The specificity-conferring code of adenylation domains in nonribosomal peptide synthetases

Torsten Stachelhaus*, Henning D Mootz and Mohamed A Marahiel

Background: Many pharmacologically important peptides are synthesized nonribosomally by multimodular peptide synthetases (NRPSs). These enzyme templates consist of iterated modules that, in their number and organization, determine the primary structure of the corresponding peptide products. At the core of each module is an adenylation domain that recognizes the cognate substrate and activates it as its aminoacyl adenylate. Recently, the crystal structure of the phenylalanine-activating adenylation domain PheA was solved with phenylalanine and AMP, illustrating the structural basis for substrate recognition.

Results: By comparing the residues that line the phenylalanine-binding pocket in PheA with the corresponding moieties in other adenylation domains, general rules for deducing substrate specificity were developed. We tested these *in silico* 'rules' by mutating specificity-conferring residues within PheA. The substrate specificity of most mutants was altered or relaxed. Generalization of the selectivity determinants also allowed the targeted specificity switch of an aspartate-activating adenylation domain, the crystal structure of which has not yet been solved, by introducing a single mutation.

Conclusions: *In silico* studies and structure–function mutagenesis have defined general rules for the structural basis of substrate recognition in adenylation domains of NRPSs. These rules can be used to rationally alter the specificity of adenylation domains and to predict from the primary sequence the specificity of biochemically uncharacterized adenylation domains. Such efforts could enhance the structural diversity of peptide antibiotics such as penicillins, cyclosporins and vancomycins by allowing synthesis of 'unnatural' natural products.

Address: Biochemie/Fachbereich Chemie, Philipps-Universität Marburg, Hans-Meerwein-Straße, D-35032 Marburg, Germany.

*Present address: Department of Biological Chemistry and Molecular Pharmacology, Harvard Medical School, 240 Longwood Avenue, Boston, MA 02115, USA.

Correspondence: Mohamed A Marahiel
E-mail: marahiel@chemie.uni-marburg.de

Key words: adenylation domain, binding pocket, nonribosomal peptide synthetase, signature sequence, substrate specificity

Received: 11 March 1999
Revisions requested: 13 April 1999
Revisions received: 27 April 1999
Accepted: 4 May 1999

Published: 29 June 1999

Chemistry & Biology August 1999, 6:493–505
http://biomednet.com/elecref/1074552100600493

 ISSN 1074-5521

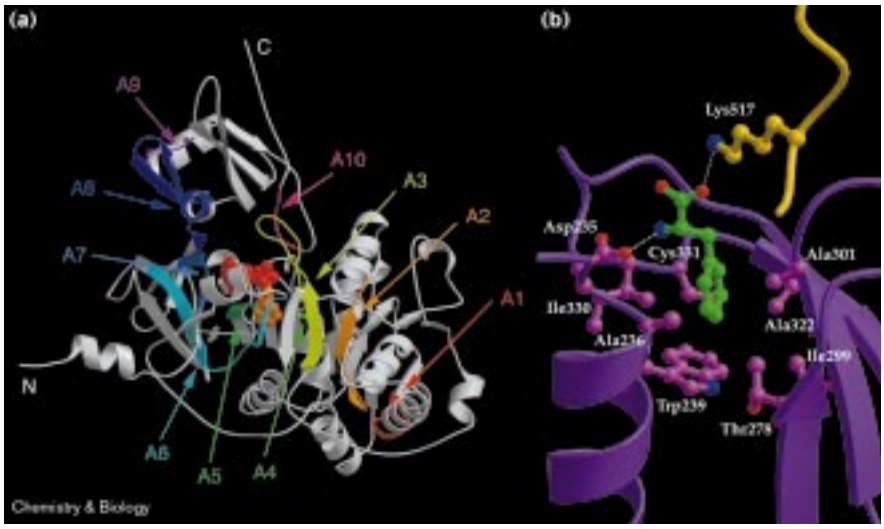
Introduction

Nonribosomally synthesized peptides represent a large group of structurally complex metabolites that are manufactured from amino, hydroxy and carboxy acid monomers by large multifunctional enzymes, termed nonribosomal peptide synthetases (NRPSs) [1–4]. Although their actual biological roles in the producing organisms (primarily bacilli, actinomycetes and filamentous fungi) are often obscure, a remarkable variety of pharmacological properties can be linked to such naturally occurring peptides. Important antibiotics synthesized by NRPSs include penicillins and cephalosporins, vancomycins, as well as the immunosuppressant cyclosporin A [1–3]. Interestingly, the same class of NRPSs also assembles several virulence-conferring siderophoric peptides, such as mycobactin and yersiniabactin [5,6].

Primary structure, size and complexity of a peptide product are dictated by the number and organization of iterated modules, which constitute the NRPS template [1–4]. Each module activates its cognate amino acid in a two-step reaction using a pair of closely coupled domains. An adenylation (A) domain selects the cognate amino acid from the pool of available substrates and generates the corresponding aminoacyl adenylate using ATP. The aminoacyl moiety is then covalently tethered to the sulfhydryl group of a phosphopantetheinyl (Ppant) prosthetic group on an adjacent thiolation domain (also called peptidyl carrier protein, PCP). Thiolation domains are post-translationally modified by Ppant transferases, which add the Ppant group [7]. The nascent peptidyl chain grows unidirectionally under the catalytic control of a condensation (C) domain each time an upstream peptidyl-S-Ppant donor is attacked by a monomeric aminoacyl-S-Ppant nucleophile [8]. Modifying the incorporated monomers (e.g. by epimerization or *N*-methylation) or the peptide backbone (e.g. by acylation, glycosylation or heterocylization) can further functionalize the synthesized peptide product. These tailoring reactions are catalyzed by specialized domains or by fusion to polyketide synthase (PKS) modules [1–3].

Although several biochemical studies have shown that the A domains are the specificity-mediating 'gate-keeper' units of the repeated modules [9–11], the selectivity-conferring elements have remained unclear. This has changed recently, with the elucidation of the crystal

Figure 1



Structural basis for phenylalanine activation by PheA (from Conti *et al.* [13]). (a) The ribbon diagram of PheA shows how the A domain is folded into a large amino-terminal and a smaller carboxy-terminal subdomain. The AMP (red) and phenylalanine (orange), bound at the interface between both subdomains, are shown in space-filling presentation. The locations of the highly conserved core motifs are indicated : A1, LTYxEL; A2, LKAGxAYVPID; A3, LAYxxYTSGTTGxPKG; A4, FDxS; A5, NxYGPTE; A6, GELxlxGxGLARGYW; A7, YKTGDQ; A8, GRxDxQVKIRGx-RVELEEVE; A9, LPxYMIP; and A10, NGKIDR (using single-letter amino-acid code where x is any amino acid). (b) The phenylalanine-specific binding pocket consists of ten residues. Asp235 and Lys517 mediate electrostatic interactions (dotted lines) with the α-amino and α-carboxylate groups of phenylalanine, whereas the sidechain specificity pocket is surrounded on one side by Ala236, Ile330 and Cys331, and on the other side by Ala322, Ala301, Ile299 and Thr278. Both sides are separated by the indole ring of Trp239 at the bottom of the pocket. This architecture allows PheA to accommodate L-Phe (green) and D-Phe with no significant change in conformation.

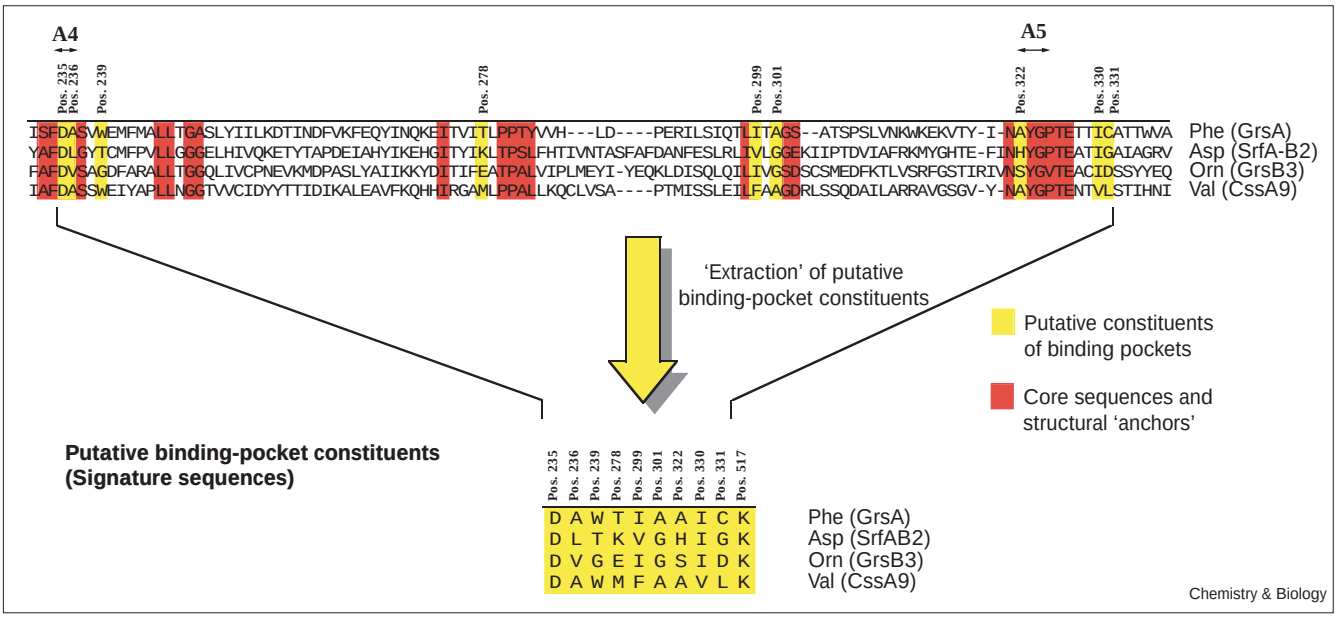
structures of two members of this superfamily of adenylate-forming enzymes. First, the structure of firefly luciferase of *Photinus pyralis* was reported [12], which confirmed the existence of a third scaffold for the ATP-dependent formation of aminoacyl adenylates, in addition to those found in class I and class II aminoacyl tRNA synthetases. Subsequently, the structure of the A domain of the peptide synthetase GrsA (termed PheA) was solved complexed with phenylalanine and AMP (Figure 1a) [13]. This second structure provided, for the first time, fundamental insight into the structural basis of substrate recognition and activation (Figure 1b).

