Minireview

Towards prediction of cognate complexes between the WW domain and proline-rich ligands

Aaron Einbond, Marius Sudol*

Department of Biochemistry, The Mount Sinai School of Medicine, One Gustave L. Levy Place, New York, NY 10029-6547, USA

Received 27 February 1996

Abstract The WW domain is a structured protein module found in a wide range of regulatory, cytoskeletal, and signaling molecules. Its ligands contain proline-rich sequences, some of which show a core consensus of XPPXY that is critical for binding. In order to gain a better understanding of the molecular and biological functions of WW domains, we decided to predict their cognate ligands by searching databases for proteins containing the XPPXY consensus. Using several axioms that take into account evolutionary conservation and functional similarity, we have identified four groups of proteins representing candidate ligands that signal through known or unknown WW domains. These include viral Gag proteins, sodium channels, interleukin receptors, and a subgroup of serine/threonine kinases. In addition, we proposed that dystrophin and β-dystroglycan bind through the WW-XPPXY link and that interference with this interaction could result in muscular dystrophy. Our study provides guidelines for experiments to probe the molecular and biological functions of the WW domain-ligand connection. Should these predictions be proven empirically, the results may have important ramifications for basic research and medicine.

Key words: Protein-protein interaction; Proline-rich motif; Viral-Gag; Sodium channel; Liddle's syndrome; Interleukin receptor; Serine/threonine kinases; SCID, severe combined immunodeficiency; Muscular dystrophy

1. Background and rationale

The WW domain is a protein module found in a wide range of structural, regulatory, and signaling molecules [1]. It has been found in 20 different proteins [2,3] including dystrophin; YAP, an adaptor-like protein that interacts with the SH3 domain of the Src-family protein tyrosine kinase Yes; ORF1, a human protein containing a Ras GTPase activating domain; Ess1, a yeast protein essential for cell division; Rsp5, a yeast protein that is involved indirectly in regulation of transcription; FE65, a brain-specific adaptor protein; proteins that bind to formins – the latter is a group of gene products encoded by the mouse *limb deformity* locus; and several other proteins from human, rat, yeast, nematode, and tobacco [2,4,5].

*Corresponding author. Fax: (1) (212) 996-7214.
E-mail: M_Sudol@smtplink.mssm.edu

Abbreviations: ENaC, epithelial sodium channel; IL-2R, Interleukin-2 receptor; MAP kinase, mitogen-activated protein kinase; MAPKAP2, MAP kinase activated protein kinase; P, proline; SH, Src homology; W, tryptophan; WBP, WW domain-binding protein; Y, tyrosine; YAP, Yes-associated protein of 65 kDa; X, any amino acid.

The domain was identified as a result of our search for substrates and regulators of the Yes proto-oncogene product. Yes is a non-receptor protein tyrosine kinase of the Src family. It is expressed at elevated levels in cerebellar Purkinje cells and in lung, kidney, and intestinal epithelial cells; this pattern of expression suggests that Yes could be involved in secretion and regulation of ion transport [6]. Based on results of Hirai and Varmus [7], we reasoned that by identifying molecules that interact with Yes in vivo, we should be able to characterize the mode of action of the Yes protein in normal and viral-Yes transformed cells. YAP (Yes-associated protein) was isolated as one of the proteins that binds the SH3 (Src homology 3) domain of Yes; its sequence contains an SH3-binding motif with the core consensus PXXP (P, proline; X, any amino acid, not necessarily the same) reported by Schreiber and colleagues [8–12]. YAP is phosphorylated in vivo by an unknown serine kinase, indicating that it may form a novel link between the serine/threonine and tyrosine signaling pathways [8]. Initial evidence for the WW domain came from comparison of the sequences of human, mouse, and chicken YAP. This comparison revealed that the mouse protein contains an insert not found in the human and chicken proteins, with significant similarity to a sequence upstream in all three proteins. Computer-aided analysis of this sequence, together with the results of biochemical and structural studies suggested that this region represented a new protein domain. One of the distinguishing features of the domain is the presence of two highly conserved tryptophan residues (W), hence the name: the WW domain [1]. Its length was set at 38 amino acids, the length of the mouse insert [2].

Since we proposed that the WW domain of YAP might mediate protein-protein interaction [1,2], we performed a functional screen of a cDNA expression library to identify its ligand. Two putative binding proteins were isolated and named WBP-1 and WBP-2 [13]. Sequence comparison of the two ligands revealed that they share a proline-rich region that binds strongly and specifically to the WW domain of YAP. The region consists of a five-amino acid sequence, PPPPY (the PY motif), which is perfectly conserved between WBP-1 and WBP-2 and occurs three times in WBP-2 [13]. By sequentially replacing each of these five positions with alanine for in vitro binding assays, a preliminary consensus of XPPXY (P, proline; Y, tyrosine; X, any amino acid, not necessarily the same) was established as necessary for binding [13]. This consensus is distinct from the SH3-binding motif PXXP. Consistent with this observation, WBP-1 does not bind to the arbitrarily chosen SH3 domains of Yes, Fyn, Abl, and GAP [13]. Considering that the XPPXY motif shows a significant level of specificity in binding to at least a subset of WW domains, and

SSDI S0014-5793(96)00263-3

knowing that as a proline-rich sequence it must bind rapidly and tightly [14], we think that XPPXY may represent an important signaling site utilized by many proteins.

Since a consensus of XPPXY is critical for binding to the WW domain of YAP, we propose that ligands to other WW domains can be predicted by searching databases for proteins containing the consensus. However, there are several problems inherent in this simple proposal: First, the motif is so small that the chance of its random occurrence in a sequence database is high. In fact, the number of occurrences of the motif found in the search is similar to that expected to appear at random. Therefore, many of the protein sequences retrieved will probably be insignificant because the presence of the motif in these proteins will have been the result of coincidence, not of evolutionary conservation. Some of the XPPXY motifs could be embedded in the proteins and may not be able to play the role of a surface-exposed module mediating protein-protein interaction. Second, the core consensus established for binding to the WW domain of YAP, with the 'alanine scan,' may only cover selected subsets of known WWs – perhaps those most closely related to the WW domain of YAP. Other variants of the consensus (XPPX [any aromatic residue] or XPPXXY) are conceivable (Sudol and Bougeret, unpublished results from phage display library screens). Indeed, polyproline regions lacking tyrosines, have recently been found in formin isoforms to bind to WW domain-containing proteins in vitro [4] (Fig. 5C). The consensus is also limiting because it does not take into account the contribution of flanking amino acids that may dictate or modulate the ligand-ligate interaction [15–17].

