FEBS Letters 414 (1997) 532–536 FEBS 19212

Prediction of substrate-specific pockets in cyclosporin synthetase

Holger Husi[a], Kurt Schörgendorfer[b], Günter Stempfer[b], Paul Taylor[a], Malcolm D. Walkinshaw[a,*]

[a]*Structural Biochemistry Group, Department of Biochemistry, The University of Edinburgh, Michael Swann Building, King's Buildings, Edinburgh EH9 3JR, UK*
[b]*Biochemie GmbH, 6330 Kufstein-Schaftenau, Austria*

Received 28 July 1997; revised version received 13 August 1997

Abstract Amino acid sequence comparisons between domains of cyclosporin synthetase have been used to identify regions of the sequence which are responsible for the recognition and binding of the individual amino acids. Using a limited set of selection rules it was possible to identify three amino acid positions in the subdomain sequences which are responsible for amino acid specificity. Homology with the firefly luciferase protein shows that these three key residues are close to each other and line the surface of a putative specific substrate binding pocket located on the amino acyl-adenylation subdomain. These results allow us to predict a large number of cyclosporin synthetase mutants which could be used to synthesise alternative cyclosporin-like peptides.

Key words: Cyclosporin A synthetase; Comparative sequence homology alignment; Active site prediction

1. Introduction

Cyclosporin synthetase (CySyn) is the most complex of a constantly rising number of peptide synthetases known so far [1]. It is a single chain polypeptide consisting of 15281 amino acids with a deduced molecular mass of 1.69 MDa [2]. The protein is produced by the fungus *Tolypocladium inflatum Gams* for the non-ribosomal synthesis of the undecapeptide cyclosporin A (CsA) (Fig. 1) [3]. CsA is used clinically as an immunosuppressant to prevent organ rejection following transplant operations [4].

The sequence of CySyn shows a repeating structure with 11 homologous modules responsible for the incorporation of each of the 11 amino acids in the product (Fig. 1) and a C-terminal 'twelfth domain' putatively used for the cyclisation reaction of CsA.

The bacterial antibiotics including gramicidin S, tyrocidine, surfactin and bacitracin, and the fungal peptides HC-toxin, enniatin A and cyclosporin A are all synthesised by multifunctional enzymes which have been characterised and reviewed [1,5]. Comparison of the amino acid sequences of the different synthetases defined a multidomain architecture [2]. The 11 large modules of CySyn are responsible for the adenylation, thioesterification and (for seven of the amino acids) *N*-methylation, of each amino acid in the growing chain. It has been shown that the first domain in CySyn is responsible for the synthesis of D-Ala[8] (Fig. 1) and that the peptide product is synthesised in a sequential and stepwise fashion [3].

This paper describes the identification of the site-specific amino acid binding pockets by dissecting the cyclosporin synthetase protein into its various domains followed by cross-comparing these subdomains with each other.

2. Materials and methods

2.1. Homology searches and protein alignment

Initial CySyn subdomain homology searches were done using a BLASTP search engine [6] on a non-redundant protein sequence database (http://www.genome.ad.jp). Protein sequences of acyl-adenylation proteins and acyl carrier proteins were extracted from the GENPEPT database (http://www.genome.ad.jp) using the program MPSRCH [7]. These sequences were used as an input to the PHD program [8] and used for the secondary structure prediction. The program SSS_align [9] incorporated the secondary structural information with sequence information to give a good alignment of the ACP subdomains and the amino acyl-adenylation subdomains with the secondary structures of their respective template proteins.

2.2. Molecular modelling

The molecular graphics program WITNOTP (A. Widmer, unpublished) was used to generate pictures of the luciferase structure showing the location of the amino acids responsible for providing specificity of amino acid substrate binding to the amino acyl-adenylation domain.

3. Results and discussion

3.1. Subdomain architecture of cyclosporin synthetase

A comparison of bacterial and fungal synthetases has led to the identification of core modules [1], which have some degree of evolutionary conservation [10]. Most important was the finding that each domain within the different modules of various synthetases can act independently and can perform consecutive sequential steps like multienzyme complexes [11]. Regions corresponding to substrate binding pockets for the synthetases have, however, not yet been identified.

