Supporting Information

Wallrapp et al. 10.1073/pnas.1300632110

SI Methods

Sequence Analysis. In December 2010, we gathered sequences contained in the following six Pfam families (1): Terpene_synth, Terpene_synth_C, Prenyltrans, TRI5, Polyprenyl_synt, and SQS_PSY. After removing redundant sequences, all-against-all pairwise BLAST e-values (2) were computed, and the resulting network was visualized at different e-value cut-offs in Cytoscape 2.8.2 (3). Based on sequence clustering and structural information, two superfamilies were defined: isoprenoid synthase type I and isoprenoid synthase type II. Sequences and associated metadata (including species information, catalytic specificity, and available crystal structures) were stored in the Structure-Function Linkage Database (SFLD) (4). Each superfamily is characterized by a multiple sequence alignment (MSA) and a Hidden Markov Model (HMM) based on the functionally and structurally characterized members of the superfamily. Superfamilies are classified further into subgroups, which also are characterized by an MSA and HMM. Among the subgroups in the isoprenoid synthase type I superfamily is the polyprenyl transferase-like subgroup of *E*-polyprenyl transferases (*E*-PTs), which by May 2011 consisted of 5,839 sequences. Automated protocols are used to add new sequence data from the National Center for Biotechnology Information Protein database on a regular basis.

Networks in this article were generated using Pythoscape (5). In the networks, nodes represent sequences, and edges correspond to sequence similarities that have BLAST e-values that are more significant than a specified cutoff. Because BLAST e-values are unidirectional (i.e., the e-value between sequence A and B is not identical to the e-value between sequence B and A), only the worst hit between two sequences was used. Networks are visualized using the yFiles organic layout provided within Cytoscape. Proteins and crystal structures applied in this study are identified by their Enzyme Function Initiative target ID (EFI-ID).

Protein Expression and Purification. ***Expression for targets in pNIC-Bsa4.*** The pNIC28-Bsa4 (6) vector containing the insert was transformed into BL21(DE3) *Escherichia coli* containing the pRIL plasmid (Stratagene) and was used to inoculate an overnight culture containing 25 ug/mL kanamycin and 34 ug/mL chloramphenicol. The culture was allowed to grow overnight at 37 °C in a shaking incubator. Then 1 mL of the overnight culture was used to inoculate 1 L of PASM-5052 autoinduction medium (7) containing 100 ug/mL kanamycin and 34 ug/mL chloramphenicol. The culture was placed in a LEX 48 airlift fermenter (Harbinger Biosciences) and incubated for 4 h at 37 °C and then overnight at 22 °C. The culture was harvested and pelleted by centrifugation and stored at −80°C until time of use.

Purification of targets in pNIC-Bsa4. Cells were resuspended in lysis buffer [20 mM Hepes (pH 7.5), 500 mM NaCl, 20 mM imidazole, and 10% (vol/vol) glycerol] and lysed by sonication. The lysate was clarified by centrifugation at 35,000 × *g* for 30 min. Clarified lysate was loaded onto an AKTAxpress FPLC (GE Healthcare). Lysate was loaded onto a 1-mL HisTrap FF column (GE Healthcare), washed with 10 column volumes of lysis buffer, and eluted in buffer containing 20 mM Hepes (pH 7.5), 500 mM NaCl, 500 mM imidazole, and 10% (vol/vol) glycerol. The purified sample was loaded onto a HiLoad S200 16/60 PR gel filtration column (GE Healthcare), which was equilibrated with SECB gel filtration buffer [20 mM Hepes (pH 7.5), 150 mM NaCl, 10% glycerol, and 5 mM DTT]. Peak fractions were collected and allowed to incubate with 1 mg of TEV-protease overnight at 4 °C. TEV-protease and uncleaved protein were removed by passing the over 1 mL of Ni Sepharose High Performance (GE Healthcare). Protein was analyzed by SDS/PAGE, snap frozen in liquid nitrogen, and stored at −80°C.

Expression for targets in CHS30. CHS30 vector, a C-terminal TEV-cleavable StrepII-6xHis-tag vector, containing the insert was transformed into BL21(DE3) *E. coli* containing the pRIL plasmid (Stratagene) and was used to inoculate an overnight culture containing 25 ug/mL kanamycin and 34 ug/mL chloramphenicol. The culture was allowed to grow overnight at 37 °C in a shaking incubator. Then 1 mL of the overnight culture was used to inoculate 1 L of PASM-5052 autoinduction medium (7) containing 100 ug/mL kanamycin and 34 ug/mL chloramphenicol. The culture was placed in a LEX 48 airlift fermenter and incubated at 37 °C for 4 h and then at 22 °C overnight. The culture was harvested and pelleted by centrifugation and stored at −80°C until time of use.

Purification of targets in CHS30. Cells were resuspended in lysis buffer [20 mM Hepes (pH 7.5), 500 mM NaCl, 20 mM imidazole, and 10% glycerol) and lysed by sonication. The lysate was clarified by centrifugation at 35,000 × *g* for 30 min. Clarified lysate was loaded onto an AKTAxpress FPLC (GE Healthcare). Lysate was loaded onto a 5-mL Strep-Tactin column (IBA), washed with five column volumes of lysis buffer, and eluted in StrepB buffer [20 mM Hepes (pH 7.5), 500 mM NaCl, 20 mM imidazole, 10% glycerol, and 2.5 mM desthiobiotin]. The eluent was loaded onto a 1-mL HisTrap FF column (GE Healthcare), washed with 10 column volumes of lysis buffer, and eluted in buffer containing 20 mM Hepes (pH 7.5), 500 mM NaCl, 500 mM imidazole, and 10% (vol/vol) glycerol. The purified sample was loaded onto a HiLoad S200 16/60 PR gel filtration column, which was equilibrated with SECB gel filtration buffer [20 mM Hepes (pH 7.5), 150 mM NaCl, 10% glycerol, and 5 mM DTT]. Peak fractions were collected, and protein was analyzed by SDS/PAGE, snap frozen in liquid nitrogen, and stored at −80°C.

Expression and Purification of targets in pSGX3. Expression and purification of targets in pSGX3 were performed according to the methods described by Sauder et al. (8) and are given in Table S1.

Protein Suspension. All the 900 series targets were in a buffer of 10 mM Hepes (pH 7.5), 150 mM NaCl, 10 mM methionine, and 10% (vol/vol) glycerol. The two 500 series targets were in 20 mM Hepes (pH 7.8), 150 mM NaCl, 10% (vol/vol) glycerol, and 5 mM DTT.

