The immunoglobulin superfamily in *Drosophila melanogaster* and *Caenorhabditis elegans* and the evolution of complexity

Christine Vogel*, Sarah A. Teichmann and Cyrus Chothia

MRC Laboratory of Molecular Biology, Cambridge CB2 2QH, UK
*Author for correspondence (e-mail: cvogel@mrc-lmb.cam.ac.uk)

Accepted 3 September 2003

Development 130, 6317-6328
Published by The Company of Biologists 2003
doi:10.1242/dev.00848

Summary

***Drosophila melanogaster* is an arthropod with a much more complex anatomy and physiology than the nematode *Caenorhabditis elegans*. We investigated one of the protein superfamilies in the two organisms that plays a major role in development and function of cell-cell communication: the immunoglobulin superfamily (IgSF). Using hidden Markov models, we identified 142 IgSF proteins in *Drosophila* and 80 in *C. elegans*. Of these, 58 and 22, respectively, have been previously identified by experiments. On the basis of homology and the structural characterisation of the proteins, we can suggest probable types of function for most of the novel proteins. Though overall *Drosophila* has fewer genes than *C. elegans*, it has many more IgSF cell-surface and secreted proteins. Half the IgSF proteins in *C. elegans* and three quarters of those in *Drosophila* have evolved subsequent to the divergence of the two organisms. These results suggest that the expansion of this protein superfamily is one of the factors that have contributed to the formation of the more complex physiological features that are found in *Drosophila*.**

Key words: Protein evolution, Cell-cell recognition, Comparative evolution, Reverse genetics

Introduction

The anatomy and physiology of an organism is determined primarily by the protein repertoire encoded in its genes and the expression patterns of these genes. This means that determining the protein repertoires of organisms makes a significant contribution to an understanding of the molecular basis of their anatomy and physiology and of why they differ between organisms.

In this paper, we describe the determination of the immunoglobulin superfamily (IgSF) repertoire in the fly *Drosophila melanogaster* and compare it with that found in the nematode *Caenorhabditis elegans*. IgSF proteins are well known for their roles in cell-cell recognition and communication – both crucial processes during embryonal development. A comparison of the functions and the size of this superfamily in the two organisms should give some idea of the nature of the changes in protein repertoires that underlie the increases in physiological complexity in the fly, for example, a more elaborate nervous system.

The IgSF repertoire in *C. elegans* was initially investigated by Hutter et al. (Hutter et al., 2000) and by Teichmann and Chothia (Teichmann and Chothia, 2000). As we show below, refinements of the genome sequence and protein predictions carried out since then have revealed additional members of the IgSF. Another smaller superfamily whose members are involved in cell adhesion processes, the cadherins, has been described previously for both the worm and fly (Hill et al., 2001).

We first describe the determination of the IgSF repertoire in *Drosophila* and of the new IgSF sequences in *C. elegans*. We then analyse the IgSF proteins common to both organisms and specific to each, in terms of their homologies and functions. In the conclusion, we discuss the implications of our results for an understanding of the role of this superfamily during the metazoan evolution and as a framework for further experimental investigation.

Materials and methods

Procedures to determine the IgSF repertoire in *Drosophila*

The complete set of predicted protein sequences of *D. melanogaster* was obtained from The Berkeley-*Drosophila*-Genome Project (The Berkeley *Drosophila* Genome Project, Sequencing Consortium, 2000). They were copied from the website at http://www.fruitfly.org/sequence/release3download.shtml. The predicted worm proteins were obtained from WormBase (Stein et al., 2001; *C. elegans* Sequencing Consortium, 1998) and from the website at http://www.wormbase.org/downloads.html. We also made some use of the predicted protein sequences of the genomes of *Anopheles gambiae* (http://www.ensembl.org/Anopheles_gambiae/) and *Caenorhabditis briggsae* (http://www.ensembl.org/Caenorhabditis_briggsae/).

The names used here for the predicted proteins are the identifiers given in FlyBase and WormBase except for those proteins with names given by experimentalists who previously determined their sequences and, in most cases, their function. These specific names start with a capital letter to denote that they refer to proteins; small letters refer to genes.

A schematic overview of the procedures used to analyse these sequences is shown in Fig. 1 and described in detail below.

The identification of proteins with IgSF domains

Domains in the sequences from fly and worm resources described above were identified using hidden Markov models (HMMs) (Krogh

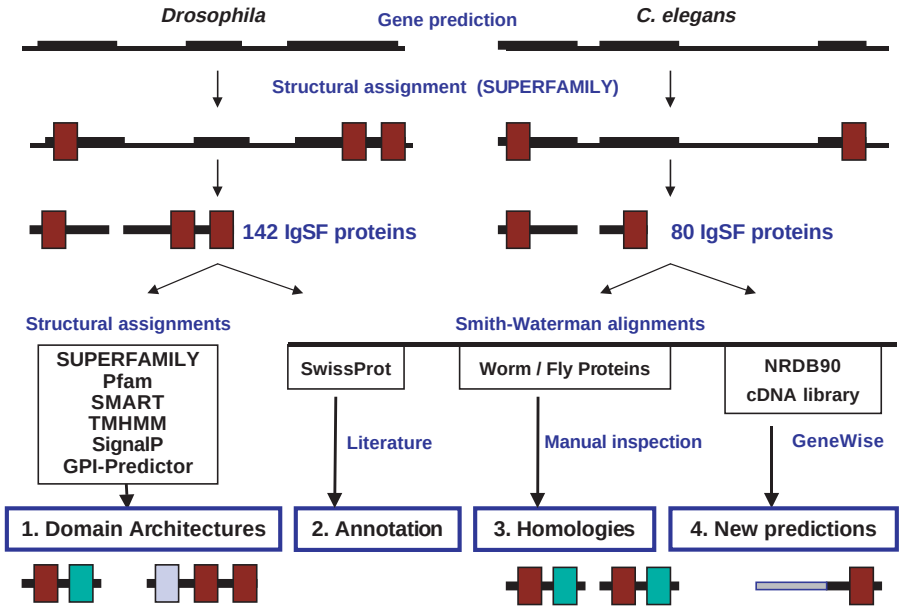

Fig. 1. Overview of the procedures to determine the IgSF repertoire in fly and worm. The genome sequence is displayed as a black line, the predicted genes are depicted as thicker lines. The thick grey line (4) represents an additional exon found with GENEWISE. Red rectangles depict predicted IgSF domains, differently coloured rectangles are domains of other superfamilies.

et al., 1994; Eddy, 1998; Karplus et al., 1998), which are probably the most sensitive automatic sequence comparison method currently available (Park et al., 1998; Madera and Gough, 2002). They are sequence profiles that, built from multiple sequence alignments, represent a family of sequences. The database SUPERFAMILY contains a library of HMMs that represent the sequences of domains in proteins of known structure (Gough et al., 2001; Gough and Chothia, 2002). These domains are whole small proteins or the regions of large proteins that are known to be involved in recombination. They are described on the Structural Classification of Proteins (SCOP) Database (Murzin et al., 1995; Lo Conte et al., 2002) where they are classified in terms of their evolutionary and structural relationships. The sequences of SCOP domains are made available through the ASTRAL database (Brenner et al., 2000; Chandonia et al., 2002) and these are used to seed the HMMs in SUPERFAMILY.

Previous to the work described here, the SUPERFAMILY HMMs were matched to the protein sequences predicted from the available genome sequences including those of *Drosophila* and *C. elegans*. The results of these matches are available from the public SUPERFAMILY database (Gough et al., 2001; Gough and Chothia, 2002). We extracted from SUPERFAMILY all *Drosophila* and *C. elegans* sequences that are matched by HMMs for IgSF domains with an expectation value score (E-value) of less than 0.01. The E-value is a theoretical value for the expected error rate. Large-scale tests show that these theoretical expectations are very close to the observed error rates. In our case, an E-value threshold of 0.01 corresponds to 1% error in the structural assignment (Gough et al., 2001).

HMM matches close to the E-value threshold were inspected by eye and judged for their correctness. In some cases they were also checked by using SMART (Schultz et al., 2000) to make domain assignments. As a result, three sequences matched with only marginally significant scores by SUPERFAMILY were rejected.

Unassigned regions of roughly 100 residues length with IgSF domains on both sides were inspected for the pattern of key residues that is a characteristic of the immunoglobulin superfamily (Chothia et al., 1988; Harpaz and Chothia, 1994). Several additional IgSF domains were detected by this procedure.

Identification of non-IgSF domains, signal sequences, transmembrane helices and GPI anchors

The proteins identified as containing one or more IgSF domains were examined for other features and domains, using six servers.

(1) The SUPERFAMILY database: the sequences matched by IgSF HMMs were examined further to see if they are also matched by HMMs for other types of domains.