The internal homology between the 11 repeating units of the CySyn sequence was determined using the program ALIGN [12] and initial boundaries between the different CsA synthetase domain modules were outlined. In order to predict the secondary structure effectively, the following procedure was used. A search of the GENPEPT database was made using the MPSRCH program [7] and 30 similar sequences for each of the modules were retrieved. These were transferred and used as input into the PHD prediction program [8,13]. The strength of this prediction algorithm is in having a wide range of related sequences which makes the boundary between regions of secondary structure more apparent. The program SSS_align [9] was then used to line up the sequence of the synthetase domains using both amino acid sequence information and secondary structure information. Comparison of the individual subdomains then allowed the dissection of the whole CySyn molecule into its modules and

*Corresponding author. Fax: (44) (131) 650 7055.

PII S0014-5793(97)01064-8

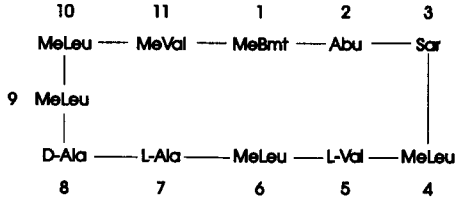

Fig. 1. Cyclosporin A, a natural undecapeptide from *Tolypocladium inflatum*.

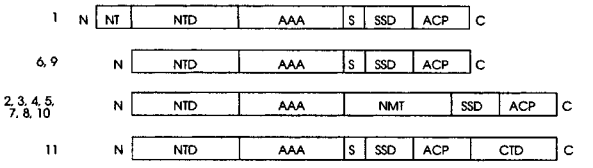

Fig. 2. Schematic diagram of the domain organisation of individual cyclosporin synthetase modules. The numbers to the left indicate the modules according to their arrangement in CySyn. Modules 1, 6, 9 and 11 all share the same sequential order including an N-terminal domain (NTD), the amino acyl-adenylation domain (AAA), a short spacer (S) followed by a synthetase-specific domain (SSD) and the acyl carrier protein homology region (ACP). In addition module 1 has an N-terminal extension (NT) and module 11 has an extra C-terminal domain (CTD). Modules 2, 3, 4, 5, 7, 8 and 10 have instead of the spacer an *N*-methyltransferase domain (NMT). Both N and C termini are depicted by N and C respectively.

domains and the definition of clear boundaries between the different modules. The results of this alignment are shown in Fig. 2 and summarised in Table 1. Amino acid identity between the domains is high and ranges between 50 and 54%.

The modular structure of CySyn can be divided essentially into two groups (Fig. 2). Group I modules (CySyn modules 1, 6, 9 and 11 which encode CsA amino acids 8, 2, 5 and 7 respectively) and the other modules where the corresponding positions in the final drug product are *N*-methylated. All CySyn modules start with an N-terminal domain (NTD) of about 460 amino acids, which may be involved in the elongation mechanism [14]. Database searches revealed no homology with other known proteins. Module 1 has a small 20 amino acid extension at its N-terminus. There is also a C-terminal domain at the end of module 11 which shows no homology with other domains.

The acyl-adenylation domain consists of about 450 amino acids and has significant homology with all peptide synthetases [1] as well as CoA synthetases and other adenylating enzymes including firefly luciferase [15]. Homology between these ATP binding domains is between 20% and 80% amino acid identity. There are 11 acyl-adenylation domains in cyclosporin synthetase and each is responsible for recognition of the cognate amino acid substrate and the formation of an activated amino acyl-adenylate by reaction of the amino acid carboxy group with bound ATP.

Seven of the 11 amino acids in CsA are *N*-methylated (Fig. 1) and each of the seven relevant CySyn modules has an additional *N*-methylation domain of about 450 amino acids at the C-terminus of the acyl-adenylation domain [2]. The other four modules have a short peptide linker of 13–37 amino acids.

In all 11 modules these domain stretches are followed by a synthetase-specific peptide with a length of approximately 54 amino acids, which could be found only in other peptide synthetases both from bacterial and fungal origin. The homology of this subdomain within the various modules varies between 50 and 70%, and the identity to other synthetases is approximately 50%. The biological function of this domain is not known.

