

Supporting Information

Identification of carotenoid biosynthetic genes

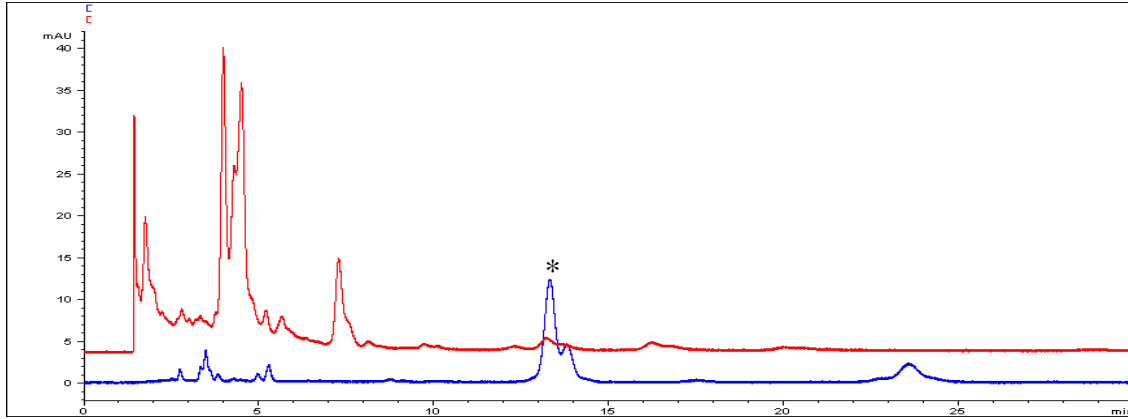


Figure S1: HPLC profiles (450 nm) of *S. tropica* crude extract (red) and partially purified isorenieratene (shown with *) from *S. coelicolor* (blue).

Isolation and characterization of sioxanthin

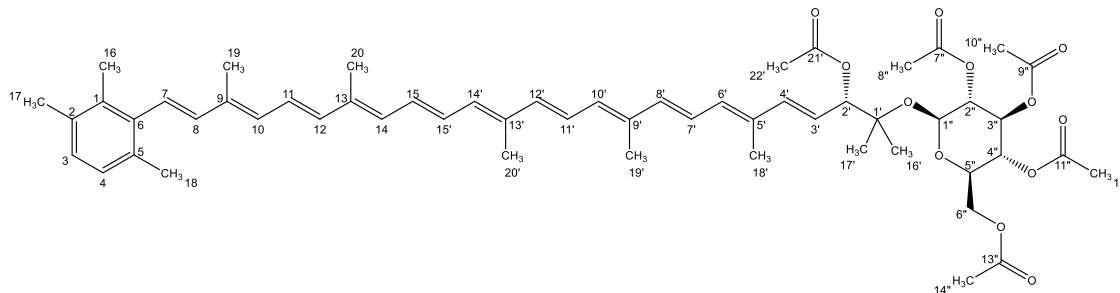


Figure S2: Penta-acetylated sioxanthin structure used in analysis and structure elucidation

The ^1H NMR spectrum has signals in the predicted regions for a glycosylated carotenoid. It shows methyl groups, including those from the added acetyl groups, in the δ 1.0-2.5 region. The structural features on the right side of the molecule, including the glycosyl and hydroxyl groups are located in the δ 3.5-6.0 region. The alkene chain is represented in the δ 6-6.75 region. Far downfield, at δ 6.95, the aromatic ring protons overlap. Carbon assignments were made using HMBC and HSQC data (Table S1)

Table S1: Carbon and proton NMR shifts and HMBC correlations (CD_3CN , 600MHz). “x” indicates atoms that could not be distinguished. Other missing values are due to overlapping signals in the spectra and are indistinguishable.

C#	δ_c	δ_H , mult.	COSY	HMBC	NOESY
1	134.7				
3	129 ^x	6.95, d (7.6) ^x		18	
4	128.1 ^x	6.93, d (7.6) ^x		18	
6	138.1				
7	126.5	6.76, d (3.5)			
16	17.1	2.21, s			
17 ^x	20.65	2.24, s		1, 3, 4, 6,	
18 ^x	20.65	2.24, s		1, 3, 4, 6	
19	12.8	2.08, s			
20	12.8				
1'	79.9				
2'	80.8	5.16, d (8.0)	5.66	1', 3', 4', 16', 21'	
3'	123.8	5.66, dd (15.6, 8.0)	6.33,	2', 5'	

			5.16		
4'	139.7	6.33	5.66	2', 3', 5', 6', 18'	
5'	135.2				
6'	133.9	6.28		7'	
7'	125.86	6.67		5'	
16'	22.3x	1.20, s		1', 2', 17'	
17'	24.0x	1.19, s		1', 2', 16'	
18'	12.7	1.89, s		4', 5', 6'	
19'	12.8	1.99		7'	
20'	12.8				
21'	170.6				
1''	95.9	4.91, d (8.0)	4.83	1'	3.82
2''	72.0	4.83, dd (9.8, 8.0)	4.91, 5.25	1'', 3'', 7''	4.99
3''	73.3	5.25, dd (9.8)	4.83, 4.99	2'', 4'', 9''	3.82
4''	69.5	4.99, dd (9.8)	5.25, 3.82	3'', 5'', 11''	4.83, 4.03
5''	72.1	3.82, ddd (8.0, 5.6, 2.5)	4.99, 4.20, 4.03	1'', 4''	4.91, 5.25, 4.20
6''	62.8	4.20x, dd (6.5, 5.2) 4.03x, dd (12.2, 2.6)	3.82 3.82	4'', 5'', 13'' 5'', 13''	3.82
7''	16.99				
8''	20.8	1.97, s		7''	
9''	170.7				
10''	21.0	2.03, s		9''	
11''	170.3				
12''	21.17	2.02			
13''	170.9				
14''	12.7	1.99, s		13''	

The structural features on the polar end of the molecule were clearly identified using the 2D NMR techniques. COSY confirmed the structure of a hexose as well as the 2'-hydroxyl and the 3',4'-desaturation (Figure). This desaturation showed correlations to the rest of the alkene chain, identifying the linkage between the sugar residue and the rest of the molecule. This linkage was confirmed by TOCSY data which identified the glycosyl group and the alkene chain as separate spin systems (Figure). The 2'-acetylated hydroxyl and the 3,4'-desaturation were confirmed by the COSY correlations as well as the splitting patterns in the ¹H NMR.

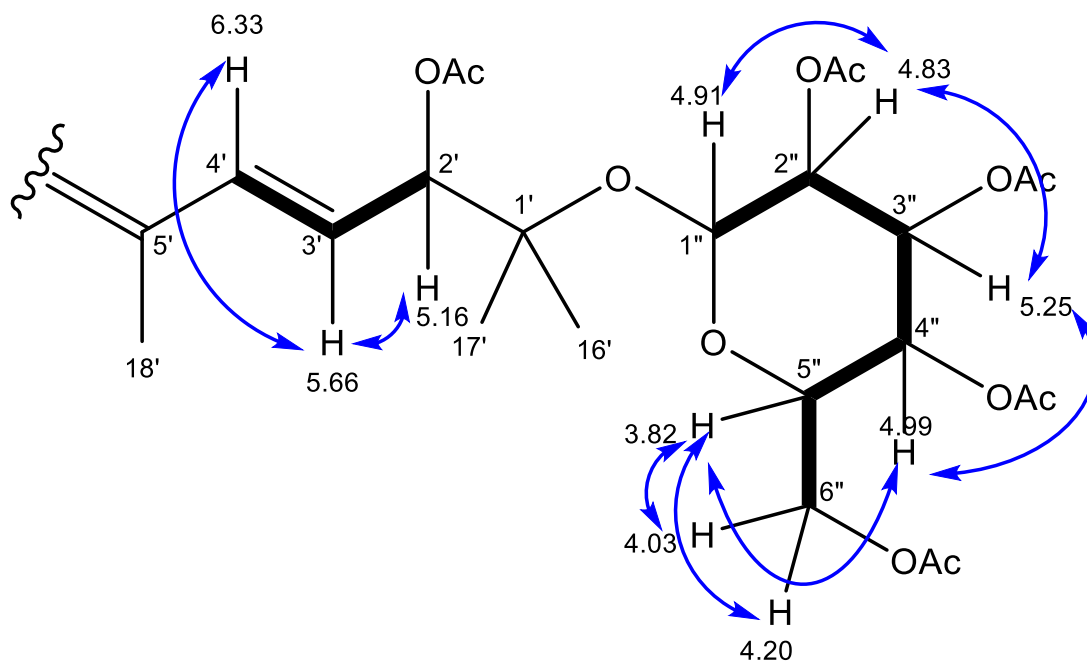


Figure S3: COSY and TOCSY correlations confirm the structural modifications on the right side of the molecule. Arrows depict COSY correlations. TOCSY spin systems are shown in bold bonds.

The linkage between the glycosyl group and the alkene chain was further confirmed through HMBC correlations (Figure). HMBC data also confirmed the presence of a total of five acetyl groups, corresponding to the five hydroxyl functional groups present in the molecule (Figure S5).

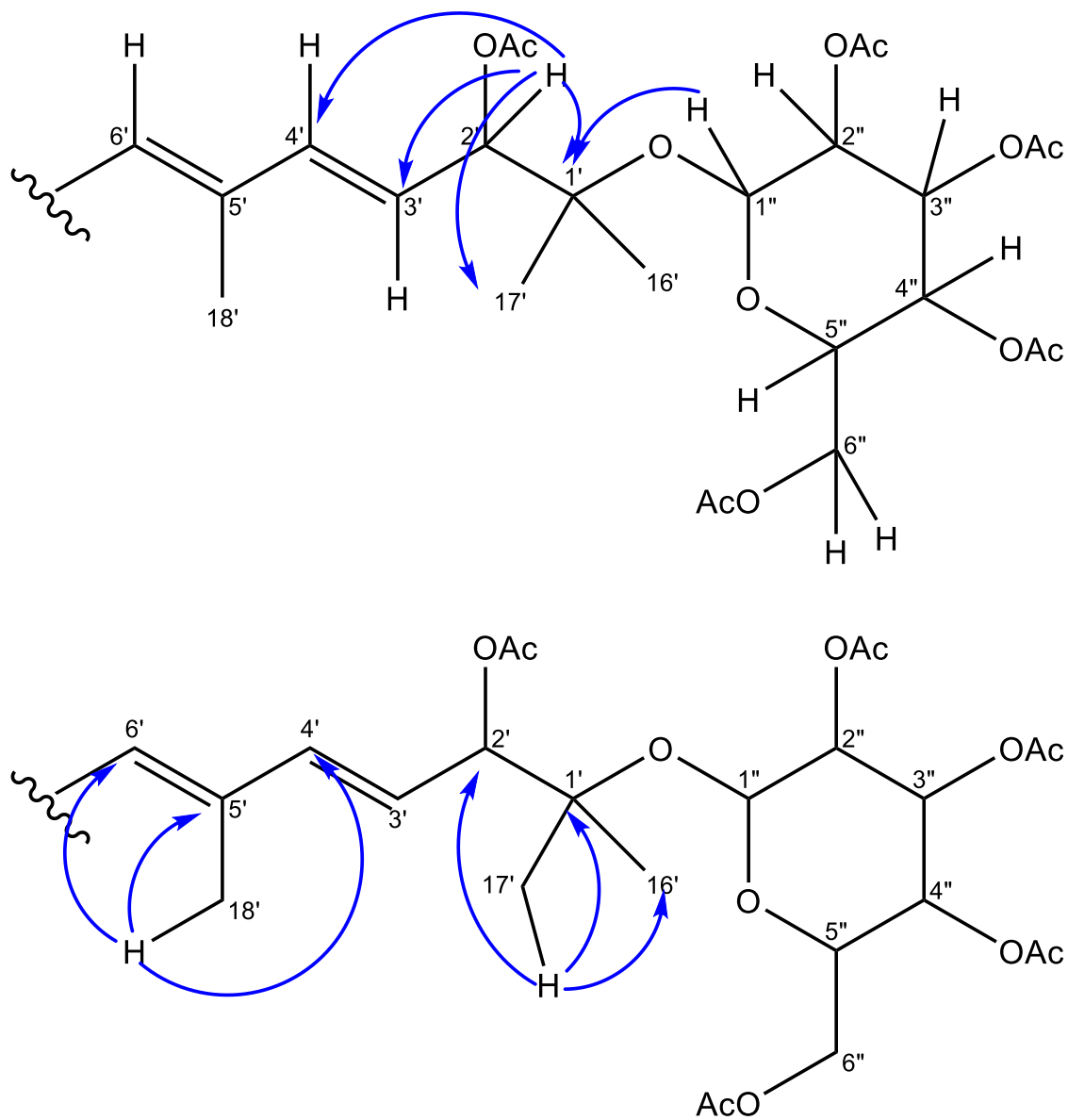


Figure S4: HMBC correlations confirming the region between TOCSY spin systems, linking the functional groups to the remainder of the molecule.