(2) The Pfam database (Bateman et al., 2002): Pfam includes HMMs for protein domains of unknown structure. The IgSF proteins were submitted to this server to see if there were any additional matches.

(3) The SMART (Schultz et al., 2000) server was used to check and extend the results of the SUPERFAMILY and Pfam HMM matches.

(4) The SignalP server (Nielsen et al., 1999) was used, with the default options for eukaryotes, to identify signal sequences.

(5) The TMHMM server (Krogh et al., 2001) was used, with default options, to identify transmembrane helices.

(6) The Predictor programme (Eisenhaber et al., 1999) was used to identify GPI anchors.

These predictions were edited manually and compared with information from the literature (see below).

The IgSF proteins are either soluble or they are attached to the membrane by a transmembrane helix or a GPI anchor. For ten proteins, the GPI Predictor (Eisenhaber et al., 1999) found sites for attachment of GPI anchors. For proteins with a transmembrane helix, the IgSF domains are always in the extracellular region. After the immunoglobulin superfamily itself, the next most abundant superfamily in IgSF proteins are fibronectin type III domains, followed by the ligand-binding domain of the LDL receptor, BPTI-like domains and protein-kinase like domains. Domains from 21 superfamilies are found in both organisms, six and 10 domain superfamilies are specific to the fly and the worm, respectively.

Revision of gene predictions

In the analyses of metazoan genome sequences, a significant fraction of the predictions made for large proteins are incomplete, particularly at their N and/or C termini (Teichmann and Chothia, 2000; Hill et al., 2001). Some of these errors can be detected if there are already experimental determinations of the predicted sequences, or of close homologues, and corrected by matching the experimental sequences to the genome using the GENEWISE procedure (see below).

To detect whether predicted protein sequences are incomplete they were matched against three sets of experimental sequences

(1) Experimentally determined IgSF proteins in the public databases. The IgSF proteins were matched to sequences in the NRDB90 sequence database (Holm and Sander, 1998) using FASTA (Pearson and Lipman, 1988) with an E-value threshold of 0.001 and a sequence identity higher than 50%. For 36 IgSF proteins, we found matches in NRDB90 that were identical in sequence but at least 30 amino acids longer than the predicted sequence.

(2) A library of some 9000 full-length *Drosophila* cDNAs (http://www.fruitfly.org/sequence/dlcDNA.shtml). For 28 IgSF proteins we found cDNAs hits that were identical in sequence but at least 30 amino acids longer than the original predicted sequence (see Tables 1-3). In these cases, it is very likely that the

Table 1. *Drosophila*-specific IgSF proteins

Cell-surface proteins I

Sequence identifier	Residues	ss tmh	Sequence matches
Beat-Ib†† CG7644*	342	ss	Beat-Ic e-104, 51%
Beat-Ic†† CG4838	534		Beat-Ib e-104, 51%
Beat-IIa†† CG14334*	454		Beat-IIb e-120, 64%
Beat-VI†† CG14064	332		Beat-Ia e-40, 40%
Dpr-1†† CG13439†	367	ss	Dpr-4 e-73, 54%
Dpr-2 CG14068‡	223		Dpr-3 e-85, 60%
Dpr-3 CG15379‡,§	253		Dpr-2 e-85, 60%
Dpr-4 CG12593‡	279		Dpr-5 e-84, 56%
Dpr-5 CG5308*	364	tmh	Dpr-4 e-84, 56%
Dpr-6 CG14162*	387	ss	Dpr-10 e-91, 56%
Dpr-7 no Flybase id‡	202		Dpr-8 e-66, 50%
Dpr-8 CT16867*	370		CG31114 e-90, 51%
Dpr-9 CG12601	338		CG31114 e-118, 96%
Dpr-10 CG32057	408	ss	Dpr-6 e-91, 56%
Dpr-11 CG31309	373	tmh	CG15183 e-91, 98%
Dpr-13 CG12557‡	171		Dpr-6 e-51, 51%
Dpr-14 CG10946*	347	ss tmh	Dpr-20 e-63, 41%
Dpr-15 CG10095*,§	795	ss	Dpr11 e-58, 45%
Dpr-16 CG12591¶	406	ss	Dpr-17 e-92, 47%
Dpr-17 CG31361*	743		Dpr-16 e-91, 47%
Dpr-18 CT34788	401	tmh	Dpr-14 e-37, 34%
Dpr-19 CG13140*	435	ss tmh	Dpr-6 e-39, 50%
Dpr-20 CG12191	525		Dpr-14 e-63, 41%
CG31114*	606	tmh	Dpr-9 e-118, 96%
CG14469	185	ss	Dpr-9 e-30, 42%**
CG15380§	190		Dpr-3 e-38, 100%
CG15183	151	tmh	Dpr-11 e-91, 98%
Three-Ig-Cluster			
CG31814	672	ss tmh	CG31646 e-109, 53%
CG14010	526	tmh	CG31646 e-92, 47%
CG14521	413	ss	CG13020 e-95, 46%
CG11320	315		CG31646 e-110, 56%
CG31708	373	ss	CG31814 e-84, 52%
CG4814	215		CG31814 e-49, 50%
CG31646	606		CG14009 e-215, 75%
CG13020*	557	ss	CG31814 e-101, 49%
Dscam†† CG17800	2019	ss tmh	CG32387 e-300, 37%
CG18630_CG7060¶	544_1114	tmh	CG32387 e-132, 39%
CG32387	1770	tmh	Dscam e-300, 37%
CG31190	2008	ss tmh	Dscam e-312, 33%
Sidestep†† CG31062	939	tmh	CG14372 e-106, 34%
CG14372‡	674		CG12950 e-167, 41%
CG12484	1162	tmh	CG12950 e-117, 37%
CG30188	1073	tmh	CG14372 e-82, 35%
CG12950*	943	ss tmh	CG14372 e-167, 41%
CG14678	283		CG14372 e-62, 39%**
Lachesin†† CG12369	359		Amalgam e-80, 36%
Faint Sausage†† CG17716	822	GPI	
Fasciclin III†† CG5803	508		
Neuromusculin†† CG8779¶	1011		
CG31431¶	550	ss tmh	
CG6490	1304	tmh	
CG15275‡	449	GPI	
CG10972	569	tmh	
CG31264*	323	tmh	
CG3624*,§	232	tmh	
CG31605	484	tmh	
CT21241*	969	tmh	
CG9211	886	ss tmh	CT23737 e-189, 44%
CT23737*	1009	ss tmh	CG9211 e-189, 44%
CG7607*	198	ss	CG14141 e-43, 51%
CG14141	147		CG7607 e-43, 51%

Cell-surface proteins II – kinases and phosphatases

Sequence identifier	Residues	ss tmh	Sequence matches
Offtrack†† CG8967	1033	ss tmh	CG8964 e-133, 53%
CG8964	433	ss tmh	Offtrack e-134, 53%
Ptp69D†† CG10975*	1464	ss	

Cell-surface proteins III – with unusual domains

Sequence identifier	Residues	ss tmh	Sequence matches
Leucine-rich proteins			
Kekkon-1†† CG12283¶	880	ss tmh	Kekkon-3 e-88, 37%
Kekkon-2†† CG4977	892	ss	Kekkon-1 e-87, 36%
Kekkon-3†† CG4192	1021		Kekkon-1 e-88, 37%
CT10486	892	tmh	CG9431 e-90, 42%
CG9431	649	ss tmh	CT10486 e-90, 42%
CG1804	836	ss tmh	CG9431 e-58, 31%
CT35992§	1797	tmh	
Other types of domain			**Domain partners**
CG17839	1206	ss tmh	[DB]
CG31714	1424	6 tmh	[HRM]

Secreted proteins

Sequence identifier	Residues	ss	Sequence matches
Amalgam†† CG2198	333		Lachesin e-80, 36%
Beat-Ia†† CG4846	427	ss	Beat-Ib e-77, 51%
Beat-IIb†† CG4135	407	ss	Beat-IIa e-120, 64%
Beat-IIIa†† CG12621	208		Beat IIIb e-83, 70%
Beat-IIIb†† CG4855	337		Beat-IIIa e-83, 70%
Beat-IIIc†† CG15138	383	ss	Beat-IIIa e-81, 61%
Beat-IV†† CG10152	413		Bea-IIIc e-55, 47%
Beat-Va†† CG10134§	253		Beat-Vb e-64, 47%
Beat-Vc†† CG14390	247		Beat-Vb e-46, 43%
Beat-Vb†† CG31298*	334	ss	Beat-Va e-63, 47%
Beat-VII†† CG14249	277		Key residue analysis
CG31970	450	ss	CG15354/5 e-46, 37%
CG15354_CG15355§	255_229	ss	CG31970 e-43, 37%
ImpL2†† CG15009*	401		
CG13992§	659	ss	
CT35293*,§	420	ss	
CG5597§	260	ss	
CG13532*	267	ss	
Unusual domain partners			**Domain partners**
Vein†† CG10491§,¶	707		EGF/Laminin
CG16974	1257	ss	Leucine-rich repeat
CG9508	823		Metalloprotease

Proteins of unknown cellular location

Sequence identifier	Residues	Sequence identifier	Residues
CG15214	288	CG14677§	841
pp-CT34321	140	CG13672§	117
CG5699	485	CG14698	107
pp-CT34320	148	CG13134§	147
pp-CT34319	93	CG31369¶	377
CG14964	1427	CG30171	3197

The entry for each sequence identifier usually represents a group of sequences that point to the same gene: the predicted protein (and potentially one or more other sequences such as the cDNA sequence), the sequence found using GENEWISE, the experimentally determined sequence or the gene prediction from the previous release of the fly genome. The sequence identifier is marked accordingly if the predicted sequence is not the longest one in the group. The sequence matches are denoted as 'match partner E-value, sequence identity'. Groups of closely related proteins are indicated by the sequence matches and their separation by spaces. ss, signal sequence; tmh, transmembrane helix; DB, disulphide bridge (domain); HRM, hormone receptor domain.

