The rsp5-domain is shared by proteins of diverse functions

Kay Hofmann, Philipp Bucher

Swiss Institute for Experimental Cancer Research, CH-1066 Epalinges s/Lausanne, Switzerland

Received 12 December 1994

Abstract A novel, unusually small, and highly conserved domain of modular intracellular proteins is described. The domain was first recognized as three repeats in the yeast rsp5 gene product and named thereafter. The rsp5 protein is thought to interact with nuclear proteins but also contains a C2 domain typical for cytoplasmic proteins. Further analyses revealed several additional occurrences of this domain in diverse protein classes, including cytoplasmic signal transduction proteins, gene products interacting with the transcription machinery, structural proteins like dystrophin, and a putative RNA helicase.

Key words: Sequence analysis; Signal transduction; Dystrophin; C2-domain; yap65; WW-domain

1. Introduction

Intracellular proteins involved in regulatory processes often contain several independent building blocks of distinct functionality besides their proper catalytic domain. These building blocks, the most widespread examples being the Src-homolgy domains SH2 and SH3 [1,2], the PH-domain [3,4] and the C2-domain [5], are usually present in a large number of otherwise unrelated sequences, frequently in multiple copies per protein. Since all these domains function in diverse sequence contexts, they are believed to form autonomous folding units, independent of specific residue contacts with the other parts of the host protein. Most of these domains have been detected in proteins from all major groups of eukaryotic organisms. They have probably been propagated by exon shuffling, which appears to be the most widely used way for an existing protein to acquire additional functionality. Most of these domains appear to mediate specific interactions with other proteins or cellular structures. The SH2-domain has been shown to bind to phosphorylated tyrosine residues [6], the SH3-domain binds specifically to certain proline-rich protein regions [2], and for the C2-domain there is emerging evidence for its role in mediating Ca^{2+}-regulated membrane attachment of the host protein [7]. Very recently, binding of the PH-domain to phosphatidylinositol-4,5-bisphosphate has been reported [8], an observation that suggests also for the PH-domain a role in protein-membrane interaction.

During a study on the distribution of C2-domains, a putative copy of this domain was detected in the yeast rsp5 gene product (genpept entry 172848), a protein of unassigned function. Since alleles of rsp5 have been reported to complement mutations in the yeast transcription factor SPT3 [9], it is likely that the rsp5 gene product interacts with this transcription factor in vivo. The most striking feature of the rsp5 protein sequence, however, is the occurrence of three copies of a 35 residue domain with a high content of aromatic and charged amino acids. Four remarkably well conserved copies of this repeat sequence are also present in a human ORF termed kiaa93 (genpept 577313), encoding an uncharacterized protein, which also contains a C2-like domain at its N-terminus and shares further regions of sequence similarity with rsp5. Starting from these seven repeat sequences, we employed the sequence profile method [10,11] to screen sequence databases for further occurrences of this domain. Members of the rsp5 domain family were found to be present in a variety of non-homologous proteins, comprising putative transcriptional regulators, proteins thought to be involved in signal transduction as well as structural proteins like dystrophin.

2. Materials and methods

Protein databases searched were SwissProt (Release 30) [12] and genpept (Release 85 + updates, November 94) [13]. Database searches with query sequences were performed using the BLAST program [14]. Profiles were constructed from aligned sequences as described in [10], using the BLOSUM45 amino acid similarity matrix [15], a gap penalty of 2.1 and a gap-extension penalty of 0.2. The profiles were used for protein database searches, using the methods described in [16]. Statistical significance of sequence-to-profile matches was assessed by regional shuffling [17] of the sequences, using a window size of 20 residues. Only scores of at least seven standard deviations above the average shuffled score were considered significant. In the course of the analysis, the profiles were iteratively refined by including significant matches into the profile-building process. For secondary structure prediction, the PHD mailserver at the EMBL [18] was used. The profiles and programs used are available on request, they can also be accessed electronically by WWW, using the resource locator 'http://ulrec3.unil.ch/'.