Crystallization, Ligand Soaking, and Collection of X-Ray Diffraction Data. Proteins were crystallized by the sitting-drop vapor diffusion method. As a rule, the protein solutions (usually 0.3 or 1 mL) were mixed with an equal volume of a precipitant solution and equilibrated at room temperature (294 K) against the same precipitant solution in clear tape-sealed 96-well INTELLI-plates (102-0001-20; Art Robbins Instruments). Crystallization was performed using either a TECAN crystallization robot (TECAN US) or a PHOENIX crystallization robot (Art Robbins Instruments) and four types of commercial crystallization screens: the WIZARD I&II screen (Emerald BioSystems), the INDEX HT and the CRYSTAL SCREEN HT (both from Hampton Research), and the MCSG screen (Microlytic).

The appearance of protein crystals was monitored either by manual inspection or using a RockImager 1000 (Formulatrix) starting within 24 h of incubation and again at weeks 1, 2, 3, 5, 8, and 12.

Before harvesting and freezing, the protein crystals were prepared as described in Table S2. Most of protein crystals were dehydrated before freezing, as described below. Each protein

sample used for crystallization contained 10% (vol/vol) glycerol, so the crystallization drops with or without protein crystals already were supplemented with glycerol, which is a good cryoprotectant. However, the 10% (vol/vol) concentration often is too low for cryoprotection. To increase the glycerol concentration and dehydrate the protein crystals at the same time, we supplemented precipitant solutions with 20–25% (vol/vol) glycerol, and the crystallization drops then were equilibrated against these glycerol-containing solutions for 2–48 h. After incubation, the crystals were harvested using cryogenic loops (Hampton Research), quickly transferred into liquid nitrogen, and stored frozen in liquid nitrogen until X-ray analysis was conducted and/or X-ray diffraction data were collected. Where necessary, the crystallization conditions were optimized manually using 24-well Cryschem sitting-drop plates (Hampton Research). The final crystallization and cryoprotection conditions related to each X-ray crystal structure are listed in Table S2.

The X-ray diffraction data for the frozen crystals were collected at 100 K on the beamline ×29A at the National Synchrotron Light Source, Brookhaven National Laboratory, Upton, NY, using a wavelength of either 0.979 Å [single anomalous difference (SAD) data] or 1.075 Å (native data). All diffraction data were processed and scaled with HKL2000 (9). The crystal structures reported here were determined either by Selenium-SAD using anomalous X-ray diffraction data for phasing or by molecular replacement using coordinates for similar structures from the Protein Data Bank (PDB) (listed in each PDB deposition as REMARK 200: STARTING MODEL) and PHASER MR software [the CCP4 program package suite (10)]. Each structure was refined using the program REFMAC (11), and the resulting models were rebuilt manually using COOT visualization and refinement software (12). The data collection and refinement statistics for these structures are listed in Tables S3–S7.

Experimental and Computational Function Annotation. ***Determination of in vitro chain length/enzyme activity assay.*** The assay mixture for in vitro determination of chain length and enzyme activity contained 35 mM Hepes (pH 7.6) 10 mM $MgCl_2$, 5 mM β-mercaptoethanol, 0.25% Triton X-100, 50 μM dimethylallyl diphosphate (DMAPP), 50 or 200 μM [^{14}C] isopentenyl diphosphate (IPP) [specific activity (SA) = 5.5 μCi/μmol] and 1–20 μg of enzyme (~0.03–0.6 nmols, assuming an average molecular mass of ~40 kDa) in a final volume of 40 μL. The reaction mixture was incubated at 30 °C for 2.5 h before being terminated by the addition of 5 μL of 0.5 M EDTA. The radiolabeled polyprenyl diphosphate products were extracted with I1-butanol (saturated with water), and solvent was removed on a SpeedVac (Thermo Scientific).

Product analysis. The polyprenyl diphosphate products were converted to the corresponding alcohols by incubating the residue with potato acid phosphatase in 200 μL of 50 mM sodium acetate buffer (pH 4.5) containing 20% (vol/vol) n-propanol, 0.1% Triton X-100, and 1 mg (2 U, SA = 2 U/mg) of acid phosphatase at 37 °C for 16 h. The incubation mixtures were extracted with hexanes, and the solvent was removed using a SpeedVac. The polyprenyl alcohols were separated by reversed-phase silica gel TLC eluted with 9:1 or 19:1 acetone:water. The radiolabeled products were detected by phosphorimaging.

Images of the TLC plates, with each enzyme identified by its GenBank identification (GI) number, are presented in Dataset S1. The isoprenoid alcohols obtained from a given incubation were well resolved into a series of bands corresponding to the addition of an increasing number of isoprene units. The chain lengths were established by comparisons with standards synthesized from sagebrush farnesyl diphosphate (FPP) synthase. This enzyme, under forcing conditions, over-elongates up to hexaprenyl (C_{30}) diphosphate when IPP is in excess (Standard in TLC 8). By aligning bands of the isoprenoid alcohols from sagebrush FPP synthase with those from GI 67866738, we obtained standards useful for chain lengths up to 10 isoprene units as C_{50} by the time-course experiment (see below).

Protein structure preparation and homology modeling. The template structures used for homology modeling were prepared with Schrodinger Protein Preparation Wizard. Missing loops and side chains were built with Prime (Schrodinger LLC). For template structures missing the S1 ligand (apo structures), we manually placed diphosphate and $Mg2^+$ ions, based on the positions in the holo structures, and adjusted the conformations of the aspartates ligating the ions. We derived the alignments from PROMALS3D (13) for each query sequence against the template with the highest sequence identity available. In addition, we manually curated the resulting alignments to guarantee the correct adjustment of key residues (i.e., the aspartate-rich motifs) in the models. All models were treated as dimers throughout the computational setup to account for residues of the interface (of the opposite chain) influencing the elongation cavity and thus eventually the product specificity.

Covalent docking. We applied covalent ligand docking on derived structures using Prime. Here, the diphosphate group of the ligand was kept frozen in the S1 active site while the tail was docked flexibly into the elongation cavity. Side chains of residues within a 5-Å vicinity of the elongation cavity were treated as conformationally flexible. The modeled ligands for each structure were DMAPP, geranyl diphosphate, FPP, geranylgeranyl diphosphate (GGPP), and farnesylgeranyl diphosphate, ranging from 5–25 carbon atoms. The covalent docking algorithm returned a list of 100 ligand-receptor models ranked by complex energy. We computed the Lennard–Jones energy (E_{LJ}) of the complex and the molecular mechanics/generalized-Born surface area (MMGBSA) binding energy for each of the top three models of each docking run, using the Surface Generalized Born (SGB) solvent model. Based on results in the test set, we exclusively applied the E_{LJ} of the complex as the scoring function for the chain-length prediction. For each enzyme, the docked ligand with the lowest relative energy score is considered the reactant of the last step of prenylation; thus the actual product of the reaction specific for this enzyme is predicted to be one C_5 unit longer.

