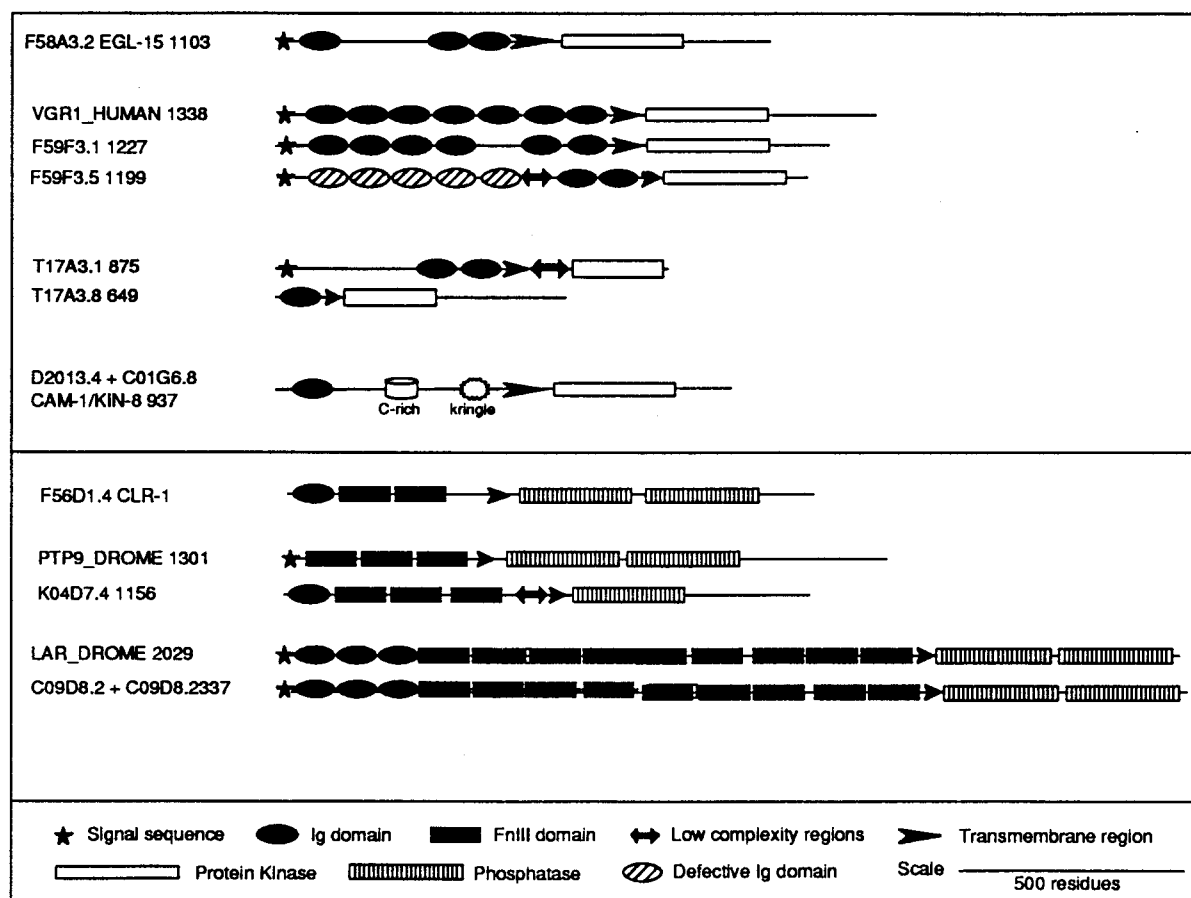


Figure 2. Tyrosine protein kinases and phosphatases. The top box contains the kinases, of which two, EGL-15 and CAM-1/KIN-8, are experimentally characterised. Four other *C. elegans* proteins have identifiable PK domains. Two of these are similar to vascular growth factor receptors. The lower box contains the phosphatases: the experimentally characterised CLR-1 and the *C. elegans* homologues of the *Drosophila* DPTP99A protein and the *Drosophila* LAR protein.

and mice (Table 1A). The VGRs have a signal sequence, seven Ig domains, a transmembrane helix and a cytoplasmic kinase domain. In F59F3.1, the last of the Ig domains cannot be found by the HMM or key residue inspection, while in F59F3.5, only two of the Ig domains are identified by the HMM and the remaining five Ig domains appear defective in their pattern of residues characteristic of immunoglobulin I set domains upon manual inspection. (The alignments of the five problematic domains with a VGR can be obtained at the website where details of our results are available.)

There are two other IgSF proteins that have PK domains. T17A3.1 has two Ig domains. T17A3.8 has a single Ig followed by a PK domain which is similar to the PK domain of the mouse kinase, megakaryocyte-associated Tyr-PK (MATK_MOUSE). MATK_MOUSE is present in the brain and may regulate the Src-family members (Klages *et al.*, 1994) (Table 1A).

Phosphotyrosine protein phosphatases

Phosphotyrosine protein phosphatases (PTPs) linked to Ig domains are likely to be cell adhesion receptors that function in a similar fashion to the protein kinase-linked receptors. One such protein, CLR-1(F56D1.4), the *C. elegans* homologue of the *Drosophila* DPTP69D phosphatase, has been shown to suppress mutations in the EGL-15 FGFR-type receptor tyrosine kinase mentioned above, so it may regulate some member of this FGF signalling pathway (Kokel *et al.*, 1998). Two proteins, K04D7.4 and C09D8.2, make matches to IgSF phosphatases in *Drosophila* (Figure 2 and Table 1A).

K04D7.4, with a C-terminal PTP domain, is similar in sequence and architecture to DPTP99A (PTP9_DROME, Figure 2), which may be involved in signal transduction and growth control in embryonic axons in the central nervous system (Tian *et al.*, 1991). C09D8.2 makes a strong match to the N-terminal part of LAR_DROME (Table 1A), which controls motor axon guidance (Krueger *et al.*, 1996). The protein adjacent to C09D8.2 on chromosome II, C09D8.1, has the same domain architecture as the C-terminal region of LAR_DROME. A GeneWise search with LAR_DROME or with the Pfam phosphatase HMM against the DNA C-terminal to C09D8.2 does not yield an extension of the protein prediction, and the EST database is also of no help in this case. However, inspection of splice site predictions and other evidence available in ACeDB make fusion of these two genes to make one entry in the *C. elegans* database a reasonable alternative to the present predictions (D. Lawson, personal communication).

Zig genes and homologues of the neural/ectodermal development factor IMP-L2

Four *C. elegans* proteins are homologous to *Drosophila* IMP-L2, a secreted neural and ectodermal

development factor (Table 1A). IMP-L2 factor is first expressed at the cellular blastoderm stage in *Drosophila* and continues to be expressed during development (Garbe *et al.*, 1993). IMP-L2 and its homologues consist of a signal sequence followed by two Ig domains, and are about 250 residues in length (see Figure 3). Three of the four *C. elegans* proteins are obviously homologues of one another (Table 1B). In addition to being 45% identical with one another, C09D7.1 and C14F5.2 are adjacent with each other on the X chromosome (Table 2 and Figure 7).

