



DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, **Published:**

LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE,

SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA,

GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

— *with international search report (Art. 21(3))*

DISCOVERY OF ENZYMES FROM THE ALPHA-KETO ACID DECARBOXYLASE FAMILY

CROSS-REFERENCES TO RELATED APPLICATIONS

[0001] This application claims priority to U.S. Provisional Application No. 62/211,439, filed August 28, 2015, the contents of which are hereby incorporated in the entirety for all purposes.

BACKGROUND OF THE INVENTION

[0002] A central goal of the bioeconomy is to reduce our dependence on petroleum through next-generation biomanufacturing. The USDA projects the industrial production of bio-based specialty chemicals to reach ~\$340 billion USD by 2025, replacing half of our needs for these products from petroleum. To accomplish this ambitious goal, non-natural biochemical pathways are needed. An example of one such pathway is the “synthetic recursive +1” carbon elongation pathway that has been developed to produce a variety of alcohol products in *E. coli*. The primary product from the current pathway is 1-butanol, with longer chain alcohols (*e.g.*, pentanol, hexanol, heptanol, and octanol) being either minor products of the pathway, or not produced at all. Yet, there is significant interest in producing long chain alcohols given their use as specialty chemicals, as well as their greater energy density when used as liquid fuels or fuel additives. The present invention fulfills these and other needs.

BRIEF SUMMARY OF THE INVENTION

[0003] In one aspect, the present invention provides an expression cassette comprising a heterologous promoter operably linked to a nucleic acid encoding an enzyme comprising 2-ketoacid decarboxylase activity, wherein the enzyme: comprises a 3-layer alpha-beta-alpha sandwich; comprises a thiamine pyrophosphate (TPP) binding pocket; has a TM align score of at least about 0.5 (*e.g.*, from about 0.5 to about 1.0, from about 0.5 to about 0.8, or from about 0.5 to about 0.9) as compared against PDB ID:2VBG; and comprises at least about a 10-fold, 25-fold, 50-fold, or 100-fold (*e.g.*, from about 10-fold to about 100,000-fold; from about 10-fold to about 10,000 fold; from about 10-fold to about 200-fold; from about 10-fold to about 100-fold;

from about 25-fold to about 100,000-fold; from about 25-fold to about 10,000 fold; from about 25-fold to about 200-fold; from about 25-fold to about 100-fold; from about 50-fold to about 100,000-fold; from about 50-fold to about 10,000 fold; from about 50-fold to about 200-fold; or from about 50-fold to about 100-fold) greater catalytic efficiency for C8 2-ketoacid substrates as compared to C3 or isoC5 2-ketoacid substrates, wherein the catalytic efficiency is measured by kcat/Km.

[0004] In some embodiments, the enzyme comprising 2-ketoacid decarboxylase activity has a TM align score of at least 0.5 (*e.g.*, from about 0.5 to about 0.8, from about 0.5 to about 0.9, or from about 0.5 to about 1.0), 0.6 (*e.g.*, from about 0.6 to about 0.8, from about 0.6 to about 0.9, or from about 0.6 to about 1.0), 0.7 (*e.g.*, from about 0.7 to about 0.8, from about 0.7 to about 0.9, or from about 0.7 to about 1.0), 0.8 (*e.g.*, from about 0.8 to about 0.9, or from about 0.8 to about 1.0), or 0.9 (*e.g.*, from about 0.9 to about 1.0) as compared against PDB ID:2VBG. In some embodiments, the substrate binding pocket of the enzyme has a solvent accessible surface area of, of about, of at least, or of at least about 5,000; 5,500; 6,000; 6,500; 7,000; 7,500; 8,000; or 8,500; 9,000; 9,500; or 10,000 Å² (*e.g.*, from about 5,000 to about 10,000 Å²; from about 5,500 to about 9,500 Å²; from about 6,000 to about 9,000 Å²; from about 6,500 to about 8,500 Å²; from about 7,000 to about 8,000 Å²; or from about 7,500 to about 8,000 Å²).

[0005] In some embodiments, the enzyme comprising 2-ketoacid decarboxylase activity comprises an amino acid sequence that is, is at least, or is at least about 14%, 15%, 16%, 17%, 18%, 19%, 20%, 21%, 22%, 23%, 24%, 25%, 26%, 27%, 28%, 29%, 30%, 31%, 32%, 33%, 34%, 35%, 36%, 37%, 38%, 39%, 40%, 41%, 42%, 43%, 44%, 45%, 46%, 47%, 48%, 49%, 50%, 51%, 52%, 53%, 54%, 55%, 56%, 57%, 58%, 59%, 60%, 61%, 62%, 63%, 64%, 65%, 66%, 67%, 68%, 69%, 70%, 71%, 72%, 73%, 74%, 75%, 76%, 77%, 78%, 79%, 80%, 81%, 82%, 83%, 84%, 85%, 86%, 87%, 88%, or 89% identical to SEQ ID NO:1. In some embodiments, the enzyme comprising 2-ketoacid decarboxylase activity comprises an amino acid sequence that is less than, or less than about, 90%, 95%, or 99% identical to SEQ ID NO:1. In some embodiments, the enzyme comprising 2-ketoacid decarboxylase activity comprises an amino acid sequence that is from about 14% to about 89% identical to SEQ ID NO:1; from about 20% to about 89% identical to SEQ ID NO:1; from about 25% to about 89% identical to SEQ ID NO:1; from about 30% to about 89% identical to SEQ ID NO:1; from about 35% to about 89%

identical to SEQ ID NO:1; from about 40% to about 89% identical to SEQ ID NO:1; from about 45% to about 89% identical to SEQ ID NO:1; from about 50% to about 89% identical to SEQ ID NO:1; from about 55% to about 89% identical to SEQ ID NO:1; from about 60% to about 89% identical to SEQ ID NO:1; from about 65% to about 89% identical to SEQ ID NO:1; from about 70% to about 89% identical to SEQ ID NO:1; from about 75% to about 89% identical to SEQ ID NO:1; or from about 80% to about 89% identical to SEQ ID NO:1.

[0006] In some embodiments, the enzyme comprising 2-ketoacid decarboxylase activity comprises an amino acid sequence that contains a mutation of a native sequence wherein the mutation is in an amino acid residue within a 2-ketoacid decarboxylase active site or in an amino acid residue having a C α within 8 Å (*e.g.*, having a C α atom within about 1, 1.5, 2, 2.5, 3, 3.5, 4, 4.5, 5, 5.5, 6, 6.5, 7, or 7.5 Å) of a 2-ketoacid decarboxylase active site, wherein the 2-ketoacid decarboxylase active site comprises the amino acids corresponding to amino acid positions 23-27, 50, 51, 56, 75, 76, 79, 82, 113, 114, 287, 352, 373-377, 379, 394, 396-399, 424-428, 430, 450-458, 461, and 532 of PDB ID:2VBG.