In order to determine which of the proteins retrieved contain the consensus as a result of evolutionary pressure, and to therefore determine which of the proteins are likely to interact with WW domain family proteins, we have looked for patterns in the primary structure and function of the proteins retrieved. The XPPXY motifs in proteins with conserved patterns are likely to mediate interaction in vivo with WW family proteins. Of course, these XPPXY-containing proteins may interact with not-yet-isolated WWs, and could be used as functional probes to retrieve new members of the WW family [3]. In addition to patterns in the topological conservation of the XPPXY motif, we have looked for possible correlations in function between proteins containing the motif and proteins containing the WW domain, again as indications of interactions that occur in vivo. Guided by these two axioms we have identified a series of putative WW ligand proteins, providing suggestions for experiments to demonstrate cognate interactions between XPPXY motif-containing and WW domain-containing proteins and to probe the biological aspects of this molecular link.

2. Methods

The Fasta program [18] was used to search the Translated Genbank database of protein sequences for the motif XPPXY. To facilitate the searches, 20 separate screens were performed, each with a different amino acid in the first position of the motif. A total of 4410 occurrences of the motif were found in the 127 887 sequences, containing a total of 39 288 428 residues, of the Genbank as of August, 1995. The actual number of proteins containing the motif is smaller since several proteins contained the motif in multiple copies. Nevertheless, considering that one would expect to find on the order of 4900 occurrences of the motif if the composition of the Genbank proteins were completely random, most of the proteins retrieved probably contain the motif as a result of *chance*. In order to begin to identify the proteins in which the motif may have a distinct, WW-related function, we selected 185 occurrences of the motif to study further (once again, the actual number of proteins containing these occurrences of the motif is smaller). This selection was made based upon the descriptions of the proteins – those selected seemed to be the best candidates for involvement in intracellular signaling. From the 185 occurrences of the motif, proteins were grouped together based upon similarity of function, interaction with similar molecules, localization to similar parts of the cell, and expression in similar organs. Many proteins were included in more than one group. Within each group the lengths of the protein sequences, the positions of the XPPXY motifs within their sequences, the exact sequences of the motifs, and relative homology of the entire proteins were compared. In addition, database searches with the Blastp program [19] were used to establish how these sub-groups related to any larger groups of proteins. For the groups of proteins in question, the most significant findings are: viral-Gag proteins, sodium channels, interleukin receptors, and serine/threonine kinases. In addition, we report the identification of the motif in dystroglycan and propose the interaction of this motif with the WW domain of dystrophin.

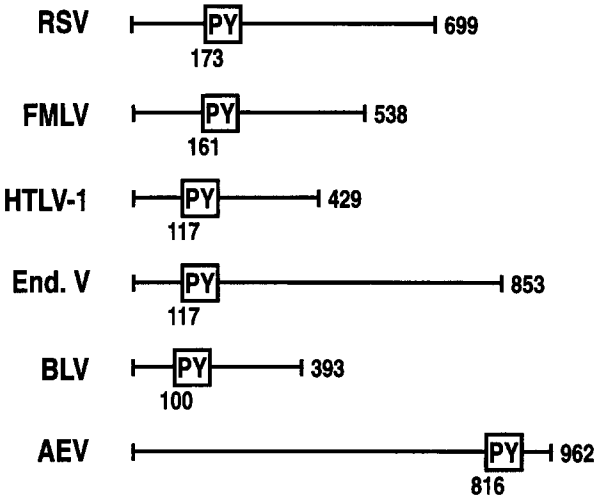
3. Protein families with proline motifs

3.1. Gag proteins

Gag proteins of retroviruses play a key role in the genesis and assembly of viral particles [20]. Because of their frequent fusion to cellularly derived sequences in viral oncogene products, Gag proteins are compelling candidates for involvement in signaling processes [21]. In fact, numerous studies have implicated Gag proteins as modulators of transforming potential for many oncogenic proteins (e.g. [22]).

The XPPXY motif was found in 13 distinct Gag proteins. Eleven of these were from various mammalian and avian retroviruses, one was a mouse cellular homolog, and one an endogenous virus-derived Gag protein in fungus. Among the collection, several Gag proteins belonged to viral oncogenes, including Crk, Erb-A and -B, Fms, Fps, and Yes. Comparison of the 13 Gag proteins in the 25-residue region, arbitrarily chosen, extending 10 residues on either side of the XPPXY motif revealed that several of them share significant homology. Excluding similar sequences from consideration, there are six distinct sequences surrounding the motif. Except for the perfect conservation of the XPPXY motif, no sequences share more than 15% identity in the 25-residue region. Even though Gags are presumably related evolutionarily, several of the Gags containing the XPPXY motif show little sequence similarity to each other in this region. This makes unlikely the possibility that XPPXY is only conserved in Gags because it is a part of a larger, well conserved region. Of the six XPPXY motifs, those from Rous sarcoma virus (RSV) and Bovine leukemia virus (BLV) contain prolines in the first and fourth positions, identical to the PY motif in WBP-1 and WBP-2, and those from Friend murine leukemia virus (FMLV) and Human T-cell leukemia virus 1 (HTLV-1) contain prolines in the fourth position but not the first (Fig. 1).

Despite the abrupt divergence of the homology of these sequences before and after the XPPXY motif, the 13 proteins demonstrate surprising similarity in the relative position of the motif within their sequences. Of the 13 Gags, those from RSV and FMLS (as well as the homologous Gags from chicken provirus RAV-O, avian sarcoma virus, and wild mouse ecotropic virus), contain the motif approx. 170 residues from the beginning of the protein; and those from HTLV-1, the *Fusarium oxysporum* endogenous virus, and BLV, (as well the homologous Gag from avian erythroblastosis virus), contain the motif between 80 and 120 residues from the beginning of the protein. Of the 13 proteins, only one of the Gag proteins from avian erythroblastosis virus neither contains the motif in one of the locations described above nor is homologous to a protein containing the motif in one of these locations (Fig. 1). A search using the Blastp program of the 25 residues surrounding the motif in RSV Gag against all available sequence databases revealed no significant homology to any Gags lacking the motif. According to this search, all reported members



Virus	Partial Sequence
RSV	CNCATATASA**PPPPY**VGSGLYPSLA
FMLV	GGPLIDLLTE**DPPPY**RDPGPPSPDG
HTLV-1	THDPPDSDPQ**IPPPY**VEPTAPQVLP
End. V	GRDPGEVLKP**SPPEY**FDGTPSKLPT
BLV	APGASAPEEQ**PPPPY**DPPAILPIIS
AEV	LIAEFSKMAR**DPPRY**LVIQGDERMH

Fig. 1. Structures and partial sequences of Gag proteins containing the XPPXY motif. Only non-homologous (less than or equal to 15% identity in the 10 residues on either side of the motif) Gags are shown. The XPPXY motif is represented by the boxed PY. The full names of the viruses and accession numbers are as follows: RSV, Rous sarcoma virus [D10652] (homologous sequences: chicken provirus RAV-O [M73497], avian erythroblastosis virus [M32090], avian sarcoma virus [J0207]); FMLV, Friend murine leukemia virus [Z11128] (homologous sequences: mouse p12 homologue (X72930), feline sarcoma virus [K01643], wild mouse ecotropic virus [M26528]; HTLV-1, human T-cell leukemia virus type 1 [D13784]/ human t-lymphotropic virus type 1 [S745562]; End. V, endogenous virus from the fungus *Fusarium oxysporum* [L34658]; BLV, bovine leukemia virus [M10987]; AEV, avian erythroblastosis virus [X52209].

of the subfamily of Gags similar to RSV contain the XPPXY motif. This finding is expected if the motif has functional significance in Gags.

