beginning near the amino termini of the SNAPs was found to contain several 34-residue repeats separated by linker regions of six amino acids. As shown in the upper portion of Fig. 1, each 34-residue repeat can be divided into five subrepeats. These SNAP subrepeats all have related sequences but can be separated into two classes by length. Four copies of a seven-residue subrepeat are present within the 34-residue repeat and are arranged in tandem pairs that are separated by a single six-residue subrepeat in the center. The repeat structure for the five SNAP sequences aligned in Fig. 1 is indicated by shading of residues that match a 'unified' consensus sequence representing both the seven-residue and six-residue subrepeats (Fig. 1). Despite substantial sequence differences among the SNAP proteins (ranging from 88% to 25% amino acid identity with α-SNAP in the aligned region), the percent identity of their 34-residue repeats to the unified consensus for the SNAP subrepeats falls within a rather narrow range as follows: α-SNAP (65%), β-SNAP (64%), dSNAP (63%), SEC17 (60%) and γ-SNAP (63%). The degree to which these comparisons indicate strict order within the SNAP sequences is indicated by our finding that a randomized version of the same α-SNAP domain sequence exhibited only 22% identity to the SNAP subrepeat consensus sequences.

Database searches using the portions of the dSNAP and α-SNAP proteins shown in Fig. 1 identified the yeast nuclear protein nuc2$^+$ as a high-ranking match and as the only protein matched by both query sequences. Another yeast nuclear protein, the cell-cycle protein CDC23, was identified as the closest match to a consensus sequence representing the SNAP seven-residue subrepeat. Both nuc2$^+$ and CDC23 are members of the TPR protein family, and the 34-residue repeats present in these proteins were found to be similar to those in the SNAPs.

The similarity between the SNAP repeats and those in the TPR proteins is shown in Fig. 1. Here, two individual TPR proteins, the yeast nuclear proteins nuc2$^+$ (Refs 3, 4) and CDC23 (Refs 3, 4), and a published consensus sequence for the TPR motif[8] are aligned with the SNAP repeat domain. Again, highlighted residues match the unified consensus sequence for the seven- and six-residue SNAP subrepeats. The aligned sequences of the nuc2$^+$ (TPRs 5–8) and CDC23 (TPRs 5–8) proteins are 43% and 40% identical to the unified consensus for the SNAP subrepeats. By contrast, a randomized sequence corresponding to the same nuc2$^+$ domain exhibited only 22% identity. The TPR consensus sequence exhibits 62% identity to the SNAP subrepeat consensus, a value similar to those observed with the repeat domains of individual SNAP proteins. Thus, a clear relationship exists between the 34-residue repeats in the SNAPs and those in established TPR proteins. Despite this, the SNAP TPRs are unusual in that they are separated by linker regions of six amino acids in length, in contrast to the tandem arrangement prevalent in other TPR proteins[3,4] (see Fig. 1).

The findings reported here may have important implications. The subrepeat structure identified in the SNAP TPRs may lead to new general insights into the structure and interactions of TPRs in all proteins in which they appear. Furthermore, the relationship of the SNAPs to the TPR proteins adds to previous work identifying NSF-related proteins with nuclear and cell-cycle functions[10,11]. Taken together, these findings suggest that similar kinds of protein interactions participate in cell functions as diverse as cell-cycle regulation and neurotransmitter release.

References

1 Südhof, T. C., De Camilli, P., Niemann, H. and Jahn, R. (1993) *Cell* 75, 1–4
2 Barinaga, M. (1993) *Science* 260, 488–489
3 Goebl, M. and Yanagida, M. (1991) *Trends Biochem. Sci.* 16, 173–177
4 Sikorski, R. S. *et al.* (1991) *Cold Spring Harbor Symp. Quant. Biol.* 56, 663–673
5 Ordway, R. W., Pallanck, L. and Ganetzky, B. (1994) *Proc. Natl Acad. Sci. USA* 91, 5715–5719
6 Whiteheart, S. W. *et al.* (1993) *Nature* 362, 353–355
7 Griff, I. C., Schekman, R., Rothman, J. E. and Kaiser, C. A. (1992) *J. Biol. Chem.* 267, 12106–12115
8 Sikorski, R. S., Boguski, M. S., Goebl, M. and Heiter, P. (1990) *Cell* 60, 307–317
9 Devereux, J., Haeberli, P. and Smithies, O. (1984) *Nucleic Acids Res.* 12, 387–395
10 Fröhlich, K. *et al.* (1991) *J. Cell Biol.* 114, 443–453
11 Thorsness, P. E., White, K. H. and Ong, W-C. (1993) *Yeast* 9, 1267–1271

RICHARD W. ORDWAY, LEO PALLANCK AND BARRY GANETZKY

Laboratory of Genetics, 445 Henry Mall, University of Wisconsin–Madison, Madison, WI 53706, USA.