A further four genes (Figure 3) also have a two-Ig domain topology, but no significant similarity to IMP-L2. The entire set of eight genes have been named zig genes and at least six of these are expressed in specific domains of the nervous system. An initial characterisation indicates that they are involved in the development of the nervous system (Hobert, 1999 <http://cpmcnet.columbia.edu/dept/gsas/biochem/labs/hobert/reseint1.html> and personal communication). ZIG-7 is the exception to this and is expressed in muscle (Hobert, 1999).

The sequence of one of these genes, zig-8 spans two predicted genes (C36F7.4 and C36F7.3). This is supported by EST data: the two pairs of ESTs, yk505g3 and yk427e11, bridge the two predictions starting within C36F7.4 at the N terminus and ending at the end of C36F7.3. The EST yk427e11.5 predicts the start site to be such that the A strand of the Ig domain in C36F7.4 is missing. The situation is further complicated by the fact that there appears to be another Ig domain-type region with residues that match the A to D strands, spanning the two N-terminal exons of C36F7.4, but these two exons are not seen together in either of the two ESTs. A schematic of the gene structure and an alignment of the two Ig domains of the zig genes is available at the website: http://www.mrc-lmb.cam.ac.uk/genomes/CE_igs.html

F54D7.4, or ZIG-7, also lacks residues at the N terminus (which is confirmed by EST data) such that four strands, half the Ig domain, are missing. One other protein, Y39E4B.8, has the same domain structure as the ZIG proteins.

Neural cell adhesion molecules

NCAM was the first neural-cell-adhesion molecule to be identified (Cunningham *et al.*, 1987). It forms homodimers with NCAMs on other neural cells and has a domain architecture of five Igs followed by two FnIIIs. The *Drosophila* homologue is called Fasciclin II (Lin & Goodman, 1994). F02G3.1, with five identified Ig and two FnIII domains, it is homologous to NCAM (Table 1A and Figure 4). The second FnIII had a significant match to the FnIII HMM, whilst the first FnIII domain could only be identified by inspection of the sequence for the key residues: these clearly had the pattern found in FnIII domains.

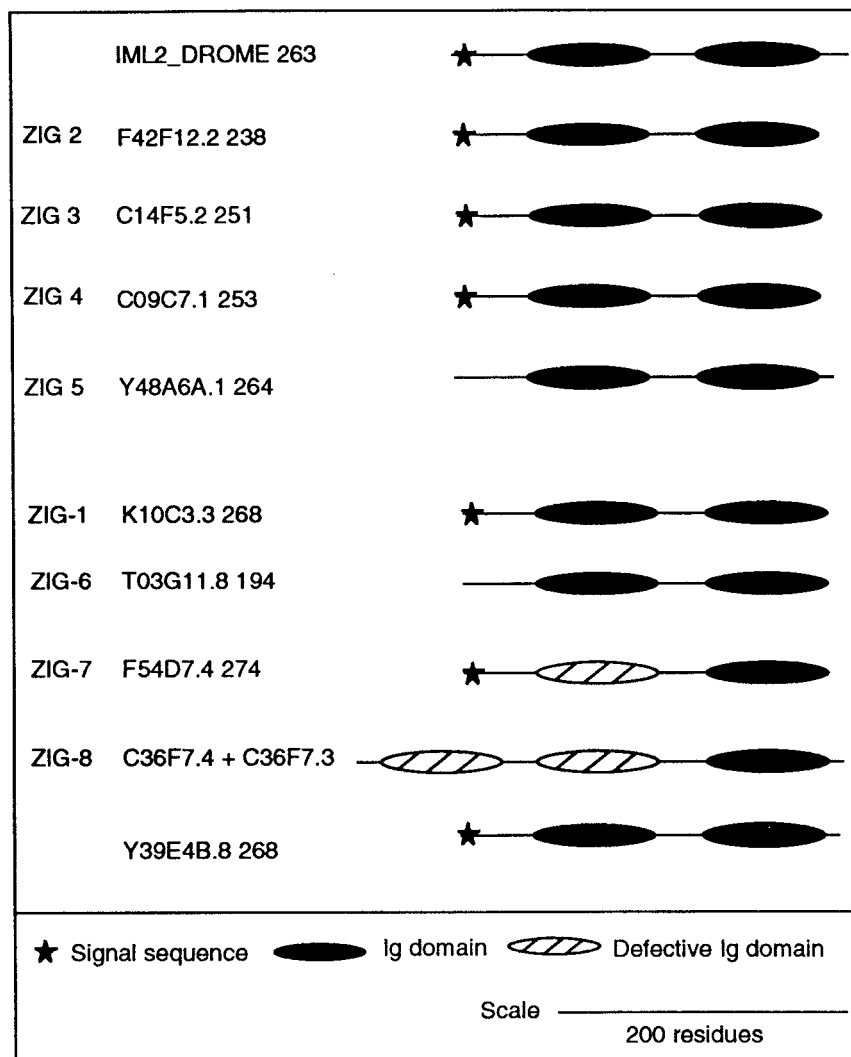


Figure 3. *zig* genes and neural/ectodermal development factor IMP-L2 homologues. Four *C. elegans* proteins are homologous to the *Drosophila* protein IMP-L2, while another three clearly have the same architecture.

Another protein that mediates homophilic cell adhesion is the *Drosophila* klingon (Butler *et al.*, 1997), which has three Ig domains followed by one FnIII domain and is homologous to C53B7.1 in *C. elegans*. F41D9.3 has the same domain architecture, and is related to the *Drosophila* wrapper protein (Nordermeer *et al.*, 1998), which also has three Igs followed by one FnIII domain. Wrapper appears to permit glial cells to have contact with commissural axons, and hence could be involved in heterophilic cell adhesion. In addition to this function, wrapper seems to be involved in regulating which glial cells survive and which are eliminated. The irregular chiasm c-roughest protein (ICCR_DROME, Ramos *et al.*, 1993) in *Drosophila* carries out an analogous function in the eye, between primary pigment cells and interommatidial cells. ICCR_DROME has five Ig domains, as does its human homologue Muc18 (Lehmann *et al.*, 1989),

but the *C. elegans* homologue K02E10.8 falls about 100 residues short of the *Drosophila* protein and has only four Ig domains (Figure 4 and Table 1A).

The human protein nephrin (Kestila *et al.*, 1998) is also a putative cell adhesion molecule, containing eight Ig domains and one FnIII domain. C26G2.1, which has six Ig domains and one FnIII domain identified by HMMs, is homologous to nephrin. The remaining two Ig domains were identified by key residue inspection; one of these is defective.