In support of our proposal, the **PPPPY** motif in the p2b region of RSV gene was shown to be required for a late step in budding after the Gag proteins have been targeted to the membrane [23]. This finding constitutes direct evidence of the biological function of the XPPXY motif. Which of the WW domain-containing proteins interacts with the PPPPY sequence of Gag remains to be determined; YAP is a potential candidate (see also Fig. 5A). The role of the PPPPY motif may be related to proposed functions of motifs of the form PXY or PXXY in endocytosis. The motifs PPGY (a version of the XPPXY motif) in lysosomal acid phosphatase [24], NPXY in the LDL receptor [25], and NPLIY in the β_2-adenergic receptor [26] have been shown to adopt reverse-turn conformations, essential as internalization signals. These signals are proposed to interact with coated pit cytoplasmic receptors [24] or adaptor proteins [25]. Although viral budding and protein internalization seem distinct, they are related processes in that they both involve localization to the membrane and result in its alteration. Perhaps the XPPXY motif in Gags and the [P/N]PXY motifs in lysosomal acid phosphatase and the LDL receptor convey signals similarly by interacting with membrane proteins containing WW or PTB/PID domains.

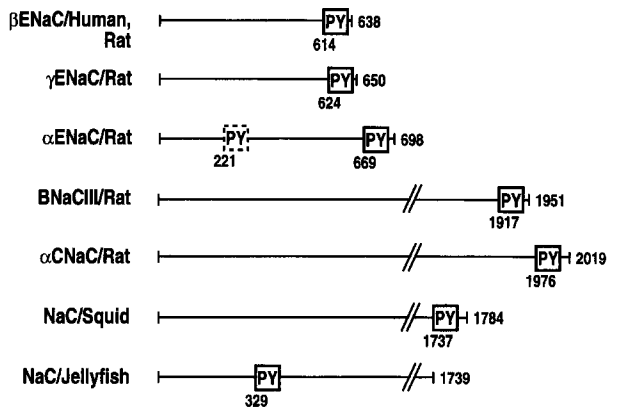
Considering that the PPPPY motif is known to bind to the WW domain of YAP, we propose that in general these Gag proteins interact, through their XPPXY motifs, with a group of WW family proteins. This is consistent with the implication of Gag proteins in modulation of transformation, resulting from the physical fusion of several of the Gag proteins to viral oncogene products, including Yes. To illustrate our proposal we present here a model of interaction between viral Yes (or Crk) oncogene product and cellular YAP protein (Fig. 5A). According to this 'double interaction' model, the WW domain of YAP binds to the PPPPY motif of Gag and prevents the normal signaling of YAP by interaction with WBP-1 or WBP-2 proteins which when in free, unbound form, could be involved in positive control of cell growth. This simple, yet attractive model is being tested experimentally.

3.2. Sodium channels

Sodium channels are membrane-spanning proteins essential for regulating sodium ion concentrations in tissues as diverse as nervous tissue and kidney epithelium. The importance of their role in metabolism is illustrated by the Mendelian form of hypertension known also as Liddle's syndrome, which results from mutations of the β-subunit of the human epithelial sodium channel [27–29]. Liddle's syndrome is one of the many genetic traits that leads to severe hypertension, in this case through excessive reabsorption of sodium in the distal nephron of the kidney [27,28]. The control of ion passage by sodium channels, or the number of sodium channels present in the membrane, may be regulated by signals from intracellular proteins.

Nine sodium channel proteins were found to contain the XPPXY motif. Of these nine, eight contained the motif between 20 and 50 amino acids from the C-terminus. The only one that did not contain the motif in this location was from the jellyfish *Cyanea capillata*; therefore its discrepancy from the pattern found in the other channels, mostly vertebrate, can possibly be explained by its evolutionary distance from them.

As before, the 25 residues surrounding the motif were compared. Eliminating from consideration similar sequences, and counting twice the one protein that contained two copies of the motif, there are seven distinct sequences surrounding the motif (Fig. 2). Besides the motif, conserved in all of them, these sequences share no more than 25% identity. Interestingly, however, there seems to be a general trend toward small hydrophobic or slightly polar residues preceding the motif and charged residues following it. Of the sodium channel XPPXY motifs, one of the two from the epithelial sodium channel α-subunit contains four prolines, as in WBP-1 and WBP-2. This occurrence of the motif has been included for completeness; however, it may not be significant since it falls in the extracellular region of the channel, a location at which signaling by the XPPXY motif is perhaps unlikely – so far there is no evidence for the presence of WW modules in extracellular compartments. The XPPXY motifs from the squid sodium channel, the β- and γ-subunits of the epithelial sodium channel, and the second motif from the α-subunit contain proline in the first but not the fourth position (Fig. 2). A Blastp program search of the 25-residue segment of the epithelial sodium channel β-subunit (arbitrarily chosen) against all available protein databases revealed that it does not contain significant similarity to any sodium channels lacking the



Protein/ Organism	Partial Sequence
βENaC/Human	SEQALPIPGT**PPPNY**DSLRLQPLDT
γENaC/Rat	PAPGSTVPGT**PPPRY**NTLRLDRAFS
αENaC/Rat	FPHPLQRLRT**PPPPY**SGRTARSGSS
αENaC/Rat	GMTPPLALTA**PPPAY**ATLGPSAPPL
BNaCIII/Rat	EKTDGSSSTT**SPPSY**DSVTKPDKEK
αCNaC/Rat	LSSSSISSTS**FPPSY**DSVTRATSDN
NaC/Squid	RANRPHSSLP**PPPSY**EESCKRQDEI
NaC/Jellyfish	ICGNASTAWN**CPPNY**LCMPGAGSNP

Fig. 2. Structures and partial sequences of sodium channel proteins containing the XPPXY motif. Only non-homologous (less than or equal to 25% identity in the 10 residues on either side of the motif) sodium channels are shown. The XPPXY motif is represented by the boxed PY (see explanation of αENaC dotted box in text). The full names of the proteins and accession numbers, if available, are as follows: βENaC/human, human epithelial sodium channel β-subunit [53] (rat ortholog [53]); γENaC/human, human epithelial sodium channel γ-subunit [X77933]; αENaC/human, human epithelial sodium channel α-subunit [X70497] (homologous to human kidney amiloride-sensitive sodium channel [L29007]); BNaCIII/rat, rat brain sodium channel III [Y00766] (homologous to rat brain sodium channel I [X03638]); αCNaC/rat, rat cardiac specific sodium channel α-subunit [M27902]; NaC/squid, *Loligo opalescens* sodium channel [L19979]; NaC/jellyfish, *Cyanea capillata* sodium channel [L15445].

