Protein Science (1998), 7:1477–1480. Cambridge University Press. Printed in the USA.

Structure and distribution of pentapeptide repeats in bacteria

ALEX BATEMAN,[1,2] ALEXEY G. MURZIN,[1] AND SARAH A. TEICHMANN[1]
[1]Medical Research Council Laboratory of Molecular Biology, Hills Road, Cambridge, CB2 2QH, United Kingdom
[2]The Sanger Centre, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SA, United Kingdom

(RECEIVED January 22, 1998; ACCEPTED March 10, 1998)

Abstract

We report the discovery of a novel family of proteins, each member contains tandem pentapeptide (five residue) repeats, described by the motif A(D/N)LXX. Members of this family are both membrane bound and cytoplasmic. The function of these repeats is uncertain, but they may have a targeting or structural function rather than enzymatic activity. This family is most common in cyanobacteria, suggesting a function related to cyanobacterial-specific metabolism. Although no experimental information is available for the structure of this family, it is predicted that the tandem pentapeptide repeats will form a right-handed β-helical structure. A structural model of the pentapeptide repeats is presented.

Keywords: β-helix; cyanobacteria; dot plot; motif; pentapeptide; structure

The complete sequence of cyanobacterium *Synechocystis* species strain PCC6803 (Kaneko et al., 1996) contains 16 proteins that have large regions consisting of a novel pentapeptide (five-residue) repeat (see Table 1). The most striking feature of this repeat is its regularity. The repeat is found in 13 to 60 tandem copies, and contains no interruption of its five-residue periodicity in all but one of these proteins. The regularity can be clearly seen in a self-self dot plot comparison of these proteins (see Fig. 1A).

The pentapeptide repeat was first found in the hglK protein from the *Anabaena* species cyanobacterium (Black et al., 1995). No other similar proteins were found at this time.

We initially defined the pentapeptide repeat protein family by searches with hidden Markov models (HMMs) (Krogh et al., 1994; Eddy, 1996) against the predicted cyanobacterial protein sequences. A single member of this family (SLR1819) can as a query with WU-blast (Altschul et al., 1990) find all the other cyanobacterial members with *p*-values lower than 10e-8. An alignment of some representative members from cyanobacterial proteins is shown in Figure 1B.

An alignment of 352 individual pentapeptide repeats was constructed and the frequency of each amino acid was calculated for each of the five positions in the repeat; these data are shown in Table 2. In the pentapeptide repeats no position is completely conserved, but the consensus of the repeats can be expressed as *A(N/D)L*X*, where *X* is any amino acid, the asterisk denotes a polar residue and the second position is either asparagine or aspartate. Proline never occurs at positions 1–3, which suggests that these residues may make main-chain hydrogen bonds, which proline could not form. The residue at position 1 has a small side-chain volume, compared with the residue at position 3 that rarely has a small side chain.

Structural model of the pentapeptide repeat

Here we predict (1) the conformation of individual repeats, (2) the number of repeats per turn of β-helix, and (3) the chirality of the β-helix and suggest a model of its structure.

Each residue in positions 1–2 is predicted to be in β-conformation with the side chains in positions 1 and 3 pointing inside the β-helix. There are only two ways to turn the main chain sharply between β-strands. In one of these, position 4 adopts an α_R conformation and position 5 is in a β-conformation $\alpha_R\beta$ turn). Alternatively, position 4 remains in a β-conformation and position 5 adopts an α_L conformation ($\beta\alpha_L$ turn). These two ways differ mainly by a 180° flip of the peptide group between positions 4 and 5. This peptide group can hydrogen bond to its counterparts in the adjacent turns of the β-helix. The formation of the hydrogen bonds is compatible with the $\beta\alpha_L$ turn, but not the $\alpha_R\beta$ turn, where it is hindered by collision of the side chain in position 4 with the main chain of the adjacent turn. Inspection of the known β-helix structures confirms the predominant occurrence of the $\beta\alpha_L$ turn in between sequential β-strands (Petersen et al., 1997).