SSSD1.1 and F39H12.a have domain architectures characteristic of cell-adhesion molecules. They both consist of a number of Ig domains followed by one FnIII domain and a transmembrane region. Because of this structural homology they can be seen as putative neural cell-adhesion molecules.

Table 2. Locations of proteins on chromosomes

Gene name	Chromosome	Start	Stop	Strand
Y71G12A_205.G	I	1499764	1506148	+
F55C7.7C	I	3356529	3377471	-
F55C7.7A	I	3356529	3388807	-
C09D1.1	I	3393073	3435307	+
C24G7.5	I	3438702	3447607	+
F54D7.4	I	4112476	4115391	-
T19B4.7	I	5004943	5010062	-
F53B6.2	I	8263861	8274902	-
C36F7.3B	I	8897228	8899913	-
C36F7.3A	I	8897228	8899913	-
C36F7.4	I	8903854	8906113	-
K10C3.3	I	9179140	9181090	+
C33F10.5	II	4787526	4794789	-
C33F10.6	II	4797515	4800726	-
F56F1.4	II	5455704	5462615	+
F22D3.6	II	6937753	6941078	-
D2013.4	II	9295454	9296285	-
C09D8.2	II	10967001	10977602	+
Y38F1A.J	II	12843358	12844862	-
ZC101.2E	II	14424565	14443938	-
ZC101.2C	II	14429644	14443938	-
ZC101.2A	II	14429644	14443938	-
ZC101.2B	II	14438732	14443938	-
ZC101.1	II	14452349	14461382	-
T17A3.8	III	143076	146290	-
T17A3.1	III	152917	156972	+
K07E12.1	III	6207567	6252881	+
ZC262.3	III	7767170	7771602	-
Y48A6A.1	III	10391708	10393858	-
Y39E4B.F	III	12347105	12353206	+
T21D12.9A	IV	232545	236857	+
T21D12.9B	IV	232545	244071	+
Y94H6A_148.D	IV	2881150	2887243	-
F28E10.2	IV	4484853	4487173	+
B0273.4A	IV	5404598	5424932	-
B0273.4B	IV	5404598	5437006	-
C18F3.2	IV	7631218	7639281	-
C18F3.3	IV	7649506	7649506	-
C27B7.7	IV	8498081	8506150	+
K04D7.4	IV	9774981	9783096	-
ZK617.1B	IV	11560437	11596252	-
ZK617.1A	IV	11560437	11596252	-
C25G4.10	IV	12264380	12270599	+
T04A11.3	IV	12274721	12278136	+
R05D8.A	V	2752136	2762248	-
F54E2.3	V	2762335	2773066	-
F54E2.4	V	2778007	2779899	-
F12F3.2	V	6064366	6075845	-
F12F3.3	V	6086947	6098648	-
H05O09.1	V	6137118	6145271	-
W06H8.C	V	6145929	6152240	-
C37C3.6B	V	7812417	7827344	+
ZK994.3	V	8461535	8467152	+
F21C10.7	V	9065690	9074946	+
Y50E8A.C	V	14655632	14657053	+
F02G3.1	X	419969	427439	+
F39H12.a	X	570804	574602	-
K02E10.8	X	2259803	2265370	-
T02C5.3	X	2448660	2455918	-
ZK377.2	X	3202491	3205497	-
ZK377.3	X	3211235	3214240	-
T03G11.8	X	4942056	4942787	-
SSSD1.1	X	6123827	6128330	-
C53B7.1	X	6597327	6604235	+
C09C7.1	X	7664671	7665555	+
C14F5.2	X	7669999	7670855	+
C18A11.7	X	7790218	7797411	-
F41D9.3	X	8126437	8131776	-
K09E2.4	X	8400490	8410551	+
M02D8.1	X	8503105	8503105	+
F15G9.4B	X	9471132	9507075	+
F15G9.4A	X	9471132	9507075	+
K09C8.5	X	10725768	10732761	-

F59F3.5	X	10735925	10741389	+
F59F3.1	X	10746027	10750862	+
F58A3.2	X	10769937	10776645	+
F48C5.1	X	11198162	11200136	-
F42F12.2	X	12086061	12086875	-
C26G2.1	X	14393582	14401167	+

Genes in bold are homologous to each other and adjacent on the chromosome.