XPPXY motif (although it does show 36% homology to a potassium channel in which the motif is not perfectly conserved).

The α-, β- and γ-subunits of the epithelial sodium channel in humans are involved in Liddle's syndrome, and the XPPXY motif of the β-subunit has been directly implicated in the disease. In one patient, a de novo mutation causing the replacement of the third proline in the **PPPNY** motif with a leucine and led to Liddle's syndrome [30]. In another case, the replacement of the terminal tyrosine in the **PPPNY** motif with histidine also correlated with the clinical diagnosis of Liddle's syndrome [31]. Other mutations of the β-subunit that have been reported to result in Liddle's syndrome all involve deleted portions of the intracellular C-terminal region, containing the **PPPNY** motif [28]. Replacement of the tyrosine with an alanine leads to an increase in the current of the mutated channel expressed in frog oocyte system, at least as great as that in a Liddle's patient [32] The same result was obtained with the α- and γ-subunits. Therefore, the XPPXY motif may perform a similar function in each subunit [32].

It seems likely from these findings that the level of sodium passage is regulated by the interaction of the XPPXY motif of the channel with a WW family protein (Fig. 5B). Such a protein has recently been identified as Nedd4, which contains three WW domains and can interact with the β- and γ-subunits of the channel [33]. Therefore, Nedd4 (or a closely related protein) may be a suppressor of channel activity [33]. In this case, mutation or absence of the XPPXY motif of the channel would prevent interaction with Nedd4, leading to increased channel current. Another possible explanation of XPPXY regulation of sodium passage is that the XPPXY motif/Nedd4 link controls the number of channels present in the membrane, either by regulating channel degradation [34], channel insertion, membrane localization, or channel internalization [32]. These data also support the proposal of a conserved function for the motif in the other sodium channels described above, which contain the motif in a similar relative position to the epithelial sodium channel subunits. Sodium conductance of these channels may be similarly controlled by WW family proteins.

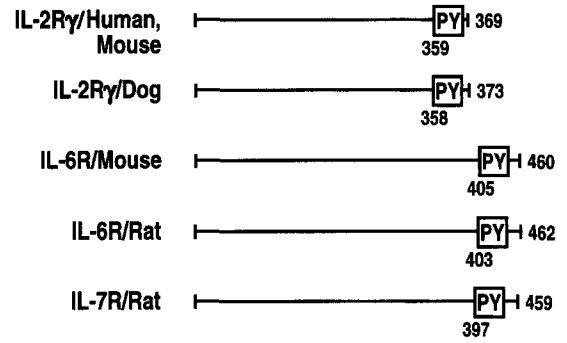
3.3. Interleukin receptors

Interleukins are hormones that, through their interaction with interleukin receptors, regulate the growth and differentiation of B cells, T cells, and other cells of the immune system, and play a central role in the mammalian immune response [35–37]. The importance of the interleukin receptors is illustrated by the X-linked disease severe combined immunodeficiency (SCID), which has been shown to result from mutations in the γ chain of the human and dog Interleukin-2 (IL-2) receptor [38–40]. Males affected by SCID suffer from severe opportunistic infections and die early in life if not treated through a bone marrow transplant [39]. Obviously essential to the function of the interleukin receptors are their intracellular regions, which must interact with cellular proteins in order to convey the signal of the interleukin hormone within the cell.

Three of the interleukin receptors, in various mammals, were found to contain an XPPXY motif in their intracellular regions. These were the human, dog, and mouse IL-2 receptor γ chain, implicated in SCID, the mouse and rat IL-6 receptor, and the mouse IL-7 receptor (sequences of other

orthologs of these receptors were not available). Similar to the Gag proteins, the XPPXY motif is conserved among the three interleukin receptors while the surrounding 25 amino acids share no more than 15% identity. One of the proteins out of the six contains prolines in the first and fourth position, like WBP-1 and WBP-2, one contains a proline in only the first, and one contains a proline in only the fourth (Fig. 3). Also like the Gag proteins, the relative position of the motif is highly conserved among the interleukin receptors. The lengths of the three proteins are similar, with the IL-2 receptor γ chain approx. 370 residues long and the IL-6 and IL-7 receptors approx. 460 amino acids long. The XPPXY motif is around amino acid 360 in IL-2R, 400 in IL-6R, and 400 in IL-7R. A Blastp search of the 25-residue segment of the IL-2 receptor γ chain against all available protein sequences revealed no significant homology to any other cytokine receptors; therefore the sub-group of interleukin receptors containing the motif does not seem to correspond to any larger group.

One possibility is that the interleukin receptors relay the signal from the interleukin hormone through the XPPXY motif in their intracellular region to a cellular WW family protein. In the case of IL-2, it is tempting to speculate that a malfunction in this process results in SCID (Fig. 5D). Of the documented mutations that lead to SCID in humans, four are single-base pair substitutions or deletions that lead to premature stop codons either in the extracellular region or the intracellular region preceding the XPPXY motif [38,39]; in either case, the mutations lead to a protein lacking an XPPXY



Protein/ Organism	Partial Sequence
IL-2Rγ/Human	SPCNQHSPYW**APPCY**TLKPET*
IL-2Rγ/Dog	SPCSQHSPYW**APPCY**TLKPEETGALIP*
IL-2Rγ/Mouse	SPCSLHSPYW**PPPCY**SLKPEA*
IL-6R/Mouse	AEKESKTTSP**PPPPY**SLGPLKPTFL
IL-6R/Rat	AEKESKTT--**SPPPY**PLGPLKPTFL
IL-7R/Rat	SPDYRDGDRN**RPPVY**QDLLPNSGNT

Fig. 3. Structures and partial sequences of interleukin recetor proteins containing the XPPXY motif. Only non-homologous (less than or equal to 15% identity in the 10 residues on either side of the motif) sodium channels are shown. The XPPXY motif is represented by the boxed PY. Asterisks denote the ends of sequences. The full names of the proteins and accession numbers are as follows: IL-2R γ/human, human interleukin-2 receptor γ-chain [D11086]; IL-2R γ/dog, dog interleukin-2 receptor γ-chain [U04361]; IL-2R γ/mouse, mouse interleukin-2 receptor α-chain [X75337]; IL-6R/mouse, mouse interleukin-6 receptor [X53802]; IL-6R/rat, rat interleukin-6 receptor [M58587]; IL-7R/rat interleukin-7 receptor [M29697].

motif. Our suggestion of a role for XPPXY signaling in SCID would not be surprising, considering the recent implication of the Jak3 kinase in the SCID syndrome [41]. In normal cells, the Jak3 kinase interacts with the C-terminus of several interleukin receptor γ chains [41]. Perhaps a WW family protein may interact with the C-terminus of IL-2 γ chain in a manner similar to Jak3 or it may simply serve as an adaptor protein between the IL-2 receptor and a tyrosine kinase.