3. Results

The initial search of the protein sequence databases with a profile, constructed from the seven different rsp5-repeat sequences of rsp5 and kiaa93, detected several proteins with highly significant profile-to-sequence matching scores. Among those sequences were all known forms of dystrophin, giving scores of more than 10 standard deviations above an average random score obtained from matching the profile to regionally shuffled dystrophin sequences. Consecutively, highly significant matches were considered to be members of the rsp5 domain family and used for the construction of refined profiles. This iterative profile refinement was continued until no additional sequences were found.

At the end of the refinement process, 30 domains occurring in 19 sequences were found, their alignment is shown in Fig. 1. Excluding obviously orthologous sequences, there remain 25 independent domains in 14 independent sequences. The majority of these sequences encode proteins of either unknown or

*Corresponding author. Fax: (41) (21) 652 6933.
E-mail: khofmann@isrec-sun1.unil.ch

SSDI 0014-5793(94)01415-9

poorly characterized function, dystrophin being a remarkable exception. However, in some cases sequence analysis of the rsp5-domain containing proteins and their domain structure (shown in Fig. 2) gave hints towards a tentative functional assignment of the protein. The following paragraph summarizes the known properties and the results of the sequence analysis of the proteins containing rsp5 domains:

The yeast rsp5 gene product (genpept 172848), briefly described above, and three more proteins form a subgroup of rsp5-domain containing sequences, characterized by the presence of an additional common domain. The N-terminus of rsp5 harbors a C2-like domain, which might be indicative of a regulated membrane attachment of the protein. The middle part of the protein contains three copies of the rsp5 domain, while the C-terminus contains a region of about 350 residues that show considerable similarity to the rat transcriptional activator URE-B1 [19]. This region, here referred to as 'URE-B1 domain' is observed in four of the proteins containing rsp5-domains and in at least eight other proteins summarized in Table 1, many of which have been either reported to be DNA-binding or participating in developmental processes. The human ORF termed kiaa93 (genpept 577313) shows a similar overall organization like the yeast rsp5 protein, it contains one C2-domain, four rsp5-domains and a C-terminal URE-B1 homology region.

It might be a homolog to rsp5 although the rsp5-domains seem to be shuffled and their flanking regions show no similarity to the respective regions in rsp5. The murine nedd4 protein (genpept 220509) is a developmentally regulated protein of unknown function [20]. Its structure resembles rsp5 and kiaa93, it contains three rsp5-domains but lacks the C2-domain, which might be contained in the not determined N-terminal part of the sequence. The *C. elegans* ORF termed f45h7.6 (genpept 577760) [21] is the fourth protein containing both rsp5 and URE-B1 homology domains. It contains two rsp5-domains but does not seem to be closely related to any of the above proteins.

A second group of sequences contains features that are typical for regulatory proteins participating in the signal transduction pathways. The yes-associated protein yap65 (genpept 517177, 517179, 434018) has been reported to bind to the SH3 domain of the tyrosine kinase yes [22]. The human and chicken sequence contain a single rsp5-domain, the reported murine sequence has a second rsp5-domain inserted into the otherwise strongly conserved sequence C-terminal of the first rsp5-domain. The SH3 binding site has been mapped to a position in the vicinity of the rsp5-domains. A chimaerin-like ORF from *C. elegans* (genpept 559419) [21] contains a rsp5-domain in the N-terminal region, a PH-domain and an extended region with similarity to chimaerin [23]. Chimaerin is a GTPase activating protein related to the breakpoint cluster protein bcr and other proteins interacting with small G-proteins.