All 11 CySyn modules have an acyl carrier protein (ACP) domain of 75 amino acids. ACP proteins are also involved in the synthesis of both polyketides and fatty acids [16], and three dimensional NMR structures have been elucidated for actinorhodin polyketide synthase ACP [17] and *Escherichia coli* fatty acid synthase ACP [18].

3.2. *Amino acid specificity pocket*

Invariant residues involved in ATP binding among a wide range of acyl adenylation proteins have been identified [19] and could also be found in all 11 CySyn domains. The cyclosporin A sequence has four occurrences of MeLeu, one occurrence of valine and one of MeVal. The mechanism for peptide synthesis is known to involve the recognition of the unmodified amino acid [20] which reacts with ATP to form an acyl-adenylate with the release of pyrophosphate. Such a P-O bond is formed between the amino acid and the AMP at each of the 11 modules. The specificity of the reaction is likely to depend

Table 1
Modular structure of cyclosporin synthetase and the substrate specificity of its activating units

Module[a]	CsA aa[b]	Cs(n) aa[c]	Length[d]	NTD[e]	AAA[f]	S[g]	NMT[h]	SSD[i]	ACP[j]
1	D-Ala	D-Ser	1104	21–481	482–935	936–973		974–1028	1029–1104
2	MeLeu	Leu	1498	1105–1567	1568–2022		2023–2471	2472–2526	2567–2602
3	MeLeu	Leu	1487	2603–3064	3065–3515		3516–3958	3959–4013	4014–4089
4	MeVal	Val	1492	4090–4550	4551–5005		5006–5450	5451–5505	5506–5581
5	MeBmt	MeLeu	1497	5582–6043	6044–6497		6498–6948	6948–7002	7003–7078
6	Abu	Ala, Thr, Val, Nva	1060	7079–7540	7541–7993	7994–8007		8008–8062	8063–8138
7	Sar		1495	8139–8601	8602–9055		9056–9502	9503–9557	9558–9633
8	MeLeu	Val	1497	9634–10095	10096–10546		10547–10999	11000–11054	11055–11130
9	Val	Nva	1072	11131–11584	11585–12036	12037–12071		12072–12126	12127–12202
10	MeLeu	Leu	1496	12203–12664	12665–13112		13113–13566	13567–13622	13623–13698
11	L-Ala	Abu	1075	13699–14162	14163–14615	14616–14641		14642–14697	14698–14773

All 11 modules were mapped against each other and other known synthetases to generate a more detailed insight of the modular structure of the cyclosporin synthetase.

[a]Module: cyclosporin synthetase module number; [b]CsA aa: amino acid incorporated in CsA at the given position; [c]Cs(n) aa: amino acid which can be found in other minor metabolites of cyclosporin at the given positions; [d]Length: total peptide length of the corresponding synthetase module; [e]NTD: N-terminal domain, its start and end positions within the whole protein are indicated; [f]AAA: amino acyl-adenylation domain; [g]S: spacer peptide; [h]NMT: *N*-methyltransferase domain; [i]SSD: synthetase-specific domain; [j]ACP: acyl carrier protein homology domain.

