Development 1994 Supplement, 27-33 (1994)
Printed in Great Britain © The Company of Biologists Limited 1994

Protein families in the metazoan genome

Cyrus Chothia

MRC Laboratory of Molecular Biology and the Cambridge Centre for Protein Engineering, Hills Road, Cambridge, CB2 2QH, UK.

SUMMARY

The evolution of development involves the development of new proteins. Estimates based on the initial results of the genome projects, and on the data banks of protein sequences and structures, suggest that the large majority of proteins come from no more than one thousand families. Members of a family are descended from a common ancestor.

Protein families evolve by gene duplication and mutation. Mutations change the conformation of the peripheral regions of proteins; i.e. the regions that are involved, at least in part, in their function. If mutations proceed until only 20% of the residues in related proteins are identical, it is common for the conformational changes to affect half the structure.

Most of the proteins involved in the interactions of cells, and in their assembly to form multicellular organisms, are mosaic proteins. These are large and have a modular structure, in that they are built of sets of homologous domains that are drawn from a relatively small number of protein families. Patthy's model for the evolution of mosaic proteins describes how they arose through the insertion of introns into genes, gene duplications and intronic recombination.

The rates of progress in the genome sequencing projects, and in protein structure analyses, means that in a few years we will have a fairly complete outline description of the molecules responsible for the structure and function of organisms at several different levels of developmental complexity. This should make a major contribution to our understanding of the evolution of development.

Key words: protein evolution, exons, introns, mosaic proteins

INTRODUCTION

An examination of the number of genes found in different organisms shows clearly how the evolution of development involves the development of new proteins:

Bacteria: *Escherichia coli*	4,000
Yeast: *Saccharomyces cerevisae*	6,000
Nematode: *Caenorhabditis elegans*	18,000
Humans	65,000

(the numbers for *C. elegans* and humans are from Wilson et al. (1994) and Fields et al. (1994) respectively). Thus, an understanding of the process by which the protein repertoire has grown and acquired new properties during the course of evolution is an essential component of a general understanding of the evolution of development.

Here I review three recent advances in our understanding of protein evolution. First, I discuss the evidence that suggests that most, or all, proteins belong to a relatively small number of families whose members are descended from a common ancestor. Second, I show that members of a family can diverge to a point where they have very few sequence identities and half their structures have different folds. Third, I describe Patthy's model for the formation of the mosaic proteins, which have played a central role the evolution of multicellular organisms. Lastly, the implications of these discoveries are discussed.

NUMBER OF PROTEIN FAMILIES

An analysis of the protein sequences and structures that were available in 1992 gave a rough estimate of the number of protein families (Chothia, 1992). The data used to make this estimate comprised (i) the initial results of the genome projects; (ii) the data bank of known protein sequences and (iii) the data bank of known protein structures. Examination of these data showed that:

(1) Of the sequences produced by each of the different genome projects, close to one third had a clear homology to an entry that was already present in the data bank of protein sequences (see Table 1). It is reasonable to assume that sequences produced by the genome project represent a random sample and so this result suggests that one third of all protein families had a representative in the sequence data bank.

(2) Of the entries in the sequence data bank, 28% matched, with a residue identity of at least 25%, the sequence of one of the entries in the protein structure data bank. (These figures come from Sander and Schneider, 1991 and personal communication quoted in Chothia, 1992, who determined the proportion of the sequences in the EMBL/SwissProt data bank that are homologous to the sequences of the proteins in the Brookhaven protein structure data bank.) The figures suggest that about one quarter of protein sequences belong to a family for which there is a known structure.

(3) The Brookhaven protein structure data bank contained

Table 1. The sequences from genome projects that are related to other previously known sequences

Source	Total number of genes	Genes related to those previously known	Reference
A. Genome Projects			
Caenorhabditis elegans			
chromosome III (part)	32	14 (44%)	Sulston et al (1992)
Yeast			
chromosome III	182	52-66 (29-36%)	Oliver et al. (1992)
chromosome IX (part)	46	15 (33%)	Barrell, Smith and Brown*
B. Large libraries of expressed genes			
Human brain			
I	~1,400	406 (~30%)	Adams et al. (1992)
II	1,531	725 (47%)	Adams et al. (1993)
Caenorhabditis elegans			
St Louis	1,517	512 (34%)	Waterson et al. (1992)
NIH	585	210 (36%)	McCombie et al. (1992)

*Unpublished results quoted in Chothia (1992).

one or more representatives of some 120 different protein families (Pascarella and Argos, 1992; Chothia, 1992).

Put together these results imply that there are some 1,500 different protein families (Table 2).