The number of repeats per turn of β-helix is determined by the mean total volume of residues in the interior positions 1 and 3. This volume is approximately the same or less than the mean total volume observed for the interior residues in the left-handed single β-helices. These helices are made of hexapeptide repeats that each contribute two residues to the interior, and they have three repeats in each turn. Using CPK models of the pentapeptide repeats, we

Reprint requests to: Alex Bateman, The Sanger Centre, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, United Kingdom; e-mail: agb@sanger.ac.uk.

Table 1. *Proteins containing pentapeptide repeats*[a]

Gene name	Length	Number of repeats	TrEMBL accession
SLR0516	166	22(66%)	Q55837
SLR0719	388	19(24%)	Q55201
SLR0967	150	19(63%)	P72857
SLR1152	331	44(66%)	P74221
SLR1519	245	37(76%)	P73963
SLR1697	574	17(15%)	P74297
SLR1819	331	60(91%)	P73709
SLR1851	162	25(62%)	P73063
SLL0183	259	18(35%)	Q55773
SLL0274	196	20(51%)	P74392
SLL0301	169	22(65%)	Q55531
SLL0414	286	24(42%)	Q55112
SLL0577	169	22(65%)	P74725
SLL1011	270	21(39%)	P73013
SLL1350	398	13(16%)	P73524
SLL1446	320	36(56%)	P74206
yisX	212	35(83%)	O06733
YYBG_BACSU	279	22(40%)	P37497
YJCF_ECOLI	430	42(49%)	P32704
MCBG_ECOLI	187	29(78%)	P05530
Q52118	295	43(73%)	Q52118

[a] The third column gives the number of repeats in the protein followed by the percentage of the total polypeptide that this represents. The final column gives the accession number of the sequence in the trEMBL database (Bairoch & Apweiler, 1997).