Experimentally determined vs. predicted product chain length of the training set, the set of 34 sequences that were screened and functionally assigned before the computational part of this study was completed ($targets_{known}$), and the set of 40 sequences for which the functional assignment was not known during the computational prediction ($targets_{blind}$) are given in Tables S8, S9, and S10, respectively.

1. Punta M, et al. (2012) The Pfam protein families database. *Nucleic Acids Res* 40(1): 290–302.
2. Altschul SF, et al. (1997) Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res* 25(17):3389–3402.
3. Cline MS, et al. (2007) Integration of biological networks and gene expression data using Cytoscape. *Nat Protoc* 2(10):2366–2382.
4. Pegg SC, et al. (2006) Leveraging enzyme structure-function relationships for functional inference and experimental design: The structure-function linkage database. *Biochemistry* 45(8):2545–2555.
5. Barber AE II, Babbitt PC (2012) Pythoscape: A framework for generation of large protein similarity networks. *Bioinformatics* 28(21):2845–2846.
6. Savitsky P, et al. (2010) High-throughput production of human proteins for crystallization: The SGC experience. *J Struct Biol* 172(1):3–13.
7. Studier FW (2005) Protein production by auto-induction in high density shaking cultures. *Protein Expr Purif* 41(1):207–234.
8. Sauder MJ, et al. (2008) *High Throughput Protein Production and Crystallization at NYSGXRC* 426:561–575.
9. Otwinowski Z, Minor W (1997) Processing of X-ray diffraction data collected in oscillation mode. *Methods Enzymol* 276:307–326.
10. Collaborative Computational Project, Number 4 (1994) The CCP4 suite: Programs for protein crystallography. *Acta Crystallogr D Biol Crystallogr* 50(Pt 5):760–763.
11. Murshudov GN, Vagin AA, Dodson EJ (1997) Refinement of macromolecular structures by the maximum-likelihood method. *Acta Crystallogr D Biol Crystallogr* 53(Pt 3): 240–255.
12. Emsley P, Cowtan K (2004) Coot: Model-building tools for molecular graphics. *Acta Crystallogr D Biol Crystallogr* 60(Pt 12 Pt 1):2126–2132.

13. Pei J, Kim B-H, Grishin NV (2008) PROMALS3D: A tool for multiple protein sequence and structure alignments. *Nucleic Acids Res* 36(7):2295–2300.
14. Chang TH, et al. (2010) Structure of a heterotetrameric geranyl pyrophosphate synthase from mint (Mentha piperita) reveals intersubunit regulation. *Plant Cell* 22(2):454–467.
15. Hosfield DJ, et al. (2004) Structural basis for bisphosphonate-mediated inhibition of isoprenoid biosynthesis. *J Biol Chem* 279(10):8526–8529.
16. Guo RT, et al. (2007) Bisphosphonates target multiple sites in both cis- and trans-prenyltransferases. *Proc Natl Acad Sci USA* 104(24):10022–10027.
17. Hsieh FL, Chang TH, Ko TP, Wang AH (2011) Structure and mechanism of an Arabidopsis medium/long-chain-length prenyl pyrophosphate synthase. *Plant Physiol* 155(3):1079–1090.
18. Tarshis LC, Proteau PJ, Kellogg BA, Sacchettini JC, Poulter CD (1996) Regulation of product chain length by isoprenyl diphosphate synthases. *Proc Natl Acad Sci USA* 93(26):15018–15023.

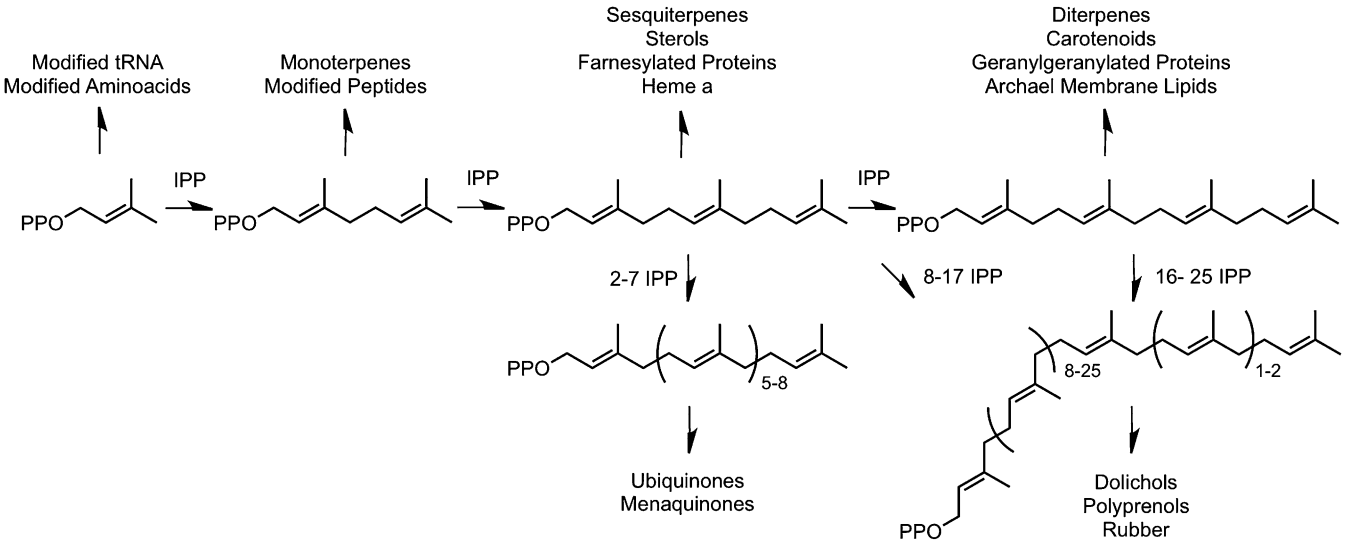

Fig. S1. Linear polyprenyltransferases serve as core for synthesis of a large variety of isoprenoids.