A third group of sequences shows features typical for structural proteins. Vertebrate dystrophin (genpept 181857) [24] is an extensively studied protein of the muscle (see [25,26] and references therein), mutations of dystrophin cause Duchenne/Becker type muscle dystrophy. It is a rod-shaped protein of 427 kDa, containing an actin binding domain at its N-terminus, several spectrin-like coiled coil repeats at the center of the molecule, a highly conserved cys-rich region and a further coiled coil region at its C-terminus, the latter probably binding to membrane associated glycoproteins. The rsp5-domain is located at the interface of the spectrin repeats and the cys-rich region, within a stretch of high sequence conservation among species. Remarkably, apo-dystrophin 1 [27], encoded by a non-muscle transcript that originates from the same gene but uses a different promoter, has its N-terminus within the rsp5-homology domain. The dystrophin-like protein utrophin (genpept 34812) [28] possesses the same overall organization. The *C. elegans* ORF zk1098.1 (genpept 297979) [29] contains two rsp5-domains at the N-terminus followed by a coiled-coil region showing similarity to the yeast myosin isoform myo2. The extreme C-terminal region consists mainly of lysine and a few other charged amino acids. The same organization, except for the lys-rich region, is shared by the yeast protein ykb2 (genpept 263498) [30].

The remaining sequences could not be assigned to one of the above groups. An ORF from rat, termed 'integrase-like protein' (genpept 57560) [31] contains a rsp5-domain at the N-terminus and a putative transmembrane region at the C-terminus. The reported sequence similarity to retroviral integrases is not confirmed by our analyses. The short yeast ORF λ-7584.2 (genpept 559927) contains also a rsp5-domain and a transmembrane helix. The yeast ess1 protein (SwissProt: ESS1_YEAST) [32], which has been reported to be possibly involved in cytokinesis or cell-separation, also contains a single copy of the rsp5-domain. The helicase-like protein from *Nicotinia sylvestris*

Table 1
Sequence in genpept containing the URE-B1 homology region

genpept Acc. No.	Description	Organism	Length	URE-B1 region	
				from	to
475516	URE-B1 (binds dynorphin promoter)	Rat	310	1	310
172848	rsp5 (contains rsp5 domains)	Yeast	809	450	809
577313	kiaa93 (contains rsp5 domains)	Human	927	600	884
220509	nedd4 (contains rsp5 domains)	Mouse	708	300	650
577760	F45H7.6 (contains rsp5 domains)	*C. elegans*	201	1	201
263499	ORF (adjacent to ykb2)	Yeast	1483	1260	1483
178745	E6-AP (binds p53 and papillomavirus E6)	Human	874	550	874
290244	hyperplastic discs protein	*Drosophila*	2894	2700	2894
55535	100 kDa protein	Rat	889	600	800
517115	ORF	Human	1050	727	1050
285983	ORF	Human	1083	727	1083
460711	ORF	Human	1992	1700	1992

The first column gives the NCBI accession number to the genpept entries, the last two columns show the approximate region of the URE-B1 homology