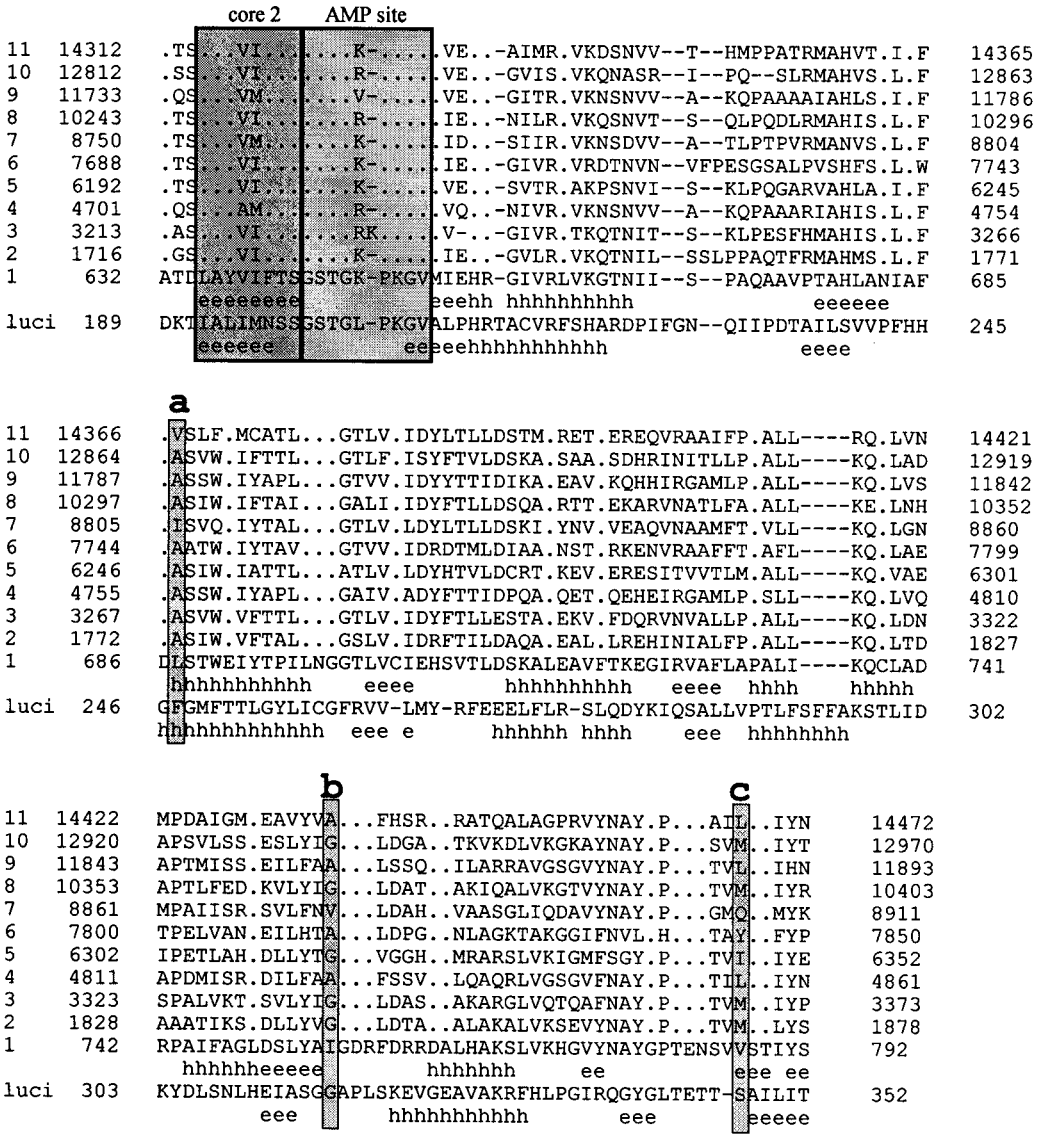

Fig. 3. Multiple amino acid sequence alignments of all 11 amino acyladenylation domains of cyclosporin synthetase and firefly luciferase. A site which was found to be homologous in other synthetases (core 2) and the AMP binding motif (AMP site) are both highlighted. Identical amino acids in all 11 CySyn domains are indicated in CySyn domains 2–11 by a period. Gaps and extended loops were introduced in the sequences for maximum alignments and are shown by a hyphen. The three sites which obey the amino acid specificity rules in order to find an active site are boxed and labelled a, b and c. The secondary structure prediction of CySyn domain 11 and the elucidated secondary structure of luciferase are included using 'h' for α-helices and 'e' for extended conformations. The amino acid positions are depicted on either side of the sequences.