[0007] In some embodiments, the enzyme comprising 2-ketoacid decarboxylase activity comprises an amino acid sequence that contains a mutation of a native sequence wherein the mutation is in an amino acid residue within a 2-ketoacid decarboxylase active site or is in an amino acid residue within 8 Å (*e.g.*, having a non-hydrogen atom within about 1, 1.5, 2, 2.5, 3, 3.5, 4, 4.5, 5, 5.5, 6, 6.5, 7, or 7.5 Å) of a 2-ketoacid decarboxylase active site, wherein the 2-ketoacid decarboxylase active site comprises the amino acids corresponding to amino acid positions 23-27, 50, 51, 56, 71, 75, 76, 79, 82, 112, 114, 287, 356, 373-377, 379, 394, 396-399, 423-428, 430, 450-458, 461, and 535 of SEQ ID NO:1.

[0008] In some embodiments, the enzyme comprising 2-ketoacid decarboxylase activity comprises an amino acid sequence that is at least 90%, 95%, or 99% identical to SEQ ID NO:1 (*e.g.*, from 90% to about 99% or more identical to SEQ ID NO:1, or from 95% to about 99% or more identical to SEQ ID NO:1). In some embodiments, the enzyme comprising 2-ketoacid decarboxylase activity comprises an amino acid sequence that contains 1, 2, 3, 4, or 5 mutations (*e.g.*, substitutions) as compared to SEQ ID NO:1, 2, or 3.

[0009] In some embodiments, the enzyme comprising 2-ketoacid decarboxylase activity comprises an amino acid sequence that contains 1, 2, or 3 mutations (*e.g.*, substitutions) in amino

acid residues corresponding to a residue selected from the group consisting of G402, M538, and F542 of SEQ ID NO:1.

[0010] In some embodiments, the mutation in an amino acid residue corresponding to a residue selected from the group consisting of G402, M538, and F542 of SEQ ID NO:1 is from a (*e.g.*, native) sequence to a residue selected from the group consisting of Val, Leu, Ile, Met, Phe, His, Gly, Ala, Thr, Tyr, and Trp.

[0011] In some embodiments, the mutation corresponds to, or is, a mutation selected from the group consisting of G402V, M538L, and F542V of SEQ ID NO:1.

[0012] In some embodiments, the enzyme comprising 2-ketoacid decarboxylase activity comprises two mutations (*e.g.*, substitutions) in amino acid residues corresponding to residues selected from the group consisting of G402, M538, and F542 of SEQ ID NO:1. In some embodiments, the enzyme comprising 2-ketoacid decarboxylase activity comprises mutations (*e.g.*, substitutions) corresponding to G402V, M538L, and F542V of SEQ ID NO:1 or comprises SEQ ID NO:3.

[0013] In some embodiments, the enzyme comprising 2-ketoacid decarboxylase activity comprises an amino acid sequence at least 90% or 99% identical (*e.g.*, from about 90% to about 99%), or identical, to SEQ ID NO:2.

[0014] In some embodiments, the enzyme comprising 2-ketoacid decarboxylase activity comprises an amino acid sequence that contains a mutation, wherein the mutation is in an amino acid residue of a 2-ketoacid decarboxylase active site residue of SEQ ID NO:2 or in an amino acid residue having a C α within 8 Å (*e.g.*, having a non-hydrogen atom within about 1, 1.5, 2, 2.5, 3, 3.5, 4, 4.5, 5, 5.5, 6, 6.5, 7, or 7.5 Å) of the 2-ketoacid decarboxylase active site, wherein the 2-ketoacid decarboxylase active site comprises the amino acids corresponding to amino acid positions 23-27, 48, 49, 54, 71, 74, 75, 78, 81, 112, 113, 286, 356, 376-380, 382, 399, 401-404, 427-432, 434, 454-462, 465, and 535 of SEQ ID NO:2.

[0015] In some embodiments, the enzyme comprising 2-ketoacid decarboxylase activity interacts with a C8 2-ketoacid substrate with a calculated interaction energy of less than -5, -5.5, -6, -6.5, -7, -7.5, -8, -8.5, -9, -9.5, -10, -10.3, -10.4, -10.5, -10.6, -10.7, -10.8, -10.9, -11, -11.5, -12, -12.5, -13, -13.5, -14, -14.5, -15, -15.5, -16, -16.5, -17, -17.5, -18, -18.5, -19, -19.5, or -20

rosetta energy units. In some embodiments, the enzyme comprising 2-ketoacid decarboxylase activity interacts with a C8 2-ketoacid substrate with a calculated interaction energy of from about -20 to about -5 rosetta energy units, from about -15 to about -8 rosetta energy units; or from about -11 to about -9 rosetta energy units.

[0016] In another aspect, the present invention provides a modified enzyme comprising 2-ketoacid decarboxylase activity, wherein the enzyme comprises: a 3-layer alpha-beta-alpha sandwich; thiamine pyrophosphate; and at least a 10-fold, 25-fold, 50-fold, or 100-fold (*e.g.*, from about 10-fold to about 100,000-fold; from about 10-fold to about 10,000 fold; from about 10-fold to about 200-fold; from about 10-fold to about 100-fold; from about 25-fold to about 100,000-fold; from about 25-fold to about 10,000 fold; from about 25-fold to about 200-fold; from about 25-fold to about 100-fold; from about 50-fold to about 100,000-fold; from about 50-fold to about 10,000 fold; from about 50-fold to about 200-fold; or from about 50-fold to about 100-fold) greater catalytic efficiency for C8 2-ketoacid substrates as compared to C3 or isoC5 2-ketoacid substrates, wherein the catalytic efficiency is measured by k_{cat}/K_m , wherein the modification comprises a mutation (*e.g.*, substitution) in the primary amino acid sequence of SEQ ID NO:1 or SEQ ID NO:2.

[0017] In some embodiments, the enzyme has a TM align score against PDB ID:2VBG of at least 0.5 (*e.g.*, from about 0.5 to about 0.8, from about 0.5 to about 0.9, or from about 0.5 to about 1.0), 0.6 (*e.g.*, from about 0.6 to about 0.8, from about 0.6 to about 0.9, or from about 0.6 to about 1.0), 0.7 (*e.g.*, from about 0.7 to about 0.8, from about 0.7 to about 0.9, or from about 0.7 to about 1.0), 0.8 (*e.g.*, from about 0.8 to about 0.9, or from about 0.8 to about 1.0), or 0.9 (*e.g.*, from about 0.9 to about 1.0). In some embodiments, the enzyme has a C α root mean squared deviation of less than 1 Å (*e.g.*, from about 0.1 Å to about 1 Å, from about 0.2 Å to about 1 Å, from about 0.3 Å to about 1 Å, from about 0.4 Å to about 1 Å, or from about 0.5 Å to about 1 Å) over at least 100 (*e.g.*, from about 100 to about 200 or more) structurally aligned residues as compared against PDB ID:2VBG.

[0018] In some embodiments, the modification comprises a mutation (*e.g.*, substitution) in an amino acid residue, wherein the amino acid residue is within a 2-ketoacid decarboxylase active site or within 8 Å (*e.g.*, having a C α atom within about 1, 1.5, 2, 2.5, 3, 3.5, 4, 4.5, 5, 5.5, 6, 6.5, 7, or 7.5 Å) of a 2-ketoacid decarboxylase active site, wherein the 2-ketoacid decarboxylase

active site comprises the amino acids corresponding to amino acid positions 23-27, 50, 51, 56, 71, 75, 76, 79, 82, 112, 114, 287, 356, 373-377, 379, 394, 396-399, 423-428, 430, 450-458, 461, and 535 of SEQ ID NO:1.