3.4. Serine/threonine kinases

Since YAP is phosphorylated in vivo by an unknown serine kinase, serine/threonine kinases are of particular interest as candidates for involvement in WW domain XPPXY ligand signaling. It was therefore appealing to find the XPPXY motif in several serine/threonine kinases, suggesting structural, and possibly functional, significance.

A total of 41 protein kinases were found to contain at least one copy of the XPPXY motif. No patterns that would indicate evolutionary conservation of the XPPXY motif have yet been identified in those described as tyrosine kinases, among which is the Abl proto-oncogene. Fourteen of the serine/threonine kinases, however, contain an XPPXY motif at a conserved position in their catalytic domains [42,43]. These serine kinases come from several organisms – including human, *Drosophila*, entamoeba, and tobacco – and act on a variety of substrates. Of the 14, five are orthologs of MAP kinase-associated protein kinase, and two, from tobacco and the ice plant, are highly similar. Therefore, there are nine distinct serine/threonine kinase sequences containing the motif. The presence of the motif in MAP kinase-associated protein kinase 2 (MAPKAP2) is interesting in itself. One of the WW family proteins is Msb1, a suppressor of mutations in the *bck1* gene encoding a MAP kinase kinase kinase, and a suppressor of *mpk1* (yeast Map kinase) mutations ([2], K. Matsumoto, personal communication). In fact, one could speculate that the ability of Msb1 to suppress *mpk1* mutations is related to an interaction of its WW domain with the XPPXY motif of MAP kinase-associated protein kinase (Fig. 5E).

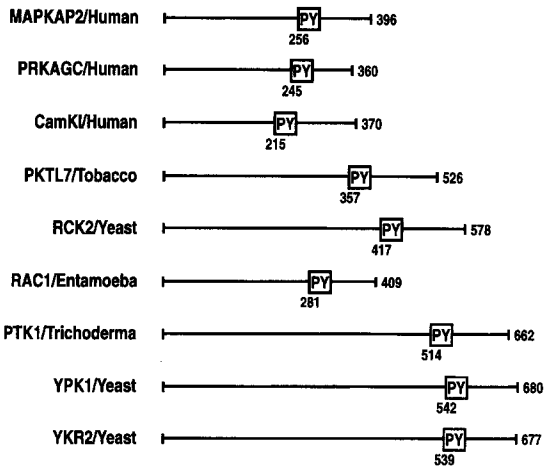
Unlike the previous entries which share relatively little homology to each other in the region directly surrounding the XPPXY, the serine/threonine kinases, as is expected from the presence of the motif in their catalytic domain, contain high homology and many conserved features in the region immediately surrounding the motif. One possible indication against the significance of the motif in serine/threonine kinases is that a Blastp search with the twenty five residue segment of MAPKAP2 showed significant homology to the catalytic domains of several kinases that lacked a perfect XPPXY motif. But this finding does not eliminate the possibility that the XPPXY has a specialized function within the catalytic domains of the kinases reported here. As is indicated in the consensus sequence (Fig. 4), only six residues are conserved in all nine serine kinases, three of which are the conserved residues of the motif. The residue in the fourth position of the motif is either phenylalanine or tyrosine for the nine serine kinases. This is perhaps an indication that the fourth residue contributes to the specificity of the motif. It is interesting to note that this region, like the corresponding region in the sodium channel proteins, contains a prevalence of hydrophobic or slightly polar residues preceding the motif and a prevalence of charged residues following the motif (Fig. 4).

The motif is consistently located between 120 and 200 residues from the C-terminus of each protein, suggesting a relatively similar position for the XPPXY motif-containing region in the overall structure of the proteins. These theoretical data, as well as the potential involvement of two already characterized WW family proteins in the MAP kinase serine phosphorylation pathway, suggests that WW family proteins could be either substrates for phosphorylation by these kinases, or adaptor molecules, of the same type as YAP.

3.5. *Dystrophin and β-dystroglycan*

Our search of the protein sequence database with the XPPXY consensus identified the β-dystroglycan receptor as a possible ligand of a WW domain. Since we have identified the WW domain in dystrophin, and it has been previously noted that the location of this domain in dystrophin is similar to that of the site at which dystrophin binds to dystroglycan [1,2], we decided to study this possble link further.

β-Dystroglycan was found to contain at least two XPPXY motifs: one is of the form PPPEY, the other SPPPY and the latter is located at the very carboxy-terminal end of the protein [44]. Recent studies implicated the very carboxy-terminal sequence containing the SPPPY motif in mediating interaction with dystrophin; however, the precise region in dystrophin that binds the polyproline region was not delineated [45]. We propose that dystrophin and β-dystroglycan bind through WW-XPPXY interaction and that interference with this interaction (through deletion of the WW domain, for example) could result in muscular dystrophy (Fig. 5F). This must be tested experimentally in transgenic mice.



Protein/Organism	Partial Sequence
MAPKAP2/Human	GVIMYILLCG**YPPFY**SNHGLAISPG
PRKACG/Human	GVLIYEMAVG**FPPFY**ADQPIQIYEK
CamKI/Human	GVIAYILLCG**YPPFY**DENDAKLFEQ
PKTL7/Tobacco	GAIMYEMLVG**YPPFY**SDDPMSTCRK
RCK2/Yeast	GCVLYTMLCG**FPPFY**DEKIDTLTEK
RAC1/Entamoeba	GILIYEMIVG**IPPFY**DDDVSLMYQK
PTK1/Trichoderma	GVLLYEMLTG**LPPFY**DENTNEMYRK
YPK1/Yeast	GVLLYEMLTG**LPPYY**DEDVPKIYKK
YKR2/Yeast	GILLYEMMTG**LPPYY**DENVPVMYKK
consensus:	G**hhhYEMLhG -PPFY** Dtth----tK

Fig. 4. Structures and partial sequences of serine/threonine kinase proteins containing the XPPXY motif. Orthologs are not shown. The XPPXY motif is represented by the boxed PY. In the consensus line, t indicates a predominately turn-like or polar position; h represents a predominantly hydrophobic position. Bold residues in the consensus line are perfectly conserved in all 9 sequences and non-bold residues are conserved in 6–8 sequences. The full names of the proteins and accession numbers are as follows: MAPKAP2/human, human MAP kinase activated protein kinase [X75346] (orthologs: rabbit [X75345], hamster [X82220], *Drosophila* [U20757]): PRKAGC/human, human testis-specific protein kinase γ-subunit [M34182]; CamKI/human, human calcium/calmodulin dependent protein kinaseI [L41816]; PKTL7/tobacco, tobacco PTKL7 protein kinase [X71057] (ortholog: ice plant [Z30329]); RCK2/yeast, yeast RCK2 protein kinase [X71065]; PTK1/Trichoderma, *Trichoderma reesei* QM9414 serine/threonine protein kinase [U05811]; YPK1/yeast YPK1 protein kinase [M21307]; YKR2/yeast, yeast YKR2 protein kinase [M24929].