*cDNA is the longest sequence in this group.

†Experimentally determined sequence is the longest in this group.

‡GENEWISE predicted sequence is the longest one in this group.

§No homologue in *A. gambiae*.

¶Sequence from *Drosophila* Release 2 is the longest one in this group.

**Borderline match: the evidence for homology between the proteins is very weak.

††Experimentally characterised sequence (trivial name).

Table 2. *C. elegans*-specific IgSF proteins

Cell-surface proteins I

Sequence identifier	Residues	ss tmh	Sequence matches
Zig-1‡‡ K10C3.3	265	ss tmh	See text
Zig-2‡‡ F42F12.2*	238	ss	
Zig-3‡‡ C14F5.2	251	ss	Zig-2 e-54, 40%
Zig-4‡‡ C09C7.1	253	ss	Zig-3 e-72, 44%
Zig-5‡‡ Y48A3A.1	260		
Zig-6‡‡ T03G11.8	194		
Zig-7‡‡ F54D7.4	255	ss	
Zig-8‡‡ Y39E4B.8	268	ss	
E04F6.9†	128	ss	E04F6.8 e-43, 57%
E04F6.8†	128		E04F6.8 e-43, 57%
Y102A11A.8†	541	ss tmh	
Y32G9A.8†	304	ss tmh	
C53B7.1	487	ss tmh	
KO9E2.4	1177	ss tmh	
T25D10.2†	231	tmh	
T19D12.7†	400	tmh	
T02C5.3	625	ss tmh	
F28D1.8†	360	tmh	
Y119C1B.9†	274	ss tmh	

Cell-surface proteins II – kinases and phosphatases

Sequence identifier	Residues	ss tmh	Sequence matches
Clr-1‡‡ F56D1.4	1442	ss tmh	
K04D7.4	1156	ss tmh	

Cell-surface proteins III – with unusual domains

Sequence identifier	Residues	ss tmh	Domain partners
F28E10.2†	279	tmh	EGF/Laminin
F48C5.1	264	ss tmh	EGF/Laminin
Y37E11AR.5†	988	ss tmh	UDP-Glycosytransferase
ZC262.3A	773	ss tmh	Leucine-rich repeat
ZK512.1*	332	tmh	Subtilisin-like domain

Secreted proteins

Sequence identifier	Residues	ss	Sequence identifier	Residues	ss
T22B11.1†	490	ss	C36F7.4B	402	ss
F22D3.4*,†	123	ss	C09E7.3†	137	ss
C25G4.11†	318	ss	C05D9.9*,†	93	ss

Proteins of unknown cellular location

Sequence identifier	Residues	Domain partners
Unusual domains		
Unc-73 F55C7.7a	2488	DBL homology domain, etc.
F21C10.7*	2541	bZIP
F22D3.6	639	Caspase-like domain
(Dig-1) K07E12.1*	13,100	
C27B7.7	1472	
H05O09.1	2735	
W06H8.3	588	
M02D8.1	197	
Y50E8A.3	151	
Y38F1A.9	109	
F12F3.2b	2808	
C24G7.5	1398	
Dim-1‡‡ C18A11.7	640	

The entry for each sequence identifier usually represents a group of sequences that point to the same gene: the predicted protein (and potentially one or more other sequences such as the cDNA sequence), the sequence found using GENEWISE or the experimentally determined sequence. The sequence identifier is marked accordingly if the predicted sequence is not the longest one in the group. The sequence matches are denoted as 'match partner E-value, sequence identity'. Groups of closely related proteins are indicated by the sequence matches and their separation by spaces. ss, signal sequence; tmh, transmembrane helix.

*No homologue in *C. briggsae*.

†The *C. elegans* protein is new to the data set compared with a previous data set (Teichmann and Chothia, 2000).

‡Experimentally characterised sequence (trivial name).

cDNAs represent the complete version of the gene or a longer splice variant.

(3) The *Drosophila* IgSF sequences were matched against those predicted for the *Anopheles gambiae* genome (http://www.fruitfly.org/sequence/dlcDNA.shtml) using Smith-Waterman alignments (Smith and Waterman, 1981).

Predicted IgSF proteins that had matched experimental versions of their sequences in NRDB, or close sequence homologues in *Anopheles* that are greater in length by at least 30 amino acids were checked using the GENEWISE program (Birney and Durbin, 2000). GENEWISE, using an HMM algorithm, tries to identify the exons in DNA that are homologous to the query protein. Because this method relies on the similarity of the two sequences, homologues with a sequence identity of more than 50% are usually required for a significant match. The homologous protein was compared with the chromosomal region containing the *Drosophila* gene and with up to 30 kb of surrounding DNA at either end of the gene. In eight cases (see Tables 1 and 3), the sequence found by GENEWISE was longer than both the original sequence and any matching cDNAs. Some *C. elegans* gene predictions were revised in a similar manner using homologues from *Caenorhabditis briggsae*. Details are described below.

In addition to these improvements in the sequences of the current FlyBase release number 3 (http://www.fruitfly.org/sequence/dlMfasta.shtml), there are 13 cases of genes predicted by the previous release, number 2, that are shorter or absent in the current release. These sequences are indicated in Tables 1 to 3.

Revision of the *C. elegans* IgSF repertoire

IgSF proteins in *C. elegans* were described previously (Hutter et al., 2000; Teichmann and Chothia, 2000). In Teichmann and Chothia (Teichmann and Chothia, 2000), 64 proteins were identified. Since then, new predictions based on revised genome sequences have been released (http://www.wormbase.org/downloads.html). These were analysed using procedures similar to those described above for *Drosophila* proteins. This resulted in a new total of 80 IgSF proteins in *C. elegans*. Of these 80, 53 are identical or nearly identical to those found in the previous work, eight are revised versions of old predictions and 19 are new (Tables 2 and 3). For the revised versions, the respective homologue in *C. briggsae* was examined and taken in one case (SSSD1.1) to improve the gene prediction using GENEWISE (Birney and Durbin, 2000).

Classification of IgSF proteins

In discussing the IgSF proteins we find that it is useful to divide them into six classes. These classes are based on broad functional similarities, although within each class the proteins also have common features in terms of domain architecture. Proteins that share a particular domain architecture belong largely, but not always, to the same cluster of closely related IgSF proteins. Details of these relationships are described in Tables 1 to 3 and the text below.

Cell surface I (see Fig. 2)

These are proteins that span the cell membrane via a transmembrane helix or are attached to the cell surface by a GPI anchor. They have an extracellular region that is exclusively, or almost exclusively, composed of IgSF and fibronectin type III (FnIII) domains, and cytoplasmic domains that are not kinases or phosphatases. Experimentally characterised proteins in this class are mainly cell-adhesion molecules that play important roles in development.

Cell surface II (see Fig. 2)

These are proteins that span the cell membrane via a transmembrane helix. They have an extracellular region that is exclusively, or almost exclusively, composed of IgSF and FnIII domains, and cytoplasmic domains that are kinases or phosphatases. All experimentally characterised proteins in this class are cell-surface receptors that bind various factors.

Cell surface III (see Fig. 2)

These are proteins that span the cell membrane via a transmembrane helix or are attached to the cell surface by a GPI anchor. They have an extracellular region that is composed of IgSF domains and a variety of different domains. Experimentally characterised proteins in this class act as signalling molecules during neural development.