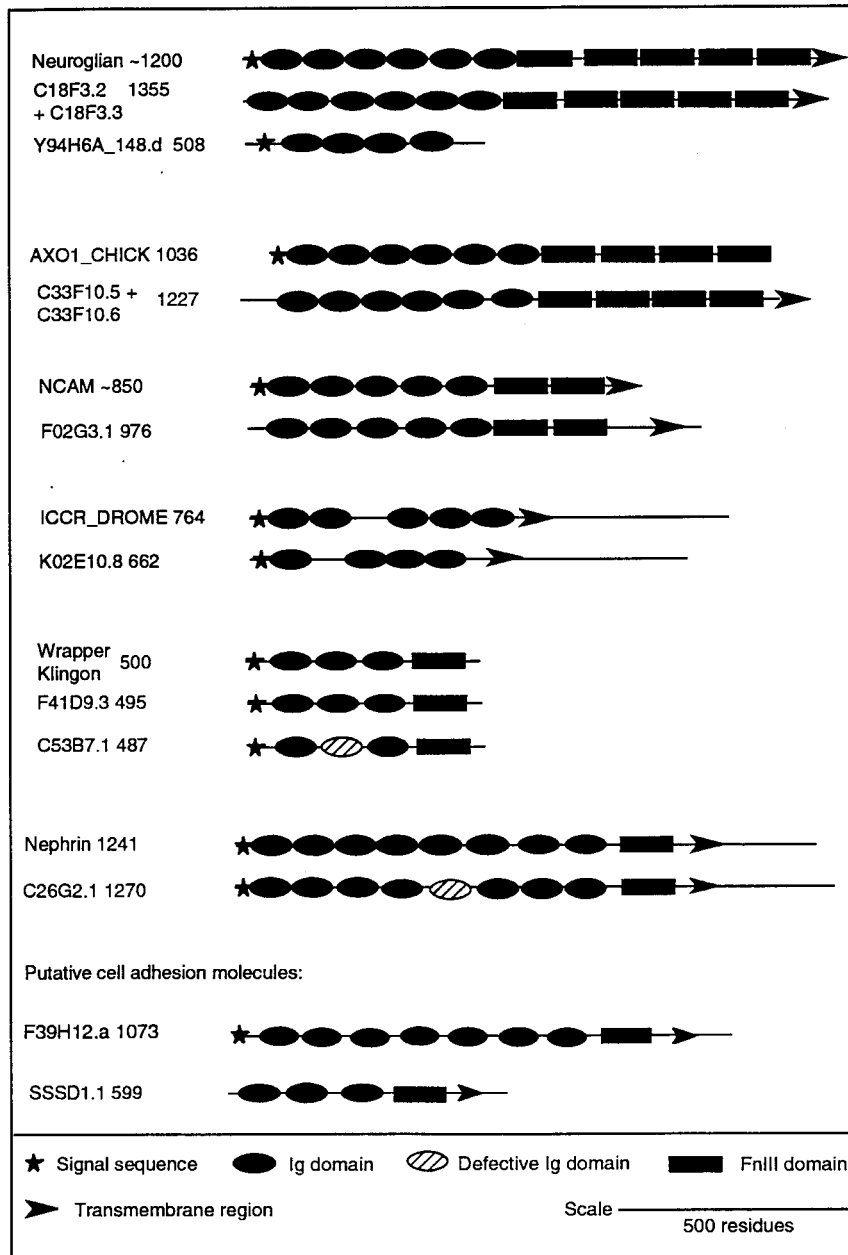


Figure 4. Neural cell adhesion molecules. Each group contains the architecture of the known neural cell adhesion protein followed by its *C. elegans* homologues. The first two groups are neuroglian and axonin, both found in many organisms and which interact with each other on adjacent cells. NCAM is also known in many organisms and forms homodimers between molecules on adjacent cells. Wrapper and klingon are *Drosophila* proteins with the same domain architecture. Wrapper is more similar to F41D9.3 while klingon is more similar to C53B7.1. ICCR is also a *Drosophila* protein, while nephryn is a human sequence.

Neuroglian and axonin homologues

Neuroglian (Bieber *et al.*, 1989) is the *Drosophila* homologue of the human cell adhesion molecule L1. It promotes neurite outgrowth by interactions with a number of proteins on both the cell surface of its own cell and on the surface of other neural cells. One of the latter type of protein is the cell adhesion molecule axonin.

The homologue to neuroglian is given by combination of two adjacent *C. elegans* proteins, C18F3.2 and C18F3.2, as well as additional amino acids found by the comparison of *Drosophila* neuroglian to the *C. elegans* DNA region in the C18F3 clone with GeneWise (Birney & Durbin, 1997). The full protein is 1355 residues long and has the same domain architecture as neuroglian. (The correct sequence can be found at the URL which accompanies this study.) The homology of this protein to *Drosophila* neuroglian, NRG_DROME, extends over the whole length with an expectation value of 0 (31% sequence identity). Y94H6A_148.d is a high-scoring neuroglian homologue (34% sequence identity, expectation value 0) that contains four Igs and is also similar to the *C. elegans* neuroglian (Table 1A). With 508 residues, it is too short to contain the six Ig/five FnIII domains that constitute a neuroglian. However, this sequence is from an unfinished region of the *C. elegans* genome, so it is possible that the true sequence is longer.

Two *C. elegans* sequences adjacent on chromosome II, C33F10.6 and C33F10.5, are homologous to different regions of axonin (Zuellig *et al.*, 1992), which has six Ig domains followed by four FnIII domains. Part of C33F10.6 is homologous to the first Ig domain of axonin. C33F10.5 is homologous to most of the remaining part of axonin. There is no axonin homologue sufficiently close to these *C. elegans* proteins for GeneWise to verify this, but inspection of splice site predictions and other evidence available in ACeDB, makes combining the two reasonable (D. Lawson, personal communication).

C. elegans IgSF proteins that contain leucine-rich repeats

Leucine-rich repeats (LRRs) mediate protein-protein interactions and there are a small number of proteins known to contain both leucine-rich repeats and an Ig domain (Nagasawa *et al.*, 1997). Two *Drosophila* proteins, KEK1 and KEK2 (Musacchio & Perrimon, 1996), each contain three LRRs and one Ig and are similar in length to ZC262.3, which also has the same architecture (Figure 5). The two splice variants of T21D12.9 contain LRRs followed by three Igs. K09C8.5 has LRRs followed by two Ig domains and a haem-dependent peroxidase domain. The functions of all these *C. elegans* and *Drosophila* proteins are unknown.

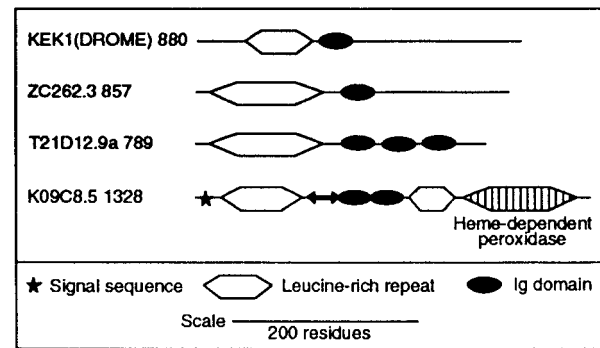


Figure 5. LRR-containing Ig superfamily members. LRRs are thought to be protein-protein interaction domains, but the specific functions of the *C. elegans* Ig superfamily members with LRR domains are unknown.

IgSF Members Without Homology

A most interesting group of Ig superfamily members in *C. elegans* are those that are uncharacterised and do not have homology to any known protein (Figure 6). There are 11 such proteins over 900 residues in length, and seven proteins under 700 residues long. Among the 11 longer proteins, there are three with long regions of low complexity sequence. F21C10.7 has many regions of low complexity besides its HSP70-CT type domain and its six Ig domains. W06H8.c and H05O09.1 also consist of low-complexity regions with nine and four Ig domains, respectively.

K09E2.4 and C27B7.7 have no low-complexity regions and a combination of Ig and FnIII domains. K09E2.4 contains three Igs followed by one FnIII and C27B7.7 has a single Ig domain with FnIII domains at the N and C termini. Neither of these patterns corresponds to one of the known domain architectures. Two of the large proteins have thrombospondin type I repeats and a single Ig domain: C37C3.6b and F53B6.2. C36C3.6b also has ten BPTI/Kunitz domains. BPTI domains are found in protease inhibitors in various tissues. The shortest of these proteins with 905 amino acids, ZC101.1, has six low-density lipoprotein receptor domains, but has a different domain architecture from characterised low-density lipoprotein receptors.

A pair of proteins that both have two FnIII domains followed by an Ig domain are almost exact replicates of each other and are adjacent to each other on chromosome IV: T04A11.3 and C25G4.10. The latter protein has an additional 500 residues at the N terminus which contain another Ig domain.