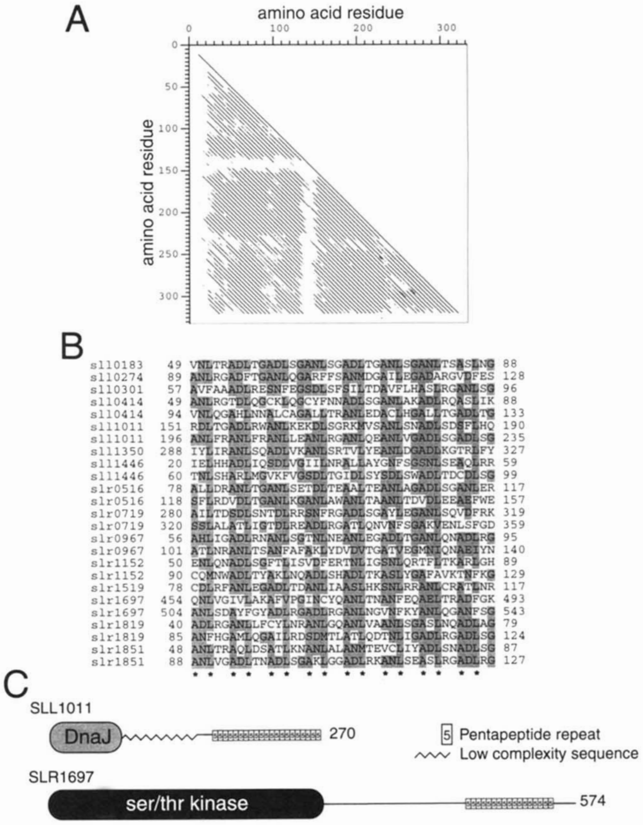
A
amino acid residue
0
100
200
300
amino acid residue
50
100
150
200
250
300
B
sll0183 49 VNLTRADLTGADLSGANLSGADLTGANLSGANLTSASLNG 88
sll0274 89 ANLRGADFTGANLQGARFFSANMDGAILEGADARGVDFES 128
sll0301 57 AVFAAADLRESNFEGSDLSFSILTDAVFLHASLRGANLSG 96
sll0414 49 ANLRGTDLQGCKLQGCYFNNADLSGANLAKADLRQASLIK 88
sll0414 94 VNLQGAHLNNALCAGALLTRANLEDACLHGALLTGADLTG 133
sll1011 151 RDLTGADLRWANLKEKDLSGRKMVSANLSNADLSDSFLHQ 190
sll1011 196 ANLFRANLFRANLLEANLRGANLQEANLVGADLSGADLSG 235
sll1350 288 IYLIRANLSQADLVKANLSRTVLYEANLDGADLKGTRLFY 327
sll1446 20 IELHHADLIQSDLVGIILNRALLAYGNFSGSNLSEAQLRR 59
sll1446 60 TNLSHARLMGVKFVGSDLTGIDLSYSDLSWADLTDCDLSG 99
slr0516 78 ALLDRANLTGANLSETDLTEAALTEANLAGADLSGANLER 117
slr0516 118 SFLRDVDLTGANLKGANLAWANLTAANLTDVDLEEAEFWE 157
slr0719 280 AILTDSDLSNTDLRRSNFRGADLSGAYLEGANLSQVDFRK 319
slr0719 320 SSLALATLIGTDLREADLRGATLQNVNFSGAKVENLSFGD 359
slr0967 56 AHLIGADLRNANLSGTNLNEANLEGADLTGANLQNADLRG 95
slr0967 101 ATLNRANLTSANFAFAKLYDVDVTGATVEGMNIQNAEIYN 140
slr1152 50 ENLQNADLSGFTLISVDFERTNLIGSNLQRTFLTKARLGH 89
slr1152 90 CQMNWADLTYAKLNQADLSHADLTKASLYGAFAVKTNFKG 129
slr1519 78 CDLRFANLEGADLTDANLIAASLHKSNLRRANLCRATLNR 117
slr1697 454 QNLVGIVLAKAFVPGINCYQANLTNANFEQAELTRADFGK 493
slr1697 504 ANLSDAYFGYADLRGADLRGANLNGVNFKYANLQGANFSG 543
slr1819 40 ADLRGANLLFCYLNRANLGQANLVAANLSGASLNQADLAG 79
slr1819 85 ANFHGAMLQGAILRDSDMTLATLQDTNLIGADLRGADLSG 124
slr1851 48 ANLTRAQLDSATLKNANLALANMTEVCLIYADLSNADLSG 87
slr1851 88 ANLVGADLTNADLSGAKLGGADLRKANLSEASLRGADLRG 127
* * * * * * * * * * * * * * * *
C
SLL1011
DnaJ
270
5 Pentapeptide repeat
Low complexity sequence
SLR1697
ser/thr kinase
574

Fig. 1. Characterization of the pentapeptide repeat family. **A:** A self-self dot plot of SLR1819 using dotter (Sonnhammer & Durbin, 1995) with a window size of 20 residues. **B:** Alignment of pentapeptide repeats from cyanobacterial proteins. Asterisks mark conserved columns. Each entry in the alignment gives the gene name followed by the number of the first residue in the alignment. The final column gives the end point in the sequence. **C:** Pentapeptide repeat proteins linked to nonhomologous domains.

found that three repeats in $\beta\beta\beta\beta\alpha_L$ conformation per turn can be packed well, whereas four repeats per turn would probably have a loose packing in the interior.

The chirality of the pentapeptide β-helix is probably determined by the packing of side chains in the interior position 3, the closest to the axis of the β-helix. These side chains are likely to adopt the same conformation and pack in a helical structure with the same chirality and pitch as the main-chain helix. In the hexapeptide repeats, the analogous position is occupied predominantly by the Cβ-branched residues: isoleucine (14%) and valine (64%). These residues adopt very similar conformations and they are tightly packed in a left-handed β-helix. In contrast, in the pentapeptide repeats (Table 2), this position is predominantly occupied by leucine (75%), then phenylalanine (13%), and methionine (4%), whereas valine and isoleucine together are found in this position in 3% of cases.