Secreted proteins (see Fig. 3)

These proteins have a variety of different domain architectures that can consist of just IgSF domains but can also include other domains, some of which are unusual. They act as intercellular messengers: secreted by one cell and interacting with cell surface receptors on other cells. Three different groups of proteins fall into this class: (1)

Table 3. IgSF proteins shared between *Drosophila* and *C. elegans*

Cell-surface proteins

Name	Sequence identifier	Residues	Sequence matches
Kirre*	CT12279	968	
Roughest*	CT13684†	767	Kirre e-144, 69%
(*C. elegans*) SYG-1*	K02E10.8	718	Kirre e-52, 26%
Wrapper*	CG10382	500	
Klingon*	CG6669	545	Wrapper e-53, 29%
	CG7166‡	467	Klingon e-42, 26%
	CG13506‡	504	Key residue analysis
	CG12274	362	Klingon e-104, 42%
(*C. elegans*)	F41D9.3b	444	Key residue analysis
Turtle*	CG15427§	1531	
	CG16857†	731	Turtle e-114, 31%
(*C. elegans*)	SSSD1.1¶	744	Turtle e-51, 27%
Echinoid*	CG12676	1332	
Fred*	CG31774	1935	Echinoid e-300, 66%
(*C. elegans*)	F39H12.4	1073	Echinoid e-79, 27%
Sticks‘n’Stones*	CG13752§	1482	
Hibris*	CG7449	1215	S'n'S e-300, 50%
(*C. elegans*)	C26G2.1	1270	S'n'S e-124, 27%
Roundabout 1*	CG13521	1395	
Roundabout 2*	CG5481	1463	Roundabout 1 e-192, 37%
Roundabout 3*	CG5423	1342	Roundabout 1 e-212, 31%
(*C. elegans*) Sax-3*	ZK377.2b	1269	Roundabout 1 e-184, 39%
Frazzled*	CG8581	1526	
(*C. elegans*) Unc-40*	T19B4.7	1415	Frazzled e-105, 26%
Sidekick*	CT16627	2223	
(*C. elegans*)	Y42H9B.2**	2294	Sidekick e-259, 30%
Neuroglian*	CT4318 †	1293	
(*C. elegans*) Lad-1*	C18F3.2	1287	Neuroglian e-115, 28%
(*C. elegans*)	Y54G2A.25**	1187	Neuroglian e-85, 27%
Fasciclin II*	CT12301	873	
(*C. elegans*)	F02G3.1c	955	Key residue analysis
D-Axonin*	CG1084	1336	(also known as Contactin)
(*C. elegans*)	C33F10.5b	1227	Contactin e-67, 24%

Cell surface – combination with unusual domains

Name	Sequence identifier	Residues	Sequence matches
LRR- protein	CG8434	1173	
(*C. elegans*)	T21D12.9b	1447	CG8434 e-87, 28%
Unc-5*,†	CG8166†	1076	
(*C. elegans*) Unc-5*	B0273.4a	947	Unc-5 e-51, 33%

Cell-surface – kinases and phosphatases

Name	Sequence identifier	Residues	Sequence matches
Heartless/FGR1*	CG7223†	785	
Breathless/FGR2*	CG32134	1052	Heartless e-215, 53%
(*C. elegans*) Egl-15*	F58A3.2	1128	Breathless e-104, 37%
PVR* (or Vgr)	CG8222	1509	PVR and F59F3.1 share
(*C. elegans*)	F59F3.1	1227	the vertebrate homologue
(*C. elegans*)	F59F3.5	1199	F59F3.1 e-300, 44%
(*C. elegans*)	T17A3.1††	1083	F59F3.5 e-239, 38%
(*C. elegans*)	T17A3.8	518	F59F3.5 e-92, 47%
(*C. elegans*)	T17A3.10**,††	352	F59F3.1 e-46, 34%

Cell-surface – kinases and phosphatases (continued)

Name	Sequence identifier	Residues	Sequence matches
(*C. elegans*) Cam-1*	C01G6.8a	928	
Nrk* (no IgSF)	CG4007-PA	724	Cam-1 e-76, sid: 29%
Ror* (no IgSF)	CG4926-PA	685	Cam-1 e-88, sid: 33%
Lar*	CG10443‡	2037	
(*C. elegans*)	C09D8.1a	2180	Lar e-300, 36%

Secreted proteins

Name	Sequence identifier	Residues	Sequence matches
VMO-I Protein	CG31619	1353	
(*C. elegans*)	F53B6.2a	1043	CG31619 e-111, 28%
Semaphorin-2a*	CG4700†	762	
(*C. elegans*)	Y71G12B.20	658	Sema-2a e-73, 30%

Extracellular matrix

Name	Sequence identifier	Residues	Sequence matches
Perlecan*	CT23996	4072	
(*C. elegans*) Unc-52*	ZC101.2e	3375	Perlecan e-195, 22%
	ZC101.1	905	Perlecan e-39, 24%
Papilin*	CG18436‡	3060	
(*C. elegans*)	C37C3.6b	1550	Papilin e-240, 28%
Peroxidasin*	CG12002	1512	
(*C. elegans*)	K09C8.5	1328	Peroxidasin e-236, 34%
(*C. elegans*)	ZK994.3	1015	Peroxidasin e-243, 42%
	CG32311	1203	
(*C. elegans*) Unc-89*	C09D1.1	6632	CG32311 e-72, 27%
(*C. elegans*) Him-4*	F15G9.4b	5198	Unc-89 e-185, 24%

Muscle proteins

Name	Sequence identifier	Residues	Sequence matches
Stretchin*	CG18255	9270	Projectin e-107, 35%
(*C. elegans*)	Y38B5A.1**	2083	Stretchin e-87, 24%
Projectin*	CG32019	8971	
(*C. elegans*) Twitchin/ Unc-22*	ZK617.1b	7158	Projectin e-300, 42%
Titin	CG1915	18074	
(*C. elegans*)	F54E2.3a	4488	Titin e-300, 31%
(*C. elegans*)	F12F3.3	3484	Titin e-54, 20%

The entry for each sequence identifier usually represents a group of sequences that point to the same gene: the predicted proteins (and potentially one or more other sequences such as the cDNA sequence), the sequence found using GENEWISE, the experimentally determined sequence or the gene prediction from the previous release of the fly genome. The sequence identifier is marked accordingly if the predicted sequence is not the longest one in the group. The sequence matches are denoted as ‘match partner E-value, sequence identity’. Groups of closely related proteins are indicated by the sequence matches and their separation by spaces. ss, signal sequence; tmh, transmembrane helix.

*Experimentally characterised sequence.

†cDNA is the longest sequence in this group.

‡Sequence from *Drosophila* Release 2 is the longest one in this group.

§Experimentally determined sequence is the longest in this group.

¶GENEWISE predicted sequence is the longest one in this group.