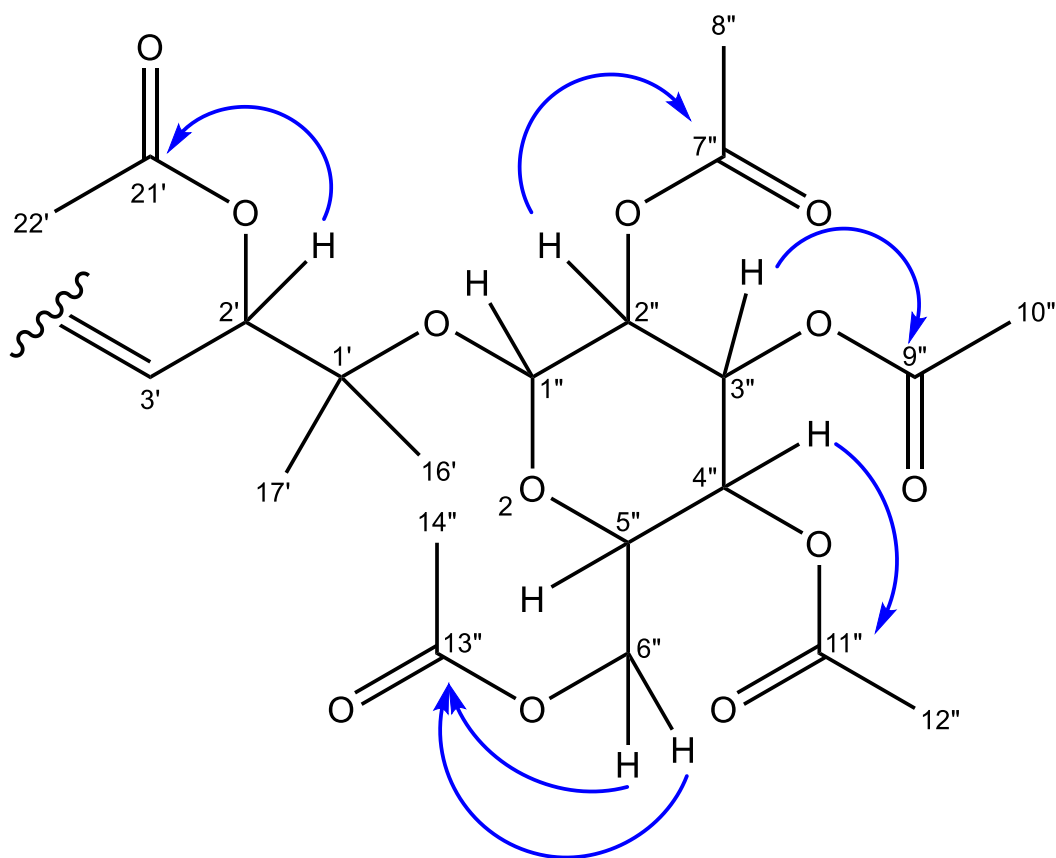


Figure S5: HMBC correlations showing correlations to the carbonyl carbons in each acetyl group.

The repetitive nature of the molecule makes it difficult to confidently assign every carbon and hydrogen atom. Though alkene and methyl groups can easily be identified, the individual isoprene units have overlapping carbon and hydrogen shifts, making them difficult to distinguish from one another. Comparisons to published NMR shifts aided in the assignment (Englert, 1995). All methyl carbons in this region have a shift of about $\delta 12.8$. The confidence in the structure of the alkene chain portion is enhanced by the UV/Vis and MS data, which adhere to the literature describing carotenoid structures (Englert, 1995).

The aromatic end of the compound was confirmed through the presence of aromatic hydrogens at carbons 3 ($\delta 6.93$) and 4 ($\delta 6.95$). Although these two positions are indistinguishable, they are consistent with published NMR shifts of a 1,2,5,6-tetrasubstituted benzene that is a

common structure element in carotenes (Figure S6)(Englert, 1995). The HMBC data show correlations to a methyl group carbon, referring to either carbon 17 or 18.

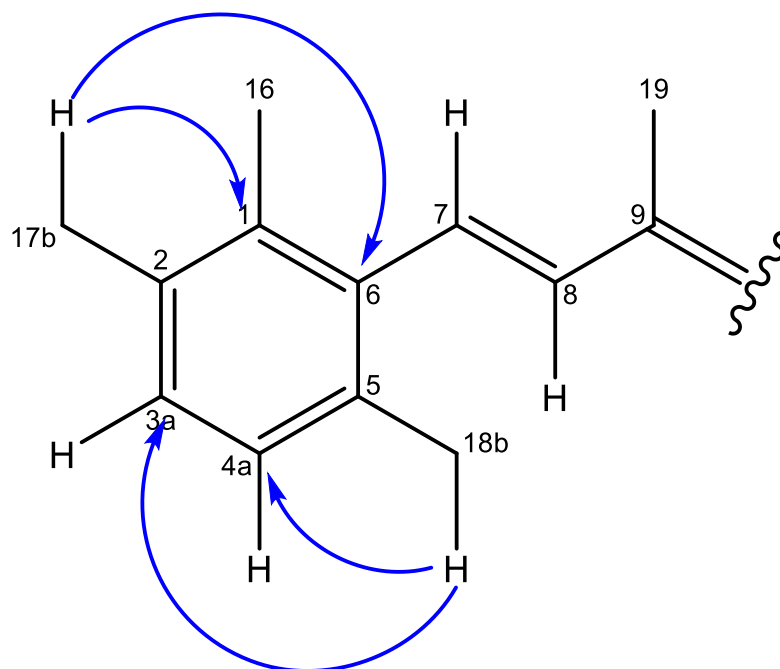


Figure S6: HMBC correlations of the aromatic ring. Letter labels refer to carbons that cannot be distinguished.

Stereochemistry of the hexose was determined through NOESY and ROESY correlations, which identified the sugar as glucose (Figure). This identification of glucose was unsurprising as D-glucose is the most common sugar known in bacterial glycosylated carotenoids (Choi *et al.*, 2013).

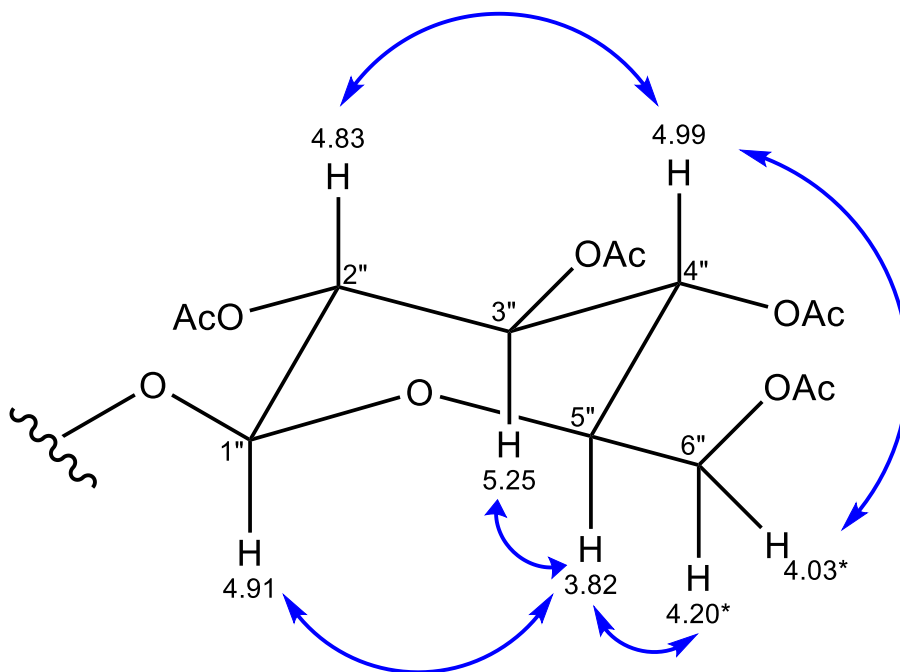


Figure S7: NOESY correlations of the sugar residue demonstrating stereochemistry.

Stereochemistry of the 2'-hydroxyl was determined through circular dichroism (CD) of the wild-type and the *strop3247* (glycosyl transferase) mutant compound. Both compounds showed the same CD spectral patterns and determined that the hydroxyl is in the *S* configuration through comparisons with previously established standards (**Error! Reference source not found.**)(Ronneberg *et al.*, 1985). The molecule showed little optical rotation, $[\alpha]^{25}_p = 0.0015$ deg ($c = 0.0008$ deg; methanol). As with CD measurements, this value was compared with the aglycone compound from the *strop3247* mutant. The mutant compound had a small and opposite rotation $[\alpha]^{25}_p = -0.0035$ deg ($c = 0.0008$ deg, methanol). The presence of an attached glucose may have an impact on the overall optical rotation of the molecule.

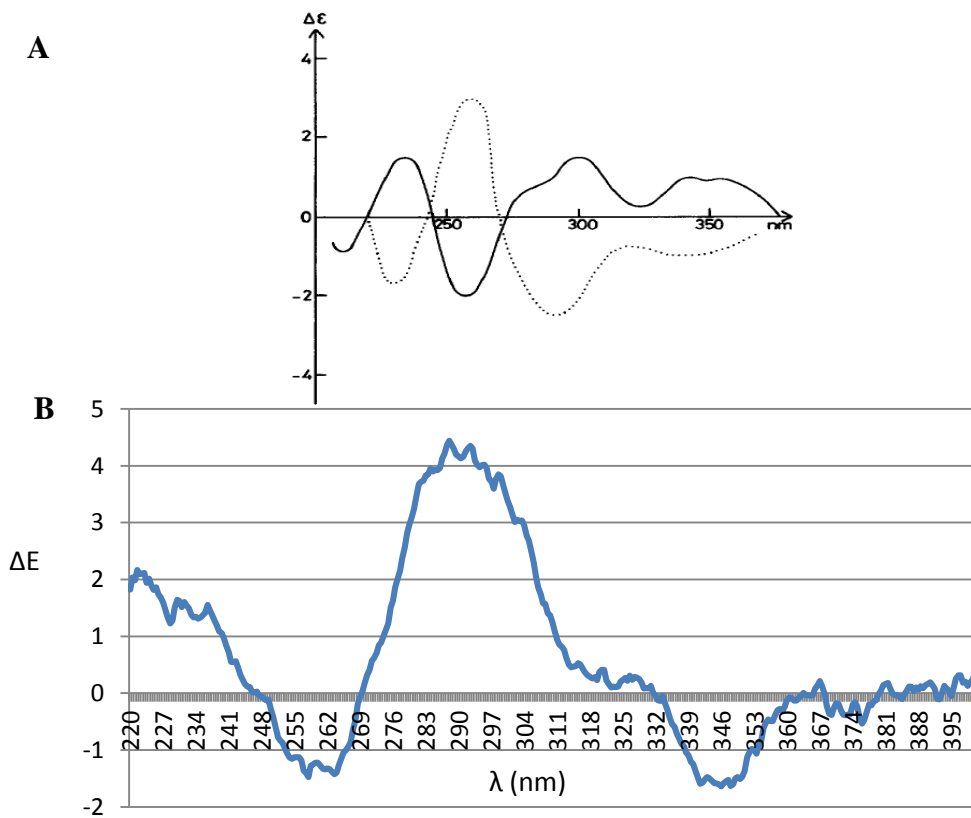


Figure S8: CD spectra of the carotenoid showing the chirality at the C-2'. A) The established CD spectra for the C-2' position in similar molecules, solid line shows *S* configuration, dotted line shows *R* configuration (adapted from Ronnenberg *et al* 1985). B) CD spectrum of the aglycone sioxanthin intermediate.