3.6. *Further considerations*

To look for further characteristics of the XPPXY motif in the proteins described, we have identified features of the first and fourth positions of the motif that may tend to expose it on the surface of a protein, and therefore allow it to interact with other proteins. One such feature is the presence of prolines in these positions, as is the case with WBP-1 and -2. Several studies have been conducted describing general structural features of proline residues and proline-rich motifs [14,46,47]. The relative rigidity of a string of four prolines is likely to force the motif to a protein's surface regardless of the surrounding sequence (G. Rose, personal communication). This is a good indication of the motif's functional significance in proteins such as Rous sarcoma virus Gag, as well as in proteins containing an additional proline in only the first or only the fourth position of the motif.

Another feature of the motif that may force it to the protein surface is the presence of charged residues in the first or fourth positions. As has been previously reported, there is a direct correlation between residue hydrophobicity and the average area of the residue buried (not exposed to the protein surface) within reported proteins [48,49]. Occurrences of the motif are therefore more likely to be on the protein surface if they contain a charged or polar residue. The presence of serine, asparagine, or arginine in the fourth position of the motif in six of the nine sodium channels indicates that the motif is more likely to be exposed on the surface of these proteins ([47], and B. Matthews, personal communication).

Related to this consideration is the idea that amino acids in the first and fourth positions may contribute to XPPXY specificity. The fact that many of the sodium channels contain serines or alanines in the fourth position of the motif while all of the serine kinases contain phenylalanines or tyrosines, for example, indicates that perhaps subtle differences in the shape of the motif lead to recognition of different WW domains. Future study of the specificity of WW domains will better illuminate this idea.

The identification of the WW domain-binding sequence APPTPPPLPP in formins, proteins involved in mouse limb and kidney development [4], indicates the extent to which different WW domains may recognize different proline-rich ligands. Once more ligands to WW family proteins have been identified, the techniques of analysis used here can be applied to each new consensus, allowing for the prediction of more WW-XPPXY cognate complexes. It is interesting that the WW domain and the SH3 domain can compete in binding the proline-rich sequence in formins in vitro [4] (Fig. 5C). The proline-rich C-terminal region of the epithelial sodium channel subunits, which contains the XPPXY motif, was originally characterized as an SH3-binding sequence, interacting with

spectrin [50,51]. Perhaps the proteins containing the WW domain and proteins containing the SH3 domain can compete similarly in vivo in binding formins and the epithelial sodium channel.

4. Concluding remarks

The primary purpose of our theoretical study was to provide guidelines for experiments to examine the molecular and biological function of the WW domain-ligand connection. The most direct experimental approaches suggested by our data would be to mutate the XPPXY motifs in representative members of Gags, sodium channels, interleukin receptors, β-dystroglycan, and serine/threonine kinases and to assess the resulting phenotypes. Should new or expected (hypertension, immunodeficency, muscular dystrophy, or attenuation of viral transformation) phenotypes be generated in transgenic animals or in relevant cell culture models, detailed studies on the cognate WW domain-containing proteins could illuminate downstream signaling events of the investigated pathways.

The three-dimensional structure of the WW domain of human YAP in complex with its peptide ligand was recently solved using NMR spectroscopy [52]. Based on this data one could try to select by computer modeling the 'best fitting' pairs of ligate and ligand and then test them experimentally. Homology modeling of the WW domain may result in general 'rules' for the structures of the XPPXY ligand.

Study of the XPPXY motif and its structural and functional relationship to the WW domain will therefore, no doubt, lead to an exponential increase in the understanding of a wide range of signaling processes and of the reasons for their malfunction in certain human diseases. Moreover, considering