The WW domain: a signalling site in dystrophin?

Duchenne and Becker muscular dystrophies are degenerative diseases caused by mutations of a single locus, the dystrophin gene (for a review see Ref. 1). The gene encodes a large molecule that belongs to a family of cytoskeletal proteins including α-actinin and β-spectrin. Alternatively spliced forms exist; the longest one consists of four domains: (1) an amino-terminal, globular actin-binding domain common to other cytoskeletal proteins, (2) 24 spectrin-like repeats forming a long rod in the middle of the molecule, (3) a cysteine-rich Ca^{2+}-binding domain, and (4) a carboxy-terminal globular domain[1] (Fig.1). Molecular analysis of the central rod-like portion of human dystrophin revealed two interruptions of the spectrin repeats and two flanking segments that appear to be hinge regions[2].

Since the flanking hinge regions are sufficiently long to form functionally independent domains, we have subjected them to sequence-database searches. Indeed, the segment following the spectrin repeats showed significant sequence similarity (probability of a chance match $<10^{-6}$, as computed using the program BLASTP, Ref. 3) to repeats in a nematode and a mouse protein (Yo61 and Nedd4 in Fig. 2). Subsequent database searches with profiles[4] and patterns[5] derived from these regions identified several other proteins that contain this novel domain, some with as many as three copies (Fig. 1). Since two strictly conserved tryptophans (W) give the strongest signal in sequence comparisons, we refer to it as the WW domain. Instrumental in the delineation of the WW domain was the cloning of the *YAP* gene, which encodes a Yes-kinase-associated protein[6]. The mouse YAP protein contains two WW domains compared with only one found in the human and chicken orthologues; the second WW domain in the mouse sequence appears to be the result of a recent duplication and thus allows the length of the domain to be estimated at about 40 amino acids.

The WW domain is often flanked by histidine-rich or cysteine-rich regions, which might bind metal ions, as in dystrophin[1]. The domain itself appears to contain β-strands (Fig. 2) grouped around four conserved aromatic positions. The presence of both a hydrophobic core and numerous charged residues (Fig. 2) is reminiscent of well-characterized protein modules involved in protein–protein interactions. Like the Src homology domains SH2 and SH3, and the pleckstrin homology (PH) domain, the WW domain occurs in a variety of molecules with no apparent common function. Despite their functional diversity, all of the proteins listed in Fig. 1 seem to be involved in cell signalling or regulation.

Dystrophin and utrophin are more than 70% identical in sequence[1]; they form tetramers via their spectrin-like repeats and are thought to have multiple