Several proteins with repetitive sequences have had their structure solved by X-ray crystallography (Yoder et al., 1993; Emsley et al., 1996; Gorina & Pavletich, 1996). A common feature of all these structures is a superhelical arrangement of secondary structures (Kobe, 1996). Hexapeptide repeats in bacterial proteins have

Table 2. *The amino acid composition of pentapeptide repeats*[a]

Position	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
1	**68.5**	2.3	0.0	0.3	0.6	1.1	0.0	2.6	0.3	1.1	0.6	0.0	0.0	0.3	0.6	**6.0**	**8.0**	**8.0**	0.0	0.0
2	0.9	0.9	**27.8**	2.3	2.6	0.0	1.13	4.3	3.7	3.4	0.3	**38.4**	0.0	1.1	3.4	3.4	2.8	1.4	0.0	2.2
3	2.0	1.4	0.0	0.0	**13.4**	0.3	0.0	1.4	0.0	**75.0**	4.2	0.0	0.0	0.0	0.0	0.0	0.6	1.7	0.0	0.0
4	**5.1**	0.6	2.8	**6.5**	1.4	2.6	2.3	4.5	4.5	1.4	1.4	5.4	0.6	**6.8**	**14.2**	**17.3**	**16.2**	3.1	0.6	2.5
5	3.4	0.0	**6.8**	**10.5**	1.7	**36.4**	2.6	0.0	**5.1**	1.4	0.0	**7.1**	0.3	5.4	**11.6**	2.6	0.3	0.0	1.7	3.1
Swiss-prot	8.7	3.6	4.5	4.9	3.6	9.4	3.6	3.3	8.6	8.1	1.3	4.1	5.2	3.7	3.7	7.1	5.9	6.2	1.0	3.0

[a] The percentage composition for each amino acid is shown for each position of the pentapeptide repeats. The bottom row shows the residue composition of SWISS-PROT (Bairoch & Apweiler, 1997) for comparison. The most commonly found residues (greater than 5%) at each position in the repeat are shown in bold type.

been found to form a left-handed β-helix (Raetz & Roderick, 1995). We suspect that the pentapeptide repeats form a superhelical structure. A model of this repeat has been made in which the pentapeptide repeats form a three sided right-handed parallel β-helix (see Fig. 2). Each pentapeptide forms one β-strand, with three pentapeptides making a single turn of the β-helix.

Given our initial assumption that the pentapeptide repeats form a parallel β-helix, we could test whether this was a plausible structure given the sequence data and the known β-helix structures. Positions 1 to 3 have a residue conservation characteristic of β-strands. To make the tight turn between strands either position 4 or 5 needs to be in a turn conformation. Examination of these two models ruled out the possibility that position 4 was the turn residue. In the model with position 4 in a turn conformation, the side chain of this position clashed with the main chain of the next turn of β-helix. A β-left conformation for position 5 leads to a sterically plausible model, in which the side chains form favorable stacking interactions. The β-left conformation is energetically unfavorable for residues other than glycine and asparagine, which agrees with the residues observed in this position (see Table 1). Our model with three repeats per turn of superhelix contains tight packing of side chains. A model of four repeats per turn could be constructed; however, the core of this structure would be loosely packed. Our final model was built using the software package modeler (Sali & Blundell, 1993), based on pentapeptide repeats of the sequence ANLSG.