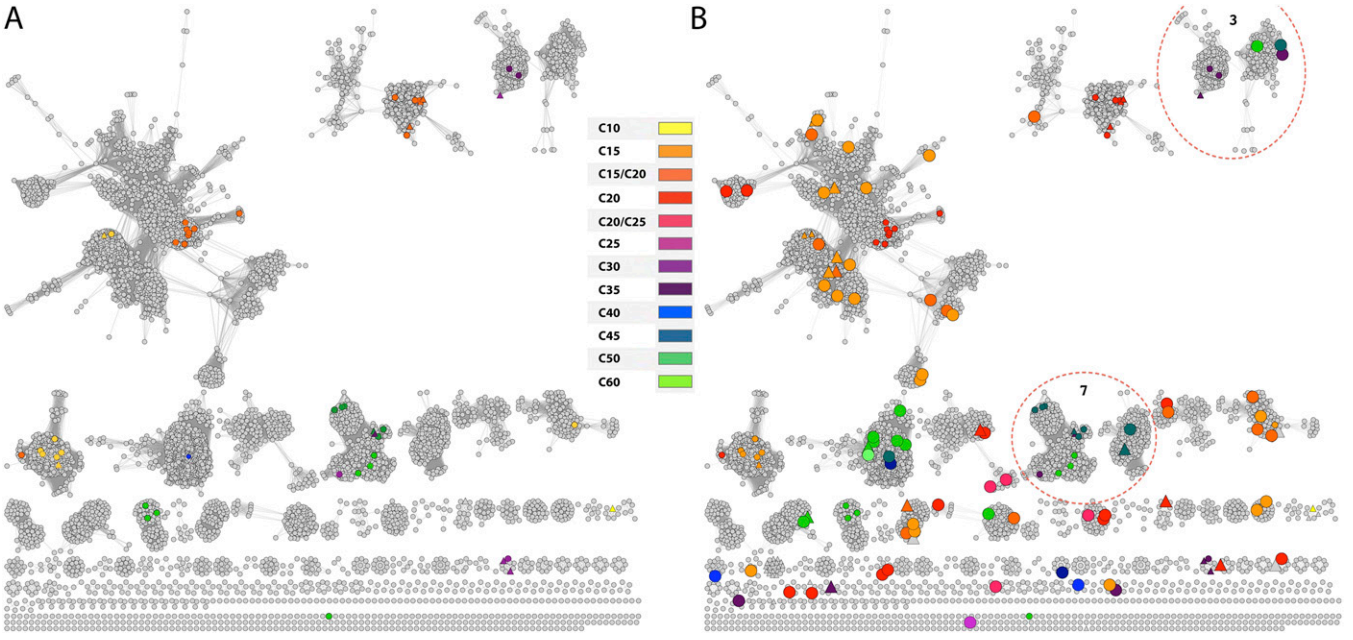

Fig. S2. Sequence similarity maps of the *E*-PTS subgroup with a BLAST e-value threshold of $1e^{-70}$. Nodes are colored according to functional assignment; triangular nodes represent crystal structures (including those with only different ligands). (*A*) Before this study, 93 sequences were annotated, and 69 crystal structures were solved. (*B*) After this study, 79 additional sequences were annotated (larger nodes), and 22 additional crystal structures were solved. Clustesr 3 and 7 are highlighted by red circles.

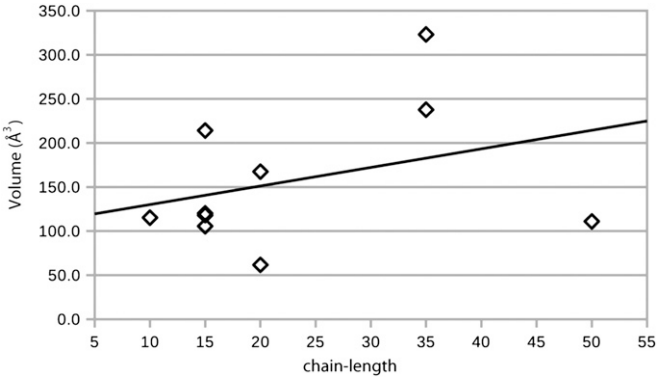

Fig. S3. Scatter plot of elongation-cavity volume vs. known product chain length of training set crystal structures.

Table S1. Protein purification vectors for the 79 E-PTSs purified in this study

EFI ID	GI	Vector	EFI ID	GI	Vector
501393	34557991	pNIC28-Bsa4	502322	15642935	pNIC28-Bsa4
501400	46579760	pNIC28-Bsa4	502349	14590908	pNIC28-Bsa4
501401	39934591	pNIC28-Bsa4	502360	55979982	pNIC28-Bsa4
501414	11499146	pNIC28-Bsa4	900386	34540567	pSGX3
501418	16800468	CHS30	900387	60682991	pSGX3
501422	28378303	CHS30	900388	15640461	pSGX3
501426	17546941	pNIC28-Bsa4	900392	53711383	pSGX3
501432	33600896	CHS30	900395	148993815	pSGX3
501449	15805956	CHS30	900396	153799383	pSGX3
501455	16081558	pNIC28-Bsa4	900397	23308797	pSGX3
501461	21227869	CHS30	900398	58337602	pSGX3
501462	28377915	pNIC28-Bsa4	900400	104774286	pSGX3
501472	29831407	CHS30	900401	125624487	pSGX3
501473	29829539	pNIC28-Bsa4	900402	15640906	pSGX3
501943	83764459	pNIC28-Bsa4	900403	29375563	pSGX3
501944	68489506	pNIC28-Bsa4	900404	52842862	pSGX3
501947	50309979	pNIC28-Bsa4	900407	118468511	pSGX3
501949	149238027	CHS30	900409	21218742	pSGX3
501952	56551751	pNIC28-Bsa4	900411	21223617	pSGX3
501954	21673095	CHS30	900416	29376566	pSGX3
501956	154483652	CHS30	900417	57833856	pSGX3
501957	254302823	CHS30	900418	16126308	pSGX3
501958	24215915	pNIC28-Bsa4	900419	29377703	pSGX3
501961	53804815	CHS30	900420	308388199	pSGX3
501964	83594313	CHS30	900421	16131077	pSGX3
501972	29840764	pNIC28-Bsa4	900422	39934115	pSGX3
501974	18310802	pNIC28-Bsa4	900423	67866738	pSGX3
501976	58578715	pNIC28-Bsa4	900425	77464324	pSGX3
501980	87120692	CHS30	900463	21225056	pSGX3
501981	154151748	pNIC28-Bsa4	900464	57238655	pSGX3
501984	83945403	pNIC28-Bsa4	900465	116334218	pSGX3
501986	63034220	pNIC28-Bsa4	900466	77862362	pSGX3
501992	126458776	CHS30	900467	116333612	pSGX3
501993	126460364	CHS30	900468	293372070	pSGX3
501995	291005007	CHS30	900469	52842540	pSGX3
501998	152991761	CHS30	900470	15645545	pSGX3
502000	300633355	CHS30	900471	19551716	pSGX3
502006	40062988	CHS30	900472	23308904	pSGX3
502008	58584994	pNIC28-Bsa4	900473	70732810	pSGX3
502009	15837263	CHS30			