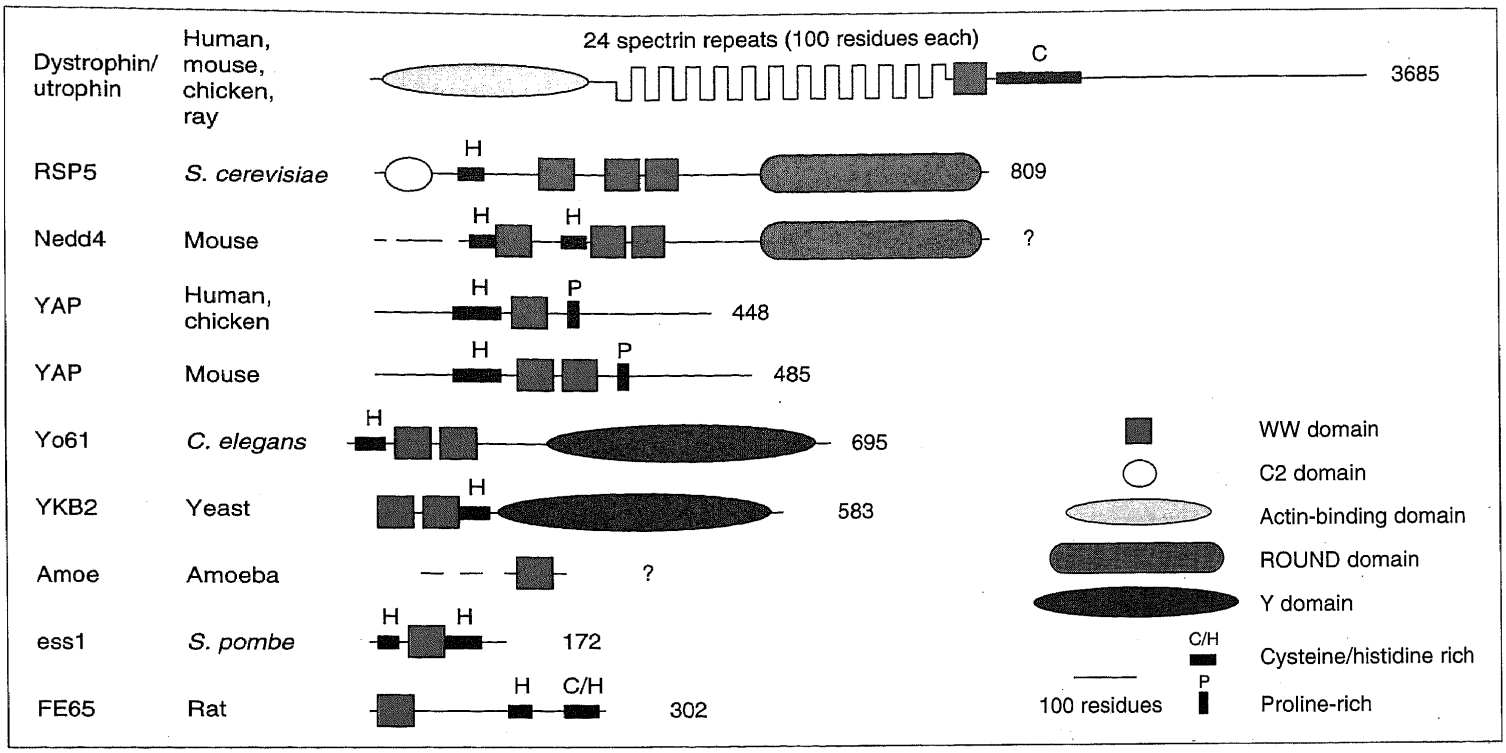

Figure 1

Modular architecture of the proteins containing the WW domain. Dashed lines denote partial sequence. Note that the 24 spectrin repeats are not drawn to scale. The C2 domain is shared with some protein kinases C, synaptotagmins and *Caenorhabditis elegans* unc-13; the ROUND domain is found at the carboxyl termini of RSP5, the oncoprotein E6-AP, UreB1, Nedd4, a hypothetical yeast protein (Ykbo) and others; the Y-domain is common to YKB2 and Yo61; and the proline-rich domain binds SH3 domains.

functions, including involvement in membrane stability, transduction of contractile forces to the extracellular environment and organization of membrane specialization[1]. YAP is a substrate of an unknown serine kinase, and it binds to the SH3 domain of the Yes oncoprotein via a proline-rich region located downstream of the WW domain[6] (Fig. 1). Mouse Nedd4 plays a role in the embryonic development and differentiation of the central nervous system[7]. Yeast *Saccharomyces cerevisiae* RSP5 is similar to Nedd4 in its molecular organization and contains an amino-terminal regulatory domain common to protein kinases C and synaptotagmins (C2-domain in the PROSITE motif database). The yeast (*Schizosaccharomyces pombe*) ess1 protein appears to be essential for growth and may be involved in cytokinesis and/or cell separation[8]. Rat FE65 is a transcription-factor activator expressed preferentially in liver; the activator domain is located within the first 232 residues of FE65 (Ref. 9), which also contain the WW domain.

The identification of the WW domain in dystrophin suggests a binding site for one of the many dystrophin-associated proteins[10]; it is closely located to the β-dystroglycan-binding site and may regulate the formation of this complex.