Although the protein sequence similarity between PheA and luciferase is only 16%, both enzymes share a highly conserved three-dimensional structure [12,13]. As shown for PheA (Figure 1a), the enzyme folds into a large amino-terminal and a smaller carboxy-terminal subdomain. The latter has been shown to be important for catalytic activity, and is rotated 94° with respect to the amino-terminal subdomain, when compared to the structure of luciferase. Because luciferase was crystallized in the absence of substrate, the rotation can be interpreted as an active conformational change during the course of substrate recognition and activation. Indeed, the active site is located at the interface between the folding domains and rotation closes the cleft between them. Most of the highly conserved core motifs (A1 to A10, Figure 1a) were found surrounding the active site where the substrates bind. This finding confirmed previous results obtained using site-directed mutagenesis and photoaffinity labeling, which indicated that these motifs are involved in ATP binding and hydrolysis [14–18].

The conclusion that shared sequence motifs play a key role in ATP binding is reasonable, because all A domains require ATP as a substrate. In contrast, the residues lining the phenylalanine-binding pocket were found to be located within a 100 amino acid (aa) stretch that possesses a lower overall similarity between different A domains (Figure 2) [13]. On the strength of sequence alignments and hybrid domain construction [19–23], this area was also proposed to confer substrate specificity on the A domains. As shown in Figure 1b, the α-amino and the α-carboxylate groups of the substrate amino acid are stabilized by electrostatic interactions with Asp235 and Lys517, respectively. These residues are located in the highly conserved core motifs A4 (Asp235) and A10 (Lys517) [1,4], with the latter being found in a loop within the carboxy-terminal subdomain (Figure 1a). The specificity pocket for the phenylalanine sidechain is bordered on one side by Ala236, Ile330 and Cys331, and by Ala322, Ala301, Ile299 and Thr278 on the other side. Both sides are appropriately separated by the indole ring of Trp239 at the bottom of the pocket. At one end, towards the viewer in Figure 1b, a water-filled channel connects the pocket with the surrounding solvent. Inspection of the PheA binding pocket shows that both D- and L-Phe can be accommodated (Figure 1b) with no significant change in conformation [13], confirming biochemical data which revealed that both stereoisomers are activated with the same catalytic efficiency [9].

We expected, because of the high similarity between peptide synthetases, that amino-acid residues that correspond to those lining the PheA binding pocket would likewise mediate substrate specificity in these enzymes.

Figure 2



Using A domain primary amino acid sequences to determine selectivity-conferring residues. Three examples (AspA from SrfA-B, OrnA from GrsB and ValA from CssA) are shown. The primary sequences between core motifs A4 and A5 (~100 aa; see Figure 1a) were aligned to the corresponding sequence of PheA (top row). Subsequent assessment of the binding-pocket constituents (yellow boxed items) was assisted by their proximity to conserved sequence motifs (red boxed items). The consecutive order of the ten constituents represents the signature sequence of an A domain. The signatures of 160 different A domains were determined in this way.

Knowledge of these residues, in turn, might allow engineering of the corresponding enzymes to alter or relax substrate specificity. The synthesis of novel peptides with modified biological and/or pharmacological properties would therefore be enabled. As a first approach to this goal, we determined and examined the binding pockets of 160 other A domains. Some general rules for deducing substrate specificity were developed on the basis of these putative recognition sites. We then mutated selectivity-conferring residues in the PheA binding pocket to test our theory. Generalization of selectivity determinants allowed us to specifically alter the substrate specificity of an aspartate-activating A domain whose crystal structure has not yet been solved.

Results and discussion

Previous studies predicted that the substrate specificity of NRPS A domains is determined by a 200 aa stretch comprising the core motifs A3 to A6 [19–23]. This segment has a lower overall similarity among different A domains (Figure 2), but it was assumed that the similarity between domains activating the same substrate should be significantly higher. Consequently, such domains should cluster together in a phylogenetic tree. This assumption was only modestly confirmed, because the determination of an ancestral relationship was significantly impaired by the influence of sequence origin [23]. This problem was partially overcome by further shortening the A domain sequences, but even in these cases the evolutionary relationship was still a substantial factor. A genuine grouping according to substrate specificity was observed only for highly related domains that probably descended from the same ancestor.

Determination of putative substrate-binding pockets

We began by examining luciferase and PheA, which, although only 16% similar in terms of their primary structure, share a highly conserved three-dimensional structure (Figure 1a) [13]. A structurally based sequence alignment revealed that 67% of the α-carbons of both enzymes are separated by less than 3 Å. Considering the much more pronounced similarity between PheA and other NRPS A domains (between 26% and 56%) [1], the conformation of their mainchains is likely to be very similar too. Substrate specificity should therefore be mediated by the nature of the residues lining the PheA binding pocket (Figures 1b and 2). Put another way, the solved structure of PheA (and luciferase) can be considered precedents for the entire superfamily of adenylating enzymes, and sequence comparisons should allow the constituents of the binding pockets of all other A domains to be determined (Figure 2).

For our initial alignments, we examined the ~100 aa stretch between core motifs A4 and A5 (Figure 2) that contains nine of the ten constituents of the PheA binding

pocket (the missing residue is the highly conserved Lys517, core motif A10, which is located in a loop within the carboxy-terminal subfolding domain) [13]. The sequences of 160 A domains were retrieved from publicly accessible databases and aligned using the program MegAlign from the DNAStar package (see the Materials and methods section). Assessment of each constituent of the proposed binding pockets was assisted by the presence of adjacent (within 2 aa residues), highly conserved sequence motifs (A4 and A5, as well as the two unnamed motifs 'TPS' and 'GE') [1], which presumably 'anchor' these moieties within the tertiary structure (Figure 2).

All sequences were then trimmed down to the constituents of their presumed binding pockets, and new sequence alignments and phylogenetic studies were performed considering only these ten residues (Figures 2,3). We postulated that if these extracted moieties represent the codon (signature sequence) of substrate specificity, then the domains that recognize the same substrate should cluster together in a phylogenetic tree. As shown in Figure 3, this novel alignment indeed revealed a clustering of A domains according to their specificity that is only slightly affected by sequence origin. In contrast to previous phylogenetic studies, bacterial and fungal sequences are not separated because of their evolutionary distances, but are found scattered over all three main branches (Figure 3). This outcome indicates that the overall concept of their structural homology is correct, and that the selected residues represent the signature for the recognition of the cognate amino acid substrate. These results can even explain some discrepancies between postulated and observed specificities, and should permit substrates of newly discovered domains to be predicted, some examples of which are shown in red boxes in Figure 3. First, Phe(TycB2) does not cluster with Phe domains, but rather with domains that activate other aromatic substrates. In fact, it was observed that, although this domain can activate phenylalanine, the preferred substrate of this domain is tryptophan [10]. Second, the molecular characterization of a NRPS template from *Bacillus licheniformis* ATCC 10716 shared highest similarity (97%) with a putative lichenysin A synthetase operon from *B. licheniformis* BNP29 [24,25]. Consequently, an asparagine specificity was proposed for the fifth module, LicB2 (and LchAB2), of the biosynthetic template. The signature sequences of these A domains match the consensus sequence of aspartate rather than asparagine domains (especially the key residue His322), however, as we will demonstrate below. Recently, the recombinant LicB2 A domain was indeed shown to activate L-Asp, but not L-Asn [25]. Third, the binding pocket of an A domain from an as yet uncharacterized peptide synthetase from *Mycobacterium tuberculosis* clusters and exactly matches the signature sequence of several phenylalanine-activating domains [26]. Fourth, sequencing and analysis of the genes involved in the biosynthesis of the vancomycin group antibiotic chloroeremomycin from *Amycolatopsis orientalis* revealed the presence of an additional peptide synthetase, Orf19, that was not addressed further in the original report [27]. The signature sequence of this open reading frame clusters with tyrosine domains and it can therefore be postulated that Orf19 is involved in the synthesis of the unusual tyrosine-based precursors recognized and incorporated by chloroeremomycin synthetases. This hypothesis is sustained by the presence of adjacent putative hydroxylases or haloperoxidases [27]. This is an example of how substrate predictions can support hypotheses regarding uncharacterized biosynthetic pathways.

Figure 3

A Domain signature sequences were used to construct a phylogenetic tree, which shows that the signatures cluster according to their domain specificity (yellow boxes). Examples are shown which demonstrate that the substrate specificity of newly discovered domains can be predicted, and that the observed rather than the postulated specificity (red boxes) determine the clustering (see examples in the text). This is also true for the mutants PheA(Thr278→Met/Ala301→Gly) and AspA(His322→Glu) (green boxes), which are specific for leucine and asparagine, respectively, in ATP–pyrophosphate exchange assays.

Up until now, the only way to identify the specificity of an unknown adenylating enzyme was biochemical characterization using the ATP–pyrophosphate exchange assay. Because previous attempts to assess A domain specificity directly from their primary sequence (using the 200 aa stretch between A3 and A6) were only 43% accurate [19], accurate predictions of NRPS module specificity were not permitted using sequence homology alone [28]. However, the new phylogenetic approach (using only the ten binding-pocket constituents shown in Figures 2 and 3) is 86% accurate (only 22 of 160 sequences surveyed are unmatched), and the true accuracy appears to be even higher (92%), because nine unmatched sequences represent single examples for unique substrates. In conclusion, for many biosynthetic systems the reliability of such predictions is already very high. For other systems, particularly in cases of organisms where only a few domains have been sequenced, it will further improve with expansion of available sequence data.