The proteins of unknown function under 700 residues in length can be divided into the three proteins over 600 residues long and four proteins under 300 residues. Two of the long proteins contain a signal sequence and one Ig domain somewhere in the protein: Y71G12A_205.g and T02C5.3. F22D3.6 is another long protein which contains a

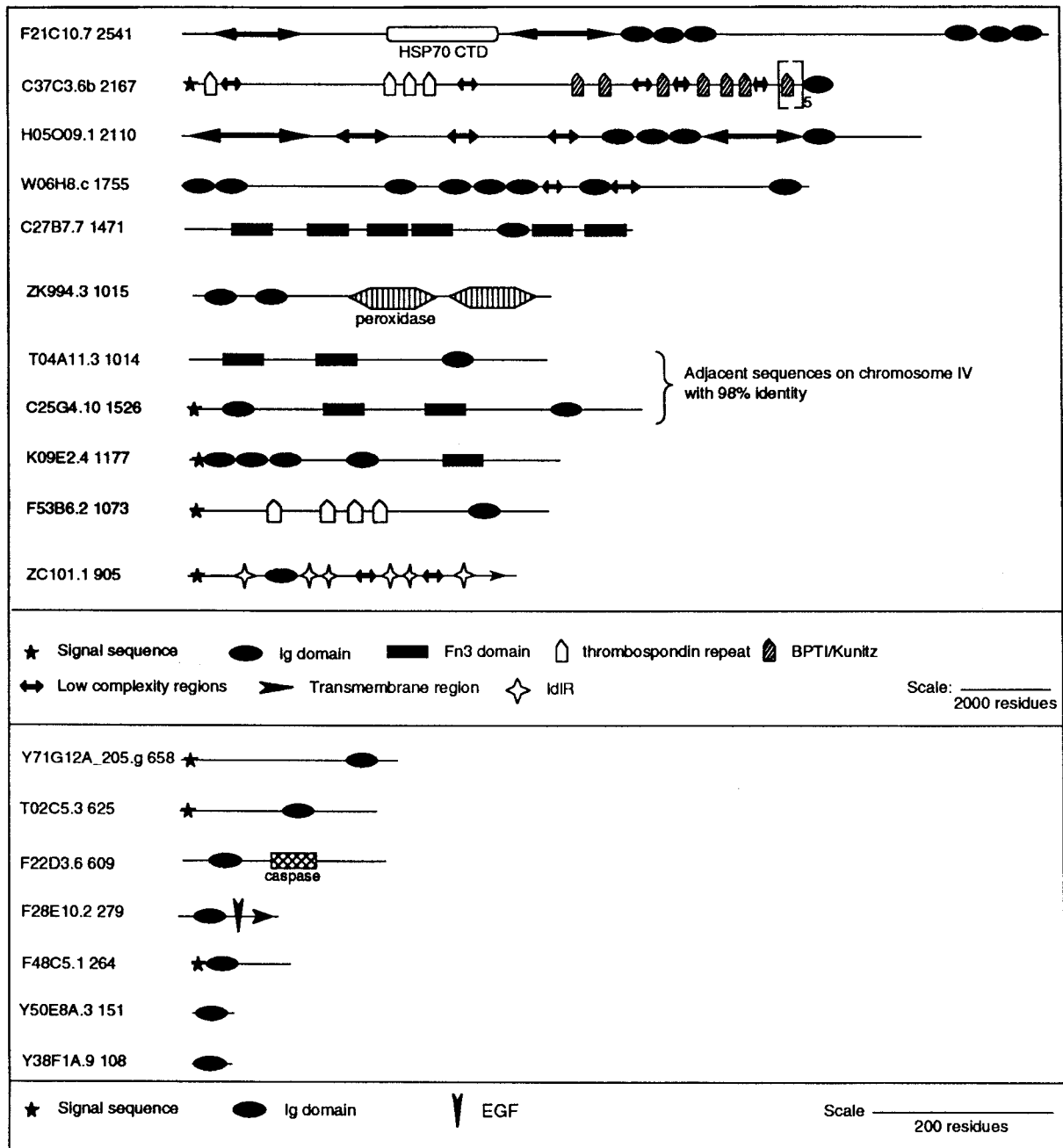


Figure 6. Proteins without homologues of known function. The 18 *C. elegans* proteins without homology to proteins of known function, divided into proteins above 900 residues in length (top) and those less than 700 residues in length (bottom).

single Ig domain, but also has a caspase domain. Three of the short proteins each contain a single Ig domain, while one also contains an EGF domain and a TM region.

Status of the IgSF assignments

From the survey of Ig superfamily members carried out here, it is obvious that there are a number of errors in gene prediction, in particular predic-

tions that separate exons belonging to a single gene. These affect at least 15 of the genes discussed here. There is evidence for six cases of two separate genes that should be joined into one: ZK377.3 and ZK377.2, C18F3.2 and C18F3.3, C36F7.4 and C36F7.3, D2013.4 and C01G6.6.8, C33F10.5 and C33F10.6, C09D8.1 and C09D8.2. One case where it is likely that three genes should become one is R05D8.a, F54E2.3 and F54E2.4. Two of these cases are based on recent independent determinations of