on the shape of the surrounding enzymatic pocket of the amino acyl-adenylation domain. This suggests that the four MeLeu domains should share identical specificity pockets which are different from all the pockets of the other domains. The sequences of the 11 aligned domains were searched for occurrences of amino acids unique to the four MeLeu domains or unique to the two valine domains. In addition to CsA, the fungus *T. inflatum* produces many other minor metabolites of the same structural type [21]. Another constraint therefore included the cross-comparison of the other naturally occurring cyclosporins (i.e. CsB to CsZ). As shown in Table 1, the only amino acid used at position 3 of cyclosporin is glycine and a D-configured residue at position 8. Other residues in the drug can vary to some extent in the incorporated amino acid [21]. The aligned sequences were searched for occurrences of the following conditions:

1. An amino acid at a given position in the aligned sequence of domains 1 (D-Ala[8]) or 7 (Sar[3]) is different from amino acids in all other domains at the same aligned position.
2. Amino acids at a given position in the aligned sequences of domains 2 (MeLeu[9]), 3 (MeLeu[10]), 8 (MeLeu[4]) and 10 (MeLeu[6]) are identical to each other and different from amino acids in all other domains at the same aligned position.
3. Amino acids at a given position in the aligned sequences of domains 4 (MeVal[11]) and 9 (Val[5]) are identical to each other and different from amino acids in all other domains at the same aligned position.

Positions in the aligned CySyn sequence which meet these criteria are likely to play an important role in the recognition

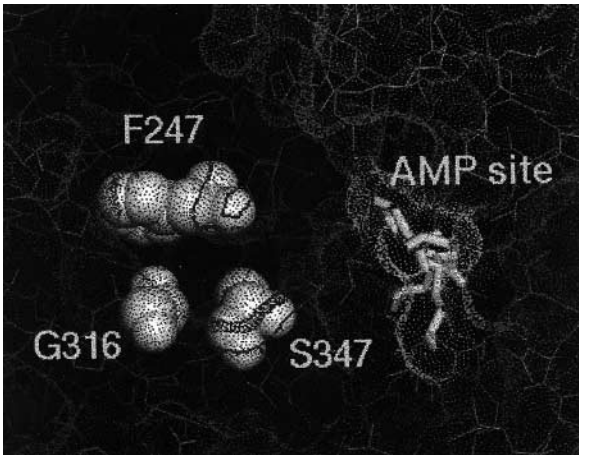

Fig. 4. A region of the X-ray structure of firefly luciferase showing the putative AMP site and the location of the three amino acids involved in substrate binding in the acyl-adenylation domains of cyclosporin synthetase which were identified using the selection rules described in the text. The dotted Connolly surface was calculated and displayed using the program WITNOTP (A. Widmer, unpublished).

and discrimination of amino acid binding to the amino acyl-adenylation domains.

Only three such amino acid positions were found in the complete sequence of the synthetase modules (Fig. 3). They are all found in the amino acyl-adenylation domain located near the ATP binding site. This strongly implies that these residues are crucially important for the recognition and binding of the amino acid substrate.

3.3. *Firefly luciferase homology with the amino acyl-adenylation domain*

The X-ray crystallographic structure of firefly luciferase was published during the course of this work [15]. This is of particular interest since this protein is known to have homology with the acyl-adenylation domain of peptide synthetases. The program SSS_align [9] was used to line up the sequence of the synthetase domain and the luciferase using both amino acid sequence information and secondary structure information. The results of this alignment are included in Fig. 3. The three amino acid positions found earlier were then mapped on the three-dimensional model of luciferase. Fig. 4 shows that the three sites are clustered together near the putative ATP binding site forming a pocket on the surface of luciferase. The three residues form an almost perfect equilateral triangle with the Cα atoms separated by between 6.1 Å and 6.9 Å. A stereo picture of these three residues in the X-ray structure of luciferase is shown in Fig. 5 (top).

After this article was submitted, the crystal structure of the N-terminal adenylation domain of gramicidin synthetase complexed with L-phenylalanine was published [23]. This 60 kDa domain has about 16% sequence identity to luciferase and superposition if the N-terminal domains of the two enzymes results in an r.m.s. deviation of 1.5 Å for over 250 pairs of equivalent Cα atoms [23]. The picture of the bound phenylalanine substrate in this structure shows that the three residues selected in our paper do indeed correspond to those forming the recognition pocket in the gramicidin synthetase structure.