Protein/species	Position	Sequences of the WW domains	Acc. No.
Dmd_Human	3052	TSVQGP**WE**RAISPN.KVP**YY**I**N**HETQTTCWDHPKMTELY	P11532
Dmd/Ray	253	TSVQGP**WE**RAISPN.KVP**YY**I**N**HQTQTTCWDHPKMTELY	M37645
Utro/Human	2813	TSVQLP**W**QRSISHN.KVP**YY**I**N**HQTQTTCWDHPKMTELF	X69086
Yap/Human	171	VP**LP**AG**WE**MAKTSS.GQR**Y**FL**N**HIDQTTTWQDPRKAMLS	X80507
Yap/Chick	169	VP**LP**P**GWE**MAKTPS.GQR**Y**FL**N**HIDQTTTWQDPRKAMLS	X76483
Yap/Mouse-1	151	VP**LP**AG**WE**MAKTSS.GQR**Y**FL**N**HNDQTTTWQDPRKAMLS	X80508
Yap/Mouse-2	218	GP**LP**D**GWE**QAMTQD.GEV**YY**I**N**HKNKTTSWLDPRLDPRF	X80508
Nedd4/Mouse-1	?	SP**LP**P**GWE**ERQDVL.GRT**YY**V**N**HESRRTQWKRPSPDDDL	D10714
Nedd4/Mouse-2	?	SG**LP**P**GWE**EKQDDR.GRS**YY**V**D**HNSKTTTWSKPTMQDDP	D10714
Nedd4/Mouse-3	?	GP**LP**P**GWE**ERTHTD.GRVFFI**N**HNIKKTQWEDPRLQNVA	D10714
RSP5/Yeast-1	228	GR**LP**P**GWE**RRTDNF.GRT**YY**V**D**HNTRTTTWKRPTLDQTE	L11119
RSP5/Yeast-2	331	GE**LP**S**GWE**QRFTPE.GRA**Y**FV**D**HNTRTTTWVDPRRQQYI	L11119
RSP5/Yeast-3	387	GP**LP**S**GWE**MRLTNT.ARV**Y**FV**D**HNTKTTTWDDPRLPSSL	L11119
YKB2_Yeast-1	1	M...SI**W**KEAKDAS.GRI**YY**Y**N**TLTKKSTWEKPKELISQ	P33203
YKB2_Yeast-2	39	LLREN**GW**KAAKTAD.GKV**YY**Y**N**PTTRETSWTIPAFEKKV	P33203
Yo61_Caeel-1	49	PSVESD**W**SVHTNEK.GTP**YY**H**N**RVTKQTSWIKPDVLKTP	P34600
Yo61_Caeel-2	94	QPQQGQ**W**KEFMSDD.GKP**YY**Y**N**TLTKKTQWVKPDGEEIT	P34600
Amoe/Amoeba	?	KMSVD**GW**KQYFTAE.GNA**YY**Y**N**EVSGETSWDPPSSLQSH	D90360
FE65/Rat	12	SD**LP**A**GW**MRVQDTS.GTY**YW**HIP.TGTTQWEPPGRASPS	X60468
ess1_Yeast	29	TG**LP**TP**W**TARYSKSKKREY**FF**N**P**ETKHSQWEEPEGTNKD	P22696
Consensus line:		LPtG**WE** ttt Gt YYhNH TtTTtW tPt	
Secondary struct.:		lllll bbb llll bbbbb llll lllllllll	

Figure 2

Alignment of selected WW domains. Protein codes are taken from the SWISSPROT database if available (Yo61, hypothetical protein from *C. elegans* chromosome III; YKB2, hypothetical protein from yeast chromosome XI; Dmd, dystrophin; Utro, utrophin; Amoe, hypothetical protein from *Acanthamoeba*). The consensus line displays conserved features (capitals, conserved amino acids; h, hydrophobic; t, turn-like or polar). Amino acids conserved in at least 60% of the sequences are shown in bold. The secondary structures were predicted using the program PHD[11] (b, β-strand; l, loop; not assigned, nearly equal preference for both β-strand and loop). All segments have a probability of less than 10^{-7} of matching the alignment by chance (computed using the MoST program[12]) except FE65. The WW domain in FE65 is, however, 38% identical to that in YAP and is therefore included.

Stop press

Recently, we became aware of independent efforts to delineate the WW domain, and would like to thank S. Kumar and P. Pavlik for communicating to us partial similarities between YAP, Nedd4 and RSP5, as well as B. André and J. Y. Springael for pointing out a mistake in the alignment. RSP5 is also called PIP1 (P. Pavlik *et al.*, unpublished) and NPI1 (C. Hein *et al.*, unpublished).