[0019] In some embodiments, the modification comprises a mutation (*e.g.*, substitution) in an amino acid residue corresponding to a residue selected from the group consisting of G402, M538, and F542 of SEQ ID NO:1. In some embodiments, the mutation (*e.g.*, substitution) in an amino acid residue corresponding to a residue selected from the group consisting of G402, M538, and F542 of SEQ ID NO:1 is a mutation from a (*e.g.*, native) sequence to a residue selected from the group consisting of Val, Leu, Ile, Met, Phe, His, Gly, Ala, Thr, Tyr, and Trp.

[0020] In some embodiments, the modification comprises two mutations (*e.g.*, substitutions) in amino acid residues corresponding to residues selected from the group consisting of G402, M538, and F542 of SEQ ID NO:1. In some embodiments, the modification comprises one or two mutations (*e.g.*, substitutions) corresponding to the mutations selected from the group consisting of G402V, M538L, and F542V of SEQ ID NO:1. In some embodiments, the modification comprises mutations (*e.g.*, substitutions) corresponding to G402V, M538L, and F542V of SEQ ID NO:1. In some embodiments, the ratio of k_{cat}/K_m against C8 2-ketoacid substrates over the k_{cat}/K_m against isoC5 2-ketoacid substrates is at least 10-fold, 25-fold, 50-fold, or 100-fold (*e.g.*, from about 10-fold to about 100,000-fold; from about 10-fold to about 10,000 fold; from about 10-fold to about 200-fold; from about 10-fold to about 100-fold; from about 10-fold to about 50-fold; from about 25-fold to about 100,000-fold; from about 25-fold to about 10,000 fold; from about 25-fold to about 200-fold; from about 25-fold to about 100-fold; from about 25-fold to about 50-fold; from about 50-fold to about 100,000-fold; from about 50-fold to about 10,000 fold; from about 50-fold to about 200-fold; or from about 50-fold to about 100-fold) greater than the native enzyme. In some embodiments, the solvent accessible surface area of the substrate binding pocket is, is about, is at least, or is at least about 5,000; 5,500; 6,000; 6,500; 7,000; 7,500; 8,000; 8,500; 9,000; 9,500; or 10,000 Å² (*e.g.*, from about 5,000 to about 10,000 Å²; from about 5,500 to about 9,500 Å²; from about 6,000 to about 9,000 Å²; from about 6,500 to about 8,500 Å²; from about 7,000 to about 8,000 Å²; or from about 7,500 to about 8,000 Å²).

[0021] In another aspect, the present invention provides a host cell comprising any of the foregoing expression cassettes or modified enzymes, wherein the host cell produces long chain

($\geq C5$) alcohols at a concentration of at least 5 mg/L, 25 mg/L, 50 mg/L, 100 mg/L, 200 mg/L, 250 mg/L, 300 mg/L, 400 mg/L, or 500 mg/L (*e.g.*, from about 5 mg/L to about 300, 400, 500, 600, 700, 800, or more mg/L, from about 10 mg/L to about 500 mg/L, from about 10 mg/L to about 400 mg/L, from about 10 mg/L to about 300 mg/L, from about 25 mg/L to about 500 mg/L, from about 25 mg/L to about 400 mg/L, from about 25 mg/L to about 300 mg/L, from about 50 mg/L to about 500 mg/L, from about 50 mg/L to about 400 mg/L, from about 50 mg/L to about 300 mg/L; from about 100 mg/L to about 500 mg/L, from about 100 mg/L to about 400 mg/L, or from about 100 mg/L to about 300 mg/L).

[0022] In some embodiments, the host cell produces heptanol as the major alcohol product or second-most major alcohol product. In some embodiments, the host cell produces hexanol as the major alcohol product or second-most major alcohol product. In some embodiments, the host cell produces at least 200 mg/L (*e.g.*, from about 200 to about 500, from about 200 to about 400, from about 250 to about 450, from about 250 to about 350, or from about 300 to about 350 mg/L) heptanol or hexanol, or a combination thereof. In some embodiments, the host cell produces octanol. In some embodiments, the host cell does not produce ethanol or propanol, or produces at least 10-fold (*e.g.*, from 10-fold to 100-fold, or more, or from 10-fold to 50-fold, or from 10-fold to 20-fold) more long chain alcohols ($\geq C5$) as compared to ethanol and/or propanol.

BRIEF DESCRIPTION OF THE DRAWINGS

[0023] **Figure 1:** Depicts an illustration of the synthetic recursive +1 pathway. This pathway employs enzymes LeuABCD from leucine biosynthesis for carbon chain elongation of 2-ketoacids. The enzyme ketoisovalerate decarboxylase (KIVD) performs the decarboxylation of 2-ketoacids and diverts carbon out from the +1 iterative cycle.

[0024] **Figure 2:** Depicts an illustration of a computationally directed genomic enzyme mining pipeline. (Left) GEOs are identified based on sequence homology to KIVD. Bioinformatics filters are used to identify a set of protein sequences from genomic databases that are predicted to be KIVD homologs and likely to be decarboxylases. Once identified, homology models are built to obtain a predicted ternary structure of each GEO. Ligand docking and design simulations are subsequently run in the presence of the target ligand (*e.g.*, C8) to evaluate the potential protein-

ligand interface energy. (Middle) A phylogenetic tree for the 239 GEOs is depicted with a bar chart above each sequence. The bar height indicates the predicted protein-ligand interface energy, the higher the bar the lower the energy. Bar height is scaled linearly relative to the lowest protein-ligand interface energy. 10 GEOs are selected for experimental characterization. (Right) The pairwise sequence identity of all 239 GEOs to each other has a mode of 20%, indicating the high level of sequence diversity within this set of sequences.

[0025] **Figure 3:** Depicts an illustration of reaction specificities for various 2 ketoacid decarboxylases. (a) Ketoisovalerate decarboxylase (KIVD) utilizes various ketoacids as substrates to produce the corresponding aldehyde. (b) Ketoacids used for *in vitro* kinetic constants in addition to 2-ketoisovaleric acid.

[0026] **Figure 4:** Depicts catalytic efficiency and specificity of characterized ketoacid decarboxylases. Specificity factor is calculated as $\frac{\log C8}{\log C5} + \frac{\log C8}{\log C3}$ where the log scaled catalytic efficiencies for C8 relative to C5 and C3 are compared. The three active GEOs and the naively selected set of decarboxylases are represented as dots and diamonds, respectively. Native KIVD and KIVD_VLV are depicted as a square and triangle, respectively. The genomic mining method and computational active site redesign approach both produced enzymes with enhanced specificity relative to native KIVD.

[0027] **Figure 5:** Depicts a molecular model of a C8 intermediate docked into KIVD (a, c) and GEO 175 (b, d). Cross sections of the space filled active site are represented in a and b where the C8 keto acid and TPP ligand are depicted. A detailed view of the amino acids within the active site are illustrated in c and d. Residues within 5 Å of C8 ligand are shown in sticks and their corresponding C-alpha are shown in spheres. Figure was generated using PyMol v1.7.4.