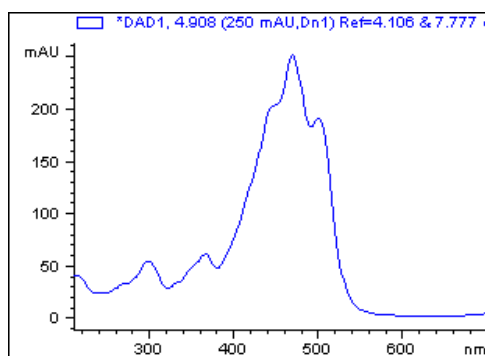


Figure S9: UV/Vis spectrum of sioxanthin showing the characteristic carotenoid peak structure.

628nicotay_2.ser
2D 1H-1H COSY
1.7mm TCI, 300K
UCSD_COSY

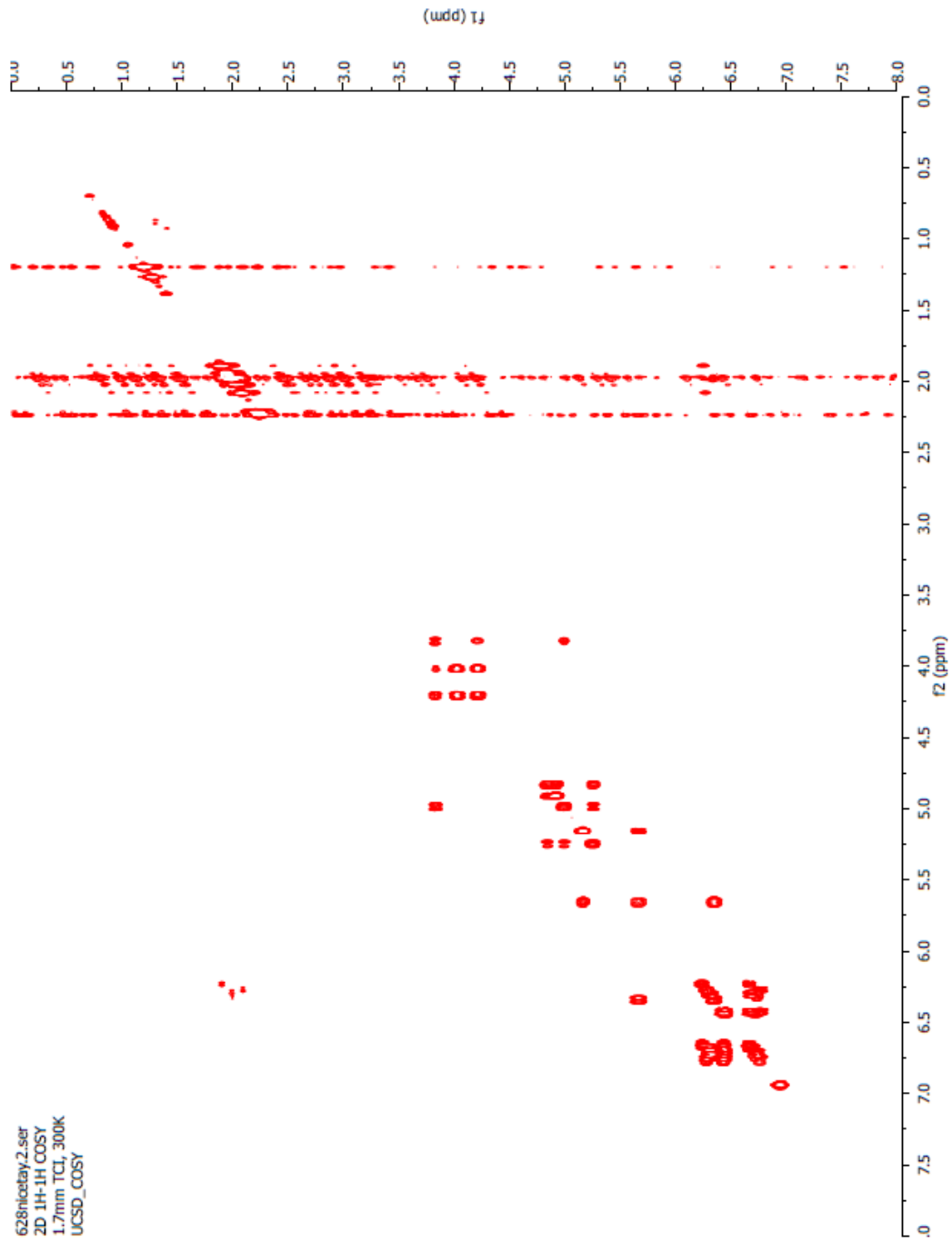


Figure S10: COSY (CD₃CN; 600MHz) spectrum of sixanthin.

628nicotay_3.ser
2D 1H-1H TOCSY with DIPSI2 spinlock, echo-antiecho
1.7 mm CP-TCL, 298K
UCSD_TOCSY

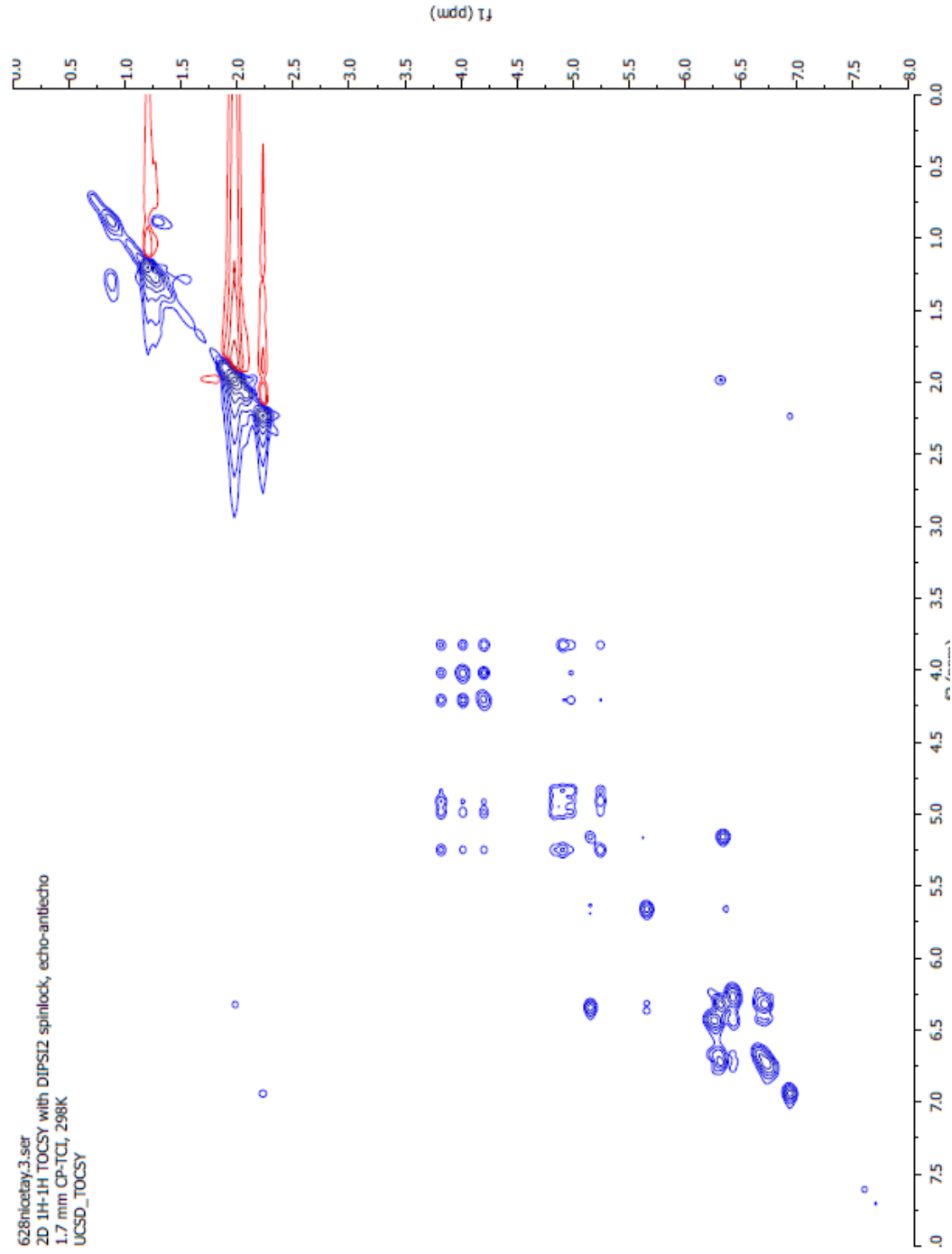


Figure S11: TOCSY (CD₃CN; 600MHz) spectrum of sixoanthin

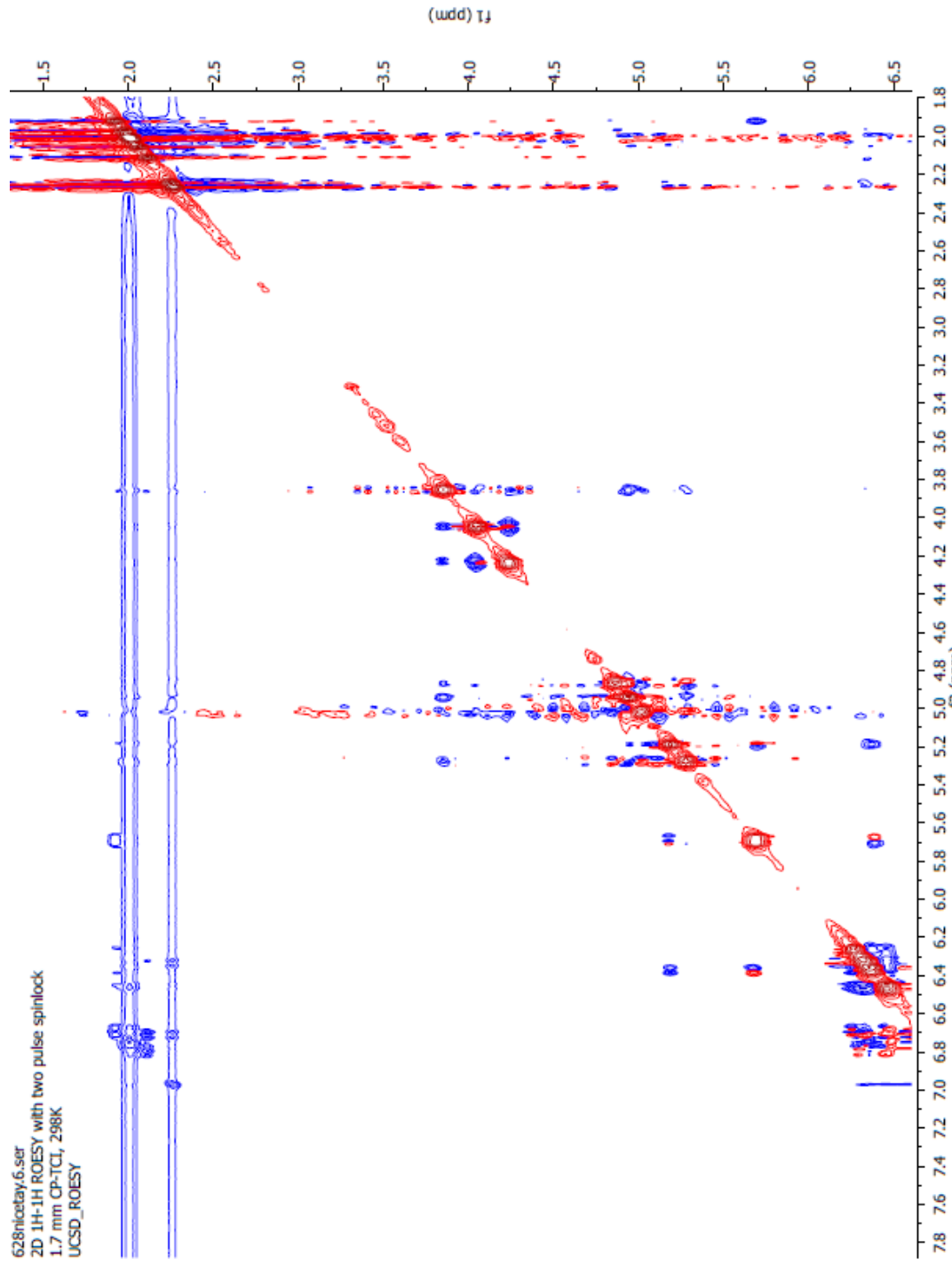


Figure S12: ROESY (CD₃CN; 600MHz) spectrum of sixanthin.