3.4. *A recognition code*

The three amino acids co-defining the putative amino acid specificity pockets in each of the of CySyn domains are given in Table 2. These residues (labelled a, b, and c in Fig. 3) were derived from the simple selection rules described above and we suggest that they provide a recognition code for the amino acid substrates of the different CySyn amino acyl-adenylation domains. The 'recognition triplet' lining the MeLeu recognition pocket consists of alanine, glycine and methionine. These residues have been modelled into the X-ray structure of luciferase and a stereo picture is given in Fig. 5 (middle). It is significant that the size of the recognition triplet is inversely proportional to the size of the cognate substrate side chain. Thus the relatively bulky leucine substrate with five non-hydrogen atoms in the side chain fits into a relatively spacious pocket with the recognition triplet contributing only five non-hydrogen side-chain atoms (one from alanine and four from methionine). In contrast when the side chain is glycine, the recognition triplet (IVQ) is rather bulky and a total of 13 side chain atoms fill the recognition pocket. Fig. 5 (bottom) shows

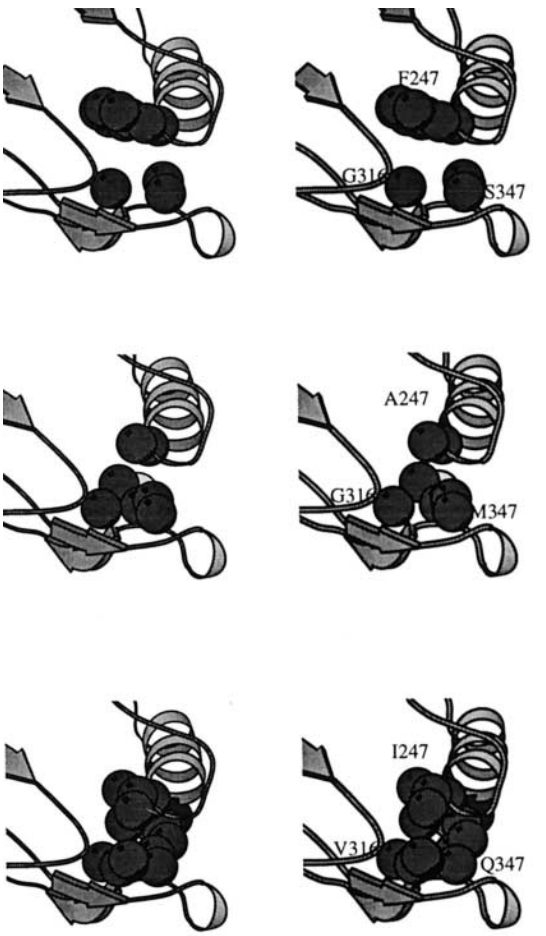

Fig. 5. Top: Stereo picture of a detail of the X-ray structure of firefly luciferase drawn using MOLSCRIPT [22] and showing the positions of the three residues defined to be most important in substrate side-chain recognition in the homologous domains of cyclosporin synthetase. Middle: As in top but showing the 'recognition triplet' for a cyclosporin leucine side-chain. The F247A and S347M mutations of the luciferase X-ray structure were carried out using the modelling program WITNOTP (A. Widmer, unpublished). Bottom: As in top but showing the 'recognition triplet' for sarcosine with the F247I, G316A and S347Q mutations.

Table 2
Amino acids involved in the specificity pockets of the various cyclosporin synthetase domains

	a	b	c
Leu	A	G	M
Val	A	A	L
Bmt	A	G	I
D-Ala	L	I	V
L-Ala	V	A	L
Gly	I	V	Q
Abu	A	A	Y
Luciferase	F	G	S

The three sites in the AAA domain (a, b and c) obeying the three substrate specificity rules (see text) have a triangular location in luciferase, creating a putative active site. The corresponding amino acids found in the luciferase molecule site are also included.

the recognition pocket for sarcosine with F247I, G316V and S347Q modelled such that the side-chains adopt their most commonly observed conformations.