Table S2. Crystallization conditions for the 19 structures of distinct polyprenyl transferases solved and applied in this study

PDB	Target ID	Ligand	Treatment of crystals before freezing	Crystallization conditions
3PDE	EFI-900467 (NYSGXRC-20032b)	Glycerol, IPP, Mg^{2+}	Dehydrated for 24 h in precipitant solution with 100 mM $MgCl_2$ and IPP powder	0.1 M sodium acetate, pH 4.5, 25% PEG 3350, 10% glycerol, vapor diffusion, sitting drop, 294K
3PKO	EFI-900465 (NYSGXRC-20019c)	Citric acid, glycerol	Direct freeze in liquid N^2	1.8 M ammonium (tri)citrate, pH 7, 10% glycerol, vapor diffusion, sitting drop, 294K
3Q1O	EFI-900470 (NYSGXRC-20030a)	IPP, Mg^{2+}, SO_4^{2-}	Crystals incubated in 2 M Li_2SO_4, 10% glycerol, 5 mM $MgCl_2$ and IPP powder for 1.5 h	0.1 M Hepes, pH 7.5, 2,000 mM ammonium sulfate, 10% glycerol, vapor diffusion, sitting drop, 294K
3QQV	EFI-900472 (NYSGXRC-20008a)	Glycerol, IPP, Mg^{2+}, unknown	Dehydrated for 24 h in precipitant solution with 100 mM $MgCl_2$ and IPP powder	0.15 M potassium bromide, pH 7.5, 30% PEG MME2000, 10% glycerol, vapor diffusion, sitting drop, 294K
3Q2Q	EFI-900471 (NYSGXRC-20015a)	Ca^{2+}, glycerol, IPP	Dry IPP was added to the drop and incubated for 24 h	0.1 M sodium acetate, pH 4.5, 30% PEG 400, 200 mM calcium acetate, 10% glycerol, vapor diffusion, sitting drop, 294K
3OYR	EFI-900420 (NYSGXRC-20011b)	Ca^{2+}, IPP, diphosphate	Dry IPP crystals added to the drop, direct freeze after 2 h of dehydration	0.1 M Mes, pH 6.0, 20% PEG 8K, 200 mM calcium acetate monohydrate, 10% glycerol, vapor diffusion, sitting drop, 294K
3TS7	EFI-501961	PO_4^{2-}	Crystals were soaked in the precipitant with 25% glycerol (vol/vol)	0.1 M Tris·HCL, 2 M ammonium phosphate monobasic, pH 8.5, vapor diffusion, sitting drop, 294K
3UCA	EFI-501974	Glycerol, PO_4^{2-}	Well B2, MCSG-1 screen, direct freeze	0.1 M Bis-Tris, 0.2 M sodium chloride, pH 5.5, 25% PEG 3350, vapor diffusion, sitting drop, 294K
3LVS	EFI-900468 (NYSGXRC-20032c)	Glycerol, PO_4^{2-}	30% glycerol was used as a cryoprotectant, short soak	20% PEG 8000, 50 mM potassium dihydrogen phosphate, 10% glycerol, vapor diffusion, sitting drop, 294K
3NF2	EFI-900463 (NYSGXRC-20006e)	SO_4^{2-}	Direct freeze	0.1 M phosphate citrate, pH 4.2, 2 M ammonium sulfate, 10% glycerol, vapor diffusion, sitting drop, 294K
3MZV	EFI-900466 (NYSGXRC-20011c)	None	Dehydrated for 2.5 h	100 mM Mes pH 6.5, 30% PEG MME 5K, 200 mM ammonium sulfate, vapor diffusion, temperature 294K
3P8R	EFI-900402 (NYSGXRC-20032e)	None	Dehydrated for 24 h	0.1 M Bis-Tris, pH 6.5, 25% PEG 3350, 200 mM ammonium acetate, 10% glycerol, vapor diffusion, sitting drop, 294K
3P8L	EFI-900403 (NYSGXRC-20032a)	None	Dehydrated for 24 h	0.1 M Bis-Tris, pH 6.5, 25% PEG 3350, 200 mM sodium chloride, 10% glycerol, vapor diffusion, sitting drop, 294K
3LOM	EFI-900469 (NYSGXRC-20026a)	PO_4^{2-}	Dehydrated for 3 h	0.1 M phosphate citrate, pH 4.2, 200 mM lithium sulfate, 20% PEG 1000, 10% glycerol, vapor diffusion, sitting drop, 294K
3P41	EFI-900473 (NYSGXRC-20027b)	Cl^-, glycerol, IPP, Mg^{2+}, diphosphate	Dry IPP crystals added to drop, direct freeze after 48 h of dehydration	0.1 M sodium cocodylate, pH 6.5, 20% PEG 8K, 200 mM magnesium acetate, vapor diffusion, sitting drop, 294K
4FP4	EFI-501993	Ger, unknown	Direct freeze in liquid N^2.	0.1 M Hepes, 10% PEG 6000, pH 7.5, 5% MPD
4F62	EFI-501980	Glycerol, SO_4^{2-}	Reservoir solution, 20% glycerol	0.1 M sodium chloride, 0.1 M Hepes, pH 7.5, 1.6 M ammonium sulfate, vapor diffusion, sitting drop, 298K
4DHD	EFI-501992	Acetate, PO_4^{2-}	Dehydrated for 3 h	0.1 M malic acid, pH 7.0, 30% PEG 3350, 10% glycerol, vapor diffusion, sitting drop, 294K
3RMG	EFI-900427	None	Direct freeze	0.1 M sodium acetate, pH 4.5, 25% PEG 3350, 10% glycerol, vapor diffusion, sitting drop, 294K

Table S3. Data collection and refinement statistics for crystals of distinct polyprenyl transferases 3PDE, 3PKO, 3Q1O, and 3QQV