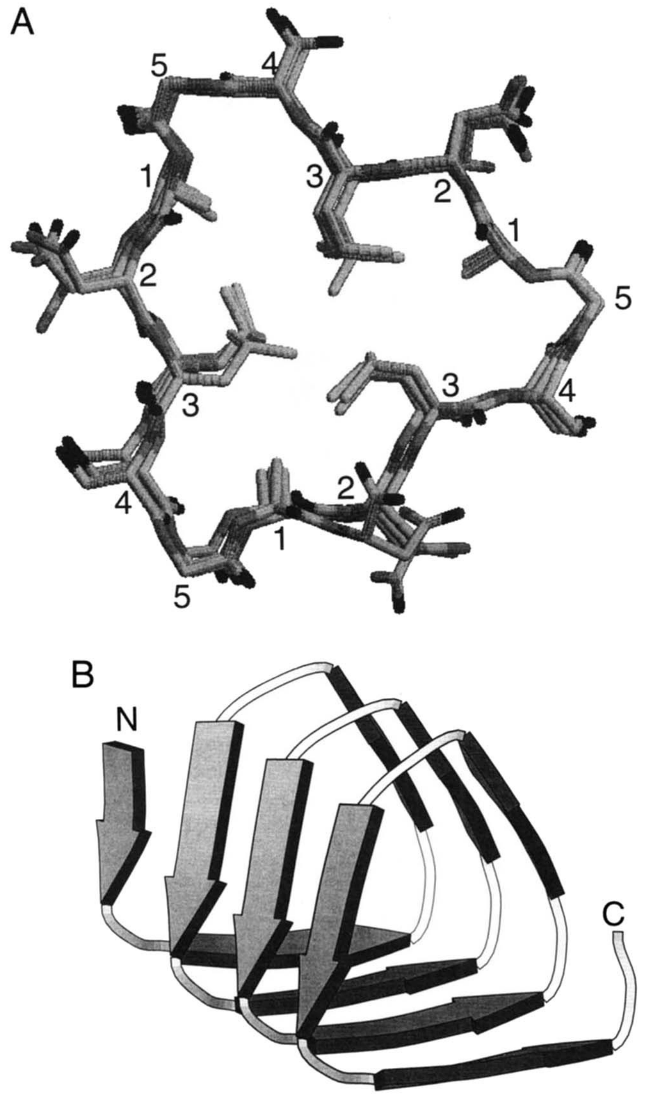

Fig. 2. A structural model of the pentapeptide repeats. **A:** A view down the superhelical axis with the carboxyl terminus nearest the reader. Each position of the repeat is numbered. This picture was produced using rasmol (Sayle & Milner-White, 1995). **B:** A molscript (Kraulis, 1991) picture of the predicted β-helix for the pentapeptide repeats. Each pentapeptide forms one side of the triangular cross section of the helix.

Our structural model contains a higher order repeat of 15 residues per turn of β-helix. However, a Fourier analysis (McLachlan & Stewart, 1976) of the pentapeptide repeat proteins does not find any evidence for a 15 residue periodicity, or any other periodicity than 5 residues. The original analysis of the pentapeptide repeats in the hglK protein only suggested a periodicity of 9 pentapeptide repeats (Black et al., 1995), but this is not found to be significant in our analysis of all pentapeptide repeat proteins. We also looked at periodicities in the left-handed β-helical hexapeptide repeat proteins (Raetz & Roderick, 1995), and did not find any higher order repeat than the hexapeptide motif. This indicates that it is not necessary to exhibit a higher order sequence repeat in order to fold into a β-helical structure. The β-helical hexapeptide repeat is known to form functional trimers. It seems possible that the pentapeptides may oligomerize in vivo. The trimerization interface of the hexapeptide repeats contains zinc ions between the β-sheets of different monomers. The monomers interact via hydrogen bonds and salt bridges in a manner lacking periodicity that could also be formed by the pentapeptide repeats. The oligomerization state of these proteins could be tested to see if this were indeed the case.

Two common features in β-helices are asparagine ladders and aliphatic stacks, where residues from one turn of helix interact with residues from adjacent turns. Position 2 of the pentapeptide repeats commonly contains asparagine that may form ladder interactions. The side chain of position 2 is on the exterior of the β-helix in our model, unlike the ladders of other β-helix structures, where the ladders are in the interior of the protein. The side chain of leucine at position three forms aliphatic stacks in our model.