Now this calculation assumes that sequence comparisons can give a complete picture of family membership. This is not the case. It has become clear, from the structures determined by X-ray crystallography and NMR, that proteins can evolve to the point where, though they continue to share the same fold, their sequence identities are no greater than that of two randomly selected sequences (a review of recent cases is found in Murzin and Chothia, 1992). This means that sequence comparisons with reasonable thresholds underestimate the extent to which proteins are related.

If we assume that the current methods of sequence comparisons can find 80% of the proteins in one family and adjust our calculations accordingly, the estimated number of protein families drops to 1000 (see Table 2). In fact, anecdotal evidence from recently determined protein structures suggest that sequence comparisons are not this efficient. Thus a conservative view of the current evidence is that the large majority of proteins come from no more than 1000 families.

Since the initial results of the genome projects, five further reports have appeared which are in accord with this estimate. Glaser et al. (1993) sequenced a 97 kb region of the *Bacillus subtilis* genome. They found 92 open reading frames; 42 of these coded for proteins homologous to entries already in data banks. Adams et al. (1993) partially sequenced the genes for about 1500 proteins from human brain. Half of these were found to have sequences homologous to those previously known. Wilson et al. (1994) reported on 2.2 megabases of contiguous nucleotide sequence from chromosome III of the nematode *Caenorhabditis elegans*. This section of the chromosome contains 483 genes of which 40% are related to previously known sequences or code for tRNAs. These groups used the current standard sequence matching techniques.

Using matching techniques that are more sensitive to the significance of low homologies than those used by previous workers and a larger data base of sequences and structures gave a higher proportion of matches in the two more recent reports. Koonin et al. (1994) re-examined the yeast chromosome III sequences. In the original report, 29-36% of the 182 sequences were reported to be similar to those previously known (Table 1; Oliver et al., 1992). Koonin et al. (1994) showed that the products of 61% of these genes have a significant similarity to an entry in the sequence data banks and 19% are similar to a protein of known three-dimensional structure. Dujon et al. (1994) reported the complete sequence of yeast chromosome XI. Of its 331 open reading frames, 67% coded for either known yeast proteins or were homologous to other proteins in the data banks.

Table 2. Calculation of the number of protein families

Proportion of genome sequences that have a related sequence in the sequence data bank:		one third*
Proportion of sequences in the data bank related to a protein of known structure:		one quarter*
Number of families represented by the known structures:		120*
Number of protein familes		
1. Assuming current sequences comparisons detect all members of protein families:	$3 \times 4 \times 120$:	1500
2. Assuming sequence comparisons detect 80% of family members:	$3 \times 0.8 \times 4 \times 0.8 \times 120$:	1000

* Values given here are for 1992 data; see text for details and references.

The concept of a protein family has a straight forward application to small and medium sized proteins that are built of just one domain. Large proteins, however, are usually built of several domains that are usually not homologous to one another but are often homologous to domains found in other quite different proteins. Two examples are the extracellular neural cell adhesion molecule and the intracellular muscle protein titin, which are built mostly from different combinations of immunoglobulin superfamily and fibronectin type III domains (see Fig. 1). Proteins of this kind can best be described as being built of components that come from different families of protein domains.

It should be emphasised that the number of 1000 families is very much smaller than the number of folds that are possible within the limitations of the physics and chemistry of proteins. It is difficult to calculate the number of possible folds but this point becomes apparent if we realise that taking just the 200 folds that are currently known, and changing the connections

between their secondary structures (within the limits of the protein chain topology rules), would give tens of thousands of different folds (A.G. Murzin, unpublished calculation).

THE EVOLUTION OF PROTEINS

In this section I discuss the general effects of mutations on the structure of proteins and how new proteins are made from combinations of old proteins.

Two particular proteins are used to illustrate certain general points. Both are members of the immunoglobulin superfamily (Fig. 1). In prokaryotes single domains that belong to this family are attached to glycosyltransferases and cellulase where they may be involved in carbohydrate recognition (Klein and Schultz, 1991; Juy et al., 1992). The eukaryote branch of the family is very large and has members that are involved in cell-cell adhesion, in the structural organisation and regulation of muscle and in the immune system. These members of the family are usually built from several domains of roughly 100 residues each and with homologous sequences and structures. In some cases these domains combine with domains from other families.

The two members of the superfamily used below to illustrate various points are the neural cell adhesion molecule (NCAM), which was mentioned above and which is involved in the development of the nervous system, and the antibody molecule (Fig. 1).

Evolution of structural diversity

The development of specialised or new functions, through the duplication of genes and their subsequent mutation, became apparent from the comparisons of the amino acid sequences of the globins (Ingram, 1961), of trypsin and chymotrypsin (Walsh and Neurath, 1964; Hartley et al., 1965) and of lysozyme and α-lactalbumin (Brew et al., 1967).