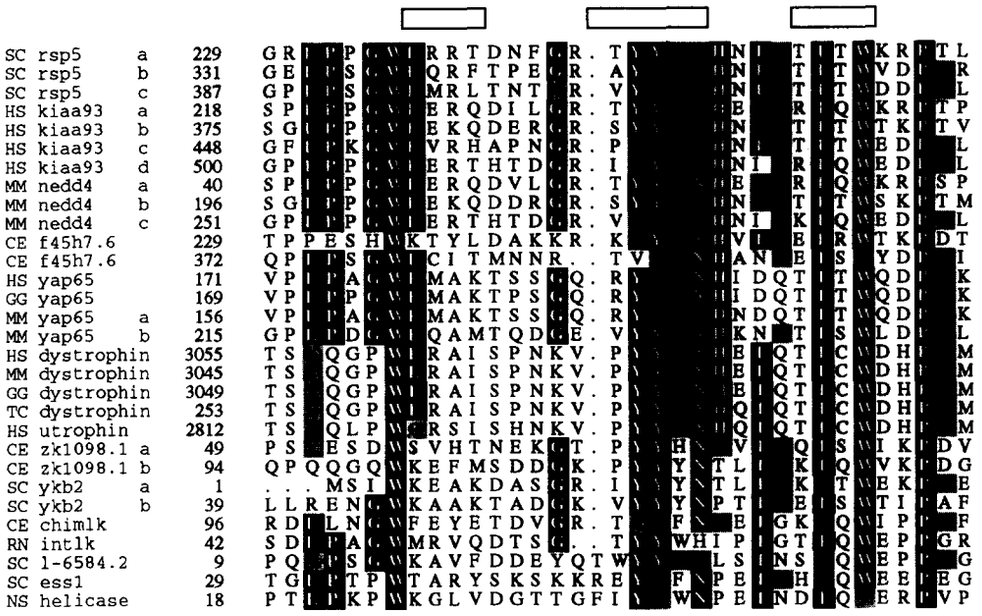

Fig. 1. A manual alignment of all 30 detected rsp5-domains is shown. Residues conserved or similar in at least 60% of all sequences are shown on black or gray background, respectively, by application of the program BOXSHADE. The first column designates the species abbreviation (HS = human, MM = mouse, RN = rat, GG = chicken, TC = *Torpedo californica*, CE = *C. elegans*, NS = tobacco, SC = yeast), the second column denominates the sequence name (see text for database accession numbers), lowercase a, b, c, or d designates the individual copies of the domain in multiple-domain proteins. The third column shows the position of the domain in the sequence. The open bars above the alignment indicate the predicted beta-sheets.

(genpept 569986) [33] closely resembles other known RNA-helicases, but has a rsp5-domain added at the extreme N-terminus.

4. Discussion

The list of sequences presented in the previous section clearly shows that the rsp5 domain occurs in a variety of proteins, probably involved in very different cellular processes. The lacking global sequence similarity of rsp5-domain bearing proteins, the multiple occurrence of the rsp5-domain in several proteins with non-homologous flanking sequences, and the common occurrence with other, previously characterized autonomous domains suggests that the rsp5-domain is a true member of the growing family of in dependent building blocks observed in intracellular proteins. It is always a difficult task to delineate the exact domain boundaries as long as three-dimensional structures are not available. In the case of the rsp5-domain, a length of 35 was chosen, corresponding exactly to the second rsp5-domain in the murine yap65 protein, which is not present in the human sequence and apparently has been inserted into (or lost from) a highly conserved sequence region. The length of 35 is also in good agreement with the observed homology region in all observed rsp5-domains. A length of only 35 residues is unusually small for an intracellular autonomous folding unit, considering the fact that disulfide bridges for stabilization of the structure are not available in this compartment; most domains characterized so far comprise about 50 to 150 residues.

The alignment of all rsp5-domains (Fig. 1) suggests a bipartite structure: two regions of sequence conservation with no observed insertions or deletions are located at the N- and C-terminal domain boundary, respectively. The central region shows less sequence conservation and a certain degree of length variation. The average amino acid composition is strongly shifted toward polar amino acids, which is not unexpected for a small folding domain that has no extensive hydrophobic core with residues buried from the solvent. Remarkable is the high content of proline. A typical rsp5 domain contains 4–5 prolines, about 10 charged residues and 4–5 aromatic residues. There are only two totally invariant residues, a tryptophane residue in the N-terminal half and a proline at the C-terminus. Three more positions are invariably occupied by aromatic residues. While the proline residues are highly enriched in the N- and C-terminal boundary regions, the charged residues are more evenly distributed over the domain. The properties of the amino acids at specific positions, together with their conservation pattern allows a prediction of their secondary or even tertiary structure. An automated secondary structure prediction using the PHD program [18] proposes two β-strands in the central region of the domain. A third beta strand in the C-terminal region is suggested by a model based on a subsignificant sequence similarity (5 standard deviations above randomized average) of the rsp5-domain to several MHC class II alpha chains, one of their three-dimensional structures being known [34]. The reported 3D-structures of other intracellular autonomous also consist mainly of β-sheet structures [2,6,35].