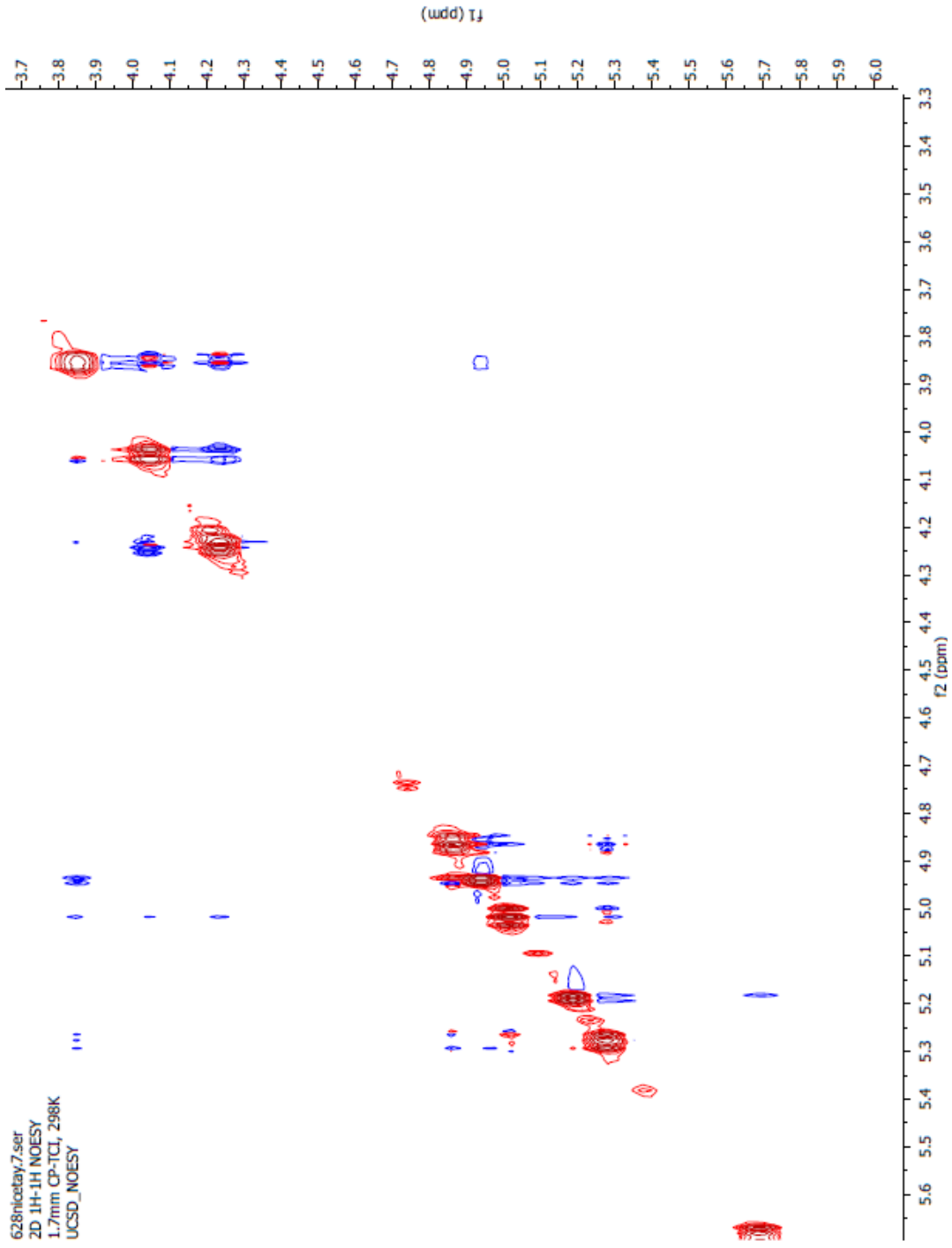


Figure S13: NOESY (CD₃CN; 600MHz) spectrum of sioxanthin.

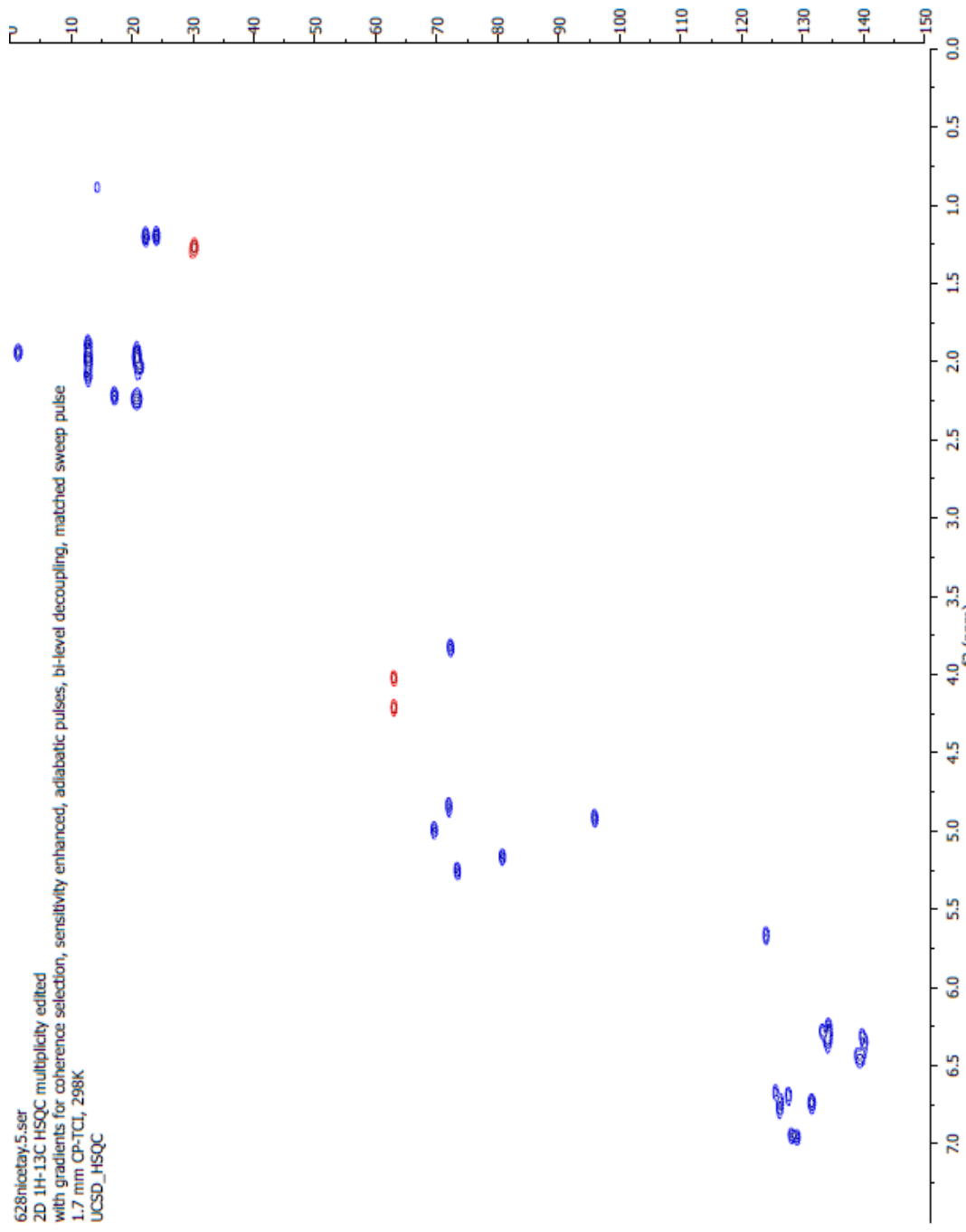


Figure S14: HSQC (CD_3CN ; 600MHz) spectrum of sioxanthin.

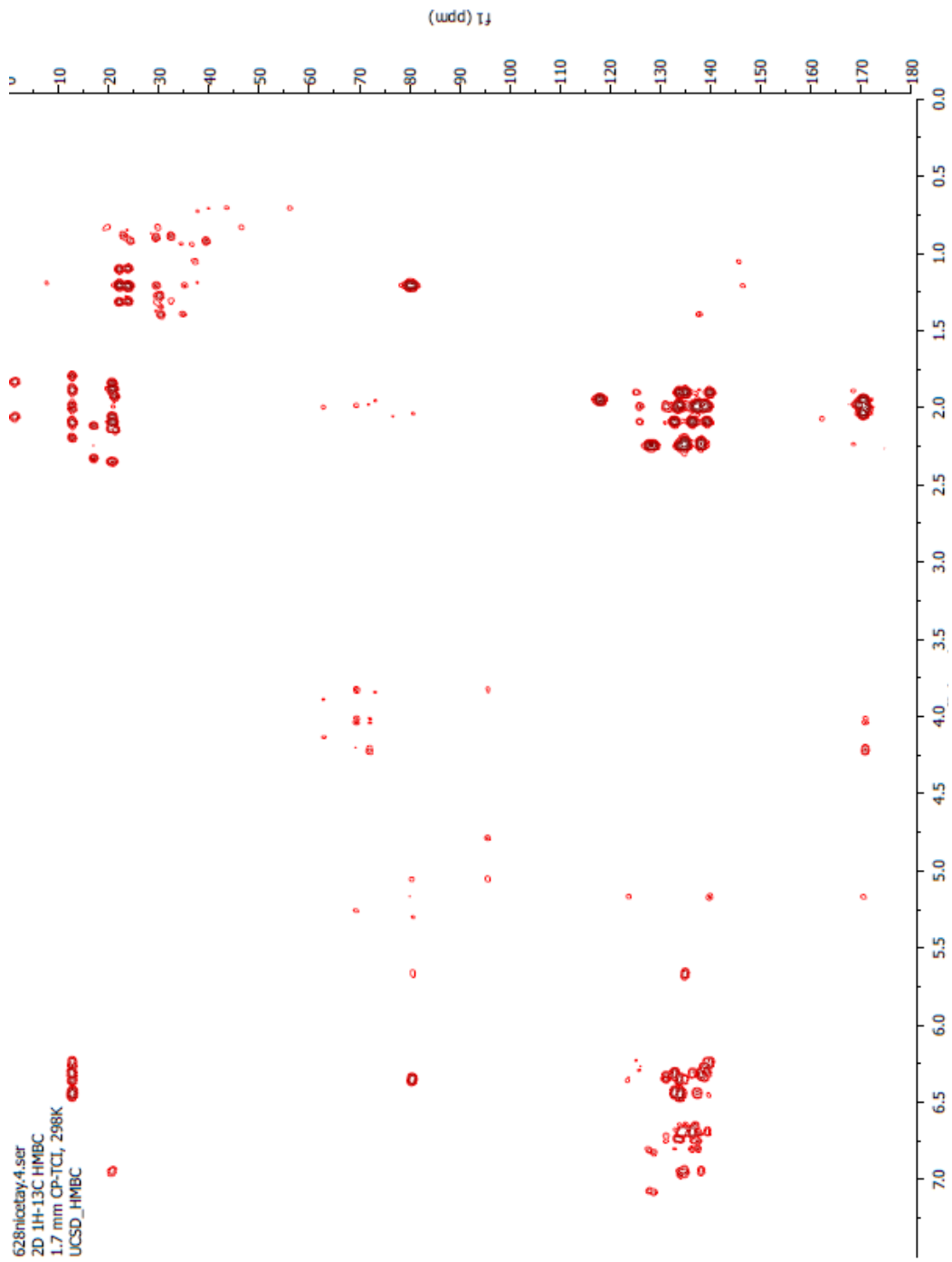


Figure S15: HMBC (CD_3CN ; 600MHz) spectrum of sioxanthin.

