FEBS Letters 374 (1995) 246–248 FEBS 16190

Divergent sequence motifs correlated with the substrate specificity of (methyl)malonyl-CoA:acyl carrier protein transacylase domains in modular polyketide synthases

Stephen F. Haydock[a], Jesús F. Aparicio[a], István Molnár[a], Torsten Schwecke[a], Lake Ee Khaw[a], Ariane König[a], Andrew F.A. Marsden[a], Ian S. Galloway[a], James Staunton[b], Peter F. Leadlay[a,*]

[a]*Cambridge Centre for Molecular Recognition, Department of Biochemistry, University of Cambridge, Tennis Court Road, Cambridge CB2 1QW, UK*
[b]*Cambridge Centre for Molecular Recognition, University Chemical Laboratory, University of Cambridge, Lensfield Road, Cambridge CB2 1EW, UK*

Received 25 August 1995

Abstract The amino acid sequences of a large number of polyketide synthase domains that catalyse the transacylation of either methylmalonyl-CoA or malonyl-CoA onto acyl carrier protein (ACP) have been compared. Regions were identified in which the acyltransferase sequences diverged according to whether they were specific for malonyl-CoA or methylmalonyl-CoA. These differences are sufficiently clear to allow unambiguous assignment of newly-sequenced acyltransferase domains in modular polyketide synthases. Comparison with the recently-determined structure of the malonyltransferase from *Escherichia coli* fatty acid synthase showed that the divergent region thus identified lies near the acyltransferase active site, though not close enough to make direct contact with bound substrate.

Key words: Acyltransferase; Structural motif; Sequence homology; Polyketide synthase; Fatty acid synthase

1. Introduction

Acyl-CoA:acyl carrier protein transferases (ATs) are key activities within fatty acid synthases [1] and polyketide synthases [2] that catalyse the transfer of acyl-CoA esters, most frequently malonyl or methylmalonyl groups, onto acyl carrier proteins (ACP). For fatty acid synthases, the extension units are derived almost wholly from malonyl-CoA, and this is also true of type II (dissociated) polyketide synthases from actinomycetes whose products are typically aromatic polyketides; and of type I (multifunctional) polyketide synthases from fungi [2]. In contrast, reduced or complex polyketides usually contain both acetate and propionate units, and are synthesised in actinomycetes by modular type I polyketide synthases, in which a different set of enzymes or 'module' (including an acyl-CoA:ACP acyltransferase) is used for every cycle of polyketide chain extension [3,4].

The modular organisation uncovered for complex polyketide synthases [3,4] prompted speculation that individual methylmalonyl-CoA:ACP transferase domains might show different stereospecificities towards the (2*R*) and (2*S*) stereoisomers of methylmalonyl-CoA, and that this might account for the different stereochemical outcome observed at the methyl-branched positions along the chain of a typical complex polyketide. It was subsequently shown that in fact all six methylmalonyltransferase domains of the erythromycin-producing polyketide synthase possess the same substrate stereospecificity, for (2*S*)-methylmalonyl-CoA, and the overall outcome must be governed by other enzyme components of the synthase [5].

However, a key question that remains, even given the availability of a high resolution X-ray crystal structure for the malonyl-CoA:ACP transferase from *Escherichia coli* fatty acid synthase [6], is the basis for the specificity of individual acyl-CoA:ACP acyltransferase (AT) domains or enzymes, for the transfer of either malonyl or methylmalonyl groups onto ACP. The nucleotide sequence is now available for the entire gene cluster encoding the modular polyketide synthase for biosynthesis of the immunosuppressant rapamycin, which contains seven propionate and seven acetate units [7]. Comparison of the sequences of these fourteen AT domains with the sequences of other ATs for fatty acid and polyketide synthases has been used here to identify divergent sequence motifs that clearly distinguish a malonyl- from a methylmalonyl-specific AT domain in a modular polyketide synthase. Although this search was conducted without reference to the possible location of the active site, a divergent sequence motif was localized in a helix which contains an invariant Gln (Gln-63 in the *E. coli* MCAT) that is in the active site [6]. However, the residues of the motif are too far removed to be in direct contact with the substrate-binding site.

2. Materials and methods

Protein sequence databanks (PIR, Swiss-Prot, OWL) were searched using the programs FASTA [8] and BLAST [9]. Sequence alignments were generated using the programs PILEUP [10] and ClustalW [11], and refined using the programs LINEUP and PRETTY [10].