References

1 Ahn, A. H. and Kunkel, L. M. (1993) *Nature Genetics* 3, 283–291

2 Koenig, M. and Kunkel, L. M. (1990) *J. Biol. Chem.* 265, 4560–4566

3 Altschul, S. F. *et al.* (1990) *J. Mol. Biol.* 215, 403–410
4 Gribskov, M. *et al.* (1987) *Proc. Natl Acad. Sci. USA* 84, 4355–4358
5 Rohde, K. and Bork, P. (1993) *CABIOS* 9, 183–189
6 Sudol, M. (1994) *Oncogene* 9, 2145–2151
7 Kumar, S. *et al.* (1992) *Biochem. Biophys. Res. Commun.* 185, 1155–1161
8 Hanes, S. D. *et al.* (1989) *Yeast* 5, 55–72
9 Duilio, A. *et al.* (1991) *Nucleic Acids Res.* 19, 5269–5274
10 Tinsely, M. J. *et al.* (1994) *Proc. Natl Acad. Sci. USA* 91, 8307–8313
11 Rost, B. and Sander, C. (1994) *Proteins* 19, 55–72
12 Tatusov, R. L., Altschul, S. F. and Koonin, E. V. *Proc. Natl Acad. Sci. USA* (in press)

PEER BORK

EMBL, Meyerhofstrasse 1, D-69117 Heidelberg, Germany and MDC, 13125 Berlin-Buch, Germany.

MARIUS SUDOL

Rockefeller University, 1230 New York Avenue, New York, NY 10021-6399, USA.

Proteasome sequences in eubacteria

The 26 S proteasome is an ATP-dependent protease that is central to the ubiquitin-dependent pathway of protein degradation[1–3] and has a highly conserved structure from slime moulds to humans. The elongated 45 nm molecule is formed by a barrel-shaped 20 S core complex and two asymmetric polar 19 S complexes[4]. The 20 S complex, also known as MCP or multicatalytic proteinase, is an ATP-independent protease consisting of four seven-membered rings that are formed by 14 related but different subunits. These fall into two families, α-type subunits, forming the outer rings, and β-type subunits, forming the inner rings of the complex. The α-type subunits are distinguished by a highly conserved amino-terminal extension of approximately 35 residues, and by a completely conserved RPxG motif. Otherwise, their sequence identity is very low and limited to only five additional positions, two of which are glycines. The β-type subunits are even more divergent and have only four invariant residues, three of which are also present in α-type subunits. The hallmark of β-type subunits is a highly variable pro-sequence[5] that may be anywhere from 4 to over 70 residues long (but appears to be missing in C7-type subunits) and that is cleaved off during proteasome assembly after an invariant glycine.

So far, proteasomes and ubiquitin have only been found in eukaryotes, where they are ubiquitous, and in the archaebacterium *Thermoplasma acidophilum*[6,7], where the 26 S complex has not been identified yet and the 20 S complex is built by only two proteins, α and β, whose sequences are approximately 30% identical to eukaryotic α- and β-type sequences[5,6]. Efforts to detect proteasomes in other prokaryotes have failed so far[8], leading to the hypothesis that a *Thermoplasma*-type organism may have been the precursor of the eukaryotic cytoplasm.

Although a variety of biochemical methods has failed to detect proteasomes in eubacteria, recent genomic sequencing has uncovered four sequences that are significantly related to β-type proteasomal subunits: HslV from *Escherichia coli*[9], LapC from *Pasteurella haemolytica*[10], HslV from *Bacillus subtilis* (Genbank Accession No. Z33639) and PrcB from *Mycobacterium leprae* (Genbank Accession No. U00017) (Figs 1, 2). Alignment with mature β-type subunits shows that of these, only PrcB contains a pro-sequence (Fig. 1). Although the sequences are only remotely related to the β-type family, three blocks of sequence (the amino-terminal 30 residues, a GxxxD motif and a GSG motif) can be aligned to the eukaryotic sequences with probabilities of occurring by chance alone of less than 10^{-15}, as computed by MACAW[11]. The alignment can be extended over almost the entire length of the eubacterial sequences, with probabilities of chance occurrence between 10^{-2} and 10^{-5} for individual ungapped blocks. In pairwise sequence comparisons, the eubacterial sequences