**The *C. elegans* protein is new to the data set compared with a previous set (Teichmann and Chothia, 2000).

††No homologue in *C. briggsae.*

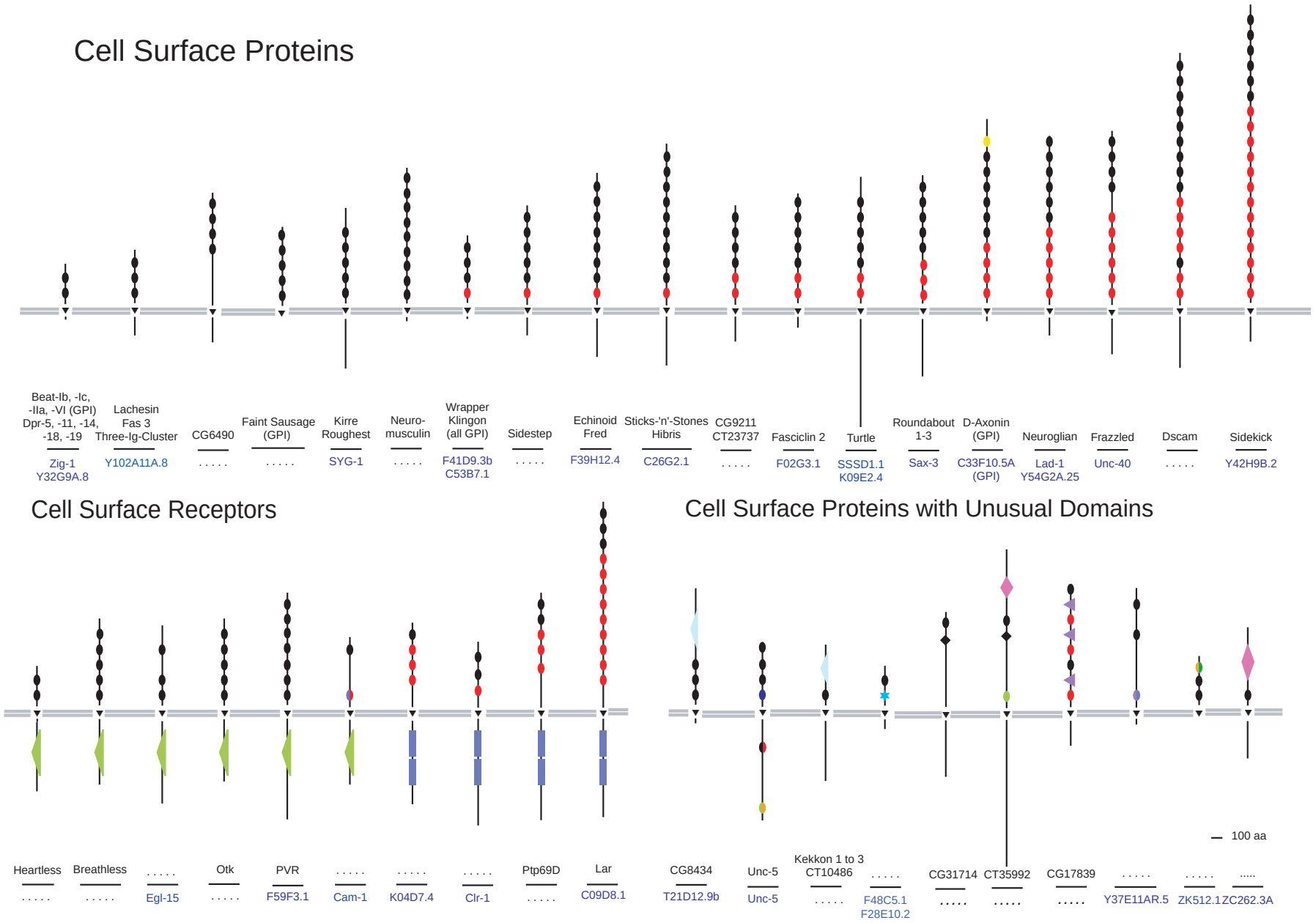

Fig. 2. Domain architectures I: cell-surface proteins, cell-surface receptors and cell-surface proteins with unusual domains. The domain architectures of IgSF proteins discussed in this work are shown as black lines representing their amino acid sequence and coloured symbols representing the domains. The legend for different domain types is at the bottom of Fig. 3. The two parallel, grey lines represent the cell membrane. Parts of proteins above the lines are extracellular, parts below the lines are intracellular. *Drosophila* proteins are in black, *C. elegans* proteins are in blue text. GPI, glycosyl-phosphatidyl-inositol anchor.

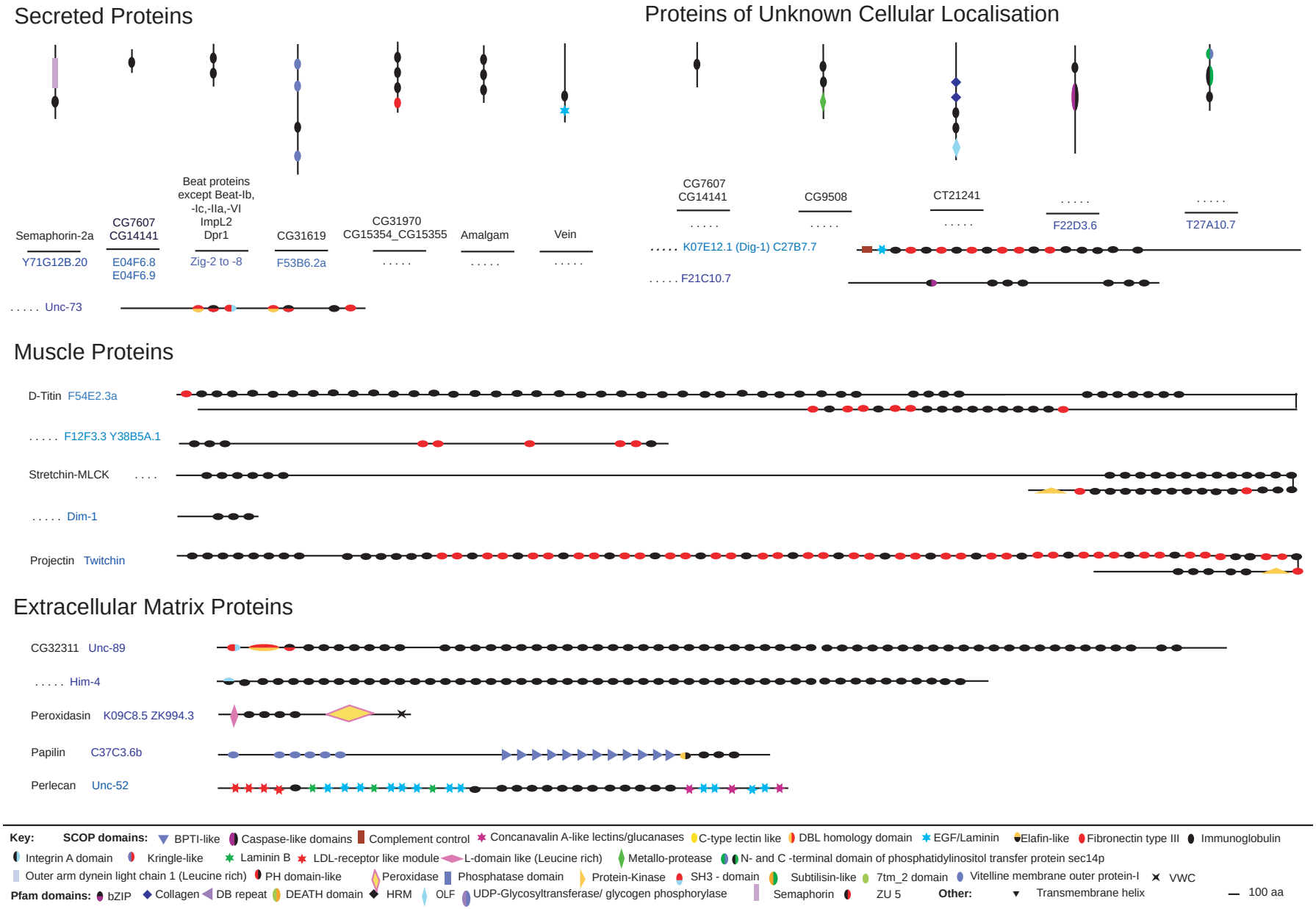

Fig. 3. Domain architectures II: secreted, extracellular matrix and muscle proteins. The domain architectures of IgSF proteins discussed in this work are shown as black lines representing their amino acid sequence and coloured symbols representing the domains. The legend for different domain types is at the bottom. *Drosophila* proteins are in black, *C. elegans* proteins are in blue text. GPI, glycosyl-phosphatidyl-inositol anchor.

proteins for which it has been shown experimentally that they are secreted; (2) proteins that have a signal sequence but no transmembrane helix or GPI anchor predicted; and (3) proteins that do not have a signal sequence, transmembrane helix or GPI anchor predicted but show sequence similarity to a proteins from (1) or (2) according to the E-value threshold described below.

Extracellular matrix proteins (see Fig. 3)

These proteins are usually rather long with more than ten IgSF domains in a row and sometimes other domains. They act in the extracellular space in cell-adhesion and cell-cell recognition processes, and thus do not have transmembrane domains or GPI anchors.

Muscle proteins (see Fig. 3)

These proteins are usually rather long with more than ten IgSF domains in a row, sometimes in combination with FnIII domains in a characteristic pattern. Some muscle proteins also have kinase domains. Experimentally characterised proteins in this class are all involved in muscle function.

All proteins were grouped into these six classes if (1) experimental work demonstrated functions characteristic to one class, (2) features in domain architecture clearly pointed towards affiliation to one class, and/or (3) the protein showed sequence similarity to a protein member of a specific class according to the E-value threshold described below. The few proteins for which none of the criteria (1), (2) or (3) apply were grouped into a 'bin' class called 'proteins of unknown cellular localisation'.

The final set of IgSF protein sequences in the two organisms have a variety of domain architectures. Figures 2 and 3 illustrate the variety of these domain architectures we found in the IgSF repertoire of fly and worm in terms of the number and kind of different domains observed in the proteins. The number of domains per protein varies from one in small signalling proteins to 68 in fly Titin. There are a few very long proteins that are in the muscle and extracellular matrix proteins classes.

Detection of relationships between IgSF proteins in *Drosophila* and *C. elegans* by sequence comparisons

In the following sections we describe and compare the IgSF proteins. To discover the relationships described below for IgSF proteins in *C. elegans* and *Drosophila*, we considered a combination of E-values for the matching sequence pairs or, for closely related proteins, sequence identities, match lengths and domain architectures. For proteins that are closely related to known structures or are very short, we also used key residue analysis (Chothia et al., 1988; Harpaz and Chothia, 1994). But before presenting this it is useful to discuss the different levels of sequence similarities that exist in these proteins and their relation to function.