[0028] **Figure 6:** Depicts results of kinetic characterization of GEO 175 and GEO 175 L376T_T240S. Kinetic constants were measured as described herein. The substrate 2-ketooctanoate carbon chain numbering referred to in the text is numbered in white. According to the molecular model of GEO 175, the double mutation L376T_T240S is predicted to recede the pocket by 1.4 Å and remove interactions between the binding pocket and carbon 8 of the ketoacid alkyl chain. This mutant was observed to decrease catalytic efficiency on C8, but has a negligible effect on shorter chain substrates.

[0029] **Figure 7:** Depicts an illustration of a screening process for identifying an engineered KIVD with altered substrate specificity. The design process can start with identifying active site residues of KIVD (depicted as spheres in panel on left). These amino acids were allowed to either remain native or sample any of eleven relatively hydrophobic amino acids. A total of 10,000 design simulations were run and the amino acids identified in the 50 lowest energy designs were used to guide construction of a small library of roughly 400 KIVD mutants from the original combinatorial space of 1011 possible active site mutations. 1200 clonal isolates were screened for activity and specificity. The KIVD_VLV mutant was selected and subsequently purified for in vitro kinetic constant characterization.

[0030] **Figure 8:** Depicts results of *in vivo* alcohol production of the synthetic recursive +1 pathway with native KIVD, KIVD_VLV, and GEO 175. Cells were incubated for forty hours in microaerobic conditions in a defined media as described in materials and methods. Each assay was performed in triplicate and titers reported if all three samples had observed product production above the limit of quantitation (5 mg/L).

[0031] **Figure 9:** Depicts toxicity as a function of alcohol chain length. The optical density of engineered *E. coli* strains after a 7-hour growth was measured as a function of the alcohol concentration added to growth media at the beginning of incubation.

[0032] **Figure 10:** Depicts an overlay of active sites between GEO 175 and BFD. Residues that are different between the proteins are shown in sticks and their corresponding alpha carbons are shown in spheres.

[0033] **Figure 11:** Depicts a screening process for identifying engineered KIVD with altered substrate specificity. Pertinent screening results and candidate mutants are shown (graphs) which contributed directly to finding KIVD_VLV. Colorimetric assays (graphs) measured approximate enzyme activity of screened mutants for ketoisovalerate (IsoC5), 2-ketobutyrate (C4), 2-ketovalerate (C5), 2-ketohexanoate (C6), and 2-ketooctanoate (C8). Higher activity is shown by larger downwards slope; greater specificity relative to KIVD was observed in mutants with reduced activity for IsoC5, C4, and C5, and only slightly affected or greater activity for C6 and C8. First round high-throughput colorimetric screening of single mutants identified candidates such as G402V and M538L/F542V. Mutations from each region that showed potential in altering the specificity of KIVD were combined to produce multi-region mutants

through site-directed mutagenesis using the plasmids of KIVD single region mutants as templates. These KIVD multi-region mutants, such as G402V M538L F542V (KIVD__VLV) were then screened a second time with the colorimetric assays. Enzyme kinetics was then measured for mutant KIVD__VLV.

[0034] **Figure 12:** Depicts results of Michaelis-Menten curve fitting of tested enzymes with C8 substrate.

[0035] **Figure 13:** Depicts results of Michaelis-Menton curve fitting of tested enzymes with C5 substrate.

[0036] **Figure 14:** Depicts results of Michaelis-Menten curve fitting of tested enzymes with C3 substrate.

[0037] **Figure 15:** Depicts results of Michaelis-Menten curve fitting of tested enzymes with isoC5 substrate.

[0038] **Figure 16:** Depicts pdb codes and accession numbers of GEOs modeled as listed on the phylogenetic tree depicted in Figure 2. Sequences were obtained from homologous sequence search in the non-redundant database using HMMER3.

[0039] **Figure 17:** Depicts protein sequences of various 2-ketoacid decarboxylases.

[0040] **Figure 18:** Depicts the native sequence and various engineered substitutions at various amino acid positions of KIVD used to probe for altered specificity for long chain 2-ketoacids. Amino acids in parentheses denote potential mutations due to degenerate codons used for mutagenesis.

[0041] **Figure 19:** Depicts activity data for various GEOs tested against C3, C5, C8, and isoC5 substrates and calculated Rosetta energy values for bound TPP-C8 ligand.

[0042] **Figure 20:** Depicts results of a structural alignment of GEOs and KIVD. Active site residues are denoted with a black box under the residue letter.

[0043] **Figure 21:** Depicts a profile of active site residues and preferred amino acid substitutions at such residues for 2-ketoacid decarboxylases described herein. A 5 Å cut-off was utilized to define active-site residues. The active site residues are defined as those protein

residues found in the model to have any non-hydrogen atoms within 5 Å of any of the atoms in the TPP-C8 ligand. The model of KIVD_VLV was generated by threading the KIVD sequence onto the crystal structure of 2VBG (88% identical in sequence to KIVD), making the three mutations, and subsequently docking in the model of the C8 substrate.

[0044] **Figure 22:** Depicts a profile of active site residues and preferred amino acid substitutions at such residues for 2-ketoacid decarboxylases described herein. An 8 Å cut-off was utilized to define active-site residues. The active site residues are defined as those protein residues found in the model to have any non-hydrogen atoms within 8 Å of any of the atoms in the TPP-C8 ligand. The model of KIVD_VLV was generated by threading the KIVD sequence onto the crystal structure of 2VBG (88% identical in sequence to KIVD), making the three mutations, and subsequently docking in the model of the C8 substrate.

[0045] **Figure 23:** Depicts pairwise identities for 2-ketoacid decarboxylases described herein.

[0046] **Figure 24:** Depicts the contents of a sequence alignment file in aln format.

DEFINITIONS

[0047] As used in this specification and the appended claims, the singular forms "a," "an," and "the" include plural reference unless the context clearly dictates otherwise.

[0048] The term "nucleic acid" or "polynucleotide" refers to deoxyribonucleic acids (DNA) or ribonucleic acids (RNA) and polymers thereof in either single- or double-stranded form. Unless specifically limited, the term encompasses nucleic acids containing known analogues of natural nucleotides that have similar binding properties as the reference nucleic acid and are metabolized in a manner similar to naturally occurring nucleotides. Unless otherwise indicated, a particular nucleic acid sequence also implicitly encompasses conservatively modified variants thereof (*e.g.*, degenerate codon substitutions), alleles, orthologs, SNPs, and complementary sequences as well as the sequence explicitly indicated. Specifically, degenerate codon substitutions may be achieved by generating sequences in which the third position of one or more selected (or all) codons is substituted with mixed-base and/or deoxyinosine residues (Batzner *et al.*, *Nucleic Acid Res.* **19**:5081 (1991); Ohtsuka *et al.*, *J. Biol. Chem.* **260**:2605-2608

(1985); and Rossolini *et al.*, *Mol. Cell. Probes* 8:91-98 (1994)). The term nucleic acid is used interchangeably with gene, cDNA, and mRNA encoded by a gene.