Later it was found that this process can go beyond the point where the diverged proteins have sequences similarities that are easily recognisable. In such cases, the common evolutionary origin only becomes clear from the similarities in the details of the three dimensional structures of the proteins. A recent striking example is the discovery that the muscle protein actin, hexokinase and the ATPase domain of the heat shock cognate protein, HSC70, belong to the same protein family. Actin and the HSC70 domain have, respectively, 375 and 386 residues. Of these, 240 have the same conformation though only 39 have identical side chains (Flaherty et al., 1991).

In general, proteins respond to changes in sequence by changes in structure (Lesk and Chothia, 1980; Chothia and Lesk, 1987). As the stability of proteins is low: the folded forms are 5-15 Kcals more stable than the denatured forms, the structural changes produced by mutations can involve only small changes in energy. This means, in turn, that structural changes in evolution occur through a series of small incremental steps. Though the individual changes are small, their cumulative effects can be large. An example of this can be seen when we examine the structural differences of the variable and constant domains in antibodies.

Antibodies are built from two types of domains: the variable (V), which form the antigen binding site and the constant (C) which are involved in various effector functions (Fig. 1). V and C domains evolved from a common ancestor, which probably had a structure intermediate between the two (Harpaz and Chothia, 1994). They have structures in which the polypeptide chain runs back and forth to form two β-sheets which pack face to face (Fig. 2). Although different V domains have small differences in structure, the two β-sheets that form the core of their structures is very largely conserved (Fig. 3A). Similarly the two β-sheets that form the core of different C domains are also conserved (Fig. 3B).

A comparison of the conserved core of V domains with that of C domains shows that there is a central region common to both but around this they have quite different structures. This is shown graphically in Fig. 3C where the conserved cores of the V and C domains are shown superimposed.

The structural differences can be measured quantitatively. Typically, V and C domains have 110 and 100 residues respec-

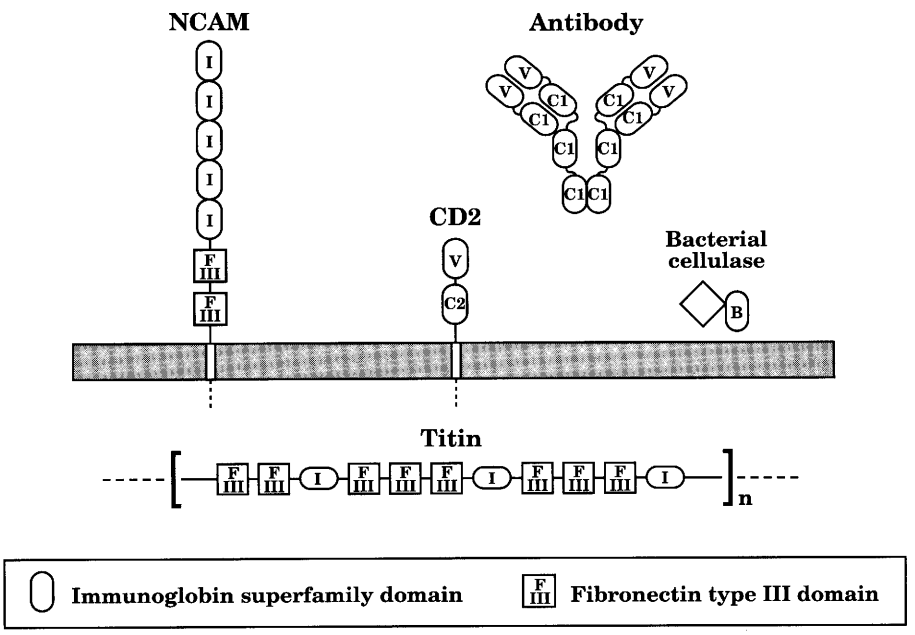


Fig. 1. Structures of five members of the immunoglobulin superfamily. Immunoglobulin superfamily domains contain approximately one hundred residues and have homologous sequences and structures. Groups of domains, that have structures more similar to one another, than they are to other members of the family, are grouped into sets. Antibodies are built of V and C1 set domains (Williams and Barclay, 1988). The CD2 adhesion molecule is built from V and C2 set domains (Jones et al., 1992) whereas the cell surface neural cell adhesion molecule (NCAM) contains five I set domains (Harpaz and Chothia, 1994) and two type III fibronectin domains that are not members of the immunoglobulin superfamily. (Some forms of the protein also contain a muscle specific domain (MSD, see Fig. 6) a transmembrane helix and a cytoplasmic domain) The muscle protein titin is also built from I set and fibronectin domains but in a different arrangement to that seen in the NCAMs. Bacterial cellulase has one B set domain linked to a non-homologous catalytic domain (Juy et al., 1992).