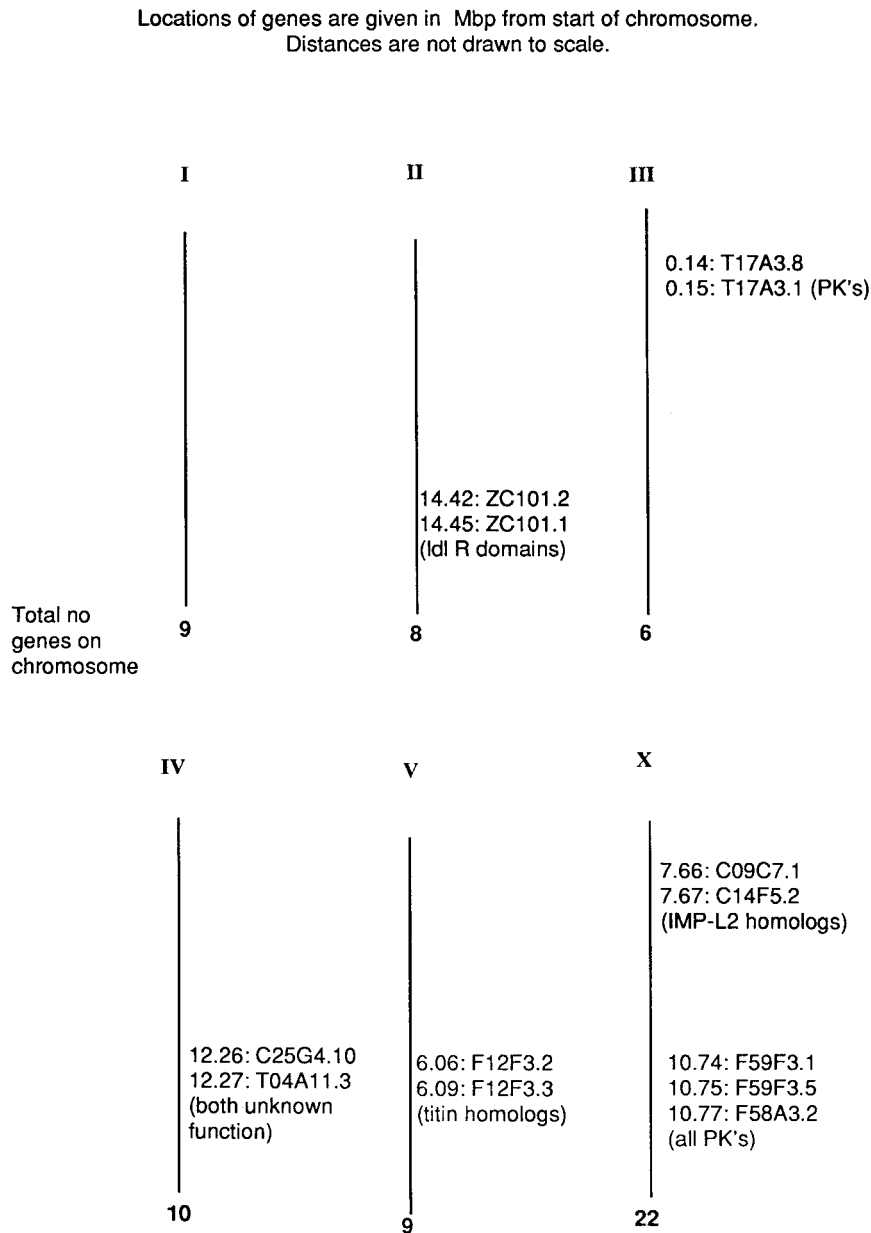


Figure 7. Distribution of Ig superfamily genes on the six chromosomes. Numbers attached to each chromosome indicate the number of Ig superfamily genes detected on that genome in this work. Genes adjacent on chromosomes are indicated by their names. The X chromosome has the highest density of Ig genes.

their gene structure: those for the *sax-3* gene (Zallen *et al.*, 1998) and the *cam-1/kin-8* gene (Forrester *et al.*, 1999). It is clear that accurate prediction of long genes is difficult without knowledge of a close homologue.

There was one case, that for C18F3.2 and C18F3.3, where the program GeneWise (Birney & Durbin, 1997) was able to improve on the prediction in the *C. elegans* database. This case involved the use of the *Drosophila* protein neuroglian which is 31% identical with the *C. elegans* protein. Other cases failed to produce a result, probably because the non-*C. elegans* homologues are too distant for the program to identify the exons reliably.

The EST database provided evidence for combining the C36G7.4 and C36F7.3 gene predictions.

There are some sequence regions, those that are hatched in the diagrams, where one part of the region has all the key residues for an Ig domain, but the remainder of what should be the Ig domain is lacking the characteristic residue types at the key positions. This could be because they are pseudogenes. Alternatively, they could be due to sequencing errors or misprediction of exon boundaries. It is worth noting that in the human V segment loci, which have been sequenced and studied for their expression patterns, about two-thirds of the sequences are actually expressed. The remaining one-third are pseudogenes.

Distribution of immunoglobulin superfamily members across the six chromosomes

Ig superfamily genes are found on all six chromosomes (I, II, III, IV, V and X) (see Table 2 and Figure 7). It is clear from this that the X chromosome has the most Ig superfamily members (22), and chromosome III the least (6).

Proteins with sequence similarity that are very close to each other on the chromosome could either be recent duplicates of one another, or be kept in proximity with one another for regulatory reasons. Such clusters are starred in Table 1B and in bold in Table 2. Two of the four proteins related to the growth factor IMP-L2, which consist of just two Ig domains, are adjacent to each other starting from 7.66 Mbp on the X chromosome, as are three of the PK domain proteins at 10.7 Mbp. Two other protein kinases are adjacent to each other at 0.14 Mbp on chromosome II, while three protein kinase homologues are close to each other on chromosome I. The two proteins with low-density lipoprotein receptor repeats are at 14.4 Mbp on the same chromosome. The two other titin homologues F12F3.2 and F12F3.3 are adjacent on the same chromosome 3 Mbp downstream. The two proteins of unknown function that have a large number of identities (98%), C25G4.10 and T04A11.3, are adjacent on chromosome IV and look like a recent gene duplication.

Immunoglobulin superfamily proteins in *C. elegans* belong to the I set

The members of the IgSF have very diverse structures. However, the inspection of the first structures to be determined showed that they could be grouped into structural "sets" (Williams & Barclay, 1988). Members of a given set have peripheral regions whose conformations are often the same even when their sequence identities are low. In the original classification of the IgSF domains three sets were identified: V, C1 and C2; the first two on the basis of known structures and the third on the basis of certain sequence characteristics that did not fit the known structures. Subsequent determination of "C2" structures showed that the sequence characteristics, as originally defined, are shared by proteins that actually form two structurally distinct sets: what is now a redefined C2 set and a new I set (Harpaz & Chothia, 1994). The standard I set molecule has a structure that combines certain structural features that are characteristic of V and C1 sets, and these suggest that the I set is the ancestor of the V, C1 and C2 sets.

The IgSF members described here were detected by a Hidden Markov model for I set proteins or by the presence of the I set key residue pattern in their sequences. The *C. elegans* protein sequences were also searched by HMMs for the V, C1 and C2 sets. These three models did match some of the *C. elegans* sequences that were matched by the I set model, but in all cases their scores are lower or

insignificant. Thus, the Ig domains described in this survey are all more closely related to the I set than they are to the other three IgSF sets.

Conclusions

We describe here 64 *C. elegans* genes which together contain 488 I set IgSF domains. Only 21 of these genes had been characterised previously. This work should be seen as an initial view of the immunoglobulin superfamily repertoire in *C. elegans*. Some gene definitions will almost certainly change as more experimental data becomes available. Some new genes may be detected in the 1% of the genome whose sequence is still unfinished. However it is very likely that the analysis described here encompasses the large majority of the IgSF members present in *C. elegans*. Thus, this work provides a basis for comparison with the IgSF repertoires in other genomes, and for experiments on function and precise structure of these proteins.

Note added in proof

After this paper went to press two papers appeared with results related to those reported here. Popovici *et al.* (1999) also detected the four IsSF tyrosine kinase receptors reported here. Hakeda *et al.* (2000) report results that support the combination of the two predicted genes R05D8.a and F54E2.3 and note its homology the *Drosophila* muscle protein kettin.

Acknowledgements

We thank Mark Diekhans for all his help in implementing SAM-T98 with constraints. Daniel Lawson (Sanger Centre) contributed his expertise on gene finding in discussions about difficult gene definitions. We are grateful to Steven Jones and Mark Gerstein for data, and Alex Bateman and Ewan Birney for helpful discussions about Hidden Markov models and GeneWise. We thank Jonathan Hodgkin for advice and for putting us in contact with Elizabeth Ryder and Oliver Hobert, to whom we are grateful for communicating their unpublished results. We are grateful to Annette Lenton for help with the Figures. S.A.T. thanks the Boehringer Ingelheim Fonds.

References

- Bairoch, A. & Apweiler, R. (1999). The SWISS-PROT sequence data bank and its supplement TrEMBL in 1999. *Nucl. Acids Res.* **27**, 49-54.
- Bateman, A. (1997). Evolution of the immunoglobulin superfamily, PhD thesis, University of Cambridge.
- Bateman, A., Jouet, M., MacFarlan, J., Du, J. S., Kenwick, S. & Chothia, C. (1996). Outline structure of the human L1 cell adhesion molecule and the sites where mutations cause neurological disorders. *EMBO J.* **15**, 6050-6059.