	GGPP synthase	phytoene synthase	phytoene desaturase	lycopene cyclase	C-1',2' hydratase	C-3',4' desaturase	carotene ketolase	Carotene beta-hydroxylase	glucosyltransferase	acyltransferase	desaturase
	CrtE	CrtB	CrtI	CrtY	CruF	CrtD	CrtO		CruC	CruD	CrtU
<i>Meiothermus ruber</i>	392	378	2710	382	383	386	381	290	385	384	
<i>Salinispora tropica</i>	4440	4441	4439	2408	3246	3248	1758	2694	3247	3486	241
%identity (aa level)	32	30	31	37	36	33	30	31	42	31	

Figure S16: BLAST analysis of genes from *Meiothermus ruber* carotenogenesis pathway against the *S. tropica* genome lead to the identification of missing sioxanthin biosynthetic genes in *terp1*. Numbers refer to gene locus and BLAST score.

Structure elucidation of mutant compounds

The purified mutant extracts, with the exception of the *strop0241* mutant, provided sufficient material for high resolution mass spectrometry (HR-MS). Mutant extracts were characterized using HR-MS, which provided a molecular formula (Table). This information, along with UV/Vis (describing the length of the chromophore) (Table) and HPLC retention time (indicating the presence or absence of polar functional groups) data allowed for the predicted structures (**Error! Reference source not found.**). Although the purified compounds from the *strop0241* mutant were visibly orange, they had very low yields. Masses obtained from this mutant were not consistent with carotenoid molecules and were likely due to impurities.

Table S2: UV/Vis and MS data gave rise to the predicted structures of carotenoids from mutant bacterial extracts.

Compound	Source	UV/Vis	%III/II	Measured mass (M+H)	Formula	Error (ppm)
41	wild-type	446, 472, 504	84			
42	pentacetate sioxanthin			937.5088	C ₅₆ H ₇₂ O ₁₂	1.0
43	<i>strop2408</i> mt	446, 470, 502	89	555.4563	C ₄₀ H ₅₈ O	0.5
44	<i>strop3246</i> mt	440, 462, 492	83	533.4145	C ₄₀ H ₅₃	0.3
45	<i>strop3247</i> mt	446, 472, 504	233	607.4140	C ₄₂ H ₅₄ O ₃	1.8
46	<i>strop3248</i> mt	436, 456, 488	88	551.4252	C ₄₀ H ₅₄ O	0.9

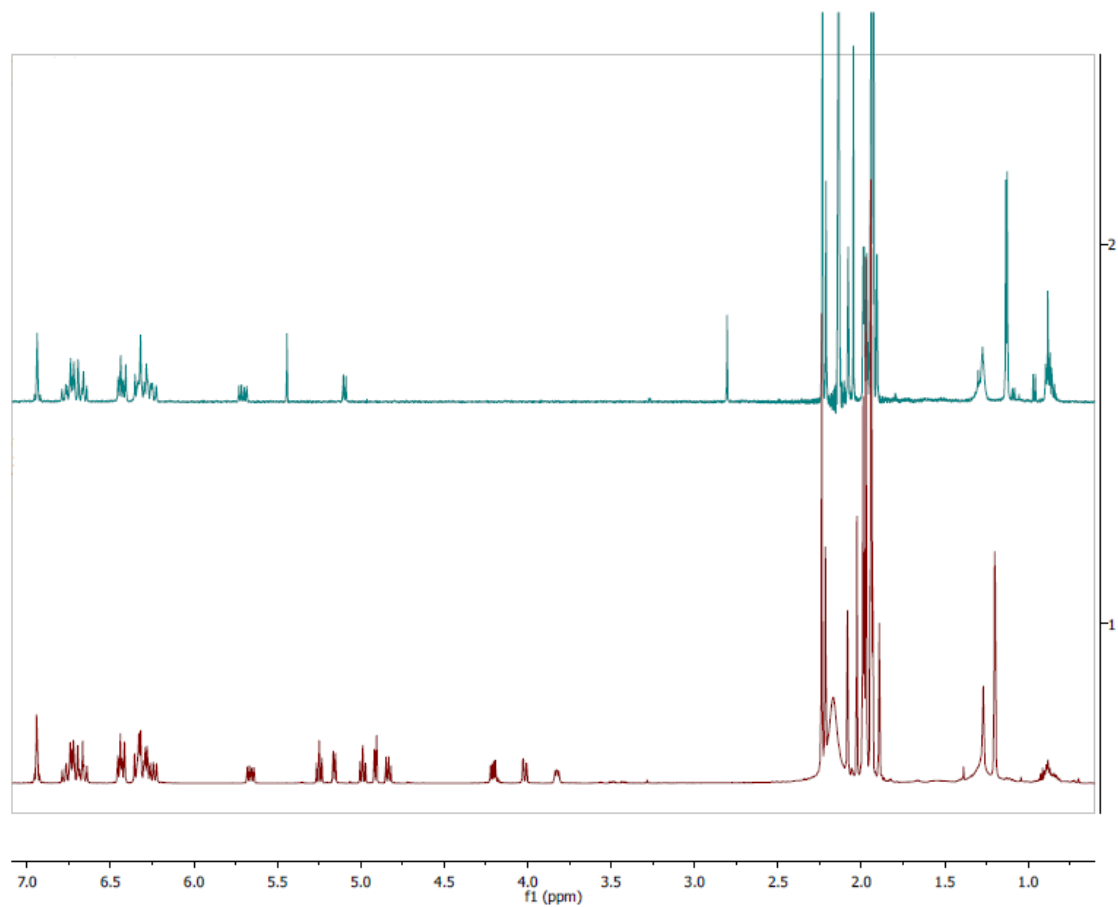


Figure S17: ¹H NMR (CD₃CN; 600MHz) spectrum of the *strop3247* mutant compound (**10**) (blue) compared with the spectrum of the wild-type sioxanthin acetate derivative (Figure S2) (red). Glucose signals are found in the δ3.5-5.5 ppm region. The NMR confirms that the structure (45) lacks the glucose functional group, but is in all other ways identical to the wild-type structure

Sioxanthin biosynthetic pathway in other bacteria

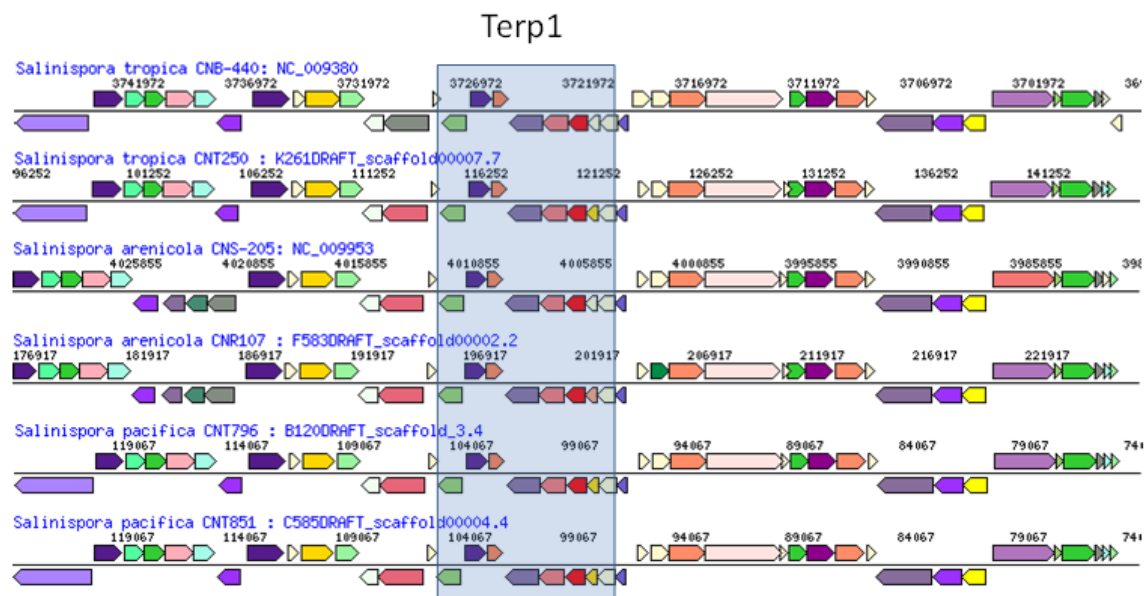


Figure S18: Alignments of *terp1* genes and gene neighborhoods from genomes of selected strains of *S. tropica*, *S. arenicola*, and *S. pacifica* showing conservation of genome organization across species.

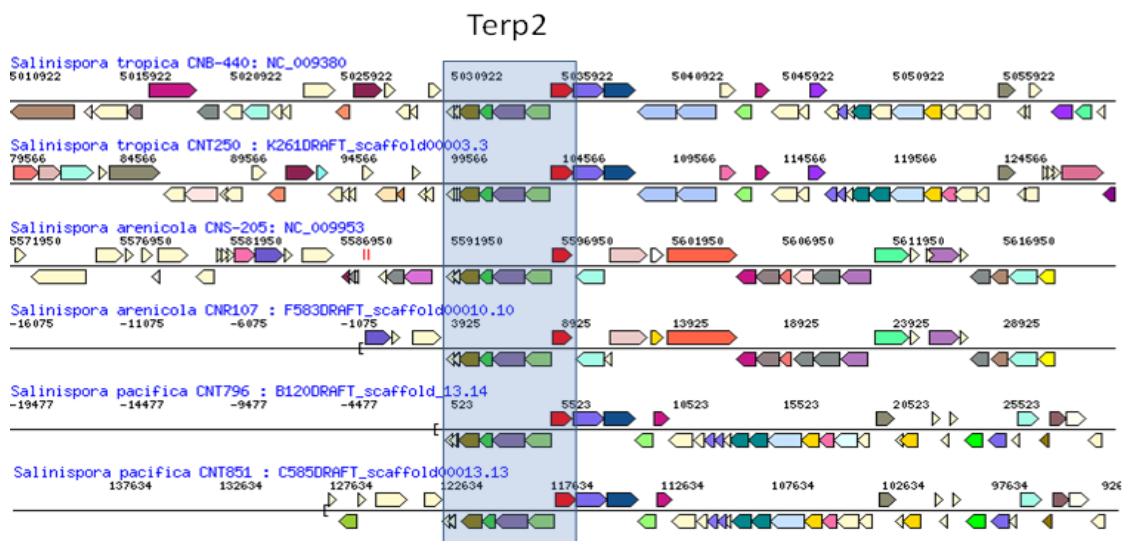


Figure S19: Alignments of *terp2* genes and gene neighborhoods from genomes of selected strains of *S. tropica*, *S. arenicola*, and *S. pacifica* showing conservation of genome organization across species.