By definition, all the proteins considered here contain at least one IgSF domain and are therefore homologous in at least that region. However, relationships at this basic level are not very informative. What is of more use are relationships that imply some functional annotation. We tried, therefore, to identify by sequence comparisons clusters of closely related IgSF proteins whose members are likely to have been produced by relatively recent gene duplication events and to have similar functions. To do this we first determined the extent to which indications of affiliation to one of the six functional classes can be detected from comparison of sequences. We took the 58 *Drosophila* IgSF proteins whose function has been experimentally characterised and allocated them to one of the six functional classes described in the last section. The 58 proteins were then matched to each other using the Smith-Waterman algorithm (Smith and Waterman, 1981). The scores in terms of E-value and sequence identity made by each of the matched pairs were examined.

For protein pairs whose sequence identities are greater that 40%, their close relationship is obvious. But for those where it is smaller than 40%, a statistical measure such as the E-value is much more reliable for inference of homology than sequence identity (Brenner et al., 1998). For those pairs that have E-values lower than 10^{-20} we plot the results shown in Fig. 4. Matches that occur between proteins in the same functional class and those that occur between proteins in different classes are distinguished. It clearly shows that most of matches with an E-value lower than 10^{-35} are between proteins within the same functional classes. The exceptions, where proteins of different functional classes match with E-values lower than 10^{-35}, arise from two clusters. The Beat proteins cluster has 14 members of which four are cell-surface class I proteins and ten are secreted proteins. Lachesin and Amalgam are two closely related proteins the first of which is a cell surface class I protein and the second is in the secreted proteins class.

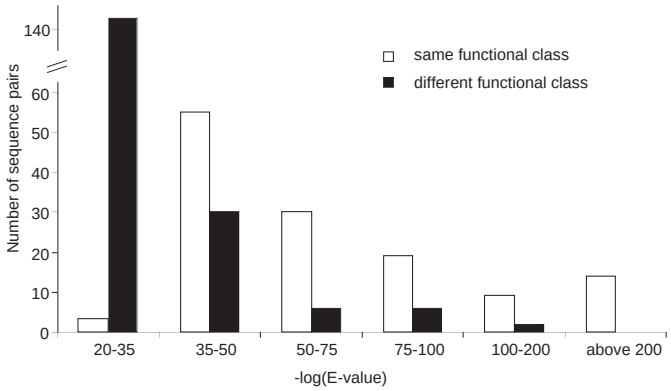

Fig. 4. E-value distribution. The histogram shows the frequency distribution of E-values between pairs of experimentally characterised IgSF proteins in *Drosophila*. The *x*-axis displays bins of the negative decadic logarithm of the E-value. White columns, proteins of the same class; black columns, proteins of different classes.

We then examined protein pairs whose match scores have E-values larger than 10^{-35} and sequence identities of less that 40%. When the cut-off parameters were slightly loosened (E-value cut-off of 10^{-30} or sequence identity cut-off of 30%), only very few more matches between proteins of the same functional classes appeared. When the cut-off parameters were further loosened, we only found matches between proteins of different functional classes.

Thus, the matches made between the 58 *Drosophila* proteins suggest that sequences with identities of 40% or greater or E-values below 10^{-35} belong to the same functional class. Note that the match region covered more than 50% of the length of both proteins. (It should be noted that not all proteins within a functional class match each other with a score less that 10^{-35}. This means that only positive results are significant; a negative one just means a function cannot be implied by sequence comparisons.)

All the IgSF proteins meeting these conditions were then grouped into clusters of closely related, homologous proteins using a single linkage algorithm: a protein qualifies as a member of a cluster if it matches at least one of the other cluster members within the above mentioned thresholds. All clusters were inspected by eye to ensure accuracy, and a few clusters were split into separate clusters based on domain architectures and inter-domain connections of subgroups of proteins within the cluster, as described below. We used these clusters to assign uncharacterised proteins that were homologous to characterised proteins to the six functional classes.

Results and discussion

The immunoglobulin superfamily repertoires in *Drosophila* and *C. elegans*

The calculations described above identified 142 IgSF proteins

in *Drosophila* and 80 proteins in *C. elegans*. We have ignored different splice variants. Those proteins known to have splice variants are represented by the longest sequence known to us. The two sets of proteins were compared in terms of their domain architectures, sequence similarities (percent identity and E-value), key residues and inter-domain connecting regions. Similarities between *Drosophila* and *C. elegans* proteins detected by these criteria would imply their presence in their common ancestor. Lack of evidence would suggest either the evolution of the protein beyond the criteria described above subsequent to their divergence or, possibly, its loss in one of the two organisms since their divergence. In Table 1, we list the 106 proteins in *Drosophila* that appear to be not closely related to those in *C. elegans* (see below). In Table 2, we list the 45 proteins in *C. elegans* that appear to be not closely related to those in *Drosophila*. In Table 3, we list the 36 *Drosophila* proteins and the 35 from *C. elegans* that are closely related to each other according to the criteria described above.

Drosophila and *Anopheles gambiae* (mosquito) diverged from their common ancestor some 250 million years ago. Of the 142 *Drosophila* proteins, 128 have a clear orthologue in *Anopheles*: i.e. the *Drosophila* and *Anopheles* homologues match each other with scores better than those they made to any other protein. A similar situation applies to *C. elegans*: *C. elegans* and *C. briggsae* diverged some 40 million years ago. Here, eight IgSF proteins in *C. elegans* lack an orthologue in *C. briggsae*. The existence of clear orthologues is good evidence that the matching proteins are not pseudo-genes. The absence of a match, however, does not necessarily mean that the sequence is a pseudo-gene. This may arise from incomplete predictions, the loss of the protein in *Anopheles* or *C. briggsae*, or its recent formation in *Drosophila* or *C. elegans.*

Prior to this work, 58 *Drosophila* and 22 *C. elegans* proteins had been identified by experimental work and assigned a function. All but 25 of the other 84 *Drosophila* and the 58 *C. elegans* IgSF proteins have been assigned to one of the six functional classes defined above. Those not classified, 12 in *Drosophila* and 13 in *C. elegans*, are placed in a class termed ‘proteins of unknown cellular localisation’ (see Tables 1 and 2).

The assignments to these functional classes have been made on the basis of sequence homology and/or the presence or absence of signal sequences and transmembrane helices. The problem with using the latter features is that the prediction of long protein sequences often misses out N-terminal and C-terminal regions (Teichmann and Chothia, 2000; Hill et al., 2001). Thus, we might expect that, in some cases, proteins currently placed in the secreted proteins class, because they have a signal sequence but no transmembrane helix or GPI anchor site, will be transferred to a cell surface class by subsequent discovery of a C-terminal region with one of these features. Similar revisions could well transfer proteins currently in the unknown class to the secreted or cell surface classes.

Table 4 summarises the distribution of the proteins, and clusters of closely related proteins, between the different functional classes. In both organisms, the two largest functional classes are the cell surface class I proteins (82 and 30 in fly and worm, respectively) and the secreted proteins class (22 and 12 proteins) many of whose members have important roles during development. These proteins form three-quarters of the *Drosophila* IgSF repertoire and half of that in *C. elegans*. The average size of the two clusters in *Drosophila* is larger than in *C. elegans*. The other four functional classes have similar numbers of fly and worm proteins. As mentioned above, these numbers are likely to be modified when more accurate data become available, but any such changes are unlikely to change the general result.

Table 4. Distribution across functional classes

	Proteins		Clusters	
	Drosophila	*C. elegans*	*Drosophila*	*C. elegans*
Cell surface I	82	31	30	21
Cell surface II	7	10	6	4
Cell surface III	11	7	6	7
Secreted proteins	23	8	13	8
Extracellular matrix	4	7	4	4
Muscle	3	4	3	3
Unknown	12	13	12	9
Total	142	80	74	56

Overview of the number of proteins and clusters of homologous proteins in the different functional classes.

Drosophila IgSF proteins

The IgSF repertoire in *Drosophila* comprises 142 proteins. Of these, 89 belong to one of 18 clusters that contain two or more closely related proteins that have totally or largely been produced by gene duplication. This means that half the repertoire in the fly, i.e. 89-18=71 proteins, have been produced by gene duplication. Some proteins have been duplicated only once, some several times. In some instances the duplications have been followed by the loss or gain of domains. The six largest clusters are Defective Proboscis extension Response (DPR) proteins (23 members), the Beat proteins (14), the Three-IgSF-Cluster (8), Sidestep (6), Kekkons (6) and Wrapper/Klingon (5) clusters. Another six clusters have only two or three members (see Tables 1 and 3).