- Bateman, A., Birney, E., Durbin, R., Eddy, S. R., Finn, R. D. & Sonnhammer, E. L. L. (1999). Pfam 3.1: 1313 multiple alignments and profile HMMs match the majority of proteins. *Nucl. Acids Res.* **27**, 260-262.
- Benian, G. M., L'Hernault, S. W. & Morris, M. E. (1993). Additional sequence complexity in the muscle gene *unc-22* and its encoded protein twitchin of *Caenorhabditis elegans*. *Genetics*, **134**, 1097-1104.
- Benian, G. M., Tinley, T. L., Tang, X. & Borodovsky, M. (1996). The *Caenorhabditis elegans* gene *unc-89*, required for muscle M-line assembly, encodes a giant modular protein composed of Ig and signal transduction domains. *J. Cell Biol.* **132**, 835-848.
- Benson, D. A., Boguski, M. S., Lipman, D. J., Ostell, J., Ouellette, B. F., Rapp, B. A. & Wheeler, D. L. (1999). GenBank. *Nucl. Acids Res.* **27**, 12-17.
- Bieber, A. J., Snow, P. M., Hortsch, N., Patel, N. H., Jacobs, J. R., Traquina, Z. R., Schilling, J. & Goodman, C. S. (1989). *Drosophila* neuroglian: a member of the immunoglobulin superfamily with extensive homology to the vertebrate neural adhesion molecule L1. *Cell*, **59**, 313-323.
- Birney, E. & Durbin, R. (1997). Dynamite: a flexible code generating language for dynamic programming methods used in sequence comparison. *ISMB*, **5**, 56-64.
- Brenner, S. E., Chothia, C. & Hubbard, T. J. P. (1998). Assessing sequence comparison methods with reliable structurally identified evolutionary relationships. *Proc. Natl Acad. Sci. USA*, **95**, 6073-6078.
- Butler, S. J., Ray, S. & Hiromi, Y. (1997). Klingon, a novel member of the *Drosophila* immunoglobulin superfamily, is required for the development of the R7 photoreceptor neuron. *Development*, **124**, 781-792.
- Chan, S. S.-Y., Zhen, H., Su, M.-W., Wilk, R., Killeen, N. T., Hedgecock, E. M. & Culotti, J. G. (1996). *unc-40*, a *C. elegans* homolog of DCC (deleted in colorectal cancer), is required in motile cells responding to *unc-6* netrin cues. *Cell*, **87**, 187-195.
- Cunningham, B. A., Hemperly, J. J., Murray, B. A., Prediger, E. A., Brackenbury, R. & Edelman, G. M. (1987). Neural cell adhesion molecule: structure, immunoglobulin-like domains, cell surface modulation and alternative RNA splicing. *Science*, **236**, 799-806.
- Devore, D. L., Horvitz, H. R. & Stern, M. J. (1995). An FGF receptor signalling pathway required for the normal cell migration of the sex myoblasts in *C. elegans* hermaphrodites. *Cell*, **83**, 611-620.
- Eddy, S. R. (1996). Hidden Markov models. *Curr. Opin. Struct. Biol.* **3**, 361-365.
- Forrester, W. C., Dell M., Perens, E. & Garriga, G. (1999). A *C. elegans* Ror receptor tyrosine kinase regulates cell motility and asymmetric cell division. *Nature*, **400**, 881-885.
- Garbe, J. C., Yang, E. & Fristron, J. W. (1993). IMP-L2: an essential secreted immunoglobulin family member implicated in neural and ectodermal development in *Drosophila*. *Development*, **119**, 1237-1250.
- Gerstein, M. (1997). A structural census of genomes: comparing bacterial, eukaryotic and archaeal genomes in terms of protein structure. *J. Mol. Biol.* **274**, 562-576.
- Hakeda, S., Endo, S. & Saigo, K. (2000). Requirements of kittin, a giant muscle protein highly conserved in overall structure in evolution, for normal muscle function, viability, and flight activity of *Drosophila*. *J. Cell Biol.* **148**, 101-114.
- Harpaz, Y. & Chothia, C. (1994). Many of the immunoglobulin superfamily domains in cell-adhesion molecules and surface receptors belong to a new structural set which is close to that containing variable domains. *J. Mol. Biol.* **238**, 528-539.
- Hodgkin, J. A., Horvitz, H. R. & Brenner, S. (1979). Non-disjunction mutants of the nematode *C. elegans*. *Genetics*, **91**, 67-94.
- Holm, L. & Sander, C. (1998). Removing near-neighbour redundancy from large protein sequence collections. *Bioinformatics*, **14**, 423-429.
- Karplus, K., Barrett, C. & Hughey, R. (1998). Hidden Markov models for detecting remote protein homologies. *Bioinformatics*, **14**, 846-856.
- Kenny, P. A., Liston, E. M. & Higgins, D. G. (1999). Molecular evolution of immunoglobulin and fibronectin domains in titin and related muscle proteins. *Gene*, **232**, 11-23.
- Kestial, M., Lenkkeri, U., Mannikko, M., Lamerdin, J., McCready, P., Putaala, H., Ruotsalainen, V., Morita, T., Nissinen, M., Herva, R., Kashtan, C. E., Peltonen, L., Homborg, C., Olsen, A. & Tryggvason, K. (1998). Positionally closed gene for a novel glomerular protein: nephrin: is mutated in congenital nephrotic syndrome. *Mol. Cell*, **4**, 575-582.
- Klages, S., Adam, D., Class, K., Fargnoli, J., Bolen, J. B. & Penschallow, R. C. (1994). Ctk: a protein-tyrosine kinase related to Csk that defines an enzyme family. *Proc. Natl Acad. Sci. USA*, **91**, 2597-2601.
- Koga, M., Take-Uchi, M., Tameishi, T. & Oshima, Y. (1999). Control of DAF-7 TGF- β expression and neuronal process development by a receptor tyrosine kinase KIN-8 in *Caenorhabditis elegans*. *Development*, **126**, 5387-5398.
- Kokel, M., Borland, C. Z., DeLong, L., Horvitz, H. R. & Stern, M. J. (1990). *clr-1* encodes a receptor tyrosine phosphatase that negatively regulates an FGF receptor signaling pathway in *Caenorhabditis elegans*. *Genes Dev.* **12**, 1425-1437.
- Krueger, N. X., Vactor, D. V., Wan, H. I., Gelbart, W. M., Goodman, C. S. & Saito, H. (1996). The transmembrane tyrosine phosphatase DLAR controls motor axon guidance in *drosophila*. *Cell*, **84**, 611-622.
- Krogh, A., Brown, M., Mian, I. S., Sjolander, K. & Haussler, D. (1994). Hidden Markov models in computational biology: applications in protein modelling. *J. Mol. Biol.* **235**, 1501-1531.
- Labeit, S. & Kolmerer, B. (1995). Titins: giant proteins in charge of muscle ultrastructure and elasticity. *Science*, **270**, 203-296.
- Lehmann, J. M., Riethmueller, G. & Johnson, J. (1989). MUC18, a marker of tumor progression in human melanoma, shows sequence similarity to the neural cell adhesion molecules of the immunoglobulin superfamily. *Proc. Natl Acad. Sci. USA*, **86**, 9891-9895.
- Leung-Hagesteijn, C. J., Spence, A. M., Stern, B. D., Zhou, Y. W., Su, M.-W., Hedgecock, E. M. & Culotti, J. G. (1992). *unc-5*, a transmembrane protein with immunoglobulin and thrombospondin type-I domains, guides cell and pioneer axon migrations in *C. elegans*. *Cell*, **71**, 289-299.
- Lin, D. M. & Goodman, C. S. (1994). Ectopic and increased expression of fasciclin II alters motoneuron growth cone guidance. *Neuron*, **13**, 507-523.
- Musacchio, M. & Perrimon, N. (1996). The *Drosophila* kekkon genes: novel members of both the leucine-rich repeat and immunoglobulin superfamilies expressed in the CNS. *Dev. Biol.*, **178**, 63-76.