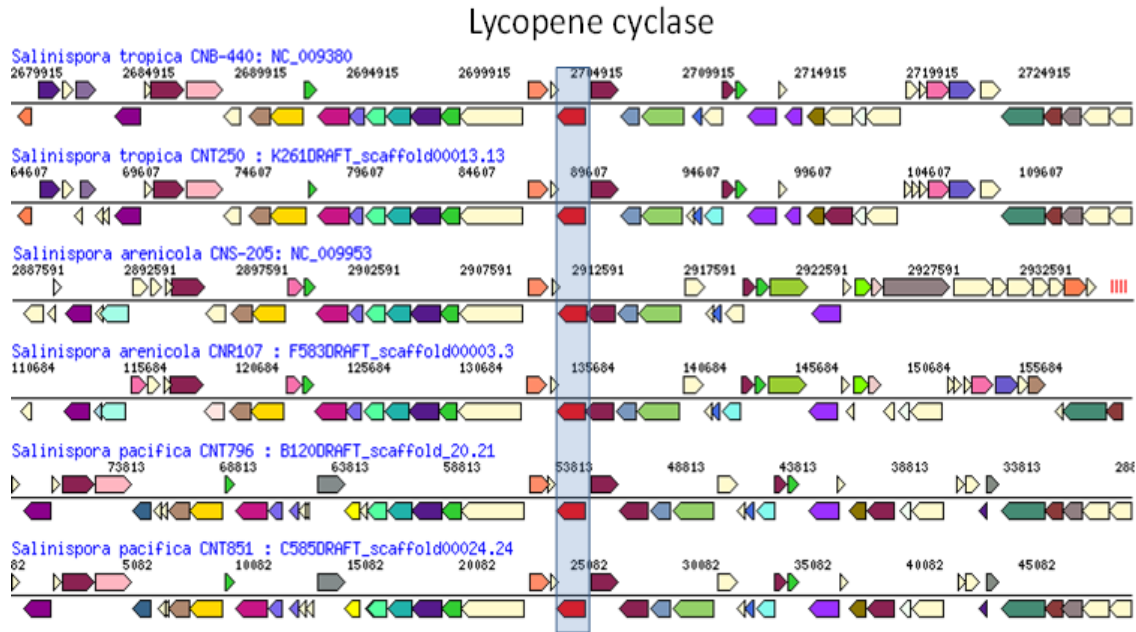


Figure S20: Alignments of lycopene cyclase genes and gene neighborhoods from genomes of selected strains of *S. tropica*, *S. arenicola*, and *S. pacifica* showing conservation of genome organization across species.

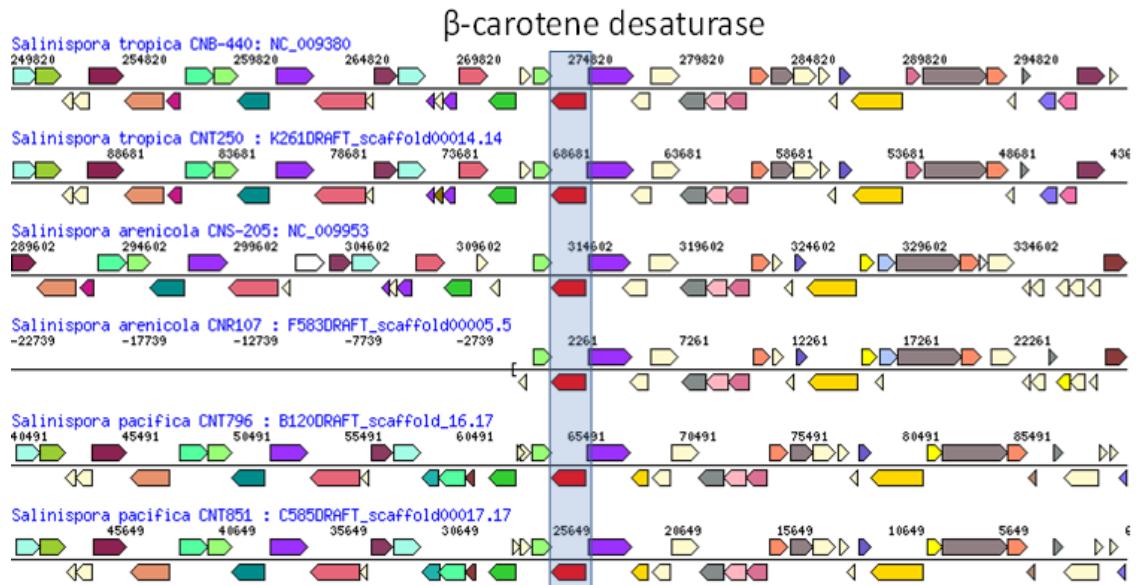


Figure S21: Alignments of β -carotene desaturase genes and gene neighborhoods from genomes of selected strains of *S. tropica*, *S. arenicola*, and *S. pacifica* showing conservation of genome organization across species.

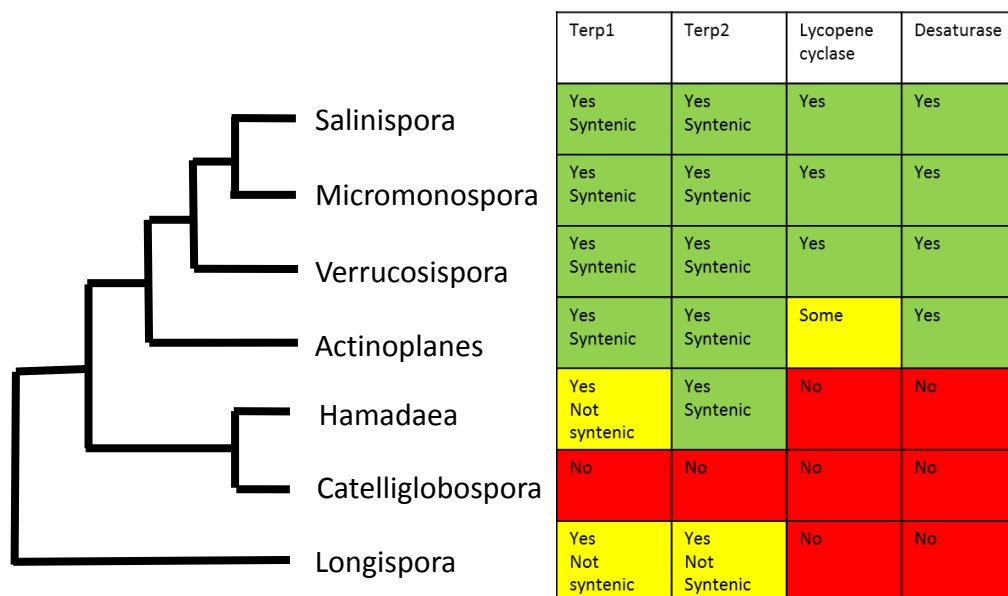


Figure S22: Summary of the presence of sixxanthin pathway genes in other Micromonosporaceae genera. The dendrogram shows relative relationships of the Micromonosporaceae genera while the table describes the presence of genes in the region and whether they are syntenic. Green shows complete and syntenic gene regions, red shows absence of genes, and yellow shows either genes present but not syntenic or genes not present in all species.

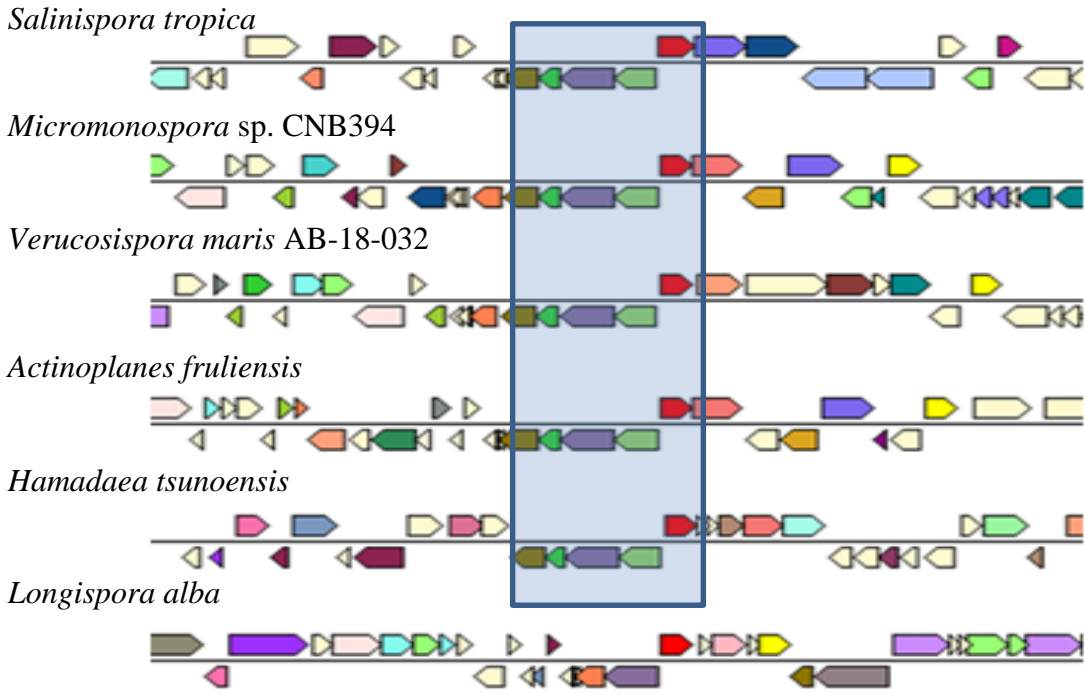


Figure S23: Comparison of the gene cluster responsible for lycopene biosynthesis (*terp2* homologs) in members of the Micromonosporaceae. Genes of a with a similar COG function are colored alike. The blue box highlights the *terp2* cluster (*strop4437-strop4441*) in *Salinispora* and genera for which this cluster is syntenic.

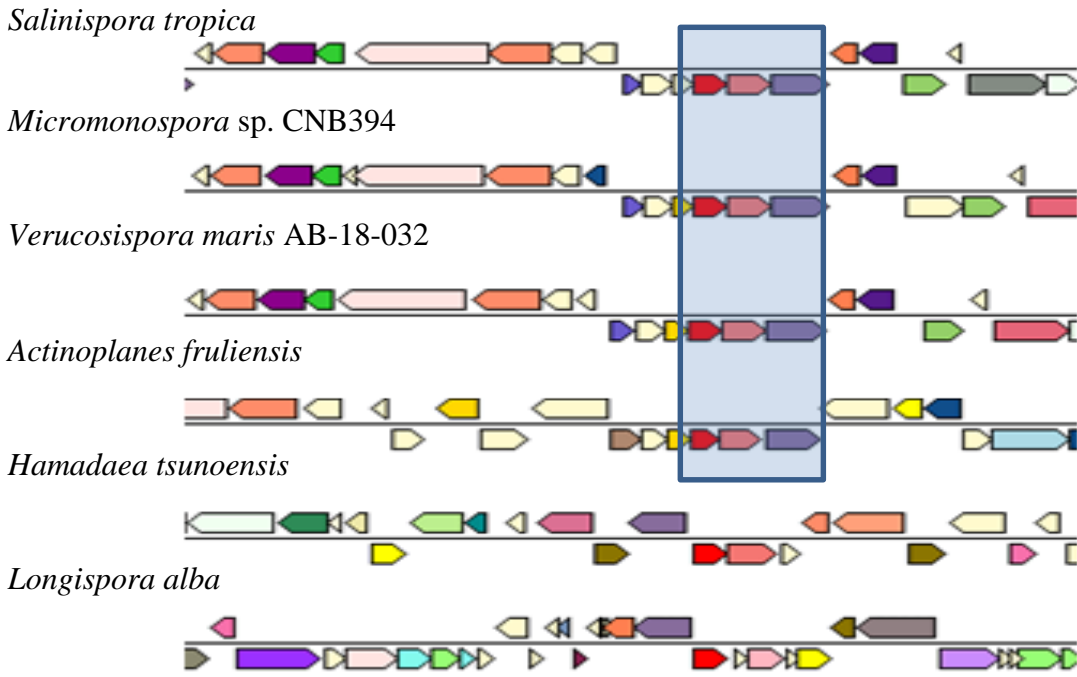


Figure S24: Comparison of the gene cluster responsible for sioxanthin glycosylation and 3', 4'-desaturation in members of the Micromonosporaceae. Genes of a with a similar COG function are colored alike. The blue box highlights the portion of the *terp1* cluster (*strop3246-strop3248*) necessary to sioxanthin biosynthesis in *Salinispora* and genera for which this cluster is syntenic.