	PDB identifier			
	3PDE	3PKO	3Q1O	3QQV
Space group	P 1	P 21 21 21	P 43 21 2	P 31 2 1
Unit cell dimension (Å)				
a	48.54	50.10	191.87	123.64
b	51.17	106.71	191.87	123.64
c	126.57	132.60	127.05	51.82
Cell angles (°)				
α	95.75	90.00	90.00	90.00
β	91.71	90.00	90.00	90.00
γ	105.31	90.00	90.00	120.00
Molecules per ASU	4	2	4	1
Solvent content	45.42	47.70	73.80	55.80
Matthew's coefficient	2.25	2.35	4.70	2.78
Ligands	Glycerol, IPP, Mg^{2+}	Citric acid, glycerol	IPP, Mg^{2+}, SO_4^{2-}	Glycerol, IPP, Mg^{2+}, unknown ligand
X-ray source	NSLS X29A	NSLS X29A	NSLS X29A	NSLS X29A
Wavelength	1.075	0.9789	0.979	1.075
Method of structure solution	MR	SAD	SAD	MR
Resolution	40.00–1.75	50.00–1.98	40.00–2.30	40.00–2.00
Resolution/refinement	40.00–1.75	50.00–1.98	40.00–2.40	40.00–2.00
Completeness (%)	97.0 (95.8)	99.9 (98.0)	100.0 (100.0)	99.9 (100.0)
I/sigma (I)	4.50 (1.30)	8.00 (1.40)	8.30 (0.6)	7.10 (0.90)
R_{sym}	0.085 (0.650)	0.088 (0.810)	0.110 (NULL)	0.100 (NULL)
R_{work} (R_{free})	17.2 (20.4)	18.5 (23.0)	21.5 (24.3)	20.9 (25.2)
R_{free} reflections (%)	3419 (3.0%)	1557 (3.1%)	2772 (3.0%)	975 (3.2%)
Average B factor	26.02	45.51	63.15	59.99
rmsd				
Bond lengths	0.011	0.012	0.007	0.008
Bond angles	1.368	1.254	1.133	1.141
Number of solvent molecules	974	258	224	146

Numbers in parenthesis are those for the high-resolution shell.

Table S4. Data collection and refinement statistics for crystals of distinct polyprenyl transferases 3Q2Q, 3OYR, 3TS7, and 3UCA

	PDB identifier			
	3Q2Q	3OYR	3TS7	3UCA
Space group	P 32 2 1	P 21 21 21	P 42 21 2	P 31 2 1
Unit cell dimension (Å)				
a	71.66	72.89	114.59	70.12
b	71.66	72.44	114.59	70.12
c	150.92	125.85	113.19	228.14
Cell angles (°)				
α	90.00	90.00	90.00	90.00
β	90.00	90.00	90.00	90.00
γ	120.00	90.00	90.00	120.00
Molecules per ASU	1	2	2	2
Solvent content	57.09	45.14	52.47	44.86
Matthew's coefficient	2.86	2.24	2.59	2.23
Ligands	Ca^{2+}, glycerol, IPP	Ca^{2+}, IPP, pyrophosphate	PO_4^{2-}	glycerol, PO_4^{2-}
X-ray source	NSLS X29A	NSLS X29A	NSLS X29A	NSLS X29A
Wavelength	0.9789	0.979	1.075	1.075
Method of structure solution	MR	SAD	MR	MR
Resolution	40.00–1.90	40.00–2.00	50.00–1.94	50.00–2.00
Reflections	36,231	45,715	56,269	45,199
Completeness (%)	99.8 (100.0)	91.9 (96.8)	94.0 (96.6)	99.8 (100.0)
I/sigma (I)	10.10 (0.90)	6.10 (1.00)	7.00 (1.10)	5.70 (1.10)
R_{sym}	0.075 (NULL)	0.074 (NULL)	0.131 (0.950)	0.132 (NULL)
R_{work} (R_{free})	18.4 (21.8)	19.9 (24.7)	22.1 (26.6)	19.1 (23.0)
R_{free} reflections (%)	1118 (3.1%)	1283 (3.1%)	1631 (3.1%)	1402 (3.1%)
Average B factor	51.09	48.83	47.28	47.28
rmsd				
Bond lengths	0.009	0.011	0.011	0.01
Bond angles	1.03	1.354	1.292	1.109
Number of solvent molecules	204	211	172	213

Numbers in parenthesis are those for the high-resolution shell.

Table S5. Data collection and refinement statistics for crystals of distinct polyprenyl transferases solved 3LVS, 3NF2, 3MZV, and 3P8R

	PDB identifier			
	3LVS	3NF2	3MZV	3P8R
Space group	P 21 21 21	P 31 2 1	C 1 2 1	P 61 2 2
Unit cell dimension (Å)				
a	54.46	68.05	123.46	79.59
b	89.46	68.05	90.25	79.59
c	132.77	147.64	80.17	215.32
Cell angles (°)				
α	90.00	90.00	90.00	90.00
β	90.00	90.00	125.76	90.00
γ	90.00	120.00	90.00	120.00
Molecules per ASU	2	1	2	1
Solvent content	53.12	53.12	49.11	58.50
Matthew's coefficient	2.62	2.62	2.42	2.96
Ligands	Glycerol, PO_4^{2-}	SO_4^{2-}	None	None
X-ray source	NSLS X29A	NSLS X29A	NSLS X29A	NSLS X29A
Wavelength	0.9791	0.9789	0.97958	0.979
Method of structure solution	SAD	SAD	SAD	SAD
Resolution	40.00–2.15	40.00–2.20	30.00–1.90	40.00–2.34
Resolution/refinement	40.00–2.15	40.00–2.20	30.00–1.90	40.00–2.50
Completeness (%)	99.8 (100.0)	96.1 (69.6)	99.9 (100.0)	99.8 (99.7)
I/sigma (I)	5.90 (0.90)	10.10 (1.00)	13.00 (3.00)	10.70 (0.90)
R_{sym}	0.085 (0.790)	0.076 (NULL)	0.064 (0.479)	0.064 (NULL)
R_{work} (R_{free})	21.8 (25.0)	23.3 (27.5)	19.4 (22.1)	23.0 (29.0)
R_{free} reflections (%)	1112 (3.1%)	652 (3.2%)	2837 (5.1%)	464 (3.2%)
Average B factor	61.88	57.07	46.95	82.5
rmsd				
Bond lengths	0.009	0.008	0.018	0.007
Bond angles	1.232	1.183	1.507	1.006
Number of solvent molecules	125	41	305	35

Numbers in parenthesis are those for the high-resolution shell.