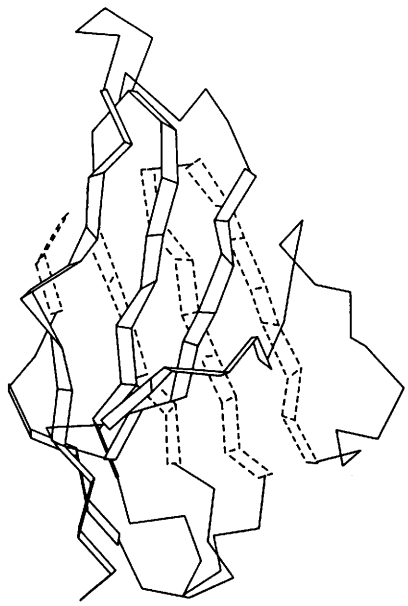


Fig. 2. The fold of the polypeptide chain in an immunoglobulin variable domain. The regions drawn as ribbons hydrogen bond to each other to form two β-sheets: one drawn in broken lines and the other in continuous lines. See also Fig. 3.

tively. Superposing individual V and C structures shows that typically, some 55 of the residues in each structure have the same conformation. That is 50% of the V domain (55/110) and 55% of the constant domain (55/100) have the same conformation but the other regions in the two proteins differ in conformations. In regions with the same conformation the proportion of residues that are identical in the V and C domains is about 15%.

The same calculation was carried out on 32 pairs of homologous proteins that come from eight protein families (Chothia and Lesk, 1986). The results are shown in Fig. 4. They demonstrate that when related proteins have 40% or more identical residues the extent of the structural changes tend to be small. But, when the sequence changes have progressed so that only 20% of the residues are identical, it is common for the structural changes to be so extensive that a core comprising only about half of each protein retains the same structure (Fig. 4).

The regions that undergo changes in conformation are minor elements of secondary structure that are on the surface, and the loops of peptide that link the major elements of secondary structure. In most proteins it is these regions that form, at least in part, the structures responsible for activity and specificity. Changes in these regions change functional properties. In the evolution of enzymes they often modify specificity; changes in the catalytic mechanisms, however, are rare but do occur occasionally (Murzin, 1993).

Evolution of mosaic proteins

The discovery that eukaryotic genes, and coding regions within genes (exons), are often separated by non-coding regions (introns) led to the proposal that new proteins can be created by shuffling exons using the intron regions for recombination (Gilbert, 1978; Blake, 1978). The current evidence is strongly against the proposal that entirely new proteins are made by shuffling individual exons (see Patthy, 1991a,b, 1994 and references therein). It is clear, however, that whole protein

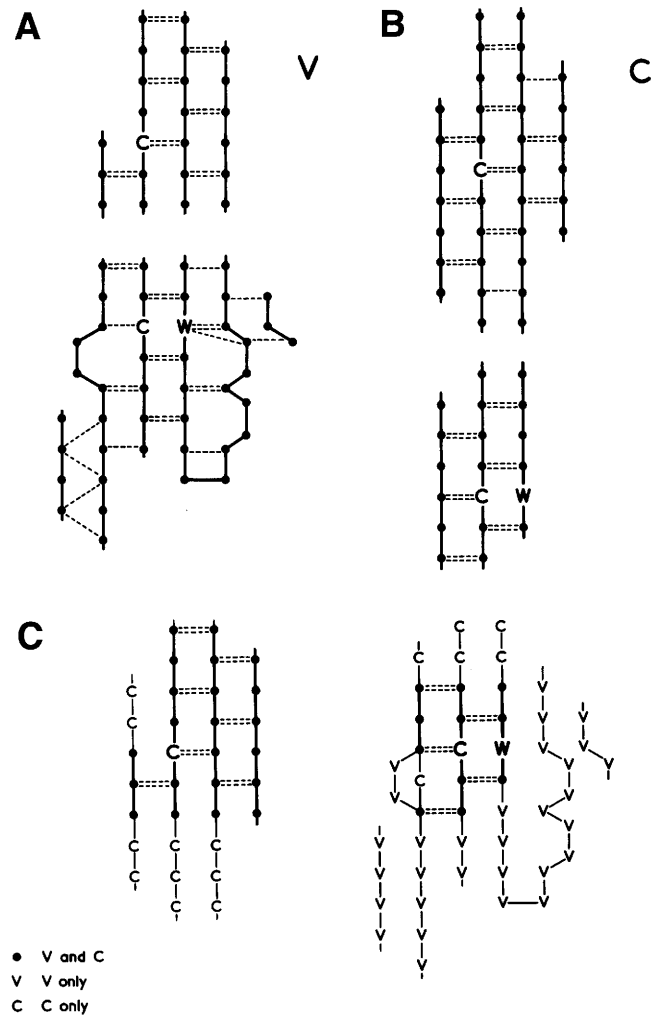


Fig. 3. The β-sheets structures in variable (V) and constant (C) domains. Small filled circles represent residues; thick lines, the polypeptide chain, and broken lines, the interchain hydrogen bonds. The sites of conserved cysteine and tryptophan residues are indicated by C and W. A shows the β-sheets structures common to all V domains and (B) shows the same for C domains. The superimposition of these structures (C) shows that only the central regions are common to both types of domains; outside this region the structures tend to be different.

domains (which may well have introns within their genes) are shuffled to create new structures.