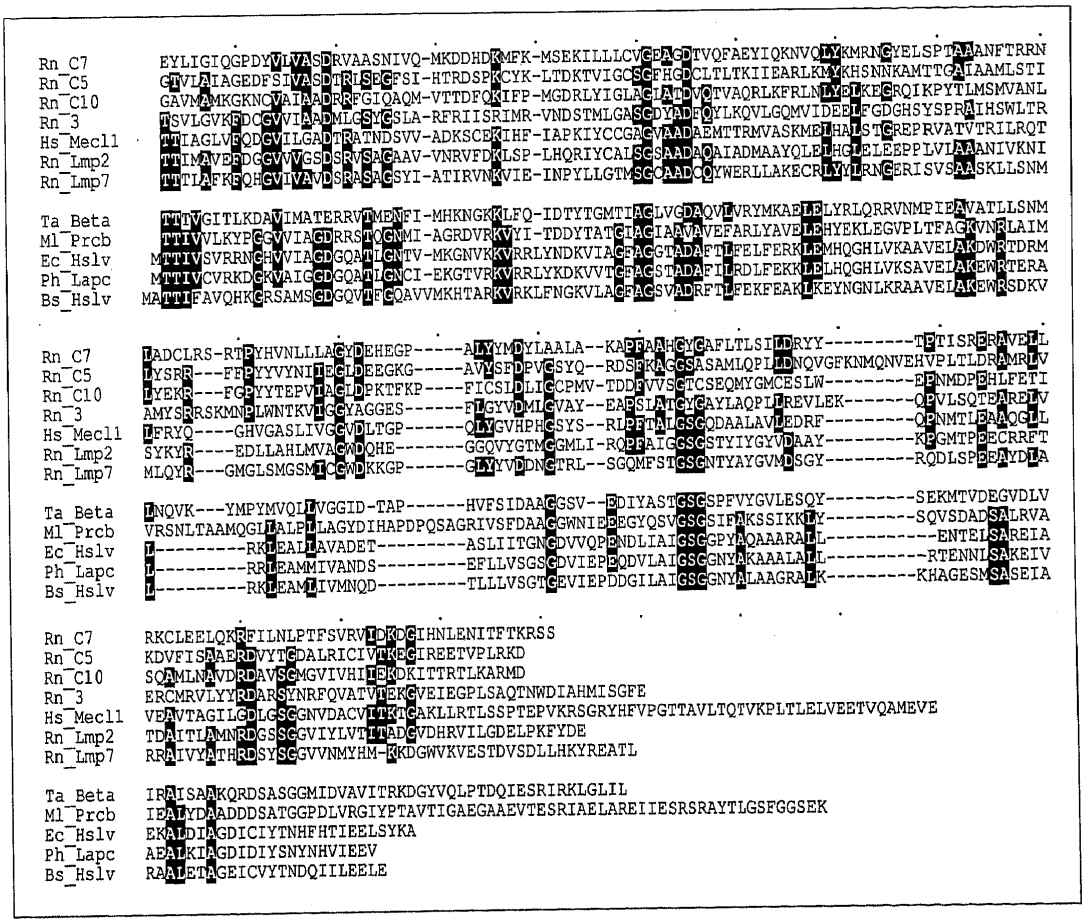

Figure 1

Alignment of 20 S β-type subunits. Eukaryotic and prokaryotic sequences are shown in separate blocks. The seven eukaryotic subfamilies are represented by rat sequences (*Rattus norvegicus*, Rn) and, where not available, by human sequences (*Homo sapiens*, Hs). The prokaryotic sequences are the β-subunit from *Thermoplasma acidophilum* (Ta), PrcB from *Mycobacterium leprae* (Ml), HslV from *Escherichia coli* (Ec), LapC from *Pasteurella haemolytica* (Ph) and HslV from *Bacillus subtilis* (Bs). Eukaryotic and archaebacterial sequences are shown in their mature form and PrcB is shown in its presumed mature form (residues 57–291). Residues in reverse type are >50% conserved within eukaryotic sequences and >75% conserved within prokaryotic sequences (a higher cutoff was necessary for prokaryotic sequences since HslV and LapC are closely related). The alignment was obtained using MACAW[11].