Many members of the large clusters have been previously identified: 20 proteins in the DPR cluster (Nakamura et al., 2002), all 14 Beat proteins (Fambrough and Goodman, 1996), Sidestep on its own (Sink et al., 2001), three Kekkons (Musacchio and Perrimon, 1996), and Wrapper and Klingon (Butler et al., 1997; Noordermeer et al., 1998). Except for the cluster of Wrapper/Klingon, all these larger clusters are in the set of *Drosophila*-specific proteins that do not have *C. elegans* orthologues. This is an example of the lineage-specific expansions of protein families described by Aravind et al. (Aravind et al., 2000).

Comments on individual proteins and protein clusters

Beat and Dpr clusters

These two clusters had been identified and their functions determined prior to this work (Fambrough and Goodman, 1996; Nakamura et al., 2002; Pipes et al., 2001). Although some of the Beat proteins have only marginal or no sequence matches, key residue analysis shows they are all related to each other. Note that some Beat proteins are attached to the cell membrane whilst others are secreted.

It proved to be difficult to reconstruct all the relationships between Dpr1 to Dpr20 described by Nakamura et al.

(Nakamura et al., 2002). In some cases, the relationships are very remote and could only be shown by key residue analysis. For some of the sequences, the gene predictions were improved using the GENEWISE procedure (see above) and the Dpr-1 homologue as the query sequence (see above and Table 1). Dpr-12 has been mentioned in the work by Nakamura et al., but it could not be found in the set of predicted proteins. Owing to its small size (56 amino acids: the size of half an Ig domain), it has been disregarded in this analysis. CG31114-PA, CG14469-PA, CG15380-PA and CG15183-PA are predicted proteins that also belong to the same cluster, but were not mentioned previously.

Dscam cluster

We were able to identify three novel Dscam-like proteins (CG18630-PA in proposed fusion with CG7060-PA, CG32387-PA and CG31190-PA). Dscam is the *Drosophila* homologue of the human Down's syndrome cell adhesion molecule (DSCAM), which is required for axon guidance (Schmucker et al., 2000). The Dscam-like proteins hence represent interesting experimental targets.

CG1084-PA

This protein has been described recently as *Drosophila* homologue of the human Contactin (Falk et al., 2002). In fact, it makes a somewhat better match to Axonin, as was also found previously for its worm orthologue C33F10.5A (Teichmann and Chothia, 2000). The differences between Axonin and Contactin are subtle, but can be important when looking at the detailed functions of the proteins: For example, Contactin is known to display heterophilic but no homophilic binding activities (Falk et al., 2002), while both were observed for Axonin (Kunz et al., 2002). Both proteins interact with members of the L1 family, e.g. NrCAM, and are involved in axon guidance.

CG15354-PA and CG15355-PA

These two proteins match the N-terminal and C-terminal halves of CG31970-PA. They are also adjacent on the chromosome. We propose a fusion of the two predictions to give one protein.

C. elegans IgSF proteins

The IgSF repertoire in *C. elegans* comprises 80 proteins. Of these 25 belong to one of seven clusters of two or more homologous worm proteins. This means that 25-7=18 proteins have been produced by gene duplication. This is only one quarter of the *C. elegans* repertoire; as we have just seen the proportion in *Drosophila* is one-half. The two largest clusters are the Zig proteins (eight members) and PVR-like kinases (five members). The other four have only two members (see Tables 2 and 3). Only 22 out of the 80 *C. elegans* protein have been identified by experiments.

Comments on individual proteins and protein clusters

Zig proteins

Only Zig-2, Zig-3 and Zig-4 have sequence matches with E-values smaller than 10^{-35}. The membership of the other sequences in this family is based on their similar domain architecture, functional roles and manual inspection of the sequence alignments (see Aurelio et al., 2003).

SSSD1.1

The SSSD1.1 sequence in Wormbase has 623 amino acid residues. Using the homologous *C. briggsae* sequence and the GENEWISE procedure, we were able to identify additional exons, which increase the length of the predicted protein to 744 residues. SSSD1.1 is probably the *C. elegans* orthologue of Turtle (see Table 3).

Proteins common and specific to *Drosophila* and *C. elegans*

Table 3 lists the proteins in the 26 clusters of closely related IgSF proteins that this work indicates as having homologues in *Drosophila* and *C. elegans*. These contain in all 36 proteins from *Drosophila* and 35 from *C. elegans*, i.e. a quarter of those in the first organism and just under half of those in the second.

Previous work had proposed putative orthologues for the *Drosophila* proteins DPTP9 (K04D7.4), Lar (C09D8.1), PTP6 (F56D1.4), ImpL2 (C14F5.2, F42F12.2, Y48A6A.1), Kirre (K02E10.8, now SYG-1), Neuroglian (C18F3.2/3) and Klingon/Wrapper (F41D9.3b). Details of these, and the relationships found in this work are described in Table 3.

The cell surface class I has been mentioned above as the largest class in both organisms and as one of the two classes with large expansions in the fly. This is also true for the subset of those proteins common to both organisms: *Drosophila* has 21 while *C. elegans* has 12 proteins in the 11 clusters of the cell surface class I. There is only one cluster in this functional class, Neuroglian, where there are more members in the worm than in the fly (two and one, respectively). The clusters in the other functional classes have similar contributions from the two organisms with one exception. The exception is the PVR cluster of kinases, which has one member from *Drosophila* but five from *C. elegans*. An expansion of the cluster of kinases in *C. elegans* has been reported before (Rubin et al., 2000).

In both organisms, the number of proteins in the two largest functional classes, the cell surface class I and secreted proteins class, is higher for the organism-specific proteins than in the shared set described above: in the worm, 13 proteins are in these two functional classes and have a *Drosophila* homologue, while 25 proteins in these two classes are worm-specific. In the fly, this relationship is even stronger: 25 cell surface class I and secreted proteins have homologues in *C. elegans*, whereas more than three times as many or 82 proteins in these classes are fly specific. That means that, in addition to the expansion of fly proteins that have homologues in the worm, both organisms also developed a large set of organism-specific proteins, with again a larger expansion in the fly. Proteins of these classes play major roles in cell adhesion processes, and are most likely to contribute to the formation of fly specific characteristics.

Supplementary database

We have deposited information on each of the IgSF proteins described in this analysis in an interactive, supplementary database that can be found at http://www.mrc-lmb.cam.ac.uk/genomes/FlyGee/. The information includes: alternative protein identifiers or experimental names, sequence homologies, structural annotation in terms of domains, transmembrane helices and signal sequences, the amino acid sequence and extensions of the gene predictions using NRDB90 or cDNA data, or references to literature. The

database can be queried using keywords or protein identifiers. Each hit can include several sequences that all represent or point to the same protein: the predicted protein, other sequences such as a matching cDNA sequence, or the sequence found using GENEWISE, an experimentally determined sequence and/or the gene prediction from the previous release of the fly genome.

Conclusions

We have identified 142 IgSF proteins in *Drosophila*, described their domain architecture, and obtained an indication of the type of function that many of the novel proteins are involved in. We have also extended the work that was previously carried out on IgSF proteins in *C. elegans*. These results should be of use in the experimental characterisation of these proteins. Experiments, in turn, will refine or correct results reported here.

Some 26 clusters of closely related IgSF proteins are common to the two organisms and members of these clusters were present prior to the divergence of worm and fly. However, three-quarters of the *Drosophila* repertoire and half the *C. elegans* repertoire have emerged since their divergence. This means that a significant fraction of pathways involving the IgSF proteins in the much simpler organism, *C. elegans*, are not a subset of those in *Drosophila* but different. We also pointed to the particular expansion of two functional classes, many of whose members are involved in cell adhesion processes that play important roles during development. Relative to *C. elegans*, the greater size of the *Drosophila* IgSF repertoire, and the particular nature of many of its proteins, must be one of the contributing factors responsible for, for example, the formation of a more complex cellular structure in *Drosophila*.

The larger number of IgSF proteins in *Drosophila* contrasts with a smaller total number of genes: the current counts are 13,639 genes in *Drosophila* and 19,537 genes in *C. elegans* (Clamp et al., 2003). Some superfamilies in an organism expanded to improve its adaptation to its environment but without substantial increase in physiological complexity. Such changes in the protein repertoire could be called ‘conservative protein family expansions’. One example is the large expansion of two chemoreceptor families in the worm as compared with the fly (Robertson, 1998). Expansion of other superfamilies can lead to the evolution of organisms of higher complexity. This process could be called ‘progressive protein family expansions’. One example are the expansions of signal transduction domain superfamilies in the metazoan worm as compared with the unicellular baker’s yeast (Chervitz et al., 1998). Another example, described here, is the expansion of the IgSF superfamily in *Drosophila* compared with that of *C. elegans*.

The general validation of this simple distinction between conservative and progressive protein family expansions will require a fuller investigation of the relationship between the size and function of protein superfamilies in organisms of different complexity.