Intronic recombination has been involved in the formation of cell adhesion molecules, cell surface receptors, proteins of the body fluids, extracellular matrix proteins and proteins involved in the structural organisation and regulation of muscle. These have been described as mosaic proteins because they are built of several discrete domains or modules that are the same type or a few different types. (Modules of the same type have homologous amino acid sequences and structures). The neural cell adhesion molecules (NCAMs) are an example of such proteins: they have different forms but all have an extracellular portion that contains seven domains (modules): five are members of the immunoglobulin superfamily and two are fibronectin type III domains (Fig. 1). Some forms also contain the small muscle specific domain (MSD).

A model for the evolution of mosaic proteins has been

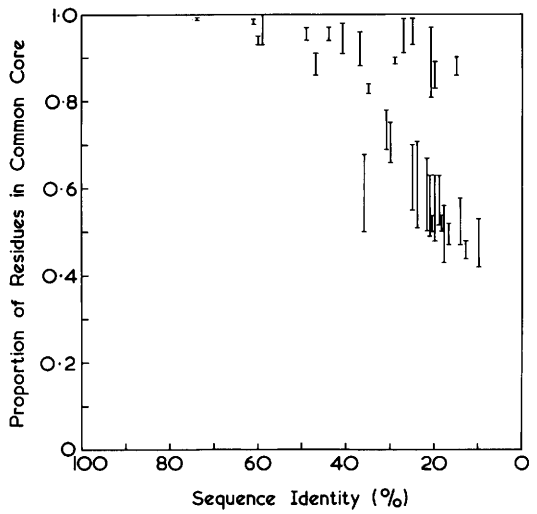


Fig. 4. For 32 pairs of homologous proteins from eight protein families, this plot shows the proportion of their structures that have the same fold, as a function of their sequence identity. If two proteins with n_1 and n_2 residues have c residues with the same fold (see Fig. 3), the proportions with the same fold are c/n_1 and c/n_2. If, of the c residues that have the same fold, b are identical, the sequence identity is 100b/c %. (This figure is adapted from Chothia and Lesk (1986) where details are given for each of the structural comparisons.)

proposed by Patthy (1991b, 1994) and is shown in Fig. 5. Starting from a gene for a cytoplasmic protein, the model proposes the following series of steps that lead to the formation of extra-cellular mosaic proteins: (i) the gain of a secretory peptide; (ii) insertions of introns at the beginning and end of the gene to form a proto-module; (iii) its duplication, and (iv) intronic recombination with other types of modules.

An intron has a "phase" that is defined by the position of the break it makes in a codon:

One implication of Patthy's model is that the introns at the

Intron phase				0		1	2	
Intron position				I		I	I	
Codons	|	T	C	A	|	G	G	A |
Residues			Ser				Gly	

beginning and end of modules must have the same phase: if they did not, duplicated genes, and other genes gained by recombination, would read out of phase. For reasons to do with the mechanisms of intron insertion and splicing, it is expected that the introns used to create mosaic proteins will be phase 1 (see Patthy, 1994).

In Fig. 6 the position and phase of the introns in the gene that codes for the extra-cellular domains of chicken NCAM (Owens et al., 1987; Prediger et al., 1988) are shown. As expected from Patthy's model, each module is separated at its exact boundary by a phase 1 intron.

Introns also occur within modules and in the NCAM gene there are one, two or three within each of the extra-cellular modules (Fig. 6). They occur at non-homologous positions and have different phases: 0, 1 or 2. The introns in domains 1 - 3 and 5 - 6 seem to be insertions that have not been useful in the evolution or function of the protein: they are probably just evolutionary "noise".

Multi-domain proteins can be formed by recombination without the help of introns: this occurs in bacteria. Inspection of the gene structures of the mosaic proteins involved in the formation of the metazoa, however, shows that all, or almost all, of those that are currently known arose through intronic recombination (Patthy, 1991b, 1994). Indeed Patthy (1994) has argued that the development of spliceosomal introns and the accumulation of a critical mass of module types made a crucial contribution to the formation of metazoa and their radiation.

Proteins with developmental isoforms

The product of a gene that is built of several exons can be modified by having alternative patterns of splicing that result in certain exons, and therefore peptides or domains, being expressed in some forms of the protein but not in others. These changes in sequence produce changes in functional properties, They produce versions of a protein that are specific to different types of cell types or to different stages of development (Smith et al., 1989; Maniatis, 1991).