Table S6. Data collection and refinement statistics for crystals of distinct polyprenyl transferases 3P8L, 3LOM, 3P41, and 4FP4

	PDB identifier			
	3P8L	3LOM	3P41	4FP4
Space group	P 21 21 21	P 43 21 2	P 31 2 1	P 1 21 1
Unit cell dimension (Å)				
a	53.86	94.78	48.36	46.96
b	100.53	94.78	48.36	115.42
c	122.66	148.35	208.59	53.14
Cell angles (°)				
α	90.00	90.00	90.00	90.00
β	90.00	90.00	90.00	110.84
γ	90.00	90.00	120.00	90.00
Molecules per ASU	2	2	1	2
Solvent content	51.20	48.91	43.65	41.38
Matthew's coefficient	2.52	2.41	2.18	2.10
Ligands	None	PO_4^{2-}	Cl^-, glycerol, IPP, Mg^{2+}, pyrophosphate	Geranylgeranyl, UNX, UNL
X-ray source	NSLS X29A	NSLS X29A	NSLS X29A	NSLS X29A
Wavelength	0.979	0.9791	1.075	1.075
Method of structure solution	SAD	SAD	MR	MR
Resolution	40.00–1.90	40.00–2.30	40.00–1.76	40.00–2.00
Resolution/refinement	40.00–2.00	40.00–2.30	20.00–1.76	38.57–2.00
Completeness (%)	99.9 (99.6)	99.9 (100.0)	97.2 (78.8)	99.6 (99.3)
I/sigma (I)	8.20 (0.90)	6.00 (2.20)	6.30 (1.50)	7.90 (1.80)
R_{sym}	0.068 (NULL)	0.096 (0.550)	0.088 (0.690)	0.064 (0.62)
R_{work} (R_{free})	20.2 (24.6)	22.2 (26.7)	19.5 (22.3)	18.6 (22.8)
R_{free} reflections (%)	1397 (3.1%)	973 (3.2%)	918 (3.2%)	1115 (3.2%)
Average B factor	54.98	61.49	34.75	61.70
rmsd				
Bond lengths	0.01	0.008	0.012	0.008
Bond angles	1.231	1.218	1.417	1.342
Number of solvent molecules	168	135	142	99

Numbers in parenthesis are those for the high-resolution shell.

Table S7. Data collection and refinement statistics for crystals of distinct polyprenyl transferases 4F62, 4DHD, and 3RMG

	PDB identifier		
	4F62	4DHD	3RMG
Space group	P 32 2 1	C 2 2 21	P 21
Unit cell dimension (Å)			
a	96.49	61.40	44.62
b	96.49	89.27	111.41
c	156.39	149.54	83.38
Cell angles (°)			
α	90.00	90.00	90.00
β	90.00	90.00	92.56
γ	120.00	90.00	90.00
Molecules per ASU	2	1	2
Solvent content	58.91	52.06	56.20
Matthew's coefficient	2.99	2.56	2.81
Ligands	Glycerol, SO_4^{2-}	Acetate, PO_4^{2-}	None
X-ray source	NSLS X29A	NSLS X29A	NSLS X29A
Wavelength	0.979	1.075	0.979
Method of structure solution	MR	MR	SAD
Resolution	41.06–2.10	70.00–1.65	50.00–2.20
Resolution/refinement	41.00–2.10	50.00–1.65	50.00–2.20
Completeness (%)	100.0 (100.0)	93.8 (100.0)	99.9 (100.0)
I/sigma (I)	10.3 (NULL)	9.90 (3.8)	9.10 (1.20)
R_{sym}	NULL (0.751)	0.117 (0.360)	0.070 (NULL)
R_{work} (R_{free})	17.8 (20.3)	18.2 (21.3)	21.50 (25.8)
R_{free} reflections (%)	1897 (4.0%)	1430 (3.1%)	1120 (3.1%)
Average B factor	53.50	27.39	66.60
rmsd			
Bond lengths	0.007	0.009	0.007
Bond angles	1.012	1.260	0.915
Number of solvent molecules	157	316	69

Numbers in parenthesis are those for the high-resolution shell.

Table S8. Training set: Experimentally determined vs. predicted product chain length

				Prediction (flexible)	
GI	PDB ID	Experiment	Source	$E_{bind/MMGBSA}$	E_{LJ}
158979013	3KRF	C_{10}	(13)	C_{15}	C_{15}
24111799	1RQI	C_{15}	(14)	C_{15}	C_{15}
116333612	3PDE	C_{15*}	EFI	C_{20}	C_{15}
15645545	3Q1O	C_{15}/C_{20*}	EFI	C_{20}	C_{15}
70732810	3P41	C_{15}/C_{20*}, $C_{15}^{\dagger}$	EFI	C_{15}	C_{20}
218766585	2E8W	C_{20}	(15)	C_{25}	C_{25}
23308904	3QQV	C_{20*}, $C_{20}^{\dagger}$	EFI	$C_{\geq 30}$	C_{20}
319443461	3AQ0	C_{35}	(16)	$C_{\geq 30}$	$C_{\geq 30}$
3915686	1UBW	C_{35}/C_{40}	(17)	$C_{\geq 30}$	$C_{\geq 30}$
16126352	3OYR	C_{50}-C_{60*}	EFI	$C_{15}^{\ddagger}/C_{\geq 30}^{\S}$	$C_{20}^{\ddagger}/C_{\geq 30}^{\S}$

Energy functions are $E_{bind/MMGBSA}$ and E_{LJ} in application with flexible protein side chains.
*IPP:DMAPP = 4:1.
†IPP: DMAPP = 1:1.
‡Crystal structure of 3OY.
§Relaxed crystal structure of 3OYR.

Table S9. Experimentally determined, predicted, and sequence-derived (TrEMBL) product chain length of *E*-PTS sequences of set targets$_{known}$