The specificity-conferring code of A domains

From clusters of domains activating the same substrate (Figure 3), the consensus sequences of various substrate-binding pockets can be determined (Table 1). These signature sequences can be interpreted as the 'code' of NRPSs [29], and, as is true for the well known ribosomal precedent, this code appears to be degenerate. For instance, at least four different signature sequences were defined for leucine-activating domains, three for valine and two for cysteine (Table 1), and we anticipate that multiple strategies also exist for other substrates.

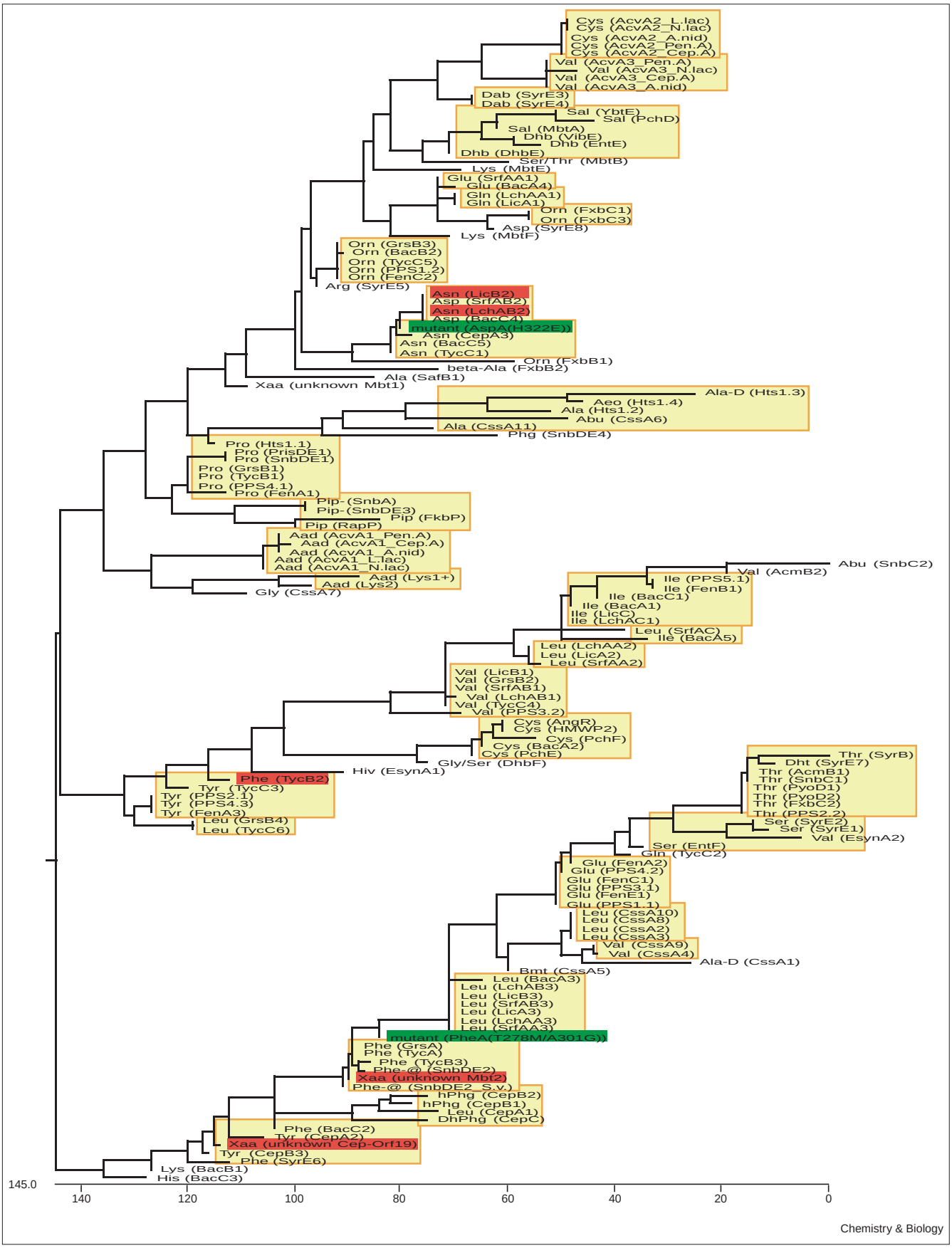

Cys (AcvA2_L.lac)
Cys (AcvA2_N.lac)
Cys (AcvA2_A.nid)
Cys (AcvA2_Pen.A)
Cys (AcvA2_Cep.A)
Val (AcvA3_Pen.A)
Val (AcvA3_N.lac)
Val (AcvA3_Cep.A)
Val (AcvA3_A.nid)
Dab (SyrE3)
Dab (SyrE4)
Sal (YbtE)
Sal (PchD)
Sal (MbtA)
Dhb (VibE)
Dhb (EntE)
Dhb (DhbE)
Ser/Thr (MbtB)
Lys (MbtE)
Glu (SrfAA1)
Glu (BacA4)
Gln (LchAA1)
Gln (LicA1)
Orn (FxbC1)
Orn (FxbC3)
Asp (SyrE8)
Lys (MbtF)
Orn (GrsB3)
Orn (BacB2)
Orn (TycC5)
Orn (PPS1.2)
Orn (FenC2)
Arg (SyrE5)
Asn (LicB2)
Asp (SrfAB2)
Asn (LchAB2)
Asp (BacC4)
mutant (AspA(H322E))
Asn (CepA3)
Asn (BacC5)
Asn (TycC1)
Orn (FxbB1)
beta-Ala (FxbB2)
Ala (SafB1)
Xaa (unknown Mbt1)
Ala-D (Hts1.3)
Aeo (Hts1.4)
Ala (Hts1.2)
Abu (CssA6)
Ala (CssA11)
Phg (SnbDE4)
Pro (Hts1.1)
Pro (PrisDE1)
Pro (SnbDE1)
Pro (GrsB1)
Pro (TycB1)
Pro (PPS4.1)
Pro (FenA1)
Pip-(SnbA)
Pip-(SnbDE3)
Pip (FkbP)
Pip (RapP)
Aad (AcvA1_Pen.A)
Aad (AcvA1_Cep.A)
Aad (AcvA1_A.nid)
Aad (AcvA1_L.lac)
Aad (AcvA1_N.lac)
Aad (Lys1+)
Aad (Lys2)
Gly (CssA7)
Abu (SnbC2)
Val (AcmB2)
Ile (PPS5.1)
Ile (FenB1)
Ile (BacC1)
Ile (BacA1)
Ile (LicC)
Ile (LchAC1)
Leu (SrfAC)
Ile (BacA5)
Leu (LchAA2)
Leu (LicA2)
Leu (SrfAA2)
Val (LicB1)
Val (GrsB2)
Val (SrfAB1)
Val (LchAB1)
Val (TycC4)
Val (PPS3.2)
Cys (AngR)
Cys (HMWP2)
Cys (PchF)
Cys (BacA2)
Cys (PchE)
Gly/Ser (DhbF)
Hiv (EsynA1)
Phe (TycB2)
Tyr (TycC3)
Tyr (PPS2.1)
Tyr (PPS4.3)
Tyr (FenA3)
Leu (GrsB4)
Leu (TycC6)
Thr (SyrB)
Dht (SyrE7)
Thr (AcmB1)
Thr (SnbC1)
Thr (PyoD1)
Thr (PyoD2)
Thr (FxbC2)
Thr (PPS2.2)
Ser (SyrE2)
Ser (SyrE1)
Val (EsynA2)
Ser (EntF)
Gln (TycC2)
Glu (FenA2)
Glu (PPS4.2)
Glu (FenC1)
Glu (PPS3.1)
Glu (FenE1)
Glu (PPS1.1)
Leu (CssA10)
Leu (CssA8)
Leu (CssA2)
Leu (CssA3)
Val (CssA9)
Val (CssA4)
Ala-D (CssA1)
Bmt (CssA5)
Leu (BacA3)
Leu (LchAB3)
Leu (LicB3)
Leu (SrfAB3)
Leu (LicA3)
Leu (LchAA3)
Leu (SrfAA3)
mutant (PheA(T278M/A301G))
Phe (GrsA)
Phe (TycA)
Phe (TycB3)
Phe-@ (SnbDE2)
Xaa (unknown Mbt2)
Phe-@ (SnbDE2_S.v.)
hPhg (CepB2)
hPhg (CepB1)
Leu (CepA1)
DhPhg (CepC)
Phe (BacC2)
Tyr (CepA2)
Xaa (unknown Cep-Orf19)
Tyr (CepB3)
Phe (SyrE6)
Lys (BacB1)
His (BacC3)
145.0
140
120
100
80
60
40
20
0
Chemistry & Biology

Table 1

The selectivity-conferring code of A domains.