Alternative patterns of splicing occur in the expression of the NCAM gene. For the extracellular domains it involves a ten residue exon bracketed by the two introns within domain 4 and the exons that form the small muscle specific domain (MSD) domain (Fig. 6). (Outside this region of the NCAM gene there are additional exons that code for a trans-membrane peptide and a cytoplasmic domain. These also have alternative splicing patterns.)

Alternative patterns of splicing are often involved in the specialisation or in modification of function. Thus the NCAM forms that express the MSD domain are specific for muscle cells (Dickson et al. 1987). It has been proposed that the expression of the peptide in domain 4 changes the function of NCAM from a molecule that promotes morphological plasticity to one that maintains stable cell-cell contacts (Doherty et al., 1992).

Expression of the additional exons do not affect the integrity of the NCAM domain structure. The extra peptide in domain

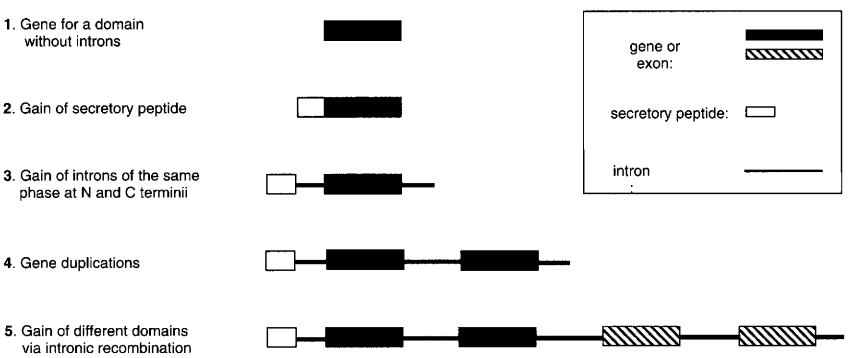
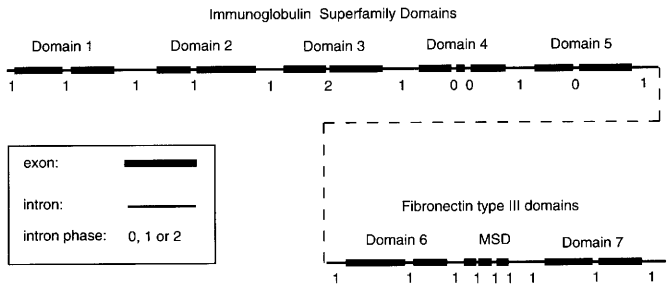


Fig. 5. The model put forward by Patthy (1991b, 1994) for the evolution of mosaic proteins (see text).



Exons and introns in the region of the gene that codes for the extra-cellular domains of neural cell adhesion molecules (NCAMs)

Fig. 6. The structure of the part of the chicken neural cell adhesion molecule (NCAM) gene that codes for the extra-cellular domains (Owens et al., 1987; Prediger et al., 1988). Note that the exon and intron regions are NOT drawn to scale. This region of the gene codes for five immunoglobulin superfamily modules or domains (numbered 1 to 5) and two type III fibronectin domains (6 and 7). Forms of the protein found in muscle also contain the small muscle specific domain MSD domain (see text).

4 is inserted in a peptide loop on the surface of the protein. The MSD domain is inserted between the two fibronectin domains (Harpaz and Chothia, unpublished data).

CONCLUSION: PROTEIN STRUCTURES AND THE EVOLUTION OF DEVELOPMENT

The evolution of development has involved the development of new proteins. The evidence reviewed here points to most current proteins being the descendants of no more than 1000 ancestors. The process by which these descendants were produced involved gene duplications followed by mutations and, for large proteins, gene fusion. In metazoa, gene fusion has largely occurred through intronic recombination of the genes for whole protein domains.

The relatively small number of protein families in biology does not arise from any intrinsic physical or chemical limitation on number of protein folds but from history. In the earliest stages of evolution, there seems to have developed a set of proteins with a range of functional properties that was sufficiently wide for it to be much easier to evolve new proteins by the duplication, modification and recombination of old proteins than by the ab initio invention of new ones.

The view of protein evolution presented here suggests a programme for understanding the molecular basis of the evolution of development. This involves: (1) the complete description of the sequences that form organisms at different levels of developmental complexity; (2) the classification of the sequences into families, and (3) the determination of the evolutionary history of the appearance of new properties within protein families.