[0049] The term “**gene**” means the segment of DNA involved in producing a polypeptide chain. It may include regions preceding and following the coding region (leader and trailer) as well as intervening sequences (introns) between individual coding segments (exons).

[0050] A “**promoter**” is defined as an array of nucleic acid control sequences that direct transcription of a nucleic acid. As used herein, a promoter includes necessary nucleic acid sequences near the start site of transcription, such as, in the case of a polymerase II type promoter, a TATA element. A promoter also optionally includes distal enhancer or repressor elements, which can be located as much as several thousand base pairs from the start site of transcription.

[0051] An “**expression cassette**” is a nucleic acid construct, generated recombinantly or synthetically, with a series of specified nucleic acid elements that permit transcription of a particular polynucleotide sequence in a host cell. An expression cassette may be part of a plasmid, viral genome, or nucleic acid fragment. Typically, an expression cassette includes a polynucleotide to be transcribed, operably linked to a promoter. In some cases, the promoter is heterologous to the polynucleotide to be transcribed. In some cases, the expression cassette is heterologous to the host cell in which it resides.

[0052] The term “**amino acid**” refers to naturally occurring and synthetic amino acids, as well as amino acid analogs and amino acid mimetics that function in a manner similar to the naturally occurring amino acids. Naturally occurring amino acids are those encoded by the genetic code, as well as those amino acids that are later modified, *e.g.*, hydroxyproline, γ -carboxyglutamate, and O-phosphoserine. Amino acid analogs refers to compounds that have the same basic chemical structure as a naturally occurring amino acid, *i.e.*, an α carbon that is bound to a hydrogen, a carboxyl group, an amino group, and an R group, *e.g.*, homoserine, norleucine, methionine sulfoxide, methionine methyl sulfonium. Such analogs have modified R groups (*e.g.*, norleucine) or modified peptide backbones, but retain the same basic chemical structure as a naturally occurring amino acid. “Amino acid mimetics” refers to chemical compounds having a structure that is different from the general chemical structure of an amino acid, but that functions in a manner similar to a naturally occurring amino acid.

[0053] There are various known methods in the art that permit the incorporation of an unnatural amino acid derivative or analog into a polypeptide chain in a site-specific manner, *see, e.g.*, WO 02/086075.

[0054] Amino acids may be referred to herein by either the commonly known three letter symbols or by the one-letter symbols recommended by the IUPAC-IUB Biochemical Nomenclature Commission. Nucleotides, likewise, may be referred to by their commonly accepted single-letter codes.

[0055] “**Polypeptide**,” “**peptide**,” and “**protein**” are used interchangeably herein to refer to a polymer of amino acid residues. All three terms apply to amino acid polymers in which one or more amino acid residue is an artificial chemical mimetic of a corresponding naturally occurring amino acid, as well as to naturally occurring amino acid polymers and non-naturally occurring amino acid polymers. As used herein, the terms encompass amino acid chains of any length, including full-length proteins, wherein the amino acid residues are linked by covalent peptide bonds.

[0056] The term “**heterologous**” as used in the context of describing the relative location of two elements, refers to the two elements such as nucleic acids (*e.g.*, promoter and protein encoding sequence), a nucleic acid and a host cell, a protein (*e.g.*, a 2-ketoacid decarboxylase, or portion thereof) and a host cell or reaction mixture, or a protein and a ligand that are not naturally found in the same relative positions. Thus, a “heterologous promoter” of a gene refers to a promoter that is not naturally operably linked to that gene. Similarly a host cell containing a heterologous expression cassette can refer to an expression cassette that is not naturally found in that host cell.

[0057] A “**3-layer alpha-beta-alpha sandwich**” refers to the 3-layer (aba) Sandwich architecture according to the CATH classification (version 4.0.0). Proteins containing such architecture have a Rossmann fold topology. For enzymes that have not been characterized by X-ray crystallography, NMR spectroscopy, or other experimental means, the presence or absence of a 2-layer alpha beta alpha sandwich can be assessed by homology modeling. Exemplary methods of homology modeling can include the use of one or more of the following programs, program codes, scripts, and/or parameters, *e.g.* in their default settings,: RosettaCM, SWISS-

MODEL, MODELLER, 3D-JIGSAW, ModPipe, Fugue, HHPred, i-Tasser, IntFOLD, M4T, ModWeb, Phyre2, RaptorX, Robetta, HHsuite, and the like, such as derivative works thereof.

[0058] A “**thiamine pyrophosphate (TPP) binding pocket**” refers to a binding pocket of an enzyme that productively binds TPP, such that the TPP bound by the enzyme can perform a ketoacid decarboxylation reaction. The presence of a TPP binding pocket can be detected by detecting bound TPP, or a radical intermediate thereof (*e.g.*, via mass spectrometry, electron paramagnetic resonance (EPR) spectroscopy, or the like). The presence of a TPP binding pocket can also be inferred on the basis of detection of TPP mediated catalysis by the enzyme. In some cases, the presence of a TPP binding pocket can be determined from an experimentally determined structure of the enzyme (*e.g.*, determined by X-ray crystallography). In some cases, the presence of a TPP binding pocket can be determined by homology modeling using any of the homology modeling methods described herein. In some cases, the presence of a TPP binding pocket can be inferred from the presence of sequence data indicative of a TPP binding domain. For example, a comparison of a candidate sequence to the Pfam database (*e.g.*, Pfam 28.0) can reveal the presence or absence of a TPP binding pocket.

[0059] A “TM align score” refers to a normalized structural similarity score comparing two 3-dimensional protein structures (*e.g.*, experimentally determined or homology model structures) calculated by version 2012/05/07 of the TM-align algorithm (TMalign.f). *See*, Y. Zhang & J. Skolnick, *Nucleic Acids Research*, 33: 2302-2309 (2005); Y. Zhang & J. Skolnick, *Proteins*, 57: 702-710 (2004); and J. Xu & Y. Zhang, *Bioinformatics*, 26, 889-895 (2010). Where one or both structures contain multiple chains of substantially the same structure (*e.g.*, due to non-crystallographic symmetry or due to the availability of an ensemble of homology models), the TM align score refers to the lowest score between the one or more chains of a first structure and the one or more chains of a second structure.

[0060] “Solvent accessible surface area” or “SASA” in reference to a substrate binding pocket of an enzyme described or claimed herein refers to a SASA calculated using RosettaScripts with Rosetta version c2bf8f674e7b416b5b756630d7ccd5d64c57512c 2015-05-04 14:47:46 -0700 from [git@github.com:RosettaCommons/main.git](https://github.com/RosettaCommons/main.git). The SASA can be calculated from a structure file (PDB) with a bound ligand. The structure file can be an experimentally determined structure

or a structure determined by any one or more of the homology modeling methods described herein.