GI	Experiment		Predicted	Template	Type	Sequence. identity	TrEMBL
	IPP:DMAPP = 4:1	IPP:DMAPP = 1:1					
77862362	C_{50}/C_{55}		C_{25}	3MZV	Apo	100	C_{50}
293372070	C_{15}		C_{15}	3LVS	Apo	100	C_{15}
19551716	C_{45}/C_{50}		C_{25}	3Q2Q	Holo	100	C_{20}
21225056	C_{15}/C_{20}	C_{15}	C_{20}	3NF2	Apo	100	Polyprenyl synthase
15640906	C_{15}/C_{20}		C_{15}	3P8R	Apo	100	C_{15}
29375563	C_{15}/C_{20}	C_{15}	C_{15}	3P8L	Apo	100	C_{15}
52842540	C_{20}[a]	C_{20}	C_{20}	3LOM	Apo	100	C_{15}
57238655	C_{15}/C_{20}	C_{15}	C_{20}	3NPK	Apo	100	C_{15}
116334218	C_{45}/C_{50}		$C_{\geq 30}$	3PKO	Apo	100	C_{20}
8272412	C_{15}		C_{15}	3LVS	Apo	95	C_{15}
77464324	C_{50}/C_{55}		C_{25}	3MZV	Apo	78	C_{20}
67866738	C_{50}		C_{25}	3MZV	Apo	53	C_{50}
39934115	C_{50}		C_{25}	3MZV	Apo	52	C_{40}
125624487	C_{20}	C_{15}	C_{20}	1RTR	Apo	51	C_{15}
16126308	C_{15}		C_{20}	3LVS	Apo	47	C_{15}
118468511	C_{20}		C_{20}	3QQV	Holo	46	Polyprenyl synthase
148993815	C_{20}	C_{20}	C_{20}	1RTR	Apo	46	FPP/GGPP synthase family
16131077	C_{45}/C_{50}		C_{25}	3MZV	Apo	46	C_{40}
15640461	C_{40}/C_{45}		C_{20}	3OYR	Holo	45	C_{40}
104774286	C_{15}		C_{15}	3PDE	Holo	44	C_{20}
21223617	C_{15}/C_{20}	C_{15}	C_{25}	3NF2	Apo	44	Polyprenyl synthase
29376566	C_{50}/C_{55}		C_{25}	3PKO	Apo	44	C_{35}
52842862	C_{60}/C_{65}		C_{25}	3MZV	Apo	44	C_{40}
21227869	C_{15}/C_{20}	C_{20}	C_{25}	1WY0	Apo	43	C_{10}
21218742	C_{20}/C_{25}		C_{20}	3QQV	Holo	41	C_{20}
58337602	C_{15}		C_{15}	3PDE	Holo	40	C_{20}
34540567	C_{15}/C_{20}	C_{15}	C_{15}	1WY0	Apo	39	Polyprenyl synthase
57833856	C_{20}/C_{25}	C_{25}	C_{25}	3QQV	Holo	39	Polyprenyl synthase
60682991	C_{50}		C_{15}	1WY0	Apo	39	FPP/GGPP synthase family
29348670	C_{50}/C_{55}		C_{25}	3MZV	Apo	38	C_{40}
153799383	C_{15}	C_{10}	$C_{\geq 30}$	3QQV	Holo	36	FPP/GGPP synthase family
39934591	C_{20}	C_{20}	C_{20}	3Q1O	Holo	36	C_{20}
53711383	C_{50}/C_{55}		$C_{\geq 30}$	3MZV	Apo	36	C_{40}
23308797	C_{20}/C_{25}	C_{20}	C_{15}	3QQV	Holo	34	C_{20}

Table S10. Experimentally determined, predicted, and sequence-derived (TrEMBL) product chain length of *E*-PTS sequences of set targets$_{blind}$

GI	Experiment		Predicted	Template	Type	Sequence identity (%)	TrEMBL
	IPP:DMAPP = 4:1	IPP:DMAPP = 1:1					
53804815	C_{15}		C_{15}	3TS7	Apo	100	C_{15}
18310802	C_{15}		C_{15}	3UCA	Apo	100	C_{15}
55979982	C_{20}		C_{15}	1WMW	Apo	100	C_{20}
15642935	C_{20}		C_{20}	2FTZ	Apo	100	C_{15}
34557991	C_{15}		C_{15}	3Q1O	Holo	56	C_{20}
28378303	C_{15}		C_{15}	3PDE	Holo	56	Polyprenyl synthase
15837263	C_{15}		C_{15}	1RQI	Holo	54	C_{15}
87120692	C_{15}		C_{15}	1RQI	Holo	53	C_{15}
28377915	C_{30}		C_{20}	3PKO	Apo	52	C_{35}
33600896	C_{15}		C_{20}	1RQI	Holo	51	C_{15}
291005007	C_{15}		C_{20}	3NF2	Apo	50	C_{10}
152991761	C_{15}		C_{15}	3Q1O	Holo	49	C_{15}
300633355	C_{15}		C_{20}	1RQI	Holo	48	C_{15}
17546941	C_{15}		C_{15}	1RQI	Holo	47	C_{15}
29831407	C_{45}		$C_{\geq 30}$	3Q2Q	Apo	47	Polyprenyl synthase
46579760	C_{15}		C_{15}	3PDE	Holo	46	C_{20}
56551751	C_{15}/C_{20}	C_{15}	C_{15}	1RQI	Holo	45	Polyprenyl synthase
16800468	C_{15}		C_{15}	3PDE	Holo	43	FPP/GGPP synthase family
68489506	C_{20}	C_{20}	C_{20}	2E8W	Holo	43	FPP/GGPP synthase family
40062988	C_{20}	C_{20}	C_{25}	3P8R	Apo	43	C_{20}
50309979	C_{25}	C_{25}	C_{25}	2E8W	Holo	42	FPP/GGPP synthase family
149238027	C_{20}	C_{20}	C_{20}	2E8W	Holo	42	FPP/GGPP synthase family
154483652	C_{15}		C_{15}	3PDE	Holo	42	FPP/GGPP synthase family
254302823	C_{15}		C_{15}	1RQI	Holo	42	C_{10}
83945403	C_{15}/C_{20}	C_{15}	C_{20}	1RQI	Holo	40	C_{15}
63034220	C_{20}	C_{20}	C_{20}	3KRF	Holo	40	C_{20}
11499146	C_{35}/C_{40}		C_{25}	3OYR	Holo	38	C_{40}
15805956	C_{40}		C_{25}	3OYR	Holo	38	Polyprenyl synthase
58578715	C_{55}/C_{60}		C_{25}	3OYR	Holo	37	C_{40}
83764459	C_{15}/C_{20}	C_{15}/C_{20}	C_{25}	2E8W	Holo	36	C_{20}
83594313	C_{20}/C_{25}	C_{20}	C_{20}	3PDE	Holo	36	C_{15}
154151748	C_{20}	C_{20}	C_{25}	3OYR	Holo	36	C_{20}
58584994	C_{15}		C_{15}	3PDE	Holo	36	C_{20}
29829539	C_{15}/C_{20}	C_{15}	C_{20}	3OYR	Holo	35	Polyprenyl synthase
21673095	C_{20}	C_{20}	C_{20}	3OYR	Holo	34	Isoprenyl synthase
16081558	C_{35}/C_{40}		C_{20}	3OYR	Holo	33	C_{10}/C_{15}
29840764	C_{40}-C_{60}		C_{15}	3PDE	Holo	32	C_{15}
126458776	C_{20}		C_{20}	3OYR	Holo	32	Polyprenyl synthase
24215915	C_{40}		C_{25}	3OYR	Holo	29	C_{20}
126460364	C_{35}		C_{25}	3AQ0	Holo	29	Polyprenyl synthase

Other Supporting Information Files

Dataset S1 (PDF)