The fact that several rsp5-domain containing sequences have been determined from genomic DNA allows an analysis of the correspondence of the domain boundaries with the exon/intron boundaries. This correspondence turned out to be very weak, in most cases the rsp5-domain is part of a considerably larger exon with varying amount of non-rsp5 related sequence on both sides. Moreover, these exons are generally not reading-frame preserving. These observations suggest frequent intron-

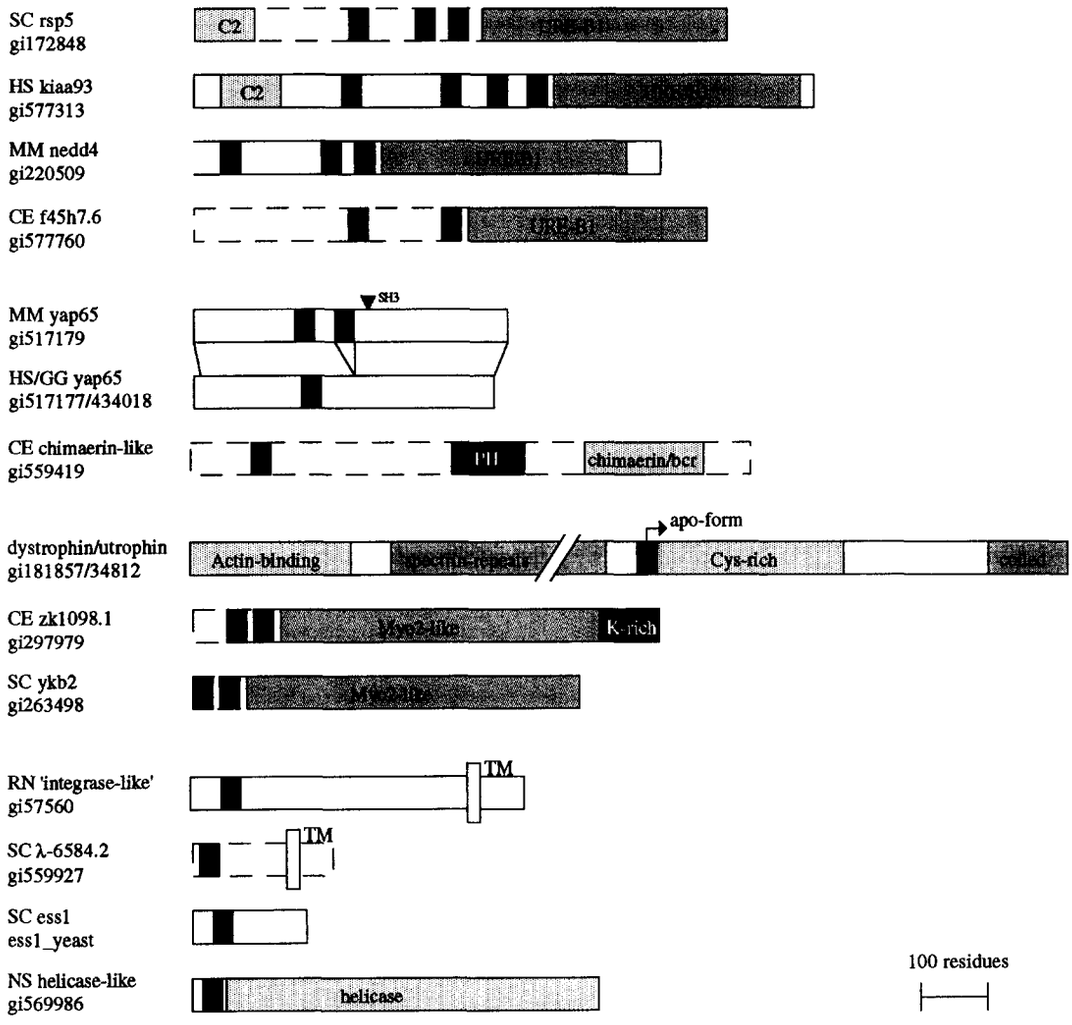

Fig. 2. The domain structure of all sequences containing rsp5-domains. The length of the sequence is drawn to scale, dashed frames indicate that the protein sequences are derived from hypothetical gene models of genomic sequences. Frames open on one side indicate incomplete sequence data. The gi-numbers are the NCBI accession numbers of the genpept sequences. Black boxes indicate rsp5-domains. Open bars labeled 'TM' designate putative transmembrane domains, the arrowhead labeled 'SH3' indicates the putative site of SH3-binding, the arrow labeled 'apo-form' indicates the N-terminus of apo-dystrophin 1.

loss processes since the exon shuffling that lead to the widespread distribution of rsp5-domains. The dystrophin gene is the only one known having an intron inserted into an rsp5-domain. The position of the intron insertion is approximately in the center of the domain. Interestingly, this intron contains a promoter that gives rise to an alternative transcripts from the dystrophin gene, the apo-dystrophin 1 [27]. This transcript, containing only the C-terminal half of the rsp5-domain is found in major amounts in non-muscle tissues, in contrast to the main dystrophin gene product, its function is completely unknown.

As the most important issue remains the question of the biological function of the rsp5-domain. The region of the rsp5-domain in dystrophin has, on the basis of its high content of proline, been referred to as a 'hinge region' [25]. However, this is unlikely to be the main function of the general rsp5-domain since in many protein this domain is observed at the extreme N-terminus, a position normally not requiring hinges. The high-proline stretch at the N-terminus of the domain resembles, but does not fulfill, the consensus patterns for SH3-binding [36]. Almost all copies of the rsp5-domain possess multiple consensus patterns for phosphorylation. However, since their position is not conserved, this makes a common mechanism of phosphorylation-dependent conformational change less likely.

The occurrence of rsp5-domains in both regulatory and structural proteins suggest at least two possible interpretations. The rsp5-domain might act as a target site of regulatory processes like e.g. phosphorylation, thus adding a regulation mechanism to the structural proteins it occurs in. On the other hand, a mechanism offering cytoskeletal attachment to regulatory proteins would also be conceivable.

After the preparation of this manuscript we learned that the rsp 5-domain is identical to the WW-domain reported recently [37]. We also note that the URE-B1 homology region of Table 1 is named 'round domain' in that reference.

Acknowledgements: This work was supported by grant 31-37678.93 from the Swiss National Science Foundation.