3. Results and discussion

3.1. Identification of divergent sequence motifs

Inspection of the aligned amino acid sequences of the fourteen AT domains from the rapamycin-producing polyketide synthase of *Streptomyces hygroscopicus* [7] allowed us to identify a region where the seven AT domains thought to incorporate acetate units share a common motif that differs significantly from the sequence of the AT domains thought to be responsible for incorporation of propionate units (Fig. 1). Inclusion in this alignment of the AT- domains of the erythromycin-producing polyketide synthase, of which the first (AT0) is

*Corresponding author. Fax: (44) (1223) 333345.
E-mail: pfl10@bioc.cam.ac.uk

Abbreviations: ACP, acyl carrier protein; AT, acyl-CoA:ACP acyltransferase; PKS, polyketide synthase.

SSDI 0014-5793(95)01119-6

AT	MOTIF	ACTIVE SITE
	ACETATE ATs	
Rap02	**ETGYA**QPALFAL**Q**V**A**L**F**G**LL** - 11aa - GHSVG	
Rap05	**ETGYA**QPALFAL**Q**V**A**L**F**G**LL** - 11aa - GHSVG	
Rap08	**ETGYA**QPALFAL**Q**V**A**L**F**G**LL** - 11aa - GHSVG	
Rap09	**ETGYA**QPALFAL**Q**V**A**L**F**G**LL** - 11aa - GHSVG	
Rap11	**ETGYA**QPALFAL**Q**V**A**L**F**G**LL** - 11aa - GHSVG	
Rap12	**ETGYA**QPALFAM**Q**V**A**L**F**G**LL** - 11aa - GHSVG	
Rap14	D**TLYA**QAGIFAME**A**A**L**F**G**LL** - 11aa - GHSIG	
Ave 2	Q**TRYA**QPALFAF**Q**VALHR**LL** - 12aa - GHSLG	
MSAS	SSDRVQILTYVM**Q**IGLSA**LL** - 11aa - GHSVG	
MCAT	K**T**WQTQPALLTASV**A**LYRVW - 12aa - GHSLG	
	PROPIONATE ATs	
Rap01	**RVDVV**QPASWAV**M**V**S**L**A**A**V**W - 11aa - GHSQG	
Rap03	**RVDVV**QPASWAV**M**V**S**L**A**A**V**W - 11aa - GHSQG	
Rap04	**RVDVV**QPASWAV**M**V**S**L**A**A**V**W - 11aa - GHSQG	
Rap06	**RVDVV**QPASWRM**M**V**S**L**A**A**V**W - 11aa - GHSQG	
Rap07	**RVDVV**QPASWAV**M**V**S**L**A**A**V**W - 11aa - GHSQG	
Rap10	**RVDVV**QPASWAV**M**V**S**L**A**A**V**W - 11aa - GHSQG	
Rap13	**RVDVV**QPASWAM**M**V**S**L**A**A**V**W - 11aa - GHSQG	
Ery 1	**RVDVV**QPVMFAV**M**V**S**L**A**SMW - 11aa - GHSQG	
Ery 2	**RVDVV**QPVLFAV**M**V**S**L**A**RLW - 11aa - GHSQG	
Ery 3	**RVDVV**QPVLFAV**M**V**S**L**A**ELW - 11aa - GHSQG	
Ery 4	**RVDV**LQPVLFSI**M**V**S**L**A**ELW - 11aa - GHSQG	
Ery 5	**RVDVV**QPALFAV**M**V**S**L**A**ALW - 11aa - GHSQG	
Ery 6	**RVDVV**QPVLFSV**M**V**S**L**A**RLW - 11aa - GHSQG	
Ole 5	**RVDVV**QPALWAV**M**V**S**L**A**RTW - 11aa - GHSQG	
Ole 6	**RVDVV**QPALWAV**M**V**S**L**A**RTW - 11aa - GHSQG	
Nem 6	**RADVV**QPVLFSV**M**V**S**L**A**ALW - 11aa - GHSQG	
Nem 9	**RADVV**QPVLFSV**M**V**S**L**A**ALW - 11aa - GHSQG	
Sor 5	**RVDVV**QPALFAV**M**V**S**L**A**ALW - 11aa - GHSQG	
MAS	GI**D**K**V**QPAVFAVQV**A**L**A**ATM - 11aa - GHSMG	
	STARTER ATs	
Ery 0	**RVEVV**QPALFAVQT**S**L**A**ALW - 11aa - GHSIG	
Ave 0	**RVDVV**QPTLFAV**M**I**S**L**A**ALW - 11aa - GHSLG	