[0061] As used herein, “2-ketoacid active site,” in the context of the claims can refer to amino acid residues corresponding to amino acid positions: (i) 23-27, 50, 51, 56, 75, 76, 79, 82, 114, 287, 373-377, 379, 394, 396-399, 423-428, 430, 450-458, and 461 of SEQ ID NO:1; (ii) 23-27, 50, 51, 56, 72, 75, 76, 79, 82, 113, 114, 287, 352, 373-377, 379, 394, 396-399, 423-428, 430, 450-458, 461, and 532 of SEQ ID NO:1; (iii) 23-27, 48, 49, 54, 74, 75, 78, 81, 113, 286, 376-380, 382, 399, 401-404, 427-432, 434, 454-462, and 465 of SEQ ID NO:2; (iv) 23-27, 48, 49, 54, 71, 74, 75, 78, 81, 112, 113, 286, 356, 376-380, 382, 399, 401-404, 427-432, 434, 454-462, 465, and 535 of SEQ ID NO:2; or (v) 22-29, 32, 43, 46-50, 54, 69-79, 81, 82, 111-113, 240, 285-287, 290, 356, 360, 376-383, 398-407, 426-435, 453-466, 472, 473, 475, 477, 526-528, and 530-535 of SEQ ID NO:2.

[0062] Amino acid positions corresponding to any one or more of the foregoing residues can be determined by structural alignment to a template structure (*e.g.*, a structure of SEQ ID NO:1 or 2). One or both structures used in the structural alignment can be experimentally determined or determined by homology modeling using any one or more of the homology modelling methods described herein. Exemplary structural alignment methods and/or programs include, but are not limited to, TM-align, LSQMAN, Fr-TM-align, DALI, DaliLite, CE, CE-MC, and the like. Additional structural alignment tools useful in determining active site amino acids include, but are not limited to, those described in web.archive.org/web/20151221055428/https://en.wikipedia.org/wiki/Structural_alignment_software.

[0063] As used herein, “rosetta energy units” in the context of ligand binding interaction energy refers to an interaction energy between a bound ligand and a protein calculated using Rosetta Enzyme Design style constraints. Program code for calculating rosetta energy units is provided herein.

DETAILED DESCRIPTION OF THE INVENTION

I. Introduction

[0064] In the current synthetic recursive +1 pathway, the enzymes LeuA, LeuB, LeuC, and LeuD (collectively known as LeuABCD) are recruited to recursively elongate 2-ketobutyrate into longer chain 2-ketoacids.^{3, 4} These 2-ketoacids are then converted to aldehydes by ketoisovalerate decarboxylase⁶ (KIVD) and subsequently to alcohols by alcohol dehydrogenase³ (ADH). (Figure. 1) In this engineered pathway, the primary determinant for the final product chain length is the number of cycles through LeuABCD.⁷ (Figure 1)

[0065] Previous efforts to engineer the product specificity of this pathway have focused on broadening the specificity of LeuA and KIVD. When the specificity of LeuA and KIVD are broadened, branched C₈ products are produced.³ When only the specificity of LeuA is broadened, additional cycles through the pathway are carried out, resulting in a mixture of C₂-C₈ linear alcohols;⁴ however, the primary product is still 1-butanol. In order to enhance long chain alcohol production, the inventors have focused on discovering a ketoacid decarboxylase specific for long chain ketoacids. By replacing KIVD with an enzyme specific for long chain ketoacids, the short chain substrates are able to reenter the +1 iteration cycle until conversion to long chain ketoacids, which are then decarboxylated and reduced into the corresponding long chain alcohols.

II. Methods

[0066] Provided herein are methods of identifying 2-ketoacid decarboxylase enzymes having an altered specificity, or having a desired endogenous specificity. In some cases, the 2-ketoacid decarboxylases have or are altered to have improved specificity for or catalytic efficiency against long chain ($\geq$ C5) ketoacids and/or provide or are altered to provide an increased amount or concentration of long chain ($\geq$ C5) alcohols when present in a synthetic recursive +1 pathway in a host cell or reaction mixture, *e.g.*, as compared to SEQ ID NO: 1.

[0067] The methods can include mining known protein sequences to identify enzymes having homology to a known 2-ketoacid decarboxylase template, at least a subset of which can be predicted to have a desired endogenous specificity or can be predicted to be able to be redesigned to have an altered specificity. Protein sequences having homology to a template enzyme

sequence (*e.g.*, SEQ ID NO:1) can, *e.g.*, be identified using a hidden markov model algorithm such as that implemented in HMMER3 or using a BLAST-type sequence similarity search. Sequences can be filtered by removing sequences that are highly similar in sequence, *e.g.*, using CD-HIT. Exemplary cut-offs for highly similar sequences can include a cut-off which removes sequences having at least about 50%, 60%, 70%, 80%, 90%, 95%, or 99% identity to the template sequence.

[0068] Structural features of sequences identified as having sequence homology to the template sequence can then be assessed. For example, a homology model of candidate hits can be generated, *e.g.*, using RosettaCM, SWISS-MODEL, MODELLER, 3D-JIGSAW, ModPipe, Fugue, HHPred, i-Tasser, IntFOLD, M4T, ModWeb, Phyre2, RaptorX, Robetta, HHSuite, and the like. Homology models so generated can be scored for three dimensional structural similarity to the structure of the template sequence. For example, a structural alignment program or methods as described herein, including but not limited to one or more of TM align, the DALI server (*e.g.*, DaliLite v. 3 available at ekhidna.biocenter.helsinki.fi/dali_server), or the cealign algorithm implemented in PyMOL (*e.g.*, version 1.7.6 available at www.pymol.org) can be used to assess structural similarity by calculating a TM align score, Z-score, or an RMSD value respectively.

[0069] In some embodiments, a TM align score is used to assess structural similarity. For example, a TM align score cut-off of at least about 0.5 to a known 2-ketoacid decarboxylase (*e.g.*, SEQ ID NO:1, SEQ ID NO:2 or any other 2-ketoacid decarboxylases described herein) can be used to filter out candidate 2-ketoacid decarboxylases from homology models of the initial sequence search results or from homology models of the sequence search results after removing highly similar sequences. Alternatively, a more stringent cut-off can be applied. For example, a TM align score of at least about 0.55, 0.6, 0.65, 0.7, 0.75, 0.8, 0.81, 0.815, 0.817, 0.85, or 0.9 can be utilized.

[0070] In some embodiments, an RMSD value as calculated by, *e.g.*, cealign or other similar methods or programs can be utilized to assess structural similarity and filter out candidate 2-ketoacid decarboxylases from homology models of the initial sequence search results or from homology models of the sequence search results after removing highly similar sequences. In exemplary embodiments, RMSD cutoffs calculated as described herein of less than about 6.7 Å,

6 Å, 5.5 Å, 5 Å, 4 Å, or 3.5 Å can be used to filter out candidate 2-ketoacid decarboxylases from homology models of the initial sequence search results or from homology models of the sequence search results after removing highly similar sequences. Alternatively, a more stringent cut-off can be applied. For example, an RMSD value of less than about 3 Å, 2.5 Å, 2 Å, or 1.6 Å can be utilized.

[0071] In some embodiments, a Z-score calculated by the DALI server is used to assess structural similarity. For example, a Z-score cut-off of at least about 2 to a known 2-ketoacid decarboxylase (*e.g.*, SEQ ID NO:1, SEQ ID NO:2 or any other 2-ketoacid decarboxylases described herein) can be used to filter out candidate 2-ketoacid decarboxylases from homology models of the initial sequence search results or from homology models of the sequence search results after removing highly similar sequences. Alternatively, a more stringent cut-off can be applied. For example, a Z-score of at least about 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, or 62 can be utilized.