References

[1] Sadowski, I., Stone, J.C. and Pawson, T. (1986) Mol. Cell. Biol. 6, 4396–4408.
[2] Yu, H., Rosen, M.K., Shin, T.B., Seidel-Dugan, C., Brugge, J.S. et al. (1992) Science 258, 1665–1668.
[3] Haslam, R.J., Koide, H.B. and Hemmings, B.A. (1993) Nature 363, 309–310.
[4] Musacchio, A., Gibson, T., Rice, P., Thompson, J. and Saraste, M. (1993) Trends Biochem. Sci. 18, 343–348.
[5] Coussens, L., Parker, P.J., Rhee, L.T., Yang-Feng, L. and Chen, E. et al. (1986) Science 233, 859–866.
[6] Waksman, G., Kominos, D., Robertson, S.C., Pant, N., Baltimore, D. et al. (1992) Nature 358, 646–653.
[7] Davletov, B.A. and Sudhof, T.C. (1993) J. Biol. Chem. 268, 26386–26390.
[8] Harlan, J.E., Hajduk, P.J., Yoon, H.S. and Fesik, S.W. (1994) Nature 371, 168–170.
[9] Eisenmann, D.M., Arndt, K.M., Ricupero, S.L., Rooney, J.W. and Winston, F. (1992) Genes Dev. 6, 1319–1331.
[10] Luthy, R., Xenarios, I. and Bucher, P., (1994) Prot. Sci. 3, 139–146.
[11] Gribskov, M., McLachlan, A.D. and Eisenberg, D. (1987) Proc. Natl. Acad. Sci. 84, 4355–4358.
[12] Bairoch, A. and Boeckmann, B. (1994) Nucleic Acids Res. 22, 3578–3580.
[13] Benson, D., Lipman, D.J. and Ostell, J. (1994) Nucleic Acids Res. 22, 3441–3444.
[14] Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990) J. Mol. Biol. 215, 403–410.
[15] Henikoff, S. and Henikoff, J.G. (1992) Proc. Natl. Acad. Sci. 89, 10915–10919.
[16] Bucher, P. and Bairoch, A., in: Proceedings of the 2nd International Conference for Intelligent Systems in Molecular Biology, AAAI press, 1994.
[17] Pearson, W.R. and Lipman, D.J. (1988) Proc. Natl. Acad. Sci. 85, 2444–2448.
[18] Rost, B., Sander, C. and Schneider, R. (1994) Comput. Appl. Biosci. 10, 53–60.
[19] Gu, J., Ren, K., Dubner, R. and Iadarola, M.J. (1994) Brain Res. Mol. Brain Res., in press.
[20] Kumar, S., Tomooka, Y. and Noda, M. (1992) Biochem. Biophys. Res. Commun. 185, 1155–1161.
[21] Wilson, R., Ainscough, R., Anderson, K., Baynes, C., Berks, M. et al. (1994) Nature 368, 32–38.
[22] Sudol, M. (1994) Oncogene 9, 2145–2152.
[23] Hall, C., Monfries, C., Smith, P., H. Lim, H., Kozma, R. et al. (1990) J. Mol. Biol. 211, 11–16.
[24] Koenig, M., Monaco, A.P. and Kunkel, L.M. (1988) Cell 53, 219–226.
[25] Koenig, M. and Kunkel, L.M. (1990) J. Biol. Chem. 265, 4560–4566.
[26] Tinsley, J.M., Blake, D.J., Pearce, M., Knight, A.E., Kendrick-Jones, J. et al. (1993) Curr. Op. Genet. Dev. 3, 484–490.
[27] Blake, D.J., Love, D.R., Tinsley, J., Morris, G.E., Turley, H. et al. (1992) Hum. Mol. Genet. 1, 103–109.
[28] Tinsley, J.M., Blake, D.J., Roche, A., Fairbrother, U., Riss, J. et al. (1992) Nature 360, 591–593.
[29] Sulston, J., Du, Z., Thomas, K., Wilson, R., Hillier, L. et al. (1992) Nature 356, 37–41.
[30] Pascolo, S., Ghazvini, M., Boyer, J., Colleaux, L., Thierry, A. et al. (1992) Yeast 8, 987–995.
[31] Duilio, A., Zambrano, N., Mogavero, A.R., Ammendola, R., Cimino, F. et al. (1991) Nucleic Acids Res. 19, 5269–5274.
[32] Hanes, S.D., Shank, P.R. and Bostian, K.A. (1989) Yeast 5, 55–72.
[33] Itadani, H., Sugita, M. and Sugiura, M. (1994) Plant Mol. Biol. 24, 249–252.
[34] Stern, L.J., Brown, J.H., Jardetzky, T.S., Gorga, J.C., Urban, R.G. et al. (1994) Nature 368, 215–218.
[35] Macias, M.J., Musacchio, A., Ponstingl, H., Nilges, M., Saraste, M. et al. (1994) Nature 369, 675–677.
[36] Lim, W.A., Richards, F.M. and Fox, R.O. (1994) Nature 372, 375–379.
[37] Bork, P. and Sudol, M. (1994) Trends Biochem. Sci. 19, 531–533.