The triplet code was originally derived without any knowledge of the possible 3D structure of the amino acyl-adenylation domains. That the positions of the three residues lie in such a clearly defined pocket near the ATP binding site of luciferase provides significant support for the original hypothesis. The structure of the N-terminal adenylation domain of gramicidin synthetase complexed with phenylalanine provides even stronger substantiation [23]. Based on this structure, eight residues lining the phenylalanine pocket have been implicated in substrate binding [23]. Some of these bind the carboxyl and amino groups of the substrate. Application of the rules, developed in our work on cyclosporin synthetase, successfully selected a subset of three of those amino acids which are likely to provide the greatest discrimination and selectivity for the different substrates.

The recognition hypothesis presented here provides new possibilities for designing cyclosporin-like products. Relatively few site-point mutations in the synthetase gene should enable the production of designed sequences in which, for example, specific leucine residues in cyclosporin could be changed to valine by a double mutation of G to A and M to L as outlined in Table 2. There is also scope for designing new amino acid specificities by using site-directed mutagenesis to engineer new triplet combinations into the recognition pocket of CySyn.

Acknowledgements: We are grateful to Dr. Andrew Coulson and Shane Sturrock for running a number of sequence alignment and homology programs and for helpful discussion. We thank Biochemie GmbH for financial support.

References

[1] Stachelhaus, T. and Marahiel, M.A. (1995) FEMS Microbiol. Lett. 125, 3–14.
[2] Weber, G., Schörgendorfer, K., Schneider-Scherzer, E. and Leitner, E. (1994) Curr. Genet. 26, 120–125.
[3] Dittmann, J., Wenger, R.M., Kleinkauf, H. and Lawen, A. (1994) J. Biol. Chem. 269, 2841–2846.
[4] Borel, J.F., Dipadova, F., Mason, J., Quesniaux, V., Ryffel, B. and Wenger, R. (1989) Pharmacol. Rev. 41, 239–242.
[5] Kleinkauf, H. and Von Döhren, H. (1996) Eur. J. Biochem. 236, 335–351.
[6] Altschul, S.F. and Gish, W. (1996) Methods Enzymol. 266, 460–480.
[7] Coulson, A.F.W., Collins, J.F. and Lyall, A. (1987) Comp. J. 30, 420–424.
[8] Rost, B., Sander, C. and Schneider, R. (1994) Comp. Appl. Biosci. 10, 53–60.
[9] Sturrock, S. (1997) PhD Thesis, Edinburgh University.
[10] Cosmina, P., Rodriguez, F., Deferra, F., Grandi, G., Perego, M., Venema, G. and Vansinderen, D. (1993) Mol. Microbiol. 8, 821–831.
[11] Stachelhaus, T., Huser, A. and Marahiel, M.A. (1996) Chem. Biol. 3, 913–921.
[12] Collins, J.F. and Coulson, A.F.W. (1990) Methods Enzymol. 183, 474–487.
[13] Rost, B. and Sander, C. (1993) J. Mol. Biol. 232, 584–599.
[14] Decrecylagard, V., Marliere, P. and Saurin, W. (1995) Life Sci. 318, 927–936.
[15] Conti, E., Franks, N.P. and Brick, P. (1996) Structure 4, 287–298.
[16] Simpson, T.J. (1995) Chem. Ind. 407–411.
[17] Crump, M.P., Crosby, J., Dempsey, C.E., Murray, M., Hopwood, D.A. and Simpson, T.J. (1996) FEBS Lett. 391, 302–306.
[18] Holak, T.A., Nilges, M., Prestegard, J.H., Gronenborn, A.M. and Clore, G.M. (1988) Eur. J. Biochem. 175, 9–15.
[19] Pavelavrancic, M., Pfeifer, E., Schroder, W., Von Döhren, H. and Kleinkauf, H. (1994) J. Biol. Chem. 269, 14962–14966.
[20] Lawen, A. and Zocher, R. (1990) J. Biol. Chem. 265, 11355–11360.
[21] Wenger, R.M. (1990) Transplant. Proc. 22, 1104–1108.
[22] Kraulis, P.J. (1991) J. Appl. Crystallogr. 24, 946–950.
[23] Conti, E., Stachelhaus, T., Marahiel, M.A. and Brick, P. (1997) EMBO J. 16, 4174–4183.