[0072] The homology models (*e.g.*, after removing models that lack significant structural similarity as assessed by, *e.g.*, TM align score, Z-score, and/or RMSD) can be assessed to identify sequences that bind a preferred substrate. For example, a covalent intermediate of a 2-ketoacid decarboxylase reaction can be modeled in the active site of the generated homology models, the complex can be relaxed computationally, *e.g.*, using Rosetta Enzyme Design, and a ligand binding energy calculated. Candidate sequences having a low ligand binding energy can be selected for *in vitro* and/or *in vivo* analysis or selected as a preferred enzyme for catalyzing decarboxylation of a substrate to produce a product that corresponds to the modeled catalytic intermediate ligand. In some embodiments, candidate sequences having a ligand binding energy of less than, or less than about, -5, -5.5, -6, -6.5, -7, -7.5, -8, -8.5, -9, -9.5, -10, -10.3, -10.4, -10.5, -10.6, -10.7, -10.8, -10.9, -11, -11.5, -12, -12.5, -13, -13.5, -14, -14.5, -15, -15.5, -16, -16.5, -17, -17.5, -18, -18.5, -19, -19.5, or -20 Rosetta energy units can be selected for *in vitro* and/or *in vivo* analysis or selected as a preferred enzyme for catalyzing decarboxylation of a substrate to produce a product that corresponds to the modeled catalytic intermediate ligand.

[0073] Methods of identifying 2-ketoacid decarboxylase enzymes having a desired altered specificity (*e.g.*, greater catalytic efficiency against long chain ($\geq C5$) ketoacids) can additionally

or alternatively include reprogramming substrate specificity of a known 2-ketoacid decarboxylase (*e.g.*, SEQ ID NO:1 or any of the GEOs described herein) using computation enzyme design methods. For example, the Rosetta Molecular Modeling Suite, or another computational design platform, can be used to design such enzymes. For example, experimentally determined structures and/or homology models of known or suspected 2-ketoacid decarboxylases can be used to explore potential active-site sequence space that would accommodate long chain ($\geq C5$) 2-ketoacids, *e.g.*, 2-ketooctanoate as substrates. In some embodiments, a modeled reaction intermediate of, *e.g.*, a C8 substrate, is docked within a predicted active site of an experimentally determined structure or homology model structure, the complex optionally minimized, and a ligand binding energy calculated. In some embodiments, ligand binding energies are sampled among a number of computationally generated enzyme mutations or ensemble conformations by allowing the identity and conformation of non-catalytic residues within the active site to change, and/or allowing backbone distances and/or angles to move during minimization.

[0074] Candidate 2-ketoacid decarboxylases can be cloned into an expression cassette, *e.g.*, under the control of a promoter such as a heterologous promoter, expressed and assayed for activity. For example 2-ketoacid decarboxylase enzymes identified by one or more of sequence mining, homology modeling, structural similarity scoring, and ligand binding analysis can be assayed for activity against one or more 2-ketoacid substrates. As another example, 2-ketoacid decarboxylase enzymes identified by virtual screening of mutants of a known 2-ketoacid decarboxylase can be assayed for activity against one or more 2-ketoacid substrates.

II. Compositions

[0075] Described herein are 2-ketoacid decarboxylases, expression cassettes encoding such 2-ketoacid decarboxylases, and host-cells and/or reaction mixtures containing such 2-ketoacid decarboxylases or expression cassettes. In some embodiments, the expression cassette is heterologous to the host cell or reaction mixture in which it resides. In some embodiments, the expression cassette contains a heterologous promoter operably linked to a polynucleotide encoding a 2-ketoacid decarboxylase. The expression cassette can, *e.g.*, comprise a heterologous promoter operably linked to a nucleic acid encoding an enzyme comprising 2-ketoacid decarboxylase activity, wherein the enzyme: comprises a 3-layer alpha-beta-alpha sandwich;

comprises a thiamine pyrophosphate (TPP) binding pocket; has a TM align score of at least 0.5 as compared against PDB ID:2VBG; and comprises at least a 100-fold greater catalytic efficiency for C8 2-ketoacid substrates as compared to C3 or isoC5 2-ketoacid substrates, wherein the catalytic efficiency is measured by kcat/Km.

[0076] In some cases, the enzyme comprising 2-ketoacid decarboxylase activity comprises an amino acid sequence that is at least 10%, 12%, 14%, 15%, 20%, 25%, 30%, 35%, 40%, 50%, 60%, 70%, 80%, 90% or 99% identical to SEQ ID NO:1. Sequence identity can be determined using methods known in the art. For example, using BLAST, HMMER3, or the like, *e.g.*, using the default settings.

[0077] In some cases, the enzyme comprising 2-ketoacid decarboxylase activity has a TM align score of at least 0.4, 0.45, 0.5, 0.55, 0.6, 0.65, 0.7, 0.75, 0.8, 0.81, 0.815, 0.817, 0.85, or 0.9; a Z-score of or of at least about 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, or 62; or an RMSD of or of less than about 6.7 Å, 6 Å, 5.5 Å, 5 Å, 4 Å, 3.5 Å, 3 Å, 2.5 Å, 2 Å, or 1.6 Å as compared against PDB ID:2VBG. In some cases, the substrate binding pocket of the enzyme has a solvent accessible surface area of, or of at least about 7,500; 8,000; 8,500; 9,000; 9,500; or 10,000 Å².

[0078] In some cases, the enzyme comprising 2-ketoacid decarboxylase activity comprises an amino acid sequence that contains a mutation of a native sequence wherein the mutation is in an amino acid residue with a Cα within 8 Å of a 2-ketoacid decarboxylase active site, wherein the 2-ketoacid decarboxylase active site comprises the amino acids corresponding to one or more of (*e.g.*, corresponding to 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, or 46 of) the following amino acid positions: (i) 23-27, 50, 51, 56, 75, 76, 79, 82, 114, 287, 373-377, 379, 394, 396-399, 423-428, 430, 450-458, and 461 of SEQ ID NO:1; (ii) 23-27, 50, 51, 56, 72, 75, 76, 79, 82, 113, 114, 287, 352, 373-377, 379, 394, 396-399, 423-428, 430, 450-458, 461, and 532 of SEQ ID NO:1; (iii) 23-27, 48, 49, 54, 74, 75, 78, 81, 113, 286, 376-380, 382, 399, 401-404, 427-432, 434, 454-462, and 465 of SEQ ID NO:2; (iv) 23-27, 48, 49, 54, 71, 74, 75, 78, 81, 112, 113, 286, 356, 376-380, 382, 399, 401-404, 427-432, 434, 454-462, 465, and 535 of SEQ

ID NO:2; or (v) 22-29, 32, 43, 46-50, 54, 69-79, 81, 82, 111-113, 240, 285-287, 290, 356, 360, 376-383, 398-407, 426-435, 453-466, 472, 473, 475, 477, 526-528, and 530-535.