The dates proposed for the completion of the most advanced of the present genome sequencing projects mean that within a few years we will have a complete description of the sequences that form organisms at several different levels of developmental complexity:

Genome project	Number of genes	Estimated date of completion
Bacteria: *Bacillus subtilus*	4,000	1997-8
Yeast: *Saccharomyces cerevisae*	6,000	1996
Nematode: *Caenorhabditis elegans*	18,000	1998
Humans	65,000	2005 (?)

A detailed knowledge of protein structure will play a central role in the interpretation of this sequence data. The relationship of strongly diverged proteins can not be recognised from sequence similarities at present but it can be seen, in many cases, from their three dimensional structures. Structural descriptions also make it easier to understand how evolutionary changes have modified the properties of proteins and produced new ones.

At the moment more than 270 protein structures are being determined each year by X-ray crystallography and NMR, and the rate is increasing (Hendrickson and Wüthrich, 1993). A significant proportion of these structures are homologous to those known from previous work. The current rate at which the structures for new families are becoming known, however, means the combination of crystallography, NMR and molecular modelling will produce, at least in outline, structures for the large majority protein families within a few years.

Thus, the combined information from the genome projects and protein structures should allow the determination of the evolutionary history of protein families and hence a description of the molecular events that have produced the evolution of development.

I thank Samantha Barré for comments on the paper.

REFERENCES

Adams, M. D., Dubnick, M., Kerlavage, A. R., Moreno, R., Kelley, J. M., Utterback, T. R., Nagle, J. W., Fields, C. and Venter, J. C. (1992). Sequence identification of 2,375 human brain genes. *Nature* **355,** 632-634.

Adams, M. D., Kerlavage, A. R., Fields, C. and Venter, J. C. (1993). 3,400 new expressed sequence identify diversity of transcripts in human brain. *Nature Genetics* **4,** 256-267.

Blake, C. C. F. (1978). Do genes-in-pieces imply proteins in pieces? *Nature* **273,** 267.

Brew, K., Vanaman, T. C. and Hill, R. L. (1967). Comparison of the amino-acid sequence of bovine α-lactalbumin and hen's egg white lysozyme. *J. Biol. Chem.* **242,** 3747-3749.

Chothia, C. (1992). One thousand families for the molecular biologist. *Nature* **357,** 543-544.

Chothia, C. and Lesk, A. M. (1986). The relation between the divergence of sequence and structure in proteins. *EMBO J.* **5,** 823-826.

Chothia, C. and Lesk, A. M. (1987) The evolution of protein structures. *Cold Spring Harbor Quant. Biol.* Vol. **LII,** 399-405.

Dickson, G., Gower, H. J., Barton, C. H., Prentice, H. M., Elsom, V. L., Moore, S. E., Cox, R. D., Quinn, C., Putt, W. and Walsh, F. S. (1987). Human muscle neural cell adhesion molecule (N-CAM): identification of a muscle-specific sequence in the extracellular domain, *Cell* **50,** 1119-1130.

Doherty, P., Moolenaar, C. E. C. K., Ashton, S. V., Michalides, R. J. A. M. and Walsh F. S. (1992). The VASE exon down regulates the neurite growth promoting activity of NCAM 140. *Nature* **356,** 791-793.

Dujon, B. Alexandaki, D., Audré, B. and 103 others. (1994). Complete DNA sequence of yeast chromosome XI. *Nature* **369,** 371-378.

Fields, C., Adams, M. D., White, O. and Venter, J. C. (1994). How many genes in the human genome? *Nature Genetics* **7,** 345-346.

Flaherty, K. M., McKay, D. B., Kabsch, W. and Holmes K. C. (1991). Similarity of the three-dimensional structures of actin and the ATPase fragment of a 70-kDa heat shock cognate protein. *Proc. Natl. Acad. Sci. USA* **88,** 5041-5045.

Gilbert, W. (1978). Why genes in pieces? *Nature* **272,** 501.

Glaser, P., Kunst, F., Arnaud, M., Coudart, M.-P., Gonzales, W., Hullo, M.-F., Ionescu, M., Lubochinsky, B., Marcelino, L., Moszer, I., Presecan, E., Santana, M., Schneider, E., Schweizer, J., Vertes, A., Rapoport, G. and Danchin, A. (1993). Bacillus subtilis genome project: cloning and sequencing of the 97 kb region from 325° to 333°. *Mol. Microbiol.* **10,** 371-384.

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.

Hartley, B. S., Brown, J. R., Kauffman, D. L. and Smillie, L. B. (1965). Evolutionary similarities between proteolytic enzymes. *Nature* **207,** 1157-1159.

Hendrickson, W. A. and Wüthrich, K. (1993). In *Macromolecular Structures 1993.* London: Current Biology Ltd.

Ingram, V. (1961). Gene evolution and the haemoglobins. *Nature* **189,** 704-708.