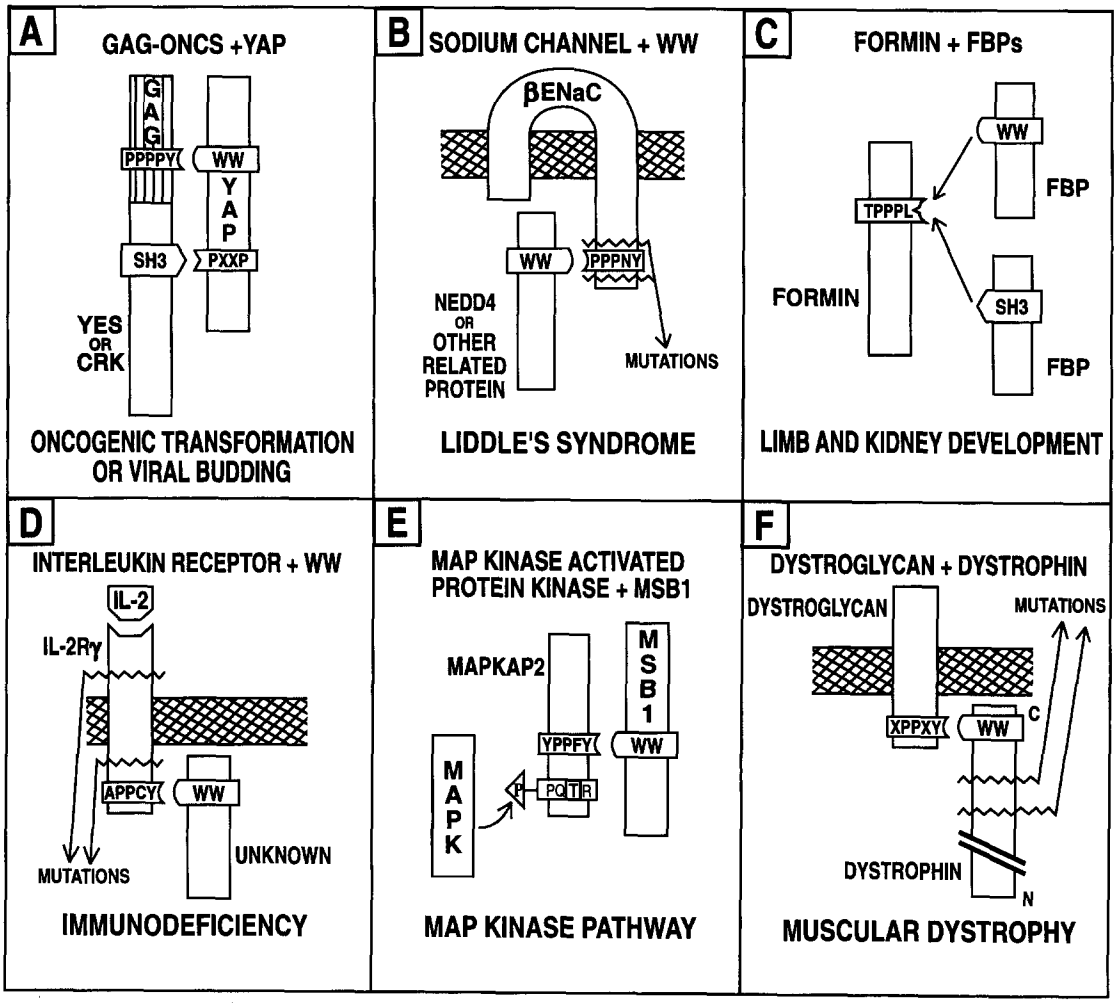


Fig. 5. Models of XPPXY-WW interaction. (A) The XPPXY motif and SH3 domain of viral Gag-Yes or Gag-Crk interact with the WW domain and SH3 binding motif of vertebrate YAP, resulting in oncogenic transformation. In normal cells the WW domain of YAP interacts with WBP-1 and/or WBP-2, which, when in free, unbound form, could be involved in positive control of cell growth. (B) Nedd4 or another related WW family protein interacts at the membrane with the XPPXY motif of the epithelial sodium channel β-subunit. Deletion of the XPPXY motif or point mutation within the motif results in Liddle's syndrome. (C) A formin interacts with two formin binding proteins (FBP). The WW domain of one FBP and the SH3 domain of the other compete to bind to the proline-rich sequence TPPPL, resulting in the regulation of limb and kidney development. (D) An unknown WW family protein interacts with the XPPXY motif of the interleukin-2 receptor γ-chain at the membrane of a normal cell. Mutations of the extracellular or intracellular regions, which would cause deletion of the XPPXY motif, result in severe combined immunodeficiency. (E) MAP kinase phosphorylates a threonine (in the MAP kinase phosphorylation site PQTR) of MAP kinase activated protein kinase 2 (MAPKAP2), which leads to the interaction of the XPPXY motif of MAPKAP2 with the WW domain of Msb1. (F) One of the XPPXY motifs of dystroglycan interacts with the WW domain of dystrophin, localizing dystrophin to the membrane of a normal cell. Mutations resulting in deletion of the C-terminal region of dystrophin prevent this interaction, causing Duchenne muscular dystrophy. C and N denote carboxy-and amino-termini, respectively, of dystrophin. In all models, the cell membrane is represented by a cross-hatched rectangle.

that both the WW domain and the core motif of its ligand are relatively short (38 and 5 residues), one could speculate that the human syndromes that involve mutations of these modules could be treated successfully not only by gene therapy approaches, but also by low molecular weight compounds. Although peptides, even short ones composed of only 4 amino acids, are inefficient as drugs, the rigid molecular shapes represented by polyproline cores of the WW domain ligands could serve as guides for rational drug design or for choosing a repertoire of organic mimotopes for functional screens.

Acknowledgements: Special thanks are due to Drs. Peer Bork, Henry Chen, Brian Matthews, Hartmut Oschkinat, Philip Leder, George Rose, Sandy Ross, Daniela Rotin, Matti Saraste, Andrew Sparks, Michael Williamson, John Wills and all members of the Sudol lab for valuable discussions and comments on the manuscript. Dr. Bonnie Kaiser, the Rockefeller University Science Outreach Program, and the Camille and Henry Dreyfus Special Grant Program are acknowledged for their enthusiastic support. The research reviewed in this paper was supported by National Cancer Institute Grants CA45757 and CAO1605, by the Council for Tobacco Research-USA Inc. Grant 3035, by Human Frontier Science Program grant, and by the Klingenstein Award in the Neurosciences (to M.S.).