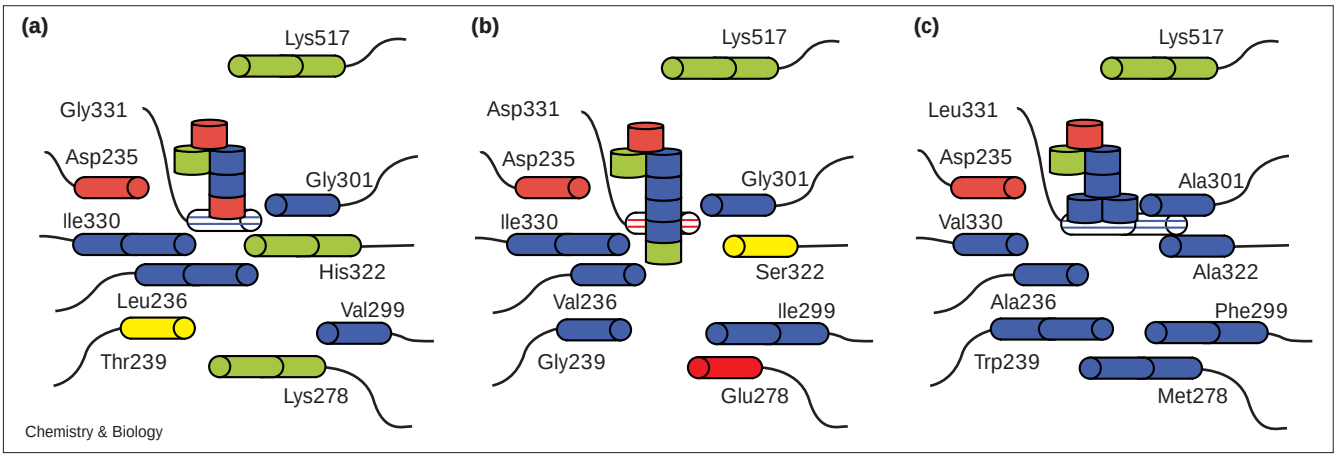
	Position											
Domain	235	236	239	278	299	301	322	330	331	517	Biosynthetic template	Similarity
Aad	E	P	R	N	I	V	E	F	V	K	AcvA	94%
Ala	D	L	L	F	G	I	A	V	L	K	CssA, Hts1	55%
Asn	D	L	T	K	L	G	E	V	G	K	BacA, CepA, Dae, Glg1, TycC	90%
Asp	D	L	T	K	V	G	H	I	G	K	BacC, SrfAB, LicB, LchAB	100%
Cys(1)	D	H	E	S	D	V	G	I	T	K	AcvA	96%
Cys(2)	D	L	Y	N	L	S	L	I	W	K	BacA, HMWP2	88%
Dab	D	L	E	H	N	T	T	V	S	K	SyrE	100%
Dhb/Sal	P	L	P	A	Q	G	V	V	N	K	EntE, DhbE, MbtA, PchD, VibE, YbtE	83%
Gln	D	A	Q	D	L	G	V	V	D	K	LicA, LchAA	100%
Glu(1)	D	A	W	H	F	G	G	V	D	K	FenA, FenC, FenE, PPS1, PPS3, PPS4	95%
Glu(2)	D	A	K	D	L	G	V	V	D	K	BacC, SrfAA	95%
Ile (1)	D	G	F	F	L	G	V	V	Y	K	BacA, BacC, LicC, LchAC	92%
Ile (2)	D	A	F	F	Y	G	I	T	F	K	FenB, PPS5	100%
Leu(1)	D	A	W	F	L	G	N	V	V	K	BacA, LicA, LchAA, LicB, LchAB, SrfAA, SrfAB	99%
Leu(2)	D	A	W	L	Y	G	A	V	M	K	CssA	100%
Leu(3)	D	G	A	Y	T	G	E	V	V	K	GrsB, TycC	100%
Leu(4)	D	A	F	M	L	G	M	V	F	K	LicA, LchAA, SrfAA	97%
Orn(1)	D	M	E	N	L	G	L	I	N	K	FxbC	100%
Orn(2)	D	V	G	E	I	G	S	I	D	K	BacB, FenC, GrsB, PPS1, TycC	98%
Phe	D	A	W	T	I	A	A	V	C	K	GrsA, SnbDE, TycA, TycB	88%
Phg/hPhg	D	I	F	L	L	G	L	L	C	K	CepB, CepC, SnbDE	80%
Pip/Pip-@	D	F	Q	L	L	G	V	A	V	K	FkbP, RapP, SnbA, SnbDE	75%
Pro	D	V	Q	L	I	A	H	V	V	K	GrsB, FenA, PPS4, SnbDE, TycB	87%
Ser	D	V	W	H	L	S	L	I	D	K	EntF, SyrE	90%
Thr/Dht	D	F	W	N	I	G	M	V	H	K	AcmB, Fxb, PPS2, PyoD, SnbC, SyrB, SyrE	91%
Tyr(1)	D	G	T	I	T	A	E	V	A	K	FenA, PPS2, PPS4	100%
Tyr(2)	D	A	L	V	T	G	A	V	V	K	TycB, TycC	80%
Tyr(3)	D	A	S	T	V	A	A	V	C	K	BacC, CepA, CepB	78%
Val(1)	D	A	F	W	I	G	G	T	F	K	GrsB, FenE, LicB, LchAB, PPS3, SrfAB, TycC	96%
Val(2)	D	F	E	S	T	A	A	V	Y	K	AcvA	94%
Val(3)	D	A	W	M	F	A	A	V	L	K	CssA	95%
Variability	3%	16%	16%	39%	52%	13%	26%	23%	26%	0%	Wobble-like positions	

This table complements Figure 3. From clusters of signature sequences derived from domains activating the same substrate, the consensus sequences for the recognition of various substrates were determined. The biosynthetic templates of origin and the overall similarity of signature sequences, which were integrated into a codon, are depicted. Variable constituents within a codon are shown (red), and proposed 'wobble'-like positions, revealing an elevated variability throughout all codons determined (≥ 36%), are indicated in cyan. Aad, δ(L-α-aminoadipic acid); Dab, 2,3-diamino butyric acid; Dhb, 2,3-dihydroxy benzoic acid; Sal, salicylate; Phg, L-phenylglycine; hPhg, 4-hydroxy-L-phenylglycine; Pip, L-pipecolinic acid; Dht, dehydrothreonine; @ indicates a modification of the residue.

The signature sequences determined could now be projected onto the crystal structure of PheA, in order to estimate the general appearance of putative substrate-binding pockets. As examples, Figure 4 shows the estimated shape of the binding and recognition sites of three different substrates: the acidic aspartate, the basic ornithine and the hydrophobic valine. Although Asp235 and Lys517 are highly conserved in all amino-acid-activating domains (mediating key interactions with the amino and the carboxyl groups of the substrate), all of the other residues are believed to determine the specific shape of the pocket and facilitate the recognition of the substrate sidechain [13]. In all three examples, the remaining residues support this assumption, as they complement the polarity of the recognized substrate (compare with Figure 2 for the determination of the putative binding pocket constituents). In aspartate activation (signature sequence 'Asp') the basic His322 (perhaps secondarily also Lys278) facilitates the key interaction with the acidic sidechain, whereas the large sidechain of Leu236 probably closes the pocket underneath against larger substrates. In ornithine activation (signature sequence 'Orn(2)') the acidic Glu278 and Asp331 seem to be the key players for the recognition of the large, basic sidechain. In valine activation (signature sequence 'Val(3)') the entire binding pocket for the hydrophobic substrate is assembled by hydrophobic residues. The bulkiness of the substrate caused by its β branching is taken into account by the exclusive use of alanine moieties in positions 236, 301 and 322 in the upper portion of the pocket and bulky residues at the bottom, which keep the two sides of the binding pocket apart.

Another similarity between the ribosomal and nonribosomal codes, besides degeneracy, is the presence of flexible

Figure 4



A simplified representation of the proposed binding pockets of three A domains. The putative binding pocket constituents of (a) an aspartate-activating domain (SrfAB2), (b) an ornithine-activating domain (GrsB3) and (c) a valine-activating domain (CssA9), determined in Figure 2, were projected onto the binding pocket of PheA shown in Figure 1b. The assessed aliphatic (blue), polar (yellow), acidic (red) and basic (green) sidechains are shown schematically. In all cases, Asp235 and Lys517 mediate key interactions with the α-amino and α-carboxylate group of the substrate, and all other residues facilitate recognition of the substrate sidechains and (ideally) complement the polarity of the recognized substrate.

positions within certain signature sequences. As shown in Table 1, residues 278 and 299 can be thought of as 'wobble'-like positions that are highly variable throughout the signature sequences. Moreover, because nonribosomal 'codons' consist of ten residues, some other positions also show flexibility. Generally binding pockets that recognize small amino acids have more flexibility in positions close to the bottom of the binding pocket, whereas for larger substrates, positions in the top portion are a little bit more imprecise (Table 1).

Variability of binding-pocket constituents provides evidence for specificity-conferring key positions

The examples outlined above further strengthen the hypothesis that sequence comparisons with PheA allow the constituents of the binding pockets of other A domains to be determined. Furthermore, the studies summarized in Table 1 and Figure 4 imply varied significance for different positions within a binding pocket for the mediation of substrate selectivity. To verify this observation, we took a closer look at the constituent amino acids of 160 specificity-conferring signature sequences (Figure 5). According to their variability the sidechains were classified into three subgroups: 'invariant' residues (positions 235 and 517), 'moderately variant' (aliphatic) residues (positions 236, 301 and 330), and 'highly variant' residues (positions 239, 278, 299, 322 and 331). Taking into account that a meaningful link between amino-acid usage and substrate polarity can be observed only for the highly variant residues, the high variability probably reflects their importance in contributing to substrate specificity.