Jones, E. Y., Davis, S. J., Williams, A. F., Harlos, K. and Stuart, D. I. (1992). Crystal structure at 2.8 resolution of a soluable form of the cell adhesion molecule CD2. *Nature* **369,** 232-239.

Juy, M., Amit, A. G., Alzari, P. M., Poljak, R. J., Claeyssens, M., Beguin, P. and Aubert, J.-P. (1992). Three dimensional structure of a thermostable bacterial cellulase. *Nature* **357,** 89-91.

Klein, C. and Schulz, G. E. (1994). Structure of cyclodextrin glycosyltransferase refined at 2 Å resolution. *J. Mol. Biol.* **217,** 737-750.

Koonin, E. V., Bork, P. and Sander, C. (1994). Yeast chromosome III: new gene functions. *EMBO J.* **13,** 493-503.

Lesk, A. M. and Chothia, C. (1980). How different amino acid sequences determine similar protein structures: the structure and evolutionary dynamics of the globins. *J. Mol. Biol.* **136,** 225-270.

Maniatis, T. (1991). Mechanisms of alternative pre-mRNA splicing. *Science* **251,** 33-34.

McCombie, W. R., Adams, M. D., Kelley, J. M., FitzGerald, M. G., Utterback, T. R., Khan, M., Dubnick, M., Kerlavage, A. R., Venter, J. C. and Fields, C. (1992). Caenorhabditis elegans expressed sequence tags identify gene families and potential disease gene homologues. *Nature Genetics* **1,** 124-131.

Murzin, A and Chothia, C. (1992). Protein architecture: new superfamilies. *Curr. Opinion Struct. Biol.* **2,** 895-903.

Murzin, A. (1993). Can homologous proteins evolve different enzymatic activities? *Trends Biochem. Sci.* **18,** 403-405.

Oliver, S. G. van der Aart, Q. J. M., Agostoni-Carbone, M. L. and 144 others. (1992). The complete DNA sequence of yeast chromosome III. *Nature* **357,** 38-46.

Owens, G. S., Edelman, G. M. and Cunningham, B. A. (1987). Organisation of the neural cell adhesion molecule (N-CAM) gene: alternative exon usage as the basis for different membrane associated domains. *Proc Natl. Acad. Sci. USA* **84,** 294-298.

Pascarella, S. and Argo, P. (1992). A data bank merging related protein structures and sequences. *Protein. Eng.* **5,** 121-137.

Patthy, L. (1991a). Exons - original building blocks of proteins? *BioEssay* **13,** 187-192.

Patthy, L. (1991b). Modular exchange principles in proteins. *Curr. Opinion Struct. Biol.* **1,** 351-361.

Patthy, L. (1994). Introns and Exons. *Curr. Opinion Struct. Biol.* **4,** 383-392.

Prediger, E. A., Hoffman, S., Edelman, G. M. and Cunningham, B. A. (1988). Four exons encode a 93-base-pair insert in three neural cell adhesion molecule mRNAs specific for chicken heart and skeletal muscle. *Proc. Natl. Acad. Sci. USA* **85,** 9610-9620.

Sander, C. and Schneider, R. (1991). Database of homology-derived protein structures and structural meaning of sequence alignment. *Proteins* **9,** 56-68.

Smith, C. W. J., Patton, J. G. and Nadal-Ginard, B. (1989). Alternative splicing in the control of gene expression. *Annu. Rev, Genet.* **23,** 527-577.

Sulston, J., Du, Z., Thomas, K., Wilson, R., Hillier, L., Staden, R., Halloran, N., Green, P., Thierry-Mieg, J., Qiu, L., Dear, S., Coulson, A., Craxton, M., Durbin, R., Berks, M., Metzstein, M., Hawkins, T., Ainscough, R. and Waterson, R. (1992). The C. elegans genome sequencing project: a beginning. *Nature* **356,** 37-41.

Walsh, K. A. and Neurath, H. (1964). Trypsinogen and chymotrypsinogen as homologous proteins. *Proc. Natl. Acad. Sci. USA* **52,** 884-888.

Waterson, R., Martin, C., Craxton, M., Huynh, C., Coulson, A., Hillier, L., Durbin, R., Green, P., Showkeen, R., Halloran, N., Metzstein, M., Hawkins, T., Wilson, R., Berks, M., Du, Z., Thomas, K., Thierry-Mieg, J. and Sulston, J (1992). A survey of expressed genes in *Caenorhabditis elegans. Nature Genetics* **1,** 114-123.

Williams, A. F. and Barclay, A. N. (1988) The immunoglobulin superfamily - domains for surface recognition, *Ann. Rev. Immunol.* **6,** 381-405.

Wilson, R., Ainscough, R., Anderson, K. and 52 others. (1994). 2.2 Mb of contiguous nucleotide sequence from chromosome III of C. elegans. *Nature* **368,** 32-38.