[0079] Amino acid positions corresponding to any one or more of the foregoing residues can be determined by structural alignment to a template structure (*e.g.*, a structure of SEQ ID NO:1 or 2). One or both structures used in the structural alignment can be experimentally determined or determined by homology modeling using any one or more of the homology modelling methods described herein. Exemplary structural alignment methods and/or programs include, but are not limited to, TM-align, LSQMAN, Fr-TM-align, DALI, DaliLite, CE, CE-MC, and the like. Additional structural alignment tools useful in determining active site amino acids include, but are not limited to, those described in en.wikipedia.org/wiki/Structural_alignment_software.

[0080] In some embodiments, the enzyme comprising 2-ketoacid decarboxylase activity comprises an amino acid sequence that contains a mutation in an amino acid residue corresponding to a residue selected from the group consisting of G402, M538, and F542 of SEQ ID NO:1. In some embodiments, the mutation in an amino acid residue corresponding to a residue selected from the group consisting of G402, M538, and F542 of SEQ ID NO:1, and wherein the mutation is to a residue selected from the group consisting of Val, Leu, Ile, Met, Phe, His, Gly, Ala, Thr, Tyr, and Trp. In some embodiments, the mutation in an amino acid residue corresponding to a mutation selected from the group consisting of G402V, M538L, and F542V of SEQ ID NO:1. In some embodiments, the enzyme comprising 2-ketoacid decarboxylase activity comprises two or three mutations in amino acid residues corresponding to residues selected from the group consisting of G402, M538, and F542 of SEQ ID NO:1. In some cases, the enzyme comprising 2-ketoacid decarboxylase activity comprises one, two, or three of the mutations corresponding to G402V, M538L, and F542V of SEQ ID NO:1.

[0081] In some embodiments, the enzyme comprising 2-ketoacid decarboxylase activity comprises SEQ ID NO:2 or SEQ ID NO:3. In some embodiments, the enzyme is, or is at least about, 80%, 90% or 99% identical to SEQ ID NO:2 or 3. In some embodiments, the enzyme comprises 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, or more insertions, substitutions, or deletions, relative to SEQ ID NO:1, 2 or 3, or relative to any of the GEOs described herein. In some cases, the insertion, substitution, or deletion is independently an insertion, substitution, or deletion of 1, 2, 3, or 4 amino acids at a given amino acid position.

[0082] In some embodiments, , the enzyme comprising 2-ketoacid decarboxylase activity interacts with a C8 2-ketoacid substrate with a calculated interaction energy of less than, or less than about, -10, -10.3, -10.4, -10.5, -10.6, -10.7, -10.8, -10.9, -11, -11.5, -12, -12.5, -13, -13.5, -14, -14.5, -15, -15.5, -16, -16.5, -17, -17.5, -18, -18.5, -19, -19.5, or -20 Rosetta energy units.

[0083] All patents, patent applications, and other publications, including GenBank Accession Numbers, cited in this application are incorporated by reference in the entirety for all purposes.

EXAMPLES

[0084] The following examples are provided by way of illustration only and not by way of limitation. Those of skill in the art will readily recognize a variety of non-critical parameters that could be changed or modified to yield essentially the same or similar results.

Example 1:

Introduction

[0085] To rationally engineer the product profile of the synthetic recursive +1 pathway two computationally directed approaches were employed to discover a decarboxylase with the desired functional profile. The first is a new approach to mine the rapidly growing sequence databases derived from genomic sequencing. Since the vast majority of these proteins are derived from the genomic sequencing of organisms (*i.e.*, genomic enzyme orthologs: GEOs), most of them have not been experimentally characterized and neither function nor specificity is known. Previous efforts to mine GEOs for function have relied on random sampling or sequence-based bioinformatics over an entire enzyme family.^{8,9,10} While enzymes with a desired function have been found using these methods, it requires a large number of genes to be produced and experimentally characterized. To enable a more efficient sampling method, a novel computational pipeline was developed that integrates the use of bioinformatics and molecular modeling to carry out forward predictions of a GEO's ability to carry out a function of interest. Using this integrative genomic mining approach a highly diverse set of ketoacid decarboxylases capable of utilizing 2-ketooctanoate (C8) as a substrate was identified. The median activity of GEOs selected using the integrative genomic mining approach is 75-fold greater than a set of naively selected proteins from the enzyme family.

[0086] The second approach focuses on reprogramming the substrate specificity of KIVD by using computational enzyme design methods. The Rosetta Molecular Modeling Suite was used to direct the generation and screening of over 400 mutants.¹¹ This effort lead to the discovery of a mutant with >600-fold increase in specificity for C8. Finally, the most specific and active enzyme from each approach is incorporated into the synthetic recursive +1 pathway to evaluate its ability to modulate *in vivo* alcohol production. Both enzymes result in a complete switch of the alcohol production profile towards long chain alcohols, demonstrating the ability to rationally control biosynthetic product specificity for chemicals not commonly found in Nature.

Results

Integrative Genomic Mining for Enzyme Function

[0087] KIVD belongs to the thiamine pyrophosphate enzyme family that is composed of more than 17,000 sequences.¹² To identify GEOs with the desired function within this vast sequence space, bioinformatics¹³ and molecular modeling¹⁴ with functional constraints were combined to guide our exploration. (Figure 2) First, ketoacid decarboxylase GEOs were identified based on sequence similarity to the KIVD used in the current synthetic recursive +1 pathway. At the time the search was performed, 2082 sequences were found in the non-redundant sequence database as significant matches. To broadly search KIVD sequence space, redundant sequences were filtered out with a sequence identity cut-off of 90%.¹⁵ Sequences derived from eukaryotic organisms were removed in order to increase the likelihood of producing soluble proteins in *E. coli* where the synthetic recursive +1 pathway has been implemented.

[0088] The approach assumed that all potential scaffolds with the desired function should be structurally homologous to KIVD and the sequence set was further refined accordingly. Since crystal structures for most of these predicted proteins are not available for analysis, homology models were produced to obtain a predicted ternary structure. Using RosettaCM one hundred models were generated for each of the 239 GEOs.¹⁴ The lowest energy model was selected as the representative for each GEO and evaluated for structural similarity to KIVD. The TMalign algorithm¹⁶ was used to overlay the models with the native KIVD crystal structure and only those with a TMalign score of >0.5 were kept. These bioinformatics and structural filters resulted in 239 GEOs as candidates for the targeted function. (Figure 16) The pairwise identity of every sequence to each other was calculated and the mode of these values lies near 20%. (Figure 2)

Thus the selected set of 239 GEOs represents a diverse sampling of sequence space for this fold family.

[0089] To further enrich the set for GEOs likely to function on C8, a modeled reaction intermediate of the C8 substrate was docked within the predicted active site. The lowest energy model for each GEO was used as a template for docking a modeled C8 intermediate with functional constraints, which ensured a productive geometric orientation between the predicted catalytic residues in the protein and intermediate. The C8 reaction intermediate used for docking is based on the synthetic thiamine pyrophosphate (TPP) analogue from a crystal structure of KIVD (PDB ID: 2VBG), which mimics the initial nucleophilic attack by the thiazolium.