C.V. has a pre-doctoral fellowship from the Boehringer-Ingelheim Fonds. We thank Lincoln Stein, Keith Bradnam, Leyla Bayraktaroglu, Aubrey de Grey, Don Gilbert, Marc Champagne, Agnes Southgate, Birgit Eisenhaber, Bernard de Bono and Julian Gough for their help at various stages of the project.

References

Aravind, L., Watanabe, H., Lipman, D. J. and Koonin, E. V. (2000). Lineage-specific loss and divergence of functionally linked genes in eukaryotes. *Proc. Natl. Acad. Sci. USA* **97**, 11319-11324.

Aurelio, O., Boulin, T. and Hobert, O. (2003). Identification of spatial and temporal cues that regulate postembryonic expression of axon maintenance factors in the C. elegans ventral nerve cord. *Development* **130**, 599-610.

Bateman, A., Birney, E., Cerruti, L., Durbin, R., Etwiller, L., Eddy, S. R., Griffiths-Jones, S., Howe, K. L., Marshall, M. and Sonnhammer, E. L. L. (2002). The Pfam protein families database. *Nucleic Acids Res.* **30**, 276-280.

Birney, E. and Durbin, R. (2000). Using GeneWise in the Drosophila annotation experiment. *Genome Res.* **10**, 547-548.

Brenner, S. E., Chothia, C. and Hubbard, T. J. P. (1998). Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships. *Proc. Natl. Acad. Sci. USA* **95**, 6073-6078.

Brenner, S. E., Koehl, P. and Levitt, M. (2000). The ASTRAL compendium for protein structure and sequence analysis. *Nucleic Acids Res.* **28**, 254-256.

Butler, S. J., Ray, S. and 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.

***C. elegans* Sequencing Consortium** (1998). Genome sequence of the nematode *Caenorhabditis elegans*. a platform for investigating biology. *Science* **287**, 2012-2018.

Chandonia, J. M., Walker, N. S., Lo Conte, L., Koehl, P., Levitt, M. and Brenner, S. E. (2002). ASTRAL compendium enhancements. *Nucleic Acids Res.* **30**, 260-263.

Chervitz, S. A., Aravind, L., Sherlock, G., Ball, C. A., Koonin, E. V., Dwight, S. S., Harris, M. A., Dolinski, K., Mohr, S., Smith, T. et al. (1998). Comparison of the complete protein sets of worm and yeast: Orthology and divergence. *Science* **282**, 2022-2028.

Chothia, C., Boswell, D. R. and Lesk, A. M. (1988). The outline structure of the T-cell Alpha-Beta-receptor. *EMBO J.* **7**, 3745-3755.

Clamp, M., Andrews, D., Barker, D., Bevan, P., Cameron, G., Chen, Y., Clark, L., Cox, T., Cuff, J., Curwen, V. et al. (2003). Ensembl 2002: accommodating comparative genomics. *Nucleic Acids Res.* **31**, 38-42.

Eddy, S. R. (1998). Profile hidden Markov models. *Bioinformatics* **14**, 755-763.

Eisenhaber, B., Bork, P. and Eisenhaber, F. (1999). Prediction of potential GPI-modification sites in proprotein sequences. *J. Mol. Biol.* **292**, 741-758.

Falk, J., Bonnon, C., Girault, J. A. and Faivre-Sarrailh, C. (2002). F3/contactin, a neuronal cell adhesion molecule implicated in axogenesis and myelination. *Biol. Cell* **94**, 327-334.

Fambrough, D. and Goodman, C. S. (1996). The Drosophila beaten path gene encodes a novel secreted protein that regulates defasciculation at motor axon choice points. *Cell* **87**, 1049-1058.

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.

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.

Harpaz, Y. and 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.

Hill, E., Broadbent, I. D., Chothia, C. and Pettitt, J. (2001). Cadherin superfamily proteins in Caenorhabditis elegans and Drosophila melanogaster. *J. Mol. Biol.* **305**, 1011-1024.

Holm, L. and Sander, C. (1998). Removing near-neighbour redundancy from large protein sequence collections. *Bioinformatics* **14**, 423-429.

Hutter, H., Vogel, B. E., Plenefisch, J. D., Norris, C. R., Proenca, R. B., Spieth, J., Guo, C. B., Mastwal, S., Zhu, X. P., Scheel, J. et al. (2000). Cell biology: Conservation and novelty in the evolution of cell adhesion and extracellular matrix genes. *Science* **287**, 989-994.

Karplus, K., Barrett, C. and Hughey, R. (1998). Hidden Markov models for detecting remote protein homologies. *Bioinformatics* **14**, 846-856.

Krogh, A., Larsson, B., von Heijne, G. and Sonnhammer, E. L. L. (2001). Predicting transmembrane protein topology with a hidden Markov model: Application to complete genomes. *J. Mol. Biol.* **305**, 567-580.

Krogh, A., Mian, I. S. and Haussler, D. (1994). A hidden Markov model that finds genes in Escherichia-coli DNA. *Nucleic Acids Res.* **22**, 4768-4778.

Kunz, B., Lierheimer, R., Rader, C., Spirig, M., Ziegler, U. and

Sonderegger, P. (2002). Axonin-1/TAG-1 mediates cell-cell adhesion by a cis-assisted trans-interaction. *J. Biol. Chem.* **277**, 4551-4557.

Lo Conte, L., Brenner, S. E., Hubbard, T. J. P., Chothia, C. and Murzin, A. G. (2002). SCOP database in 2002: refinements accommodate structural genomics. *Nucleic Acids Res.* **30**, 264-267.

Madera, M. and Gough, J. (2002). A comparison of profile hidden Markov model procedures for remote homology detection. *Nucleic Acids Res.* **19**, 30.

Murzin, A. G., Brenner, S. E., Hubbard, T. and Chothia, C. (1995). Scop – a Structural Classification of Proteins Database for the Investigation of Sequences and Structures. *J. Mol. Biol.* **247**, 536-540.

Musacchio, M. and 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.

Nakamura, M., Baldwin, D., Hannaford, S., Palka, J. and Montell, C. (2002). Defective proboscis extension response (DPR), a member of the Ig superfamily required for the gustatory response to salt. *J. Neurosci.* **22**, 3463-3472.

Nielsen, H., Brunak, S. and von Heijne, G. (1999). Machine learning approaches for the prediction of signal peptides and other protein sorting signals. *Protein Eng.* **12**, 3-9.

Noordermeer, J. N., Kopczynski, C. C., Fetter, R. D., Bland, K. S., Chen, W. Y. and 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., 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.

Pearson, W. R. and Lipman, D. J. (1988). Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci. USA* **85**, 2444-2448.

Pipes, G. C., Lin, Q., Riley, S. E. and Goodman, C. S. (2001). The Beat generation: a multigene family encoding IgSF proteins related to the Beat axon guidance molecule in Drosophila. *Development* **128**, 4545-4552.

Robertson, H. M. (1998). Two large families of chemoreceptor genes in the nematodes Caenorhabditis elegans and Caenorhabditis briggsae reveal extensive gene duplication, diversification, movement, and intron loss. *Genome Res.* **8**, 449-463.

Rubin, G. M., Yandell, M. D., Wortman, J. R., Miklos, G. L. G., Nelson, C. R., Hariharan, I. K., Fortini, M. E., Li, P. W., Apweiler, R., Fleischmann, W. et al. (2000). Comparative genomics of the eukaryotes. *Science* **287**, 2204-2215.

Schmucker, D., Clemens, J. C., Shu, H., Worby, C. A., Xiao, J., Muda, M., Dixon, J. E. and Zipursky, S. L. (2000). Drosophila Dscam is an axon guidance receptor exhibiting extraordinary molecular diversity. *Cell* **101**, 671-684.

Schultz, J., Copley, R. R., Doerks, T., Ponting, C. P. and Bork, P. (2000). SMART: a web-based tool for the study of genetically mobile domains. *Nucleic Acids Res.* **28**, 231-234.

Sink, H., Rehm, E. J., Richstone, L., Bulls, Y. M. and Goodman, C. S. (2001). sidestep encodes a target-derived attractant essential for motor axon guidance in Drosophila. *Cell* **105**, 57-67.

Smith, T. F. and Waterman, M. S. (1981). Identification of common molecular subsequences. *J. Mol. Biol.* **147**, 195-197.

Stein, L., Mangone, M., Schwarz, E., Durbin, R., Thierry-Mieg, J., Spieth, J. and Sternberg, P. (2001). WormBase: network access to the genome and biology of Caenorhabditis elegans. *Nucleic Acids Res.* **29**, 1012-1012.

Teichmann, S. A. and Chothia, C. (2000). Immunoglobulin superfamily proteins in Caenorhabditis elegans. *J. Mol. Biol.* **296**, 1367-1383.

The Berkeley *Drosophila* Genome Project, Sequencing Consortium (2000). The genome of *Drosophila melanogaster. Science* **287**, 2185.