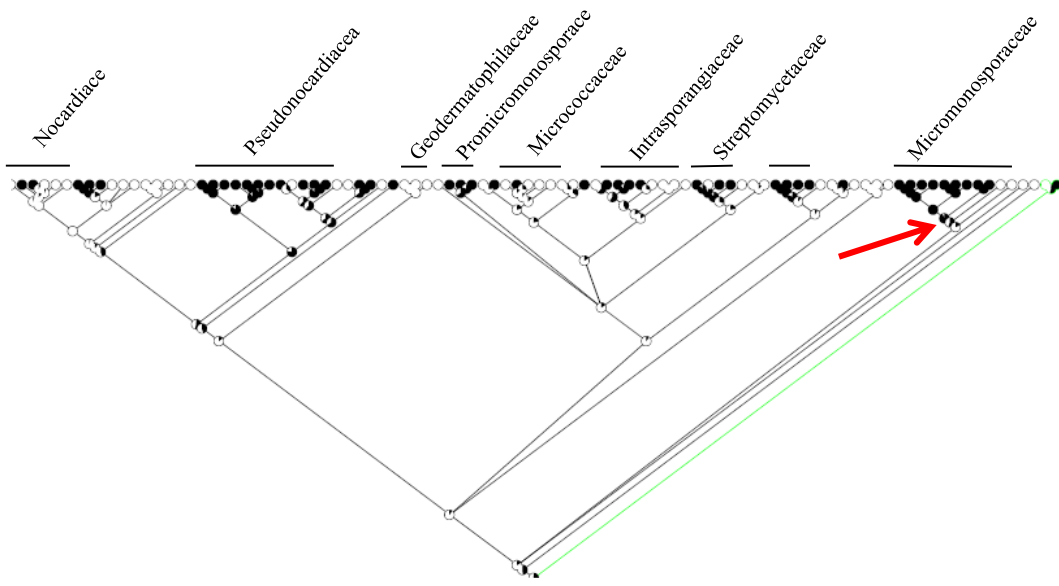


Figure S25: Character tree (maximum-likelihood using *rpoB* sequences) showing the distribution of *crtU* homologs among the actinomycetes as well as the predicted traits of ancestral species. A cyanobacterial outgroup is shown in green. Presence of a gene is indicated by a black circle and likelihood of presence in an ancestral species is shown in the proportion of black in an ancestral node. Red arrow indicates ancestor that likely acquired the gene.

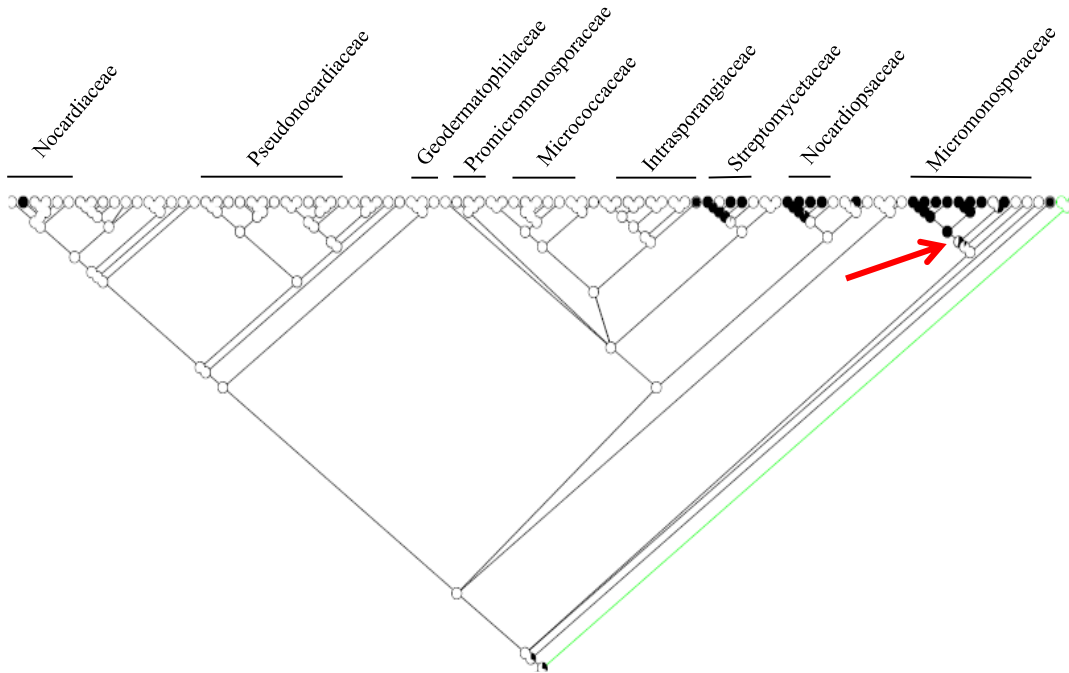


Figure S26: Character tree (maximum-likelihood using *rpoB* sequences) showing the distribution of *crtY* homologs among the actinomycetes as well as the predicted traits of ancestral species. A cyanobacterial outgroup is shown in green. Presence of a gene is indicated by a black circle and likelihood of presence in an ancestral species is shown in the proportion of black in an ancestral species. Red arrow indicates ancestor that likely acquired the gene.

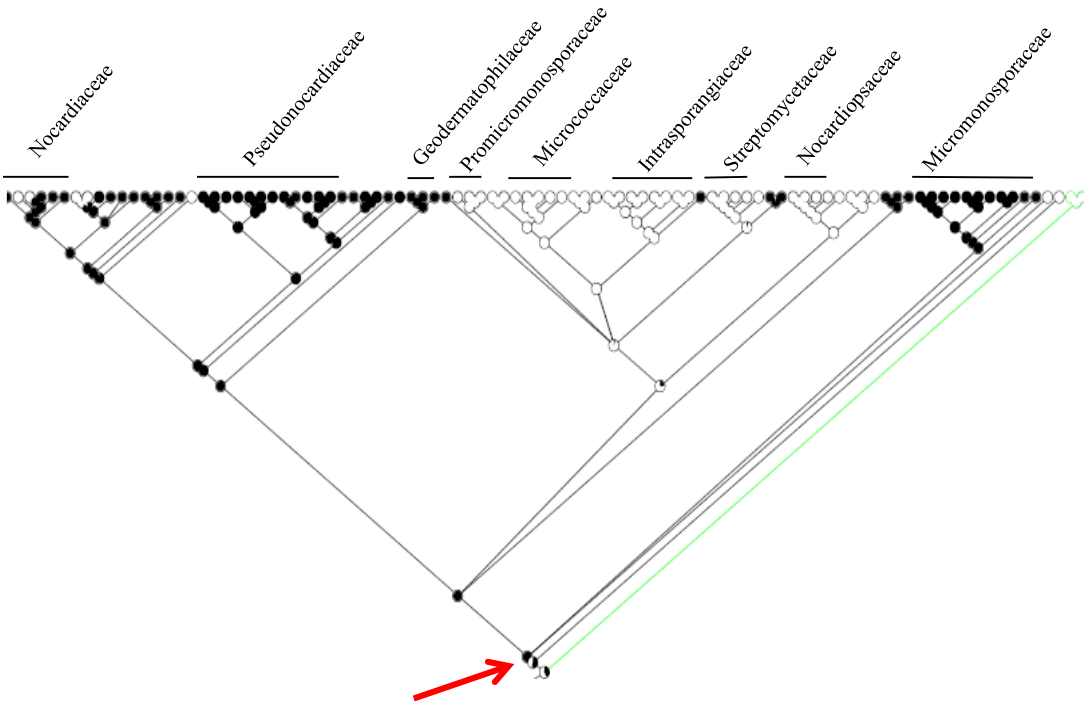


Figure S27: Character tree (maximum-likelihood using *rpoB* sequences) showing the distribution of acetyltransferase homologs among the actinomycetes as well as the predicted traits of ancestral species. A cyanobacterial outgroup is shown in green. Presence of a gene is indicated by a black circle and likelihood of presence in an ancestral species is shown in the proportion of black in an ancestral node. Red arrow indicates ancestor that likely acquired the gene.

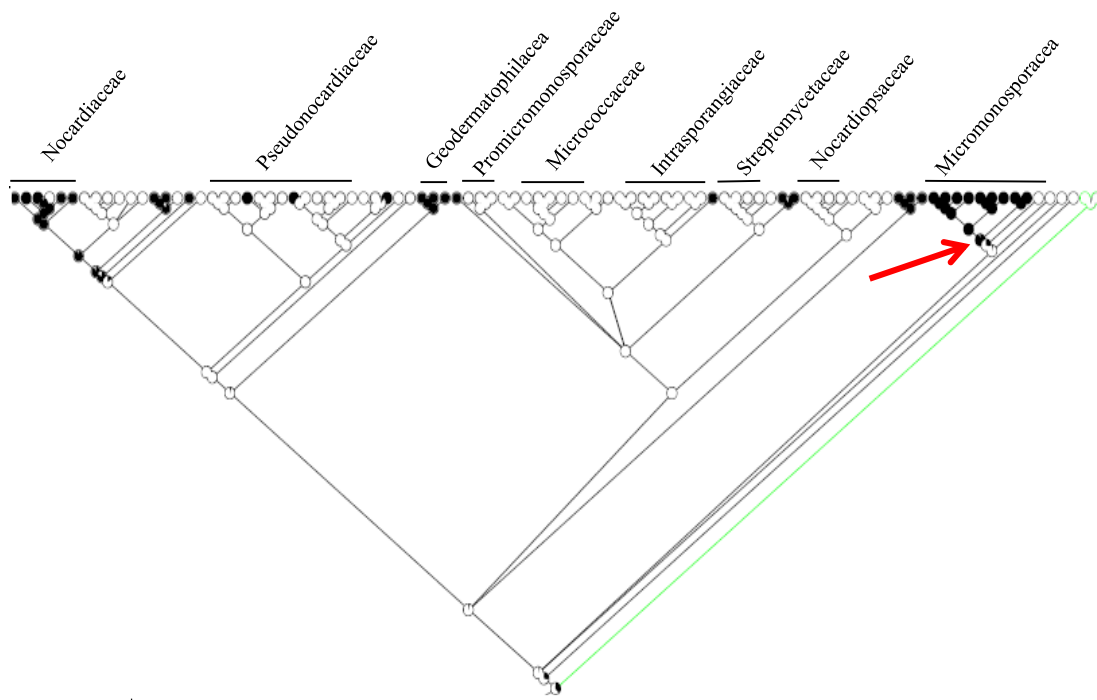


Figure S28: Character tree (maximum-likelihood using *rpoB* sequences) showing the distribution of *cruC* homologs among the actinomycetes as well as the predicted traits of ancestral species. A cyanobacterial outgroup is shown in green. Presence of a gene is indicated by a black circle and likelihood of presence in an ancestral species is shown in the proportion of black in an ancestral node. Red arrow indicates ancestor that likely acquired the gene.