Pentapeptide repeats in other bacteria

Pentapeptide repeats were found in other bacterial proteins using sequence comparison. Searches of the SWIR nonredundant database found other examples of this repeat. *Bacillus subtilis* contains two pentapeptide repeat proteins, YisX (TREMBL:O06733) and YYBG_BACSU (SWISS:P37497) (Ogasawara et al., 1994). Neither of these two proteins is clearly orthologous to the cyanobacterial proteins. In *Escherichia coli* there are also two pentapeptide repeat proteins, YJCF_ECOLI (SWISS:P32704) and McbG (SWISS:P05530) (Garrido et al., 1988; Blattner et al., 1993). The function of YJCF_ECOLI is unknown, but it has been found that McbG provides some antibiotic resistance to microcin B17. In cells producing microcin, loss of McbG causes slower growth and pronounced induction of the SOS response. The SOS response is normally induced by damage to DNA, whereas these affects are due to microcin inhibiting DNA replication. Therefore, McbG may bind to microcin and inhibit its action. In bacterium *Erwinia stewartii*, a single pentapeptide repeat protein of unknown function has been sequenced (TREMBL:Q52118).

Function and subcellular localization of pentapeptide repeat proteins

The proteins containing the pentapeptide repeats are of unknown molecular function. However, in two of the proteins the repeats are linked to other nonhomologous domains (see Fig. 1C). SLL1011 contains a DnaJ-like domain at its amino terminus. DnaJ domains associate with hsp70 heatshock proteins. SLL1011 also contains a region of low sequence complexity as defined by the SEG program with default parameters (Wootton, 1994). The cyanobacteria *Synechocystis* species contains eight different proteins with DnaJ domains. SLR1697 is a protein kinase that contains pentapeptide repeats at its carboxyl terminus. The kinase domain of SLR1697 is most similar to the kinases of other prokaryotes. These kinases share the most similarity with serine threonine kinases of eukaryotes. Both DnaJ and kinase domains are cytoplasmic, and as there is no evidence of a transmembrane region in either protein we suggest a cytoplasmic subcellular location for these two pentapeptide repeat proteins. However, for the majority of these proteins, we have one or more transmembrane spans predicted using TmPred (Hofmann & Stoffel, 1993) in the nonpentapeptide regions. Whether the pentapeptide repeats are also found in the extracellular environment is unclear.

HglK is a pentapeptide repeat containing protein that is involved in localization of heterocyst glycolipids in *Anabaena* species cyanobacterium (Black et al., 1995). The HglK protein is anchored in the membrane by four transmembrane helices at the amino terminus of the protein. In heterocysts, the specialized cells that fix nitrogen, specific glycolipids are found in the outer membrane layer. In HglK null mutants these glycolipids are synthesized but not transported to the membrane. A mutation that truncates the carboxyl terminus of the HglK protein, which contains the pentapeptide repeats, has an indistinguishable phenotype from the null mutant. This suggests that the pentapeptide repeats are a functional region of the protein. Black et al. (1995) studied the ultrastructure of the HglK null mutant by electron microscopy. Mutants in nitrogen free media develop lacunae (distended areas between thylakoid membranes), which may be due to a defect in lipid transport or a structural component.

Conclusion

The large number of these repeat proteins in cyanobacteria suggests that they could be involved in metabolism specific to cyanobacteria, such as nitrogen fixation or photosynthesis. Given the regularity of the pentapeptide repeat and its probable lack of extended loops, it seems unlikely that it will have an enzymatic function. A structural role seems more likely given its predicted extended structure. We hope that the structure and function will soon be elucidated for this unusual family of proteins

Acknowledgments

We would like to thank Andrew McLachlan for use of his Fourier analysis program. AGB is funded by a grant from the Wellcome Trust, AGM is an MRC senior fellow and SAT has a predoctoral fellowship from the Boehringer Ingelheim Fonds.

References

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol 215*:403–410.