- Nagasawa, A., Kubota, R., Umamura, Y., Nagamine, K., Wang, Y., Asakawa, S., Kudoh, J., Minoshima, S., Mashima, Y., Oguchi, Y. & Shimizu, N. (1997). Cloning of the cDNA for a new member of the immunoglobulin superfamily (ISLR) containing leucine-rich repeats. *Genomics*, **44**, 273-279.
- Nielsen, H., Engelbrecht, J., Brunak, S. & von Heijne, G. (1997). Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Protein Eng.* **10**, 1-6.
- Nordermeer, J. N., Casey, C. K., Fetter, R. D., Bland, K. S., Chen, W.-Y. & Goodman, C. S. (1998). Wrapper, a novel member of the Ig superfamily, is expressed by midline glia and is required for them to ensheath commissural axons in *Drosophila*. *Neuron*, **21**, 991-1001.
- Park, J. & Teichmann, S. A. (1998). DIVCLUS: an automatic method in the GEANFAMMER package that finds homologous domains in single- and multi-domain proteins. *Bioinformatics*, **14**, 144-150.
- Park, J., Karplus, K., Barrett, C., Hughey, R., Haussler, D., Hubbard, T. & Chothia, C. (1998). Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods. *J. Mol. Biol.* **284**, 1201-1210.
- Pearson, W. R. & Lipman, D. J. (1988). Improved tools for biological sequence comparison. *Proc. Natl Acad. Sci. USA*, **85**, 2444-2448.
- Popovici, C., Roubin, R., Coulier, F., Pontarotti, P. & Birnbaum, D. (1999). The family of *Caenorhabditis elegans* tyrosine kinase receptors: similarities and differences with mammalian receptors. *Genome Res.* **9**, 1026-1039.
- Ramos, R. G. P., Igloi, G. L., Lichte, B., Baumann, U., Maier, D., Schneider, T., Brandstatter, J. H., Frohlich, A. & Fischbach, K. F. (1993). The irregular chiasm c-roughest locus of *Drosophila*, which affects axonal projections and programmed cell death, encodes a novel immunoglobulin-like protein. *Genes Dev.* **7**, 2533-2547.
- Rogalski, T. M., Gilbert, M. M., Devenport, D. & Moerman, D. G. (1998). The dim-1 gene encodes a novel protein required for myofilament stability and unc-112 encodes a homolog of the human MIG-2 protein. *Worm Breeder's Gazette*, **15**, 23.
- Rogalski, T. M., Williams, B. D., Mullen, G. P. & Moerman, D. G. (1993). Products of the unc-52 gene in *Caenorhabditis elegans* are homologous to the core protein of the mammalian basement membrane heparan sulfate proteoglycan. *Genes Dev.* **7**, 1971-1984.
- Schultz, J., Milpetz, F. & Ponting, C. P. (1998). SMART: a simple modular architecture research tool: identification of signalling domains. *Proc. Natl Acad. Sci. USA*, **95**, 5857-5864.
- Steven, R., Kubiseski, T. J., Zheng, H., Kulkarni, S., Mancillas, J., Morales, A. R., Hogue, C. W. V., Pawson, T. & Culotti, J. (1998). unc-73 activates the Rac GTPase and is required for cell and growth cone migrations in *C. elegans*. *Cell*, **92**, 785-795.
- Teichmann, S. A., Park, J. & Chothia, C. (1998). Structural assignments to the *Mycoplasma genitalium* proteins show extensive gene duplication and domain rearrangement. *Proc. Natl Acad. Sci. USA*, **95**, 14658-14663.
- Teichmann, S. A., Chothia, C. & Gerstein, M. (1999). Advances in structural genomics. *Curr. Opin. Struct. Biol.* **9**, 390-399.
- The *C. elegans* Sequencing Consortium (1998). Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science*, **282**, 2012-2018.
- Tian, S.-S., Tsoulfas, P. & Zinn, K. (1991). Three receptor-linked protein-tyrosine phosphatases are selectively expressed on central nervous system axons in the *Drosophila* embryo. *Cell*, **67**, 675-685.
- von Heijne, G. (1992). Membrane protein structure prediction, hydrophobicity analysis and the positive-inside rule. *J. Mol. Biol.* **225**, 487-494.
- Waterston, R. H., Thomson, J. N. & Brenner, S. (1980). Mutants with altered muscle structure of *Caenorhabditis elegans*. *Dev. Biol.* **77**, 271-302.
- Williams, A. F. & Barclay, A. N. (1988). The immunoglobulin superfamily - domains for cell surface recognition. *Annu. Rev. Immunol.* **6**, 381-405.
- Wootton, J. C. & Federhen, S. (1993). Statistics of local complexity in amino acid sequences and sequence databases. *Comput. Chem.* **17**, 149-163.
- Zallen, J. A., Yi, A. & Bargmann, C. (1998). The conserved immunoglobulin superfamily member SAX-3/Robo directs multiple aspects of axon guidance in *C. elegans*. *Cell*, **92**, 217-227.
- Zuellig, R. A., Rader, C., Schroeder, A., Kalousek, M. G., Von Bohlen, F., Osterwalder, T., Ivan, C., Stoeckli, E. T., Halbach, F., Affolter, H. U., Fritz, A., Hafen, E. & Sonderegger, P. (1992). The axonally secreted cell-adhesion molecule, axonin-1: primary structure, immunoglobulin and fibronectin type III domains and glycosyl-phosphatidylinositol anchorage. *Eur. J. Biochem.* **204**, 453-463.

Edited by G. von Heijne

(Received 2 November 1999; received in revised form 27 December 1999; accepted 27 December 1999)