Fig. 1. Alignment of the sequences for the divergent motifs found in AT domains of type I polyketide synthases, and active sites. The intervening number of amino acids (aa) is indicated. Rap 1–14 indicate the different AT domains present in modules 1–14 of the PKS for rapamycin biosynthesis [7]. Ery 1–6, Ave 2, Ole 5 and 6, Nem 6 and 9, and Sor 5, indicate AT domains present in the corresponding modules of the modular PKSs for erythromycin [3,4], avermectin, oleandomycin [13], nemadectin [14] and soraphen A [15] biosynthesis respectively. MSAS represents the 6-methylsalicylic acid synthase AT [16], and MAS that of mycocerosic acid synthase [17]. The sequence of the *Escherichia coli* malonyl-CoA:acyl carrier protein transacylase (MCAT) [6] is also shown. Ery 0 and Ave 0 indicate AT domains that load the starter unit in the erythromycin [12] and avermectin PKSs respectively.

required for loading of propionyl-CoA and the other six all incorporate methylmalonyl-CoA [5,12], localised the divergent substrate-specific sequence motifs to a relatively small stretch of 20 amino acids, N-terminal of the catalytic serine residue found in the consensus sequence GXSXG common to all these enzymes (Fig. 1).

As shown in Fig. 1, the derived consensus sequence motif for those AT domains that incorporate propionate extender units is RVDVV-7-M-1-S-1-AXhW, where h represents an aliphatic hydrophobic, X is either Arg, Ser, Ala or Glu, and those residues common to all AT domains are omitted. In those AT domains that incorporate acetate extender units the same stretch of 20 amino acids has the consensus sequence ETGYA-7-Q-1-A-1-FGLL. *Rap* AT14 differs significantly from the rest, for reasons that are presently unclear. It may be seen in Fig. 1 that the two AT domain sequences published for the polyketide synthase for the macrolide oleandomycin (which contains propionate extender units only) [13] both match the propionate motif exactly. Sequence data have also been obtained (Galloway, I.S. and Leadlay, P.F., unpublished results) for the acetate-specific AT domain from the second module of the avermectin-producing polyketide synthase in *Streptomyces avermitilis*, and the sequence is an excellent match to the derived acetate-specific motif. AT sequences have also been reported for modules seven and nine of the PKS from *Streptomyces cyaneogriseus* that synthesises the avermectin analogue nemadectin [14]. These modules are both predicted to be propionate-specific, and in agreement with this both AT sequences convincingly match the propionate consensus. Finally, partial sequence is available for the recently-characterised gene cluster for the antifungal macrolide soraphen A from the gliding bacterium *Sorangium cellulosum* [15]. The sequence of the AT matched the propionate consensus in Fig. 1, in agreement with the authors' assignment of the sequenced DNA to module 5 of the PKS. It appears likely, given the overall high sequence similarity observed in all modular type I polyketide synthases so far reported, that these divergent motifs will be particularly valuable in predicting the acyltransferase specificity of newly-sequenced polyketide synthase modules.

When the acetate-specific AT domain of 6-methylsalicylic acid synthase from *Penicillium patulum*, [16] a typical type I fungal polyketide synthase, was aligned with the sequences in Fig. 1, it was seen to have a sequence SSDRV-7-QIGLSALL, a better match to the acetate-specific motif. The rat and chicken type I fatty acid synthases differ too much in sequence from the modular polyketide synthases in this region for comparison to be useful. However, mycocerosic acid synthase from *Mycobacterium tuberculosis* [17], which utilises methylmalonyl-CoA to make multiply branched fatty acids, aligns reasonably with the propionate motif as expected.

3.2. *Position of the divergent sequence motifs in relation to the AT active site*

When the sequence of the *Escherichia coli* malonyl-CoA:ACP transacylase, the crystal structure for which has been recently reported at 1.5 Å resolution, is aligned with the AT sequences from the modular polyketide synthases (Fig. 1), a reasonably good match is found with the acetate-specific consensus motif. This encouraged us to examine the crystal structure for the possible significance of these sequence motifs. It is striking that the position and the length of the motifs we have identified coincide exactly with the position and length of helix 6 (residues 59–79 inclusive) in the *E. coli* AT structure, including the polar residue Gln-63 (*E. coli* AT numbering), which is invariant in all the sequences aligned in Fig. 1, and which contributes its side-chain to the AT active site. However, closer analysis showed that the residues of the divergent sequence motifs are too far away to contribute directly to a binding site for the acyl portion of the substrate acyl-CoA. Two other single residues which, from the crystal structure, are in the AT active site, are consistently divergent in the modular type I polyketide synthase AT domains. First, residue Gln-250 (*E. coli* numbering) which acts as an H-bond acceptor for the active-site histidine, is always Gln in the AT domains specific for acetate extender units, and always Asn in those specific for propionate extender units. Second, residue Leu-93, which is adjacent in the sequence to the serine nucleophile and whose main chain carbonyl donates a H-bond to a water molecule, is always replaced by Gln in the AT domains specific for propion-

ate units. Again, no structural significance can yet be given to these observations. For the present, the value of the motifs we have identified lies in the assignment of the substrate specificity of a newly- sequenced AT domain.