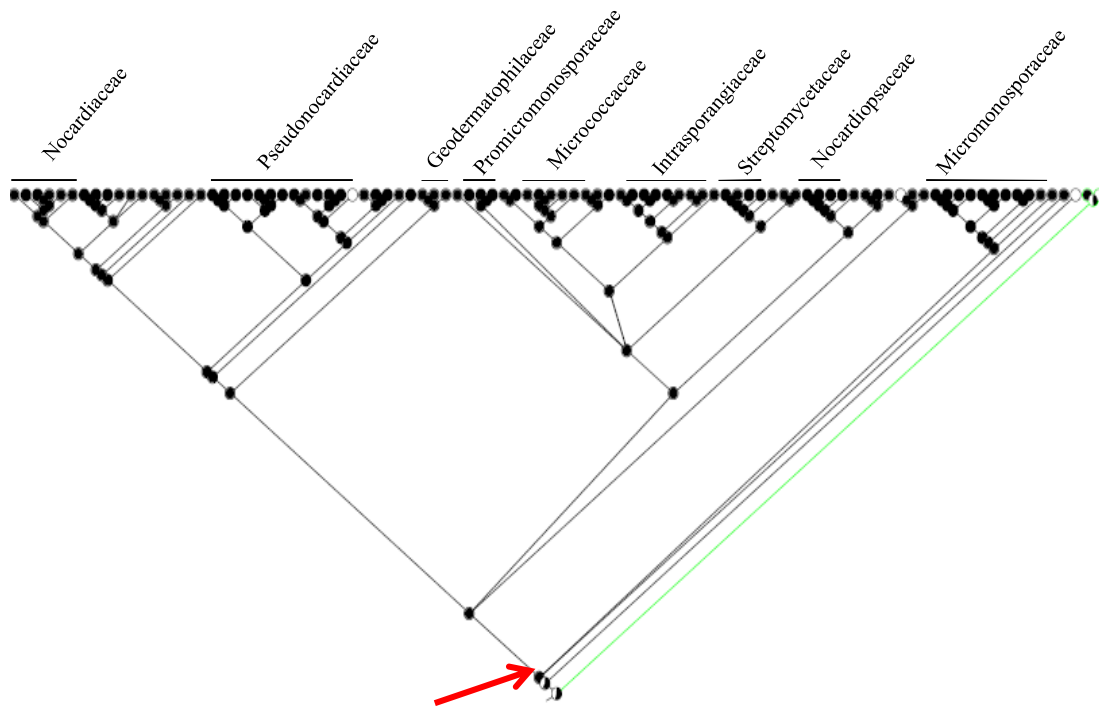


Figure S29: Character tree (maximum-likelihood using *rpoB* sequences) showing the distribution of *crtE* homologs among the actinomycetes as well as the predicted traits of ancestral species. A cyanobacterial outgroup is shown in green. Presence of a gene is indicated by a black circle and likelihood of presence in an ancestral species is shown in the proportion of black in an ancestral node. Red arrow indicates ancestor that likely acquired the gene.

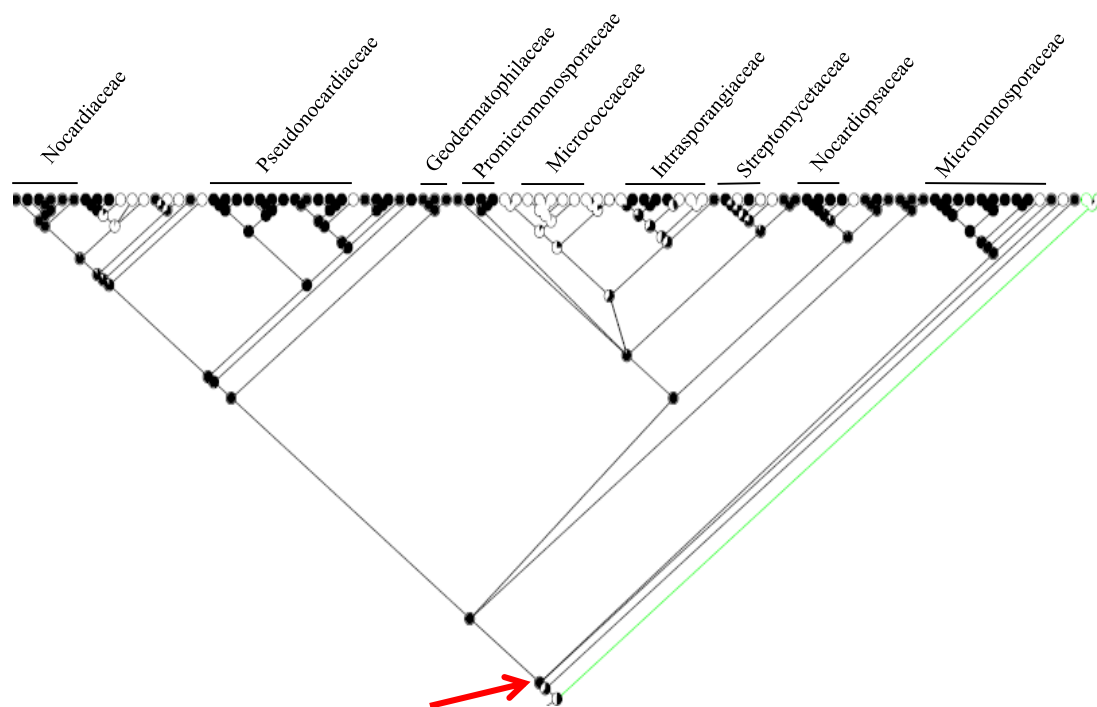


Figure S30: Character tree (maximum-likelihood using *rpoB* sequences) showing the distribution of *crtB* homologs among the actinomycetes as well as the predicted traits of ancestral species. A cyanobacterial outgroup is shown in green. Presence of a gene is indicated by a black circle and likelihood of presence in an ancestral species is shown in the proportion of black in an ancestral node. Red arrow indicates ancestor that likely acquired the gene.

Supplemental Experimental Procedures:

Inactivation of putative carotenoid biosynthetic genes

Table S3: Sequences of the tailed PCR primers used in the PCR-directed mutagenesis of *S. tropica* genes. Apramycin resistance cassette binding site is shown in bold

Primer name	Primer sequence
0241F	TTCTCAACACCGGCATATCTCCGGCAAAGTGGTGCATG ATTCCGGGGATCCGTCGACC
0241R	GGGATCGCCGGCTTCCGGTCAGTGTCCCGGGCGAGCAGT GTAGGCTGGAGCTGCTTC
2408F	CGGGCGGCCGGTCCGGCCGGTGACCCGACCATGGGTG ATTCCGGGGATCCGTCGACC
2408R	GCTGGACGATGCAGCACCCGGCCCGGCGACGGGCGGT CATGTAGGCTGGAGCTGCTTC

3244F GACTTGTCGAACGAGCCACGCCCGGTCAACCGCCGCGCAATTCCGGGGATCCGTCGACC
3244R CGTCCCACCTTCGTTCCGCCGTCCGTGGTCCAGCCCAGTGTAGGCTGGAGCTGCTTC
3245F CTGGTGCGGTGGACACCGGACGACCTCGTTCGGCGGCTGATTCCGGGGATCCGTCGACC
3245R GGGGGCTCTCCGGAGGCGGCCGGGTGGAGGGAGCGGCAGTGTAGGCTGGAGCTGCTTC
3246F CGGCCACCGCCCTGCCGAGGCGTACCCGCCAGACCGTCATTCCGGGGATCCGTCGACC
3246R CGGTGCCCCACTTCGTCGGGGGCCGGCGGTGAGACGTCGTGTAGGCTGGAGCTGCTTC
3247F CTGTTGCTGGCGGTGCTGGCCGGCGTGGCCGCGCTGACCATTCCGGGGATCCGTCGACC
3247R TGCCTCTAGACGACCGGACGGCCCCGCCAACTCAGGCGTGTAGGCTGGAGCTGCTTC
3248F GGCACACCATGGCGCGGATCGTGATCGTCGGCGCCGGGGATTCCGGGGATCCGTCGACC
3248R ATCGCGACACCGCGGGCGGGCGTCCGGCGGACCGGGCCTATGTAGGCTGGAGCTGCTTC
3249F GCCACGACACCGCGCGATCGGTGCGACGGTTGATCTGTGATTCCGGGGATCCGTCGACC
3249R TCGCGATTCCCGTGCCTCCGGTGGGCCGGGAGCGTCATGTAGGCTGGAGCTGCTTC
3250F CTAGGGATAGCGGCCGTGCGCCGCTAGAGTCATGGCGTGATTCCGGGGATCCGTCGACC
3250R CGTCGCACCGATCGCGCGGTGTCGTGGCGGTCTGGTCATGTAGGCTGGAGCTGCTTC
3251F GCACTGGCCGACTTCCTGGTGACCCGACGCGCCTGGATGATTCCGGGGATCCGTCGACC
3251R CCCCCTCCGGTCGAGGTCTGGCCGATGAGGCGGGGTCATGTAGGCTGGAGCTGCTTC
4440F TCGACGCAACTTGTAACCTAGGGTATGCTCCCCACATGATTCCGGGGATCCGTCGACC
4440R CTGGTACCTCCTCGAGAAAACGCGCCCATCATGCTCGCTGTAGGCTGGAGCTGCTTC
4441F GTTGCGTCGATTGTCGTCGAATTGAGGAGGATCGGTGATTCCGGGGATCCGTCGACC
4441R CGGTCAGCCTAGGGAAGGGCCGGGTGTGGGTGTCAACAGTGTAGGCTGGAGCTGCTTC

Purification of the S. tropica carotenoid

Purification of the primary pigment required two rounds of HPLC peak isolation using a Hewlett Packard Series II 1090 Liquid Chromatograph. HPLC traces were monitored at 450 nm and major carotenoid peaks were confirmed by examining the UV/Vis spectrum. Care was taken to keep the samples in low light conditions, in cool temperatures and under nitrogen, where possible. Crude extracts were first purified on a reversed-phase system using a Phenomenex Luna 5 µm C18 100A 250 x 10mm column. The solvent used for the wild-type and most mutant

extracts was 94.05% methanol, 5% DCM, 0.05% Et₃N, and 0.05M ammonium acetate run under isocratic conditions at 3 mL/min. The conditions were slightly altered for better separation of the *strop3246* mutant extracts and consisted of 85.45% methanol, 9.5% acetonitrile, 5% DCM, 0.05% Et₃N, and 0.05M ammonium acetate, also under isocratic conditions. The major HPLC peak of each crude extract was collected and dried under nitrogen.

Semi-pure extracts were then re-injected on the same instrument under normal phase conditions. Samples were separated on a Phenomenex Luna 5 μm Silica (2) 100A 250 x 10 mm column. The solvent system was run isocratically with a flow rate of 3 mL/min and for most samples consisted of the following: 93% hexanes, 5% DCM, 2% isopropanol. The one exception was the *strop2408* mutant extract which used a solvent system of 91% hexanes, 5% DCM, and 4% isopropanol. Purified samples were dried under nitrogen and stored at -20° C. Sample purity was confirmed via ¹H NMR. All solvents were of HPLC grade purity.

Structure elucidation of the carotenoid compounds

Purified compound was analyzed by NMR at the UC San Diego Skaggs School of Pharmacy and Pharmaceutical Sciences NMR Facility on a Varian NPA600 MHz NMR fitted with a 1.7 mm inverse detection triple resonance (H-C/N/D) cryoprobe. The wild-type sample in CD₃CN was analyzed by ¹H NMR as well as COSY, TOCSY, NOSY, HMBC, and HMQC spectroscopy. Mutant samples were analyzed by ¹H NMR. High resolution mass spectrometry was performed at the UCSD Chemistry and Biochemistry Molecular MS Facility using an Agilent 6230 APCI-TOFMS in positive mode. UV/Vis was determined on the HPLC and confirmed on an Agilent Cary60 UV-Vis in methanol. Circular dichroism was performed in distilled methanol on a Jasco 810 spectropolarimeter in quartz cells with a 1 mm path length and scanning from 220-400 nm. Optical rotation was measured on a Jasco P210 polarimeter in distilled methanol.

Supporting Information References

Choi, S.K., Osawa, A., Maoka, T., Hattan, J., Ito, K., Uchiyama, A. et al. (2013) 3-beta-Glucosyl-3'-beta-quinovosyl zeaxanthin, a novel carotenoid glycoside synthesized by *Escherichia coli* cells expressing the *Pantoea ananatis* carotenoid biosynthesis gene cluster. *Appl Microbiol Biotechnol* **97**: 8479-8486.

Englert, G. (1995) NMR Spectroscopy. In *Carotenoids: Spectroscopy*. Britton, G.L.-J., S.; Pfander, H. (ed). Basel: Birkhauser, pp. 147-259.

Ronneberg, H., Andrewes, A.G., Borch, G., Berger, R., and Liaaenjensen, S. (1985) CD correlation of C-2' substituted monocyclic carotenoids. *Phytochemistry* **24**: 309-319.