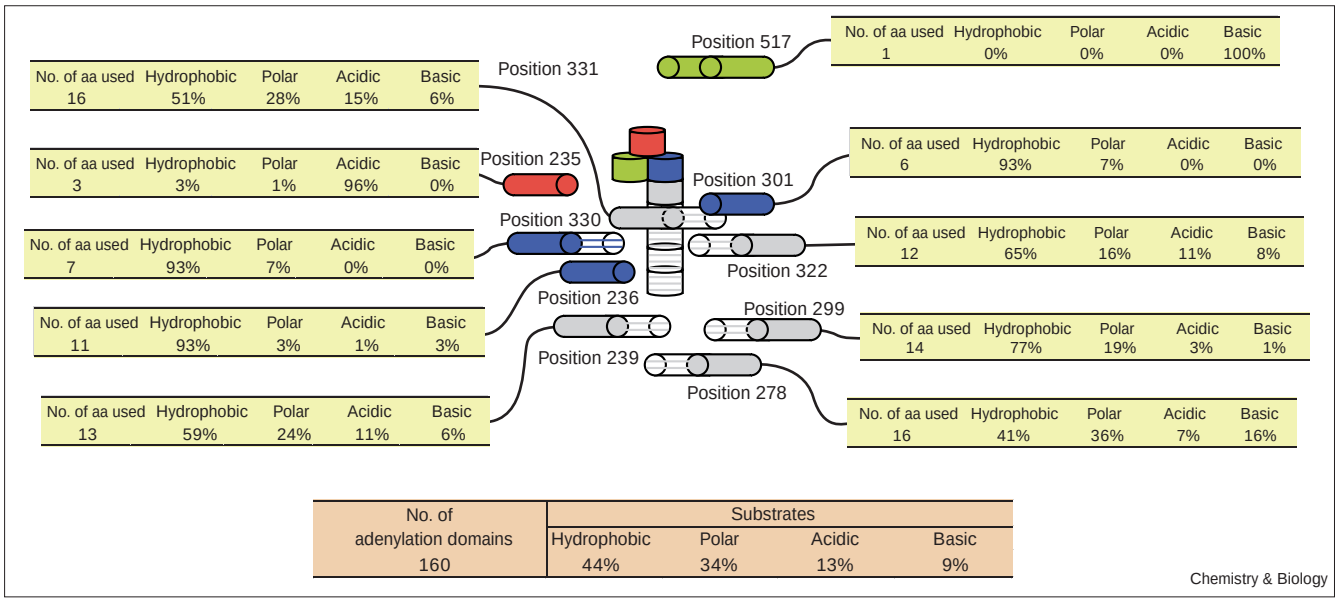
Invariant residues

Of the invariant residues, Asp235 stabilizes the α-amino group of the substrate and is therefore essential (except in domains that do not have an α-amino group — that is the α-aminoadipate activating domains of AcvA synthetases and carboxy-/hydroxy-acid-activating domains, such as luciferase, EntF and YbtE). Lys517 is strictly invariant and pairs with the α-carboxylate of the substrate as well as the O-4′ and O-5′ of the ATP/AMP ribose moiety, presumably fixing their positions in the active site and clamping the carboxy-terminal folding domain in an active conformation. Both residues therefore mediate key interactions with the α-amino and α-carboxylate group of the substrate (and not the sidechain; see Figures 1b and 3).

Moderately variant (aliphatic) residues

The residues in positions 236, 301 and 330 vary only slightly throughout all binding pockets, and the vast majority of A domains (93% of 160 domains surveyed) use aliphatic sidechains in these positions. Taking into account the low overall variability and the under-representation of charged or polar sidechains, it is very unlikely that these residues are the major elements for the discrimination and selection of different substrates. As we demonstrate below, however, these positions might modulate the catalytic activity (V_{max}) and fine-tune the specificity of the corresponding domains. It should be noted that this observation is inconsistent with previous predictions made in the cyclosporin system, which specified that two of these three residues provide the greatest discrimination and selectivity for different substrates [21]. We will discuss this point below.

Figure 5



Observed variations of amino acids that constitute substrate-binding pockets. The proposed signature sequences from 160 different A domains were investigated (compare with Figure 2). The proportional distribution of the nature of their substrates is shown in the lower table. The yellow table linked to each position displays the number of different amino acids found in that position, and the proportional occurrence of hydrophobic, polar, acidic and basic sidechains. According to these data, the ten constituent amino acids can be classified into three subgroups. Positions 235 (aspartate: acidic, red) and 517 (lysine: basic, green) are considered 'invariant'. Positions 236, 301 and 330 are only 'moderately variant'. The vast majority (93%) of the A domains examined use hydrophobic sidechains in these positions (blue). 'Highly variant' are the residues at positions 239, 278, 299, 322 and 331 (gray), which reveal the highest variability of amino-acid usage.

Highly variant residues

The residues in positions 239, 278, 299, 322 and 331 show the highest flexibility with respect to amino-acid usage. Apart from positions 278 and 299, which are generally very adaptable ('wobble'-like), we predict that they facilitate substrate specificity. Considering their relative localization within the binding pocket, and taking into account the data shown in Table 1, as well as Figure 4, one can assume that position 322 possesses a higher impact for smaller substrates, and positions 239 and 278 for larger substrates. This hypothesis is supported by the consensus sequences for aspartate (His322), glutamate (Lys239) and ornithine (Glu278) activating domains, where amino acid usage and substrate polarity appear to be perfectly linked (Table 1).

Mutational investigation of the PheA binding pocket

With all the assumptions from the *in silico* studies in mind, we set out to experimentally evaluate our predictions. First, we mutated several selectivity-conferring residues of the PheA binding pocket, and investigated the mutants for changes in the rate of ATP–pyrophosphate exchange (enzyme activity) and the activation of miscognate amino acid substrates (enzyme specificity). It should be noted that wild-type PheA possesses a significant side-specificity for the miscognate substrates tryptophan (18%) and leucine (7%) [9]. The results of these studies are summarized in Table 2 and can be generalized as follows. First, substitutions with slightly larger residues reduce the space within the pocket. Accordingly the ATP–pyrophosphate exchange rate of those mutants decreased, whereas their specificity for the cognate phenylalanine substrate increased (e.g. Thr278→Met, Thr278→Gln, Ala301→Val, Ala322→Ser and Cys331→Leu). In all these cases, the side-selectivity for the miscognate substrates tryptophan and leucine decreased by a factor of ≤ 20. Second, exactly the opposite was observed when slightly smaller residues were substituted, which increases the available space within the binding pocket. In those cases the phenylalanine-specific catalytic activity could be restored, whereas the ability to discriminate between phenylalanine and tryptophan/leucine relaxed by a factor of ≤ 5 (e.g. Ala301→Gly, Ala322→Gly and Ile330→Val). Third, mutation of the 'highly variant' residues (positions 239, 278, 322 and 331) caused an up to fivefold drop in enzyme activity (e.g. Trp239→Leu, Thr278→Met and Cys331→Leu). Moreover, the introduction of charged (or polar) sidechains changed the polarity of the binding pocket and led to inactive enzymes (e.g. Ala322→Asp, Ala322→Lys and Ala322→Asn). Fourth, in several cases the observed increase in activation of a certain miscognate substrate was predictable. For example, the PheA

Table 2

Activity and specificity of PheA mutants.

Mutant	Activity (%; wt = 100)	Substrate		Comments
		Cognate	Miscognate	
PheA (wild type)	100	L-Phe/D-Phe	Trp (18), Leu (7)	Wild type
PheA(A236L)	nd	nd	nd	Insoluble
PheA(W239G)	3	L-Phe	D-Phe (55), Val (10)	Almost inactive
PheA(W239L)	20	L-Phe/D-Phe	Ile (12), Val (10)	
PheA(T278M)	32	L-Phe/D-Phe	Leu (21)	Met278 in Leu(1) codon
PheA(T278Q)	39	L-Phe/D-Phe	none	No miscognate aa
PheA(I299T)	68	L-Phe/D-Phe	none	No miscognate aa
PheA(A301G)	100	L-Phe/D-Phe	Trp (22), Leu (15)	Gly301 in all Leu codons
PheA(A301V)	77	L-Phe/D-Phe	none	No miscognate aa
PheA(A322D)	nd	nd	nd	Inactive
PheA(A332E)	nd	nd	nd	Inactive
PheA(A332G)	100	L-Phe/D-Phe	Trp (20), Leu (10)	
PheA(A322I)	nd	nd	nd	Inactive
PheA(A322K)	nd	nd	nd	Inactive
PheA(A322N)	nd	nd	nd	Inactive
PheA(A322Q)	1	Thr/Orn	nd	Almost inactive
PheA(A322S)	30	L-Phe/D-Phe	Leu (17)	
PheA(I330V)	95	L-Phe/D-Phe	Trp (31), Tyr (1)	Val330 in Trp/Tyr codons
PheA(C331L)	26	L-Phe/D-Phe	None	No miscognate aa

nd, not determined.

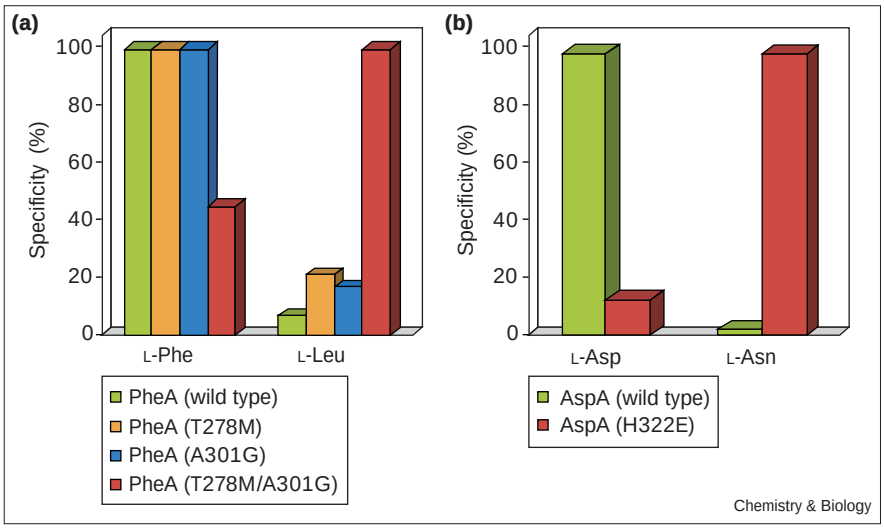
(Ile330→Val) mutant had almost wild-type activity and specificity for phenylalanine, but had an enhanced ability to activate the miscognate substrates tryptophan and tyrosine). Interestingly, all known tryptophan/tyrosine-activating A domains carry a valine in this position. A similar consensus was found for mutations towards a 'Leu(4)' codon (Table 1).

The signature sequences of 'Phe' and 'Leu(4)' (Table 1) share ~60% similarity, and the main differences involve positions 278, 301 and 322 (note: two out of the three are highly variant). Therefore, single substitutions in PheA towards 'Leu(4)' (Thr278→Met and Ala301→Gly) were predicted and found to enhance the activation of leucine (Table 2). Interestingly, for PheA(Thr278→Met), the catalytic efficiency for the side-specific activation of L-Leu remained unchanged (same V_{max}), whereas the catalytic activity for the activation of the cognate substrates (D- and L-Phe) suffered a threefold drop. In contrast, in the case of PheA(Ala301→Gly), the V_{max} for the activation of phenylalanine did not change, whereas the catalytic efficiency for the activation of the miscognate L-Leu experienced a twofold increase. All changes observed were significantly above the background level. How only three differences facilitate discrimination between phenylalanine and leucine is not yet known, but at least for position 301 one can assume that this residue discriminates against the bulkiness of CH (phenylalanine) versus CH_3 (leucine) moieties in the δ-position of the particular substrate sidechain. All A domains that activate substrates with bulky γ- or δ-positions have the smallest possible sidechain (glycine) in position 301 (Table 2).

Targeted alteration of the substrate specificity of PheA

In order to alter the substrate specificity of PheA to leucine, we constructed the double mutant PheA (Thr278→Met/Ala301→Gly). The signature sequence of this construct is about 80% similar to 'Leu(4)' (Table 1) and clusters in the phylogenetic tree close to leucine-activating domains (Figure 3; green box). In accordance with our predictions, we expected that the double mutant would activate leucine rather than phenylalanine. As shown in Figure 6a, this construct in fact preferentially activated L-Leu with a catalytic efficiency approaching or even exceeding that of the corresponding wild-type enzyme (mutant activity for L-Leu: k_{cat}/K_m = 86 mM^{-1} min^{-1}, wild-type activity for L-Phe: 69 mM^{-1} min^{-1}; Table 3). The minimal gain in L-Leu-specific activation efficiency was estimated to be about 30-fold. Surprisingly, the mutant was only slightly impaired in the activation of D-Phe, whereas the catalytic efficiency for the activation of the other stereoisomer, L-Phe, sustained a significant sevenfold loss (Figure 6a and Table 3). For wild-type PheA, in contrast, it has been shown that the polar interactions between the protein and both ligands, D- and L-Phe, are almost identical [13]. The benzyl ring of the D-Phe sidechain is rotated by about 30° relative to the L-Phe ring, which leads to a marginal displacement of the relative position of the β carbon, but not the α carbon. Further investigations (i.e. solving the crystal structure) will be required to explain the observed activation

Figure 6



Targeted alteration of the substrate specificity of PheA and AspA. The specificity of wild-type (green) and mutant (red, orange and blue) proteins was investigated using the ATP–pyrophosphate exchange assay. The applied substrates are shown on the x axis, and the maximum value obtained for each protein was set to 100%. (a) In PheA, single substitutions towards the 'Leu(4)' codon (Thr278→Met and Ala301→Gly) modestly increased the specificity for leucine, although the preferred substrate was still phenylalanine. The corresponding double mutant, however, preferentially activated leucine with a catalytic efficiency approaching that of wild-type PheA. (b) A single His322→Glu mutation in AspA was sufficient to completely alter the substrate specificity of the mutant protein from aspartate to asparagine. The observed activation pattern of the mutants coincides with the appearance of their signature sequences in the phylogenetic tree shown in Figure 3 (green boxes).

pattern (L-Leu > D-Phe >> L-Phe) of the double mutant PheA(Thr278→Met/Ala301→Gly).

Targeted alteration of the substrate specificity of an aspartate-activating domain

To further examine our ideas about specificity predictions and the impact of the 'highly variant' residues 239, 278 and 322 as selectivity-conferring key positions, we attempted to alter the substrate specificity of an A domain whose crystal structure has not yet been solved. First, we noted that the signature sequences for aspartate- and asparagine-activating domains are very similar (80%; Table 1), and the two major differences involve positions 322 (histidine versus glutamate; a highly variant residue) and 330 (valine versus isoleucine; a moderately variant residue). We chose the aspartate-activating A domain from the well-characterized surfactin biosynthetic complex [22,30–32], and tried to alter its specificity to asparagine. The wild-type AspA construct was shown to specifically activate L-Asp, but not L-Asn, with moderate activity (k_{cat}/K_m = 10 mM^{-1} min^{-1}; Figure 6b and Table 3). In contrast, the single mutant AspA(His322→Glu) demonstrated a high selectivity for L-Asn (Figure 6b), although suffering an about tenfold loss of catalytic efficiency (k_{cat}/K_m = 1 mM^{-1} min^{-1}; Table 3). Thus, for the mutant AspA(His322→Glu), a complete switch in specificity was accomplished by changing a single amino-acid residue. So far, the minimal gain in L-Asn-specific activation efficiency was estimated to be about 70-fold. We postulate that an additional Ile330→Val mutation might further modulate and improve the catalytic activity of the (now) asparagine-specific A domain. Lastly, AspA(H322→Glu) also displayed a fivefold increase in the activation efficiency of isoleucine, leucine and valine.

Do these results conflict with previous observations?

The experimental data presented are in very good agreement with the computer predictions, and support the postulates made about the relevance of 'moderately variant' and 'highly variant' positions for substrate discrimination and selectivity. However, independent *in silico* studies on cyclosporin synthetase (CssA), carried out without knowledge of the PheA structure, predicted two 'moderately variant' residues (positions 236 and 301) and only one 'highly variant' residue (position 331) as having the largest impact on the selection of the cognate substrate [21]. Only the eleven domains of the cyclosporin synthetase (and some mutants) were considered, and the corresponding alignments were searched with restrictive criteria to identify key residues involved in recognizing an amino-acid substrate [21]. For example, an amino acid residue at a given position was required to be identical for domains activating the same substrate, and different from amino acids in all other domains. As shown in Table 1, this does not necessarily have to be the case. Identification of another highly variant residue was missed, because the same amino acid (methionine) is present at the same position (278) in valine- and glycine-activating CssA domains.

The three residues identified by Husi *et al.* [21] do not disagree with our conclusion. On the contrary, their results can be explained using the general rules determined for the deduction of substrate specificity in A domains. As indicated by their proximity in the phylogenetic tree (Figure 3), most signature sequences of CssA domains (valine, leucine and Bmt [(4*R*)-(4*E*)-2-butenyl-4-methyl-L-threonine]; Table 1) are very similar. They probably descended from the same ancestor, and therefore represent

Table 3

Catalytic efficiency of wild-type and mutant enzymes.

Protein	Substrate	K_m (mM)	k_{cat} (min^{-1})	k_{cat}/K_m (mM^{-1} min^{-1})
PheA (wt)	L-Phe	0.9	62	69
	D-Phe	0.9	65	72
	L-Leu	nd	nd	nd
PheA	L-Phe	5.2	58	11
(T278M/A301G)	D-Phe	1.4	57	41
	L-Leu	0.7	60	86
Asp (wt)	L-Asp	2.9	29	10
	L-Asn	nd	nd	nd
AspA(H322E)	L-Asp	nd	nd	nd
	L-Asn	27.6	30	1

nd, not determined.

a nice example of how a few subtle mutations within the constituent residues of a binding pocket can alter substrate specificity. For example, the signatures for valine and leucine selection in CssA employ similar strategies for recognizing their cognate, aliphatic substrates and differ only in two positions (position 278, methionine versus leucine, and position 301, alanine versus glycine; Table 1). Based on the mutational analysis of PheA, one would expect that valine and leucine domains could activate both substrates (leucine and valine) but with different catalytic efficiencies. The finding that all CssA domains exhibit a significant activation of miscognate amino acids, which actually causes the formation of several minor cyclosporin A (CsA) derivatives, supports this hypothesis [33]. The composition of the immunosuppressant undecapeptide can be readily altered by increasing the concentration of a miscognate amino acid both *in vivo* and *in vitro* [34,35].

Limitations

Comprehensive computational studies allowed us to establish some general rules for the determination and prediction of A domain selectivities. Some limitations should be noted, however. First, predictions of substrate specificities might be impeded by the limited amount of sequence information available. As shown in Figure 3, the signature sequences for recognizing several substrates are currently unprecedented (e.g. glycine, histidine and lysine). Besides, NRPSs are not restricted to proteinogenic substrates, and more than 300 different compounds are known to be incorporated into nonribosomally synthesized peptides. The facet of known substrates encompasses amino, hydroxy and carboxy acids, and many of the corresponding signatures for recognizing such 'unusual' substrates are presently unduplicated as well (e.g. L-α-aminobutyric acid, Abu, Bmt, and D-α-hydroxyl isovaleric acid, Hiv). Second, specificity-conferring selection rules in a newly discovered system could be quite different from known A domains, and an accurate forecast could fail due to insufficient baseline data. For instance, only a few sequence and biochemical data sets are presently available from streptomycete NRPSs, although these organisms are recognized as prolific sources of secondary metabolites. Third, phylogenetic studies can give only a hint of a presumed substrate, and this indication has to be proven (at least) by evaluation of the amino acid usage in the signature sequence and, if possible, comparison with known NRPS codons. For example, as presented in this study, only by considering their signature sequences (especially position 322) is a differentiation between asparagine and aspartate domains possible.

We anticipate that future efforts in sequencing and biochemical characterization of novel NRPS systems will greatly improve the reliability of any kind of prediction.

Significance

We present here the deciphering of the specificity-conferring code used by adenylation domains of nonribosomal peptide synthetases (NRPSs). These enzymes catalyze the biosynthesis of small, bioactive peptides, many of which are pharmaceutically important [1–3]. The remarkable structural diversity of these low molecular weight compounds results from their being manufactured from amino, hydroxy and carboxy acid monomers. The sites of a multifunctional NRPS template that select and activate these substrates are adenylation domains, the specificity of which thus dictates the composition of the corresponding peptide product.

By comparing the sequence of PheA, the structure of which is known, with the sequences of 160 other adenylation domains, the ten residues that (putatively) form their amino-acid-binding pockets were identified. From groups of domains activating the same substrate the signature sequences ('codons') for different substrate specificities were determined. Paralleling the genetic code, the NRPS selectivity-conferring code features redundancy (e.g. there are two signature sequences for cysteine and four for leucine). We expect that accumulating sequence and biochemical data will result in discovering additional signatures for given specificities.

Using the rules, which connect signature sequence and amino-acid specificity, we were able to predict from primary sequences the substrates of several uncharacterized or unknown adenylation domains. In the future, such predictions should help to circumvent the troublesome biochemical characterizations of newly discovered NRPS clusters of unknown function (e.g. those being uncovered by whole genome sequencing projects).

Finally, the reliability and generalizability of the signature sequences was demonstrated. Mutations in all positions of the binding pocket of PheA were shown to alter

or relax its substrate specificity. By introducing two mutations guided by a signature sequence of leucine, the selectivity of PheA could be converted to leucine. Likewise, a single mutation in an aspartate-activating domain was predicted and shown to alter the specifity to asparagine. These results could enable us to rationally alter the primary structure of pharmacologically important peptide antibiotics such as penicillins, cyclosporins and vancomycins, simply by site-directed mutation of their adenylation domains. In addition to broadening the heterogeneity of a natural product, reducing a secondary side-specificity might also be desirable. Engineering of the corresponding substrate recognition sites could reduce undesired byproducts produced during the large-scale industrial production of naturally occurring compounds (e.g. cyclosporin A and vancomycin).

Materials and methods

PCR amplification and cloning of PheA and AspA mutants

All PheA mutants were constructed by site-directed mutagenesis of pPheA [9] using inverse PCR. PCR amplification of the entire plasmid was performed with the following 5′-phosphorylated oligonucleotides using the 'Expand long-range PCR' system (Boehringer Mannheim, Germany) in accordance with the manufacturer's protocol (mutated codons are bold and italicized: (1) 3′-A236: 5′-ATCAAAAGAGATGCTGGCA-3′; (2) 5′-A236L: 5′-***TTA***TCTGTATGGGAGATGTTTATG-3′; (3) 3′-W239: 5′-TACAGATGCATCAAAAGAG-3′; (4) 5′-W239G: 5′-***GGA***GAGATGTTTATGGCTTTGTTAAC-3′; (5) 5′-W239L: 5′-***TTA***GAGATGTTTATGGCTTTGTTA-3′; (6) 3′-T278: 5′-AATAACAGTGATTTCCTTTTGG-3′; (7) 5′-T278M: 5′-***ATG***CTGCCACCTACCTATGTAGTT-3′; (8) 5′-T278Q: 5′-***CAG***CTGCCACCTACCTATGTAGTT-3′; (9) 3′-I299: 5′-TAACGTTTGTATCGATAAAATAC-3′; (10) 5′-I299T: 5′-***ACT***ACAGCAGGCTCAGCTAC-3′; (11) 3′-A301G: 5′-***TCC***TGTAATTAACGTTTGTATCG-3′; (12) 3′-A301V: 5′-***AAC***TGTAATTAACGTTTGTATCG-3′; (13) 5′-A301: 5′-GGCTCAGCTACCTCGCCT-3′; (14) 3′-A322: 5′-AATGTAAGTTACTTTCTCCTTC-3′; (15) 5′-A322D: 5′-AA***GAT***TATGGCCCTACGGAAACA-3′; (16) 5′-A322E: 5′-AAT***GAA***TATGGCCCTACGGAAACA-3′; (17) 5′-A322G: 5′-AAT***GGC***TATGGCCCTACGGAAACA-3′; (18) 5′-A322I: 5′-AAT***ATT***TATGGCCCTACGGAAACA-3′; (19) 5′-A322K: 5′-AAT***AAG***TATGGCCCTACGGAAACA-3′; (20) 5′-A322N: 5′-AAT***TTG***TATGGCCCTACGGAAACA-3′; (21) 5′-A322Q: 5′-AAT***CAG***TATGGCCCTACGGAAACA-3′; (22) 5′-A322S: 5′-AAT***AGC***TATGGCCCTACGGAAACA-3′; (23) 3′-I330: 5′-AGTT-GTTTCCGTAGGGC-3′; and (24) 5′-I330V: 5′-***GTT***TGTGCGACTACATGGG-3′.

PCR products were purified using the 'QIAquick-spin' PCR purification kit (Qiagen), blunted and intramolecularly ligated in the presence of *Dpn*I. This restriction enzyme requires a methylated recognition site and therefore allowed selective decomposition of the PCR template pPheA (purified from a *dam*$^+$ strain). Standard procedures were applied for DNA manipulations [36] and the preparation of the recombinant plasmids using *Escherichia coli* strain XL1-Blue [37]. Cloning of the PCR fragments yielded the plasmids pPheA(A236L) (using oligonucleotides 1 and 2), pPheA(W239G) (3 plus 4), pPheA(W239L) (3 plus 5), pPheA(T278M) (6 plus 7), pPheA(T278Q) (6 plus 8), pPheA(I299T) (9 plus 10), pPheA(A301G) (11 plus 13), pPheA(A301V) (12 plus 13), pPheA(A322D) (14 plus 15), pPheA(A322E) (14 plus 16), pPheA(A322G) (14 plus 17), pPheA(A322I) (14 plus 18), pPheA(A322K) (14 plus 19), pPheA(A322N) (14 plus 20), pPheA(A322Q) (14 plus 21), pPheA(A322S) (14 plus 22), and pPheA(I330V) (23 plus 24). The integrity of all constructs was confirmed by sequencing using the ABI prism 310 Genetic Analyzer (ABI).

The DNA fragment encoding the aspartate-activating A domain of SrfA-B was PCR amplified from chromosomal DNA of *Bacillus subtilis* JH642 using the oligonucleotides (25) 5′-AspA: 5′-AATCCATGGCGAACGTTCGGCTGTCTG-3′, and (26) 3′-AspA: 5′-AATGGATCCGGCCAAGGCCTTGCC-3′ [22,30–32,38]. The PCR product was purified, digested with *Nco*I and *Bam*HI, and ligated into the His-tag vector pQE60, which was cut in the same manner. Cloning yielded the plasmid pAspA, which could then be used as a template for the generation of pAspA(H322E) using inverse PCR as described above: (27) 5′-AspA(H322E): 5′-TAAT***GAG***TACGGCCCGACAGAAGC-3′ (28), and 3′-AspA(H322E): 5′-ATAAATTCGGTATGTCCATAC-3′. The integrity of both constructs was confirmed by DNA sequencing.

Expression and purification of wild-type and mutant A domains

Expression of the mutant genes and purification of the His_6-tagged proteins were carried out as described previously [10,39]. Ligating into the *Bam*HI site of pQE60 results in appending the amino acid sequence 'GSRSHHHHHH' at the carboxyl terminus of each recombinant protein. As judged by SDS–PAGE [40], most proteins could be purified to apparent homogeneity using single-step Ni^{2+}-affinity chromatography; two constructs were insoluble (Table 2). Fractions containing the recombinant proteins were pooled and dialyzed against assay buffer (50 mM HEPES, pH 8.0, 100 mM sodium chloride, 10 mM magnesium chloride, 2 mM dithioerythritol (DTE) and 1 mM EDTA). After addition of 10% glycerol (w/v), the proteins could be stored at –80°C with no observable loss of activity. Protein concentrations were determined using the calculated extinction coefficients for their absorbance at 280 nm (A_{280nm}): 64,060 M^{-1} cm^{-1} for PheA and all PheA mutants except PheA(W239Xaa), 58,370 M^{-1} cm^{-1} for PheA(W239Xaa), and 39,780 M^{-1} cm^{-1} for AspA and AspA(H322E).

ATP–pyrophosphate exchange assay

The ATP–pyrophosphate exchange reaction was carried out to examine the activity and specificity of all recombinant A domains purified [8,10]. The specificity was checked with all proteinogenic amino acids, as well as L-ornithine and D-phenylalanine. Reaction mixtures contained (final volume: 100 µl): 50 mM HEPES, pH 8.0, 100 mM sodium chloride, 10 mM magnesium chloride, 2 mM DTE, 1 mM EDTA, 0–2 mM amino acid and 250 nM enzyme. The reaction was initiated by the addition of 2 mM ATP, 0.2 mM tetrasodium pyrophosphate and 0.15 µCi (16.06 Ci/mmol) of tetrasodium [^{32}P]pyrophosphate (NEN/DuPont) and incubated at 37°C for 10 min. Reactions were quenched by adding 0.5 ml of a stop mix containing 1.2% (w/v) activated charcoal, 0.1 M tetrasodium pyrophosphate and 0.35 M perchloric acid. Subsequently, the charcoal was pelleted by centrifugation, washed once with 1 ml water and resuspended in 0.5 ml water. After addition of 3.5 ml liquid scintillation fluid (Rotiscint Eco Plus; Roth), the charcoal-bound radioactivity was determined by liquid scintillation counting (LSC) using a 1900CA Tri-Carb liquid scintillation analyzer (Packard).

Computer analysis for the determination of putative substrate-binding pockets

The sequences of 160 A domains were retrieved from publicly accessable databases (NCBI, Swiss–Prot etc.). After outlining the 100-aa stretches between core motifs A4 and A5, the sequences were aligned using the program MegAlign from the DNA Star package, applying the Clustal method with default parameters. The only purpose of this step was to ease the assessment of the constituents of the proposed binding pockets. This goal was achieved by considering their positions relative to adjacent, highly conserved core motifs (A4 and A5, as well as the two unnamed motifs 'TPS' and 'GE'; structural 'anchors'). All sequences were trimmed to the constituents of their presumed binding pockets, and new sequence alignments and phylogenetic studies were performed using only these ten residues. Clustering of signature sequences according to their specificity was obtained using several

methods and sets of parameters. The particular phylogenetic tree shown in Figure 3 was constructed as in Hein [41] applying the following parameters: gap penalty 12, gap length penalty 6 and Ktuple 2.