The crystal structure also reveals that residue Arg-117, which is highly conserved in all these enzymes, is also located in the active site. Arg-117 is substituted by Trp in the AT domain that loads a propionyl group onto the erythromycin-producing polyketide synthase [3,12], and it is also substituted by Trp in an analogous AT domain that loads isobutyryl or isovaleryl groups onto the avermectin-producing polyketide synthase (Marsden, A.F.A. and Leadlay. P.F., unpublished results). It is therefore tempting to ascribe Arg-117 a role in stabilizing the carboxylate group of (methyl)malonyl-CoA. However, several AT domains are known to contain Arg-117 but not to discriminate between acetyl-CoA and malonyl-CoA as substrates [6], and further work will be required to pinpoint the exact contribution made by this active site residue.

3.3. Conclusions

There is great interest in determining how acyltransferases from modular polyketide synthases choose the extender units for incorporation into the growing polyketide chain. It has been assumed that both the level of intracellular pools of acyl-CoA esters, at different stages of mycelial development, and the substrate specificity of the acyltransferases themselves might contribute to this choice. Our identification of a localised region of sequence that is near, but not in direct contact with, the active site and which diverges according to whether the AT domain is specific for malonyl- or methylmalonyl-CoA as substrate, raises intriguing questions as to how these changes exert their effect, and also suggests how it might be possible in future to alter the specificity of a given AT by localised mutagenesis.

Acknowledgements: We thank the Wellcome Trust, the Biotechnology and Biological Research Council, and Pfizer Inc., Groton, USA for financial support, the Spanish Ministry of Education for a fellowship (to J.F.A) and the Medical Research Council for a training fellowship (to S.F.H.). We also thank Dr Z. Derewenda for kindly supplying the co-ordinates of the *E. coli* MCAT, and Dr H.A.I. McArthur and one of the referees for their helpful comments on the ms.

References

[1] Wakil, S.J. (1989) Biochemistry 28, 4523–4530.
[2] Hopwood, D.A. and Sherman, D.H. (1990) Annu. Rev. Genet. 24, 37–66.
[3] Donadio, S., Staver, M.J., McAlpine, J.B., Swanson, S.J. and Katz, L. (1991) Science 252, 675–679.
[4] Cortés, J., Haydock, S.F., Roberts, G.A., Bevitt, D.J. and Leadlay, P.F. (1990) Nature 348, 176–178.
[5] Marsden, A.F.A., Caffrey, P., Aparicio, J.F., Loughran, M.S., Staunton, J. and Leadlay, P.F. (1994) Science 263, 378–380.
[6] Serre, L., Verbree, E.C., Dauter, Z., Stuitje, A.R. and Derewenda, Z.S. (1995) J. Biol. Chem. 270, 12961–12964.
[7] Schwecke, T., Aparicio, J.F., Molnár, I., König, A., Khaw, L.E., Haydock, S.F., Oliynyk, M., Caffrey, P., Cortés, J., Lester, J.B., Böhm, G.A., Staunton, J. and Leadlay, P.F. (1995) Proc. Natl. Acad. Sci. USA 92, 7839–7843.
[8] Pearson, W.R. (1990) Methods Enzymol. 183, 63–98.
[9] Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990) J. Mol. Biol. 215, 403–410.
[10] Devereux, J., Haeberli, P. and Smithies, O. (1984) Nucleic Acids Res. 12, 387–395.
[11] Thompson, J.D., Higgins, D.G. and Gibson, T.J. (1994) Nucleic Acids Res. 22, 4673–4680.
[12] Aparicio, J.F., Caffrey, P., Marsden, A.F.A., Staunton, J. and Leadlay, P.F. (1994) J. Biol. Chem. 269, 8524–8528.
[13] Swan, D.G., Rodríguez, A.M., Vilches, C., Méndez, C. and Salas, J.A. (1994) Mol. Gen. Genet. 242, 358–362.
[14] MacNeil, D.J., Occi, J.L., Gewain, K.M., MacNeil, T., Gibbons, P. H., Foor, F. and Morin, N. (1993) in: Industrial Microorganisms: Basic and Applied Molecular Genetics (Baltz, R.H., Hegeman, G.D. and Skatrud, P., Eds.) pp. 245–256. American Society for Microbiology, Washington DC.
[15] Schupp, T., Toupet, C., Cluzel, B., Neff, S., Hill, S., Beck, J. and Ligon, J.M. (1995) J. Bacteriol. 177, 3673–3679.
[16] Beck, J., Ripka, S., Siegner, A., Schiltz, E. and Schweizer, E. (1990) Eur. J. Biochem. 192, 487–498.
[17] Mathur, M. and Kolattukudy, P.E. (1992) J. Biol. Chem. 267, 19388–19395.