References

[1] Bork, P. and Sudol, M. (1994) Trends Biochem. Sci. 19, 531–533.
[2] Sudol, M., Bork, P., Einbond, A., Katsury, K., Druck, T., Negrini, M., Huebner, K. and Lehman, D. (1995) J. Biol. Chem. 270, 14733–14741.
[3] Sudol, M., Chen, H.I., Bougeret, C., Einbond, A. and Bork, P. (1995) FEBS Letters 369, 67–71.
[4] Chan, D.C., Bedford, M.T. and Leder, P. (1996) EMBO J. 15 (in press).
[5] Fiore, F., Zambrano, N., Minopoli, G., Donini, V., Duilio, A. and Russo, T. (1995) J. Biol. Chem. 270, 30853–30856.
[6] Sudol, M. (1990) Exp. Med. 8, 94–100.
[7] Hirai, H. and Varmus, H. (1990) Mol. Cell Biol. 10, 1307–1318.
[8] Sudol, M. (1994) Oncogene 9, 2145–2152.
[9] Yu, H., Chen, J.K., Feng, S., Dalgarno, D.C., Brauer, A.W. and Schreiber, S.L. (1994) Cell 76, 933–945.
[10] Feng, S., Chen, J.K., Yu, H., Simon, J.A. and Schreiber, S.L. (1994) Science 266, 1241–1247.
[11] Ren, R., Mayer, B.J., Cicchetti, P. and Baltimore, D. (1993) Science 259, 1157–1161.
[12] Saraste, M. and Musacchio, A. (1994) Nature Struct. Biol. 1, 835–837.
[13] Chen, H.I. and Sudol, M. (1995) Proc. Natl. Acad. Sci. USA 92, 7819–7823.
[14] Williamson, M.P. (1994) Biochem. J. 297, 249–260.
[15] Rickles, R.J., Botfield, M.C., Weng, Z., Taylor, J.A., Green, O.M., Brugge, J.S. and Zoller, M.J. (1994) EMBO, J. 13, 5598–5604.
[16] Sparks, A.B., Quilliam, L.A., Thorn, J.M., Der, C.J. and Kay, B.K. (1994) J. Biol. Chem. 269, 23853–23856.
[17] Alexandropoulos, K., Cheng, G. and Baltimore, D. (1995) Proc. Natl. Acad. Sci. USA 92, 3110–3114.
[18] Pearson, W.R. and Lipman, D.J. Proc. Natl. Acad. Sci. USA 85, 2444–2448.
[19] Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990) J. Mol. Biol. 215, 403–410.
[20] Wills, J.V. and Craven, R.C. (1991) AIDS 5, 639–654.
[21] Dickson, C., Eisenman, R., Fan, H., Hunter, E. and Teich, N. (1984) in: RNA Tumor Viruses, Vol. 1 (Weiss, R., Teich, N., Varmus, H. and Coffin, J. eds.) pp. 513–648, Cold Spring Harbor, NY.
[22] Foster, D.A., Shibuya, A. and Hanafusa, H. (1985) Cell 42, 105–115.
[23] Parent, L.J., Bennett, R.P., Craven, R.C., Nelle, T.D., Krishna, N.K., Bradford Bowzard, J., Wilson, C.B., Puffer, B.A., Montelaro, R.C. and Wills, J.W. J. Virol. 69, 5455–5460.
[24] Eberle, W., Sander, C., Klaus, W., Schmidt, B., Von Figura, K. and Peters, C. (1991) Cell 67, 1203–1209.
[25] Bansal, A. and Gierasch, L.M. (1991) Cell 67, 1195–1201.
[26] Barak, L.S., Tiberi, M., Freedman, N.J., Kwatra, M.M., Lefkowitz, R.J. and Caron, M.G. (1994) J. Biol. Chem. 269, 2790–2795.
[27] McDonald, F.J., Snyder, P.M., McCray Jr., P.B. and Welsh, M.J. (1994) Am. J. Physiol. 266, L728–L734.
[28] Shimkets, R.A., Warnock, D.G., Bositis, C.M., Nelson-Williams, C., Hansson, J.H., Schambelan, M., Gill Jr., J.R., Ulick, S., Milora, R.V., Findling, J.W., Canessa, C.M., Rossier, B.C. and Lifton, R.P. (1994) Cell 79, 407–414.
[29] Lifton, R.P. (1995) Proc. Natl. Acad. Sci. USA 92, 8545–8551.
[30] Hansson, J.H., Schild, L., Lu, Y., Wilson, T.A., Gautschi, I., Shimkets, R., Nelson-Williams, C., Rossier, B.C. and Lifton, R.P. (1995) Proc. Natl. Acad. Sci. USA 92, 11495–11499.
[31] Tamura, H., Enomoto, N., Matsui, N., Sasaki, S. and Marumo, F. (1995) J. Am. Soc. Nephrol. 6, 728 (Abstr. 1205).
[32] Snyder, P.M., Price, M.P., McDonald, F.J., Adams, C.M., Volk, K.A., Zeiher, B.G., Stokes, J.B. and Welsh, M.J. (1995) Cell 83, 969–978.
[33] Staub, O., Dho, S., Henry, P., Correau, J., Ishikawa, T., McGlade, J. and Rotin, D. (1996) EMBO J. (in press).
[34] Huibregtse, J.M., Scheffner, M., Beaudenon, S. and Howley, P.M. (1995) Proc. Natl. Acad. Sci. USA 92, 2563–2567.
[35] Goodwin, R.G., Friend, D., Ziegler, R.J., Falk, B.A., Gimpel, S., Cosman, D., Dower, S.K., March, C.J., Namen, A.E. and Park, L.S. (1990) Cell 60, 941–951.
[36] Baumann, M., Baumann, H. and Fey, G. (1990) J. Biol. Chem. 265, 19853–19862.
[37] Takeshita, T., Asao, H., Ohtani, K., Ishii, N., Kumaki, S., Tanaka, N., Munakata, H., Nakamura, M. and Sugamura, K. (1992) Science 257, 379–382.
[38] Noguchi, M., Yi, H., Rosenblatt, H.M., Filipovich, A.H., Adelstein, S., Modi, W.S., McBride, O.W. and Leonard, W.J. (1993) Cell 73, 147–157.
[39] Puck, J.M., Deschenes, S.M., Porter, J.C., Dutra, A.S., Brown, C.J., Willard, H.F. and Henthorn, P.S. (1993) Hum. Mol. Genet. 2, 1099–1104.
[40] Henthorn, P.S., Somberg, R.L., Fimiani, V.M., Puck, J.M., Patterson, D.F. and Felsburg, P.J. (1994) Genomics 23, 69–74.
[41] Russell, S.M., Tayebi, N., Nakajima, H., Riedy, M.C., Roberts, J.L., Aman, M.J., Migone, T.-S., Noguchi, M., Market, M.L., Buckley, R.H., Shea, J.J. and Leonard, W.J. (1995) Science 270, 797–800.
[42] Zu, Y., Wu, F., Gilchrist, A., Ai, Y., Labadia, M.E. and Huang, C. (1994) Biochem. Biophys. Res. Commun. 200, 1118–1124.
[43] Beebe, S.J., Oyen, O., Sandberg, M., Froysa, A., Hansson, V. and Jahnsen, T. (1990) Mol. Endocrinol. 4, 465–475.
[44] Ibraghimov-Beskrovnaya, O., Milatovich, A., Ozcelik, T., Yang, B., Koepnick, K., Francke, U. and Campbell, K.P. (1993) Human. Mol. Genet. 2, 1651–1657.
[45] Jung, D., Yang, B., Meyer, J., Chamberlain, J.S. and Campbell, K.P. (1995) J. Biol. Chem. 270, 27305–27310.
[46] MacArthur, M.W. and Thornton, J.M. (1990) J. Mol. Biol. 218, 397–412.
[47] Hurtley, J.H., Mason, D.A. and Matthews, B.W. (1992) Biopolymers 32, 1443–1446.
[48] Rose, G.D., Geselowitz, A.R., Lesser, G.J., Lee, R.H. and Zehfus, M.H. (1985) Science 229, 834–838.
[49] Lesser, G.J. and Rose, G.D. (1990) Proteins 8, 6–13.
[50] McDonald, F.J. and Welsh, M.J. (1995) Biochem. J. 312, 491–497.
[51] Rotin, D., Bar-Sagi, D., O'Brodovich H., Merilainen, J., Lehto, V.P., Canessa, C.M., Rossier, B.C. and Downey, G.P. (1994) EMBO J. 13, 4440–4450.
[52] Macias, M.J., Hyvonen, M., Schultz, J., Chen, H.I., Saraste, M., Sudol, M. and Oschkinat, H. (1996) Nature (submitted).
[53] Shimkets, R.A., Warnock, D.G., Bositis, C.M., Nelson-Williams, C., Hansson, J.H., Schambelan, M., Gill Jr., J.R., Ulick, S., Milora, R.V., Findling, J.W., Canessa, C.M., Rossier, B.C. and Lifton, R.P. (1994) Cell 79, 407–414.