Bairoch A, Apweiler R. 1997. The SWISS-PROT protein sequence data bank and its supplement trEMBL. *Nucleic Acids Res 25*:31–36.

Black K, Buikema W, Haselkorn R. 1995. The hglK gene is required for localization of heterocyst-specific glycolipids in the cyanobacterium *Anabaena* sp. strain PCC7120. *J Bacteriol 177*:6440–6448.

Blattner F, Burland V, Plunkett G III, Sofia H, Daniels D. 1993. Analysis of the *Escherichia coli* genome. IV. DNA sequence of the region from 89.2 to 92.8 minutes. *Nucleic Acid Res 21*:5408–5417.

Eddy SR. 1996. Hidden Markov models. *Curr Opin Struct Biol 6*:361–365.

Emsley P, Charles I, Fairweather N, Isaacs N. 1996. Structure of bordetella pertussis virulence factor P.69 pertactin. *Nature 381*:90–92.

Garrido MdC, Herrero M, Kolter R, Moreno F. 1988. The export of the DNA replication inhibito Microcin B17 provides immunity for the host cell. *EMBO J 7*:1853–1862.

Gorina S, Pavletich N. 1996. Structure of the p53 tumor suppressor bound to the ankyrin and SH3 domains of 53BP2. *Science 274*:1001–1005.

Hofmann K, Stoffel W. 1993. TMbase: A database of membrane spanning proteins segments. *Biol Chem Hoppe-Seyler 347*:166.

Kaneko T, Sato S, Kotani H, Tanaka A, Asamizu E, Nakamura Y, Miyajima N, Hirosawa M, Sugiura M, Sasamoto S, Kimura T, Hosouchi T, Matsuno A, Muraki A, Nakazaki N, Naruo K, Okumura S, Shimpo S, Takeuchi C, Wada T, Watanabe A, Yamada M, Yasuda M, Tabata S. 1996. Sequence analysis of the genome of the unicellular *Cyanobacterium synechocystis* sp. strain PCC6803. II. Sequence determination of the entire genome and assignment of potential protein-coding regions. *DNA Res 3*:109–136.

Kobe B. 1996. Leucines on a roll. *Nat Struct Biol 3*:977–980.

Kraulis P. 1991. MOLSCRIPT: A program to produce both detailed and schematic plots of protein structures. *J Appl Crystallography 24*:946–950.

Krogh A, Brown M, Mian IS, Sjolander K, Haussler D. 1994. Hidden Markov models in computational biology. *J Mol Biol 235*:1501–1531.

McLachlan A, Stewart M. 1976. The 14-fold periodicity in alpha-tropomyosin and the interaction with actin. *J Mol Biol 103*:271–298.

Ogasawara N, Nakai S, Yoshikawa H. 1994. Systematic sequencing of the 180 kilobase region of the *Bacillus subtilis* chromosome containing the replication. *DNA Res 1*:1–14.

Petersen T, Kauppinen S, Larsen S. 1997. The crystal structure of rhamnogalacturonase A from *Aspergillus aculeatus*: A right handed parallel β helix. *Structure 5*:533–544.

Raetz C, Roderick S. 1995. A left handed parallel beta helix in the structure of UDP-N-acetylglucosamine acyltransferase. *Science 270*:997–1000.

Sali A, Blundell TA. 1993. Comparative protein modelling by satisfaction of spatial restraints. *J Mol Biol 234*:779–815.

Sayle R, Milner-White E. 1995. RASMOL: Biomolecular graphics for all. *Trends Biochem Sci 1995*:374.

Sonnhammer E, Durbin R. 1995. A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis. *Gene 167*:GC1–GC10.

Wootton JC. 1994. Sequences with "unusual" amino acid compositions. *Curr Opin Struct Biol 4*:413–421.

Yoder M, Keen N, Jurnak F. 1993. New domain motif: The structure of pectate lyase C, a secreted plant virulence factor. *Science 260*:1503–1507.