Acknowledgements

We are indebted to Christopher T. Walsh and Thomas A. Keating for critical reading of the manuscript and valuable suggestions. We are obliged to Peter Brick for providing Figure 1, and for excellent technical assistance we thank Inge Schüler. We also acknowledge the contribution of Jürgen May and Sascha Dökel, who helped with the cloning of pAspA. T.S. is recipient of an European Molecular Biology Organization fellowship, and H.D.M. acknowledges his Ph.D. fellowship from the Stiftung Stipendien-Fonds des Verbandes der Chemischen Industrie. This work was supported by the Deutsche Forschungsgemeinschaft, EG project Cell Factories and the Fonds der Chemischen Industrie.

References

1. Marahiel, M.A., Stachelhaus, T. & Mootz, H.D. (1997). Modular peptide synthetases involved in nonribosomal peptide synthesis. *Chem. Rev.* **97**, 2651-2673.
2. von Döhren, H., Keller, U., Vater, J. & Zocher, R. (1997). Multifunctional peptide synthetases. *Chem. Rev.* **97**, 2675-2705.
3. Cane, D.E., Walsh, C.T. & Khosla, C. (1998). Harnessing the biosynthetic code: combinations, permutations, and mutations. *Science* **282**, 63-68.
4. Mootz, H.D. & Marahiel, M.A. (1997). Biosynthetic systems for nonribosomal peptide antibiotic assembly. *Curr. Opin. Chem. Biol.* **1**, 543-551.
5. Quadri, L.E.N., Sello, J., Keating, T.A., Weinreb, P.H. & Walsh, C.T. (1998). Identification of a *Mycobacterium tuberculosis* gene cluster encoding the biosynthesis enzymes for assembly of the virulence-conferring siderophore mycobactin. *Chem. Biol.* **5**, 631-645.
6. Gehring, A.M., *et al.*, & Perry, R.D. (1998). Iron acquisition in plaque: molecular logic in enzymatic biogenesis of yersiniabactin by *Yersinia pestis*. *Chem. Biol.* **5**, 573-586.
7. Lambalot, R.H., *et al.*, & Walsh, C.T. (1996). A new enzyme superfamily – the phosphopantetheinyl transferases. *Chem. Biol.* **3**, 923-936.
8. Stachelhaus, T., Mootz, H.D., Bergendahl, V. & Marahiel, M.A. (1998). Peptide bond formation in nonribosomal peptide biosynthesis. *J. Biol. Chem.* **273**, 22773-22781.
9. Stachelhaus, T. & Marahiel, M.A. (1995). Modular structure of peptide synthetases revealed by dissection of the multifunctional enzyme GrsA. *J. Biol. Chem.* **270**, 6163-6169.
10. Mootz, H.D. & Marahiel, M.A. (1997). The tyrocidine biosynthesis operon of *Bacillus brevis*: complete nucleotide sequence and biochemical characterization of functional internal adenylation domains. *J. Bacteriol.* **179**, 6843-6850.
11. Dieckmann, R., Lee, Y-O., van Liempt, H., von Döhren, H. & Kleinkauf, H. (1995). Expression of an active adenylate-forming domain of peptide synthetases corresponding to acetyl-CoA-synthetases. *FEBS Lett.* **357**, 212-216.
12. Conti, E., Franks, N.P. & Brick, P. (1996). Crystal structure of firefly luciferase throws light on a superfamily of adenylate-forming enzymes. *Structure* **4**, 287-298.
13. Conti, E., Stachelhaus, T., Marahiel, M.A. & Brick, P. (1997). Structural basis for the activation of phenylalanine in the nonribosomal biosynthesis of gramicidin S. *EMBO J.* **16**, 4174-4183.
14. Gocht, M. & Marahiel, M.A. (1994). Analysis of core sequences in the D-Phe activating domain of the multifunctional peptide synthetase TycA by site-directed mutagenesis. *J. Bacteriol.* **176**, 2654-2662.
15. Hamoen, L.W., Eshuis, H., Jongbloed, J., Venema, G. & van Sinderen, D. (1994). A small gene, designated *comS*, located within the coding region of the fourth amino acid-activation domain of *srfA*, is required for competence developement in *Bacillus subtilis*. *Mol. Microbiol.* **15**, 55-63.
16. Saito, M., Hori, K., Kurotsu, T., Kanda, M. & Saito, Y. (1995). Three conserved glycine residues in valine activation of gramicidin S synthetase 2 from *Bacillus brevis*. *J. Biochem.* **117**, 276-282.
17. Pavela-Vrancic, M., Pfeifer, E., Schröder, W., von Döhren, H. & Kleinkauf, H. (1994). Identification of the ATP-binding site in tyrocidine synthetase 1 by selective modification with fluorescein 5′-isothiocyanate. *J. Biol. Chem.* **269**, 14962-14966.
18. Pavela-Vrancic, M., Pfeifer, E., van Liempt, H., Schäfer, H-J., von Döhren, H. & Kleinkauf, H. (1994). ATP binding in peptide synthetases: determination of contact sites of the adenine moiety by photoaffinity labeling of tyrocidine synthetase 1 with 2-azidoadenosine triphosphate. *Biochemistry* **33**, 6276-6283.
19. de Crécy-Lagard, V., *et al.*, & Blanc, V. (1997). Streptogramin B biosynthesis in *Streptomyces pristinaespiralis* and *Streptomyces virginae*: molecular characterization of the last structural peptide synthetase gene. *Antimicrob. Agents Chemother.* **41**, 1904-1909.
20. Elsner, A., Engert, H., Sänger, W., Hamoen, L., Venema, G. & Bernhard, F. (1997). Substrate specificity of hybrid modules from peptide synthetases. *J. Biol. Chem.* **272**, 4814-4819.
21. Husi, H., Schörgendorfer, K., Stempfer, G., Taylor, P. & Walkinshaw, M.D. (1997). Prediction of substrate-specific pockets in the cyclosporin synthetase. *FEBS Lett.* **414**, 532-536.
22. Cosmina, P., Rodriguez, F., de Ferra, F., Perego, M., Venema, G. & van Sinderen, D. (1993). Sequence and analysis of the genetic locus responsible for surfactin synthesis in *Bacillus subtilis*. *Mol. Microbiol.* **8**, 821-831.
23. Turgay, K., Krause, M. & Marahiel, M.A. (1992). Four homologous domains in the primary structure of GrsB are related to domains in a superfamily of adenylate-forming enzymes. *Mol. Microbiol.* **6**, 529-546.
24. Yakimov, M.M., Kröger, A., Slepak, T.N., Giuliano, L., Timmis, K.N. & Golyshin, P.N. (1998). A putative lichenysin A synthetase operon in *B. licheniformis*: initial characterization. *Biochim. Biophys. Acta* **1339**, 141-153.
25. Konz, D., Dökel, S. & Marahiel, M.A. (1999). Molecular and biochemical characterization of the protein template controlling biosynthesis of the lipopeptide lichenysin. *J. Bacteriol.* **181**, 133-140.
26. Cole, S.T., *et al.*, & Barrell, B.G. (1998). Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature* **393**, 537-544.
27. van Wageningen, A.M., *et al.*, & Solenberg, P.J. (1998). Sequencing and analysis of genes involved in the biosynthesis of a vancomycin group antibiotic. *Chem. Biol.* **5**, 155-162.
28. Steller, S., *et al.*, & Vater, J. (1999). Structural and functional organization of the fengycin synthetase multienzyme system from *Bacillus subtilis* b213 and A1/3. *Chem. Biol.* **6**, 31-41.
29. Lipmann, F. (1971). Attempt to map a process evolution of peptide biosynthesis. *Science* **173**, 875-884.
30. Vollenbroich, D., Kluge, B., D'Souza, C., Zuber, P., & Vater, J. (1993). Analysis of a mutant amino acid-activating domain of surfactin synthetase bearing a serine to alanine substitution at the site of carboxyl thioester formation. *FEBS Lett.* **325**, 220-224.
31. D'Souza, C., Nakano, M.M., Corbell, N. & Zuber, P. (1993). Amino-acylation site mutations in amino acid-activating domains of surfactin synthetase: effects on surfactin production and competence development in *Bacillus subtilis*. *J. Bacteriol.* **175**, 3502-3510.
32. Nakano, M.M., Marahiel, M.A. & Zuber, P. (1988). Identification of a genetic locus required for biosynthesis of the lipopeptide antibiotic surfactin in *Bacillus subtilis*. *J. Bacteriol.* **170**, 5662-5668.
33. Kleinkauf, H. & von Döhren, H. (1988). Peptide antibiotics, β-lactams, and related compounds. *CRC Crit. Rev. Biotechnol.* **8**, 1-32.
34. Lawen, A. & Traber, R. (1993). Substrate specificities of cyclosporin synthetase and peptolide SDZ214-103 synthetase. *J. Biol. Chem.* **268**, 20452-20465.
35. Lawen, A., Traber, R., Reuille, R. & Ponelle, M. (1994). *In vitro* biosynthesis of ring-extended cyclosporins. *Biochem. J.* **300**, 395-399.
36. Sambrook, J., Fritsch, E.F. & Maniatis, T. (1989). *Molecular Cloning: A Laboratory Manual (*2nd edn). Cold Spring Harbor Laboratory Press, Cold Spring Harbor.
37. Bullock, W.O., Fernandez, J.M. & Short, J.M. (1987). XL1-Blue: a high efficiency plasmid transforming *recA E. coli* strain with β-galactosidase selection. *Biotechniques* **5**, 376-379.
38. Hoch, J.A. & Mathews, J. (1973). Chromosomal location of pleiotropic sporulation mutations in *Bacillus subtilis*. *Genetics* **73**, 215-228.
39. Stachelhaus, T., Hüser, A. & Marahiel, M.A. (1996). Biochemical characterization of peptidyl carrier protein (PCP), the thiolation domain of multifunctional peptide synthetases. *Chem. Biol.* **3**, 913-921.
40. Laemmli, U.K. (1970). Cleavage of structural proteins during the assembly of the head of bacteriophage T4. *Nature* **227**, 680-685.
41. Hein, J.J. (1990). Unified approach to alignment and phylogenies. *Methods Enzymol.* **183**, 626-645.