18 Kooman-Gershman, M., Honée G., Bonnema, G. and De Wit, P. J. G. M. (1996) *Plant Cell* 8, 929–938

19 Dangl, J. L. (1994) *Curr. Top. Microbiol. Immunol.* 192, 99–118

20 Alfano, J. R. and Collmer, A. (1996) *Plant Cell* 8, 1683–1698

21 Cornelius, G. R. and Wolf-Watz, H. (1997) *Mol. Microbiol.* 23, 861–867

22 Grant, M. R. *et al.* (1995) *Science* 269, 843–846

23 Gopalan, S. *et al.* (1996) *Plant Cell* 8, 1095–1105

24 Leister, R. T., Ausubel, F. M. and Katagiri, F. (1996) *Proc. Natl. Acad. Sci. U. S. A.* 93, 15497–15502

25 Tang, X. *et al.* (1996) *Science* 274, 2060–2063

26 Scofield, S. R. *et al.* (1996) *Science* 274, 2063–2065

27 Kearney, B. and Staskawicz, B. J. (1990) *Nature* 346, 385–386

28 Ritter, C. and Dangl, J. L. (1995) *Mol. Plant-Microbe Interact.* 8, 444–453

29 Rohe, M. *et al.* (1995) *EMBO J.* 14, 4168–4177

30 Herbers, K., Conrads-Strauch, J. and Bonas, U. (1992) *Nature* 356, 172–174

31 Yang, Y. and Gabriel, D. W. (1995) *Mol. Plant-Microbe Interact.* 8, 627–631

32 Van Den Ackerveken, G., Marois, E. and Bonas, U. (1996) *Cell* 87, 1307–1316

33 Lawrence, G. J., Finnegan, E. J., Ayliffe, M. A. and Ellis, J. G. (1995) *Plant Cell* 7, 1195–1206

34 Song, W-Y. *et al.* (1995) *Science* 270, 1804–1806

35 Zhou, J., Loh, Y-T., Bressan, R. A. and Martin, G. B. (1995) *Cell* 83, 925–935

36 Zhou, J., Tang, X. and Martin, G. B. (1997) *EMBO J.* 16, 3207–3218

37 Lemaitre, B. *et al.* (1996) *Cell* 86, 973–983

38 Parker, J. E. *et al.* (1996) *Plant Cell* 8, 2033–2046

39 Century, K., Holub, E. B. and Staskawicz, B. J. (1995) *Proc. Natl. Acad. Sci. U. S. A.* 92, 6597–6601

40 Cao, H., Bowling, S. A., Gordon, S. and Dong, X. (1994) *Plant Cell* 6, 1583–1592

41 Delaney, T. P., Friedrich, L. and Ryals, J. A. (1995) *Proc. Natl. Acad. Sci. U. S. A.* 92, 6602–6606

42 Cao, H. *et al.* (1997) *Cell* 88, 57–63

43 Ryals, J. *et al.* (1997) *Plant Cell* 9, 425–439

44 Schulze-Lefert, P., Peterhaensel, C. and Freialdenhoven, A. (1997) in *The Gene for Gene Relationship in Host-Parasite Interactions* (Crute, I., Burdon, J. and Holub, E., eds), pp. 45–63, CAB International

45 Dixon, M. S. *et al.* (1996) *Cell* 84, 451–459

46 Cai, D. *et al.* (1997) *Science* 275, 832–834

47 Bent, A. F. *et al.* (1994) *Science* 265, 1856–1860

48 Mindrinos, M., Katagiri, F., Yu, G-L. and Ausubel, F. M. (1994) *Cell* 78, 1089–1099

49 Ori, N. *et al.* (1997) *Plant Cell* 9, 521–532

Cytoplasmic signalling domains: the next generation

Peer Bork, Jörg Schultz and Chris P. Ponting

Since the late 1980s, when Src-homology SH2 and SH3 domains were identified, the repertoire of non-catalytic signalling domains has increased to number over 30. As it is expected that further regulatory domains shall be found, unravelling the complex network of their interactions remains an on-going challenge.

THE EARLY YEARS of protein crystallography revealed that the fundamental unit of structure is the domain, which is a region of a polypeptide chain that is folded into a spatially distinct structural unit[1]. Later, during the explosion in protein sequence determination that occurred from the late 1970s onwards, sequence-similar domains were found frequently as units within several otherwise unrelated proteins. Such domains are termed modules and may be considered as the 'raw materials' used in the gradual acquisition of function by multidomain and multifunctional proteins during their evolution (see Ref. 2 and references therein). Modules that are similar in topology and in sequence are likely to have diverged from a single common ancestor and an 'exon shuffling' mechanism for this propagation was proposed by Gilbert in 1978 (Ref. 3). The early 1980s provided strong evidence for this proposition following the identification of numerous extracellular proteins containing arrays of modules. More than 60 extracellular modules have been named and classified[4]. Most of these three-dimensional folds are known[2].

P. Bork and **J. Schultz** are at the EMBL, Heidelberg, Germany, and at the Max-Delbrück-Center for Molecular Medicine, Berlin-Buch, Germany; **C. Ponting** is at the Oxford Centre for Molecular Sciences, University of Oxford, Fibrinolysis Research Unit, South Parks Road, Oxford, UK OX1 3RH.
Email: bork@embl-heidelberg.de
ponting@bioch.ox.ac.uk

By the end of the 1980s, powerful sequencing and analytical tools led to the identification in protein kinases of the first cytoplasmic regulatory domains. Later, it became clear that these Src-homology 2 and 3 (SH2, SH3) domains, and the subsequently identified pleckstrin-homology (PH) domain, are present in dozens of diverse proteins, and signalling via these domains became a major focus of research (reviewed in Refs 5, 6). This first generation of signalling domains was joined, in the first half of the 1990s, by a second that included WW, C1, C2, PDZ, PTB/PI, DEATH and WD40 domains each of whose three-dimensional structures and ligand-binding characteristics have recently been determined (see Table in Poster for details and for additional references).

Since 1995, over 20 new domains in signalling proteins have been identified by sequence analysis. This third generation includes the Ras-association (RA) domain, a probable ubiquitin-binding domain (UBA) and SAM, a domain that appears to form homotypic and heterotypic dimers. The ligands of the remaining domains and the tertiary structures of all third generation domains remain to be determined.

A definition of cytoplasmic signalling domains

Here, we discuss the more than 30 signalling domains that are known, and we summarise their structural and functional features. Because it is impossible to distinguish cytoplasmic signalling modules from those that primarily function in transport, protein sorting or cell-cycle regulation, we will discuss only those domains that occur in at least two proteins with different domain organisations, of which one also contains a domain with a kinase, phosphatase, cyclase, ubiquitin ligase or phospholipase enzymatic activity, or with a GTPase-activation or guanine nucleotide exchange activity. These activities are known to mediate transduction of an extracellular signal towards the nucleus in order to initiate a cellular response to the signal; an example of a signal transduction network is shown in the Poster that accompanies this article.

We are aware that our set of signalling domains is incomplete, because it lacks, for example, Ig domains and ARM, ankyrin, spectrin and tetratricopeptide repeats, that occur in several molecules

Copyright © 1997, Elsevier Science Ltd. All rights reserved. 0968–0004/97/$17.00 PII: S0968-0004(97)01084-0

known to regulate signal transduction pathways. It does, however, include FHA, LIM, PDZ, RanBD, SAM, UBA and WD40 domains, all of which have been found in both cytoplasmic and nuclear proteins, and C2 and PDZ domains that have been found in both cytoplasmic and secreted proteins (in perforin and interleukin 16, respectively). A few of these domains, such as SH3 (in yeast PAS20), PDZ (in densin-180), SPRY (in ryanodine receptors), SAM (eph-like tyrosine kinases) and IQ (in sodium channels), are found within cytoplasmic portions of transmembrane proteins.

Specificity vs. functional redundancy

Whereas it is widely accepted that most extracellular modules can perform different functions in different settings and make use of different surface regions for their interactions, cytoplasmic domains are still often believed to mediate similar functions using similar ligand-binding modes. For example, SH2, SH3 and PH domains bind phosphotyrosine-containing and specific proline-containing motifs, and phosphoinositides, respectively, and target specificities are provided by subtle variations within their single binding sites. Recent data suggest, however, that representatives of a signalling domain family may mediate different functions: for example, PH domains bind phosphoinositides, protein kinase Cs and $G_{\beta\gamma}$ subunits[7]; C1 domains bind diacylglycerol and Ras^{GTP} (Ref. 8); and C2 domains bind calcium, phospholipids and intracellular proteins[9]. It is also apparent that domains from different domain families may possess similar molecular functions: for example, SH3 and WW domains compete to bind a specific proline-rich motif in formins[10]. It should be noted that although intermolecular protein–protein interactions are the norm in signalling, the intramolecular interactions of SH2 and SH3 domains with extensions to the tyrosine kinase Src may provide a paradigm of kinase autoinhibition and other mechanisms[11].

Structural features

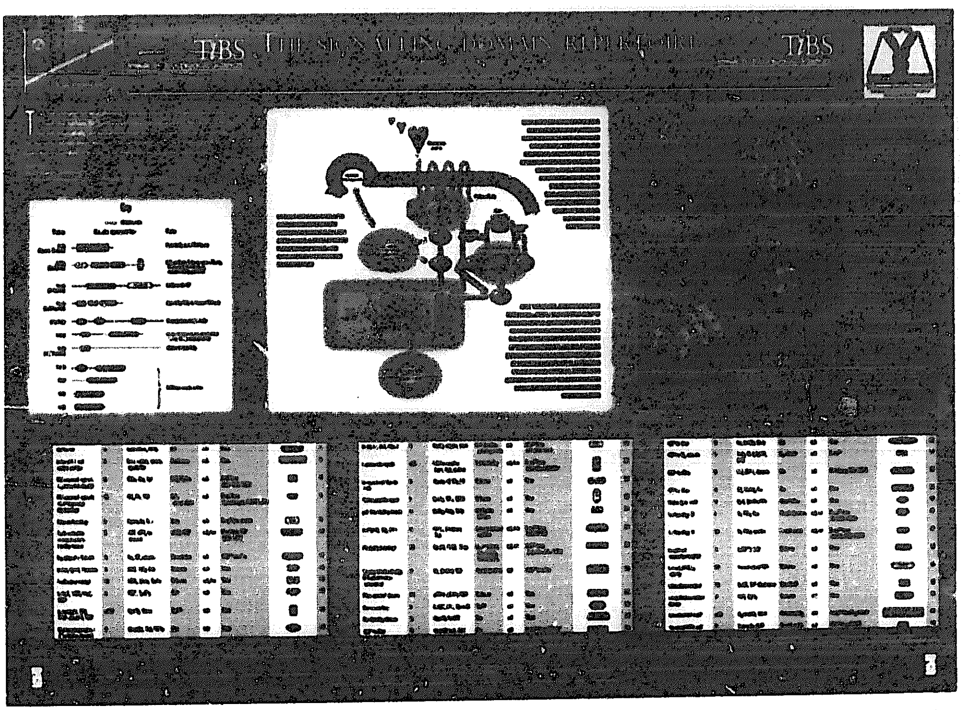
Signalling regulatory modules vary in secondary structure, fold and in size. Domain sizes vary from 30–300 residues. The smallest, OPR, is likely to compensate for its lack of a hydrophobic core by folding around Ca^{2+} ion(s), in a similar manner to C1 and LIM domains, which fold around Zn^{2+}. The largest signalling domains may yet be found to be enzymes or else to be constructed from several modules. In addition, larger domains may be composed of degenerate copies of a motif that in isolation is structurally unstable yet, as multiple repeats, represents a stable and functional unit. For example, seven-bladed β-propellers such as those that occur in $G_{\beta\gamma}$ proteins[12] contain seven repeated 'WD40 motifs' that form a globular domain. EF-hands and tetratricopeptide repeats also appear to require at least two or four copies, respectively, in order to form stable structures. Leucine-rich repeats (LRRs) that occur in numerous proteins including cyclases (see Poster) are successive α/β units that assemble as a horseshoe-like superstructure formed either by two or more proteins, each with low copy numbers of LRRs or by a large number of tandem repeats within a single protein[13].

The majority of cytoplasmic signalling modules occur as single units or else as tandem arrays. Examples of tandem repeats include two PTB domains in FE65, three PH domains in myosin X, four WW domains in the ubiquitin ligase Nedd-4, five LIMs in human pinch and even nine PDZ domains in a *Caenorhabditis elegans* open-reading frame (C52a11.4). Each tandem repeat is likely to have arisen from an internal gene duplication event via recombination. By contrast, duplicated regions of genes that encode single or multiple domains may be inserted elsewhere in the genome, thereby generating either pairs of closely-related paralogues or else modular proteins with correlated domain arrangements.

Although many closely-related paralogues exist in eukaryotic genomes, there are only a few clear examples where the arrangement of contiguous domains in more distantly-related homologues appears to be preserved: for example, all known Dbl-homologous domains are followed by PH domains. By contrast, although 66 eukaryotic proteins containing PDZ domains are known, including several with C2, LIM, PH, PTB, PX, SAM and SH3 domains, no PDZ domain has yet been identified in proteins containing either SH2 or C1 domains.

One surprising observation is that several of these domains are not uniformly positioned within polypeptide chains. Our analyses suggest that for most signalling domains, there is little preference for domains to appear in the amino-terminal, middle or carboxy-terminal thirds of proteins. However, over 60% of C1, CH, FHA or HR1 domains are present in the amino-terminal third and over 80% of protein tyrosine phosphatase or DEATH domains appear in the carboxy-terminal third. In addition, whereas there is little preference for a favoured position of Ser/Thr kinase domains within the polypeptide chain, the majority (87%) of tyrosine kinase domains are found in the carboxy-terminal third of polypeptides.

The propagation of a domain type need not occur via an insertion within the interdomain linker because entire domains



are known to be inserted into intradomain surface loops[14]. PH domains support an SH2–SH2–SH3 tridomain unit, a C1 domain and a PDZ domain, as insertions in PLCγ, in the Ser/Thr protein kinase ROCK and in syntrophins, respectively. A PH domain itself is inserted into a band 4.1-homologous domain in Mig2. Many more such examples are likely to remain undetected given that their identification by sequence analytical methods is fraught with difficulties.

Evolution

Organisms as diverse as bacteria, yeast, plants and mammals require signal transduction processes to respond to constantly-changing environments. Although all of these organisms appear to contain both Ser/Thr kinases and His kinases, bacterial signal transduction is mediated primarily by two-component His kinase sensor and response regulator systems, whereas eukaryotic signalling appears to be dominated by phosphorylation on serine or threonine.

Phosphotyrosine-mediated signalling via 'true' protein tyrosine kinases (those most similar in sequence to Src, for example) appears to be a relatively late development in eukaryotic evolution, as the *Saccharomyces cerevisiae* genome appears to lack both 'true' tyrosine kinases[15] and cytoplasmic phosphotyrosine-binding domains, such as SH2 or PTB/PI domains (only a single, divergent *S. cerevisiae* SH2 domain has been identified in the nuclear protein SPT6); current data suggest that plants also lack 'true' tyrosine kinase, SH2 and PTB/PI domains. However, the *S. cerevisiae* genome does contain several so-called 'dedicated protein tyrosine kinases' that are highly similar in sequence to Ser/Thr kinases, yet possess high specificities for tyrosine-containing polypeptides[15]. In addition, three protein tyrosine phosphatase homologues can be found in yeast.

Bacteria appear to possess many homologues of eukaryotic signalling enzymes. Kinases and phosphatases specific for Ser/Thr and/or Tyr, cyclases, and phospholipases C and D homologues have all been found in bacteria. However, of the regulatory signalling domains discussed here, only PDZ, FHA, LRR and cyclic nucleotide monophosphate binding domains have been identified in bacteria (see Poster for details and abbreviations). It is possible either that the remaining signalling domains are absent in bacteria, or else that sequence similarities between eukaryotic and bacterial domain homologues are currently undetectable. That the latter might be the case is supported by an intriguing finding[16] that *Escherichia coli* BirA contains two domains that show considerable structural similarity to an SH2–SH3 tandem domain pair. Thus, many eukaryotic signalling domains may have evolved from ancient progenitors.

Archaeal genomes are known to contain homologues of eukaryotic Ser/Thr phosphatases[17] and bacterial His kinases[18]. However, an analysis of the *Methanococcus jannaschii* genome reveals that, except for two protein kinase homologues, this organism contains no obvious homologues of the signalling domains discussed here. Viral genomes have been shown to contain homologues of eukaryotic SH2, SH3, PH, kinase and phospholipase D domains.

Compared with the information currently available for bacterial genomes, eukaryotic genomic data remain largely incomplete, and even in the only eukaryotic genome completed, signalling domain-containing proteins currently represent only 5% of all *S. cerevisiae* proteins. However, current information is sufficient to show that, in contrast to extracellular proteins, intron borders and phases are not well-defined in intracellular modular proteins. This, together with the frequent occurrence of signalling domains in lower eukaryotes, such as yeast (see Poster) that possess only very few introns, leads to the conclusion that there is little evidence that 'exon shuffling' processes[19,20] participated in the spread of intracellular signalling domains. It is also clear that the relatively ancient cytoplasmic signalling proteins, as exemplified in yeast, are less modular than the more modern extracellular 'communication' and receptor proteins of multicellular organisms, and possess a greater proportion of sequence that does not fold to compact domains.

Conclusions

Although sequence databases are currently expanding exponentially, novel domain families are being identified only at an approximately linear rate.Improvements in sequence analytical techniques have allowed identification of domain families that are considerably more divergent than in the past, yet the linear rate of discovery is likely to continue until the point at which fully-automated domain detection procedures are developed. Those domain families that are currently known already present a considerable challenge, and determinations of their structures and functions are likely to be forthcoming in the near future. The combination of bioinformatics, structure determination, and molecular and cellular biology, will continue to reveal the complex nature of the regulatory functions of signalling domains, the interaction of signalling proteins in multimolecular complexes and the networks of interacting pathways that transduce extracellular signals to the nucleus.

Acknowledgements

C. P. P. is a member of the Oxford Centre for Molecular Sciences and is a Wellcome Trust Research Fellow. J. S. is supported by a BMBF grant.

References

1 Janin, J. and Chothia, C. (1985) *Methods Enzymol.* 115, 420–430

2 Bork, P., Downing, A. K., Kieffer, B. and Campbell, I. D. (1996) *Quart. Rev. Biophys.* 29, 119–167

3 Gilbert, W. (1978) *Nature* 271, 501

4 Bork, P. and Bairoch, A. (1995) *Trends Biochem. Sci.* 20, Poster 02 (Suppl.)

5 Cohen, G. B., Ren, R. and Baltimore, D. (1995) *Cell* 80, 237–248

6 Pawson, T. (1995) *Nature* 373, 573–579

7 Shaw, G. (1996) *BioEssays* 18, 35–46

8 Brtva, T. R. *et al.* (1995) *J. Biol. Chem.* 270, 9809–9812

9 Nalefski, E. A. and Falke, J. J. (1997) *Protein Sci.* 5, 2375–2390

10 Chan, D. C., Bedford, M. T. and Leder, P. (1996) *EMBO J.* 15, 1045–1054

11 Xu, W., Harrison, S. C. and Eck, M. J. (1997) *Nature* 385, 595–602

12 Sondek, J. *et al.* (1996) *Nature* 379, 369–374

13 Kobe, B. and Deisenhofer, J. (1995) *Curr. Opin. Struct. Biol.* 5, 409–416

14 Russell, R. B. (1994) *Protein Eng.* 7, 1407–1410

15 Hunter, T. and Plowman, G. D. (1997) *Trends Biochem. Sci.* 22, 18–22

16 Russell, R. B. and Barton, G. J. (1993) *Nature* 364, 765

17 Leng, J., Cameron, A. J. M., Buckel, S. and Kennelly, P. J. (1995) *J. Bacteriol.* 177, 6510–6517

18 Rudolph, J. and Oesterhelt, D. (1995) *EMBO J.* 14, 667–673

19 Patthy, L. (1996) *Matrix Biol.* 15, 301–310

20 Bork, P. (1996) *Matrix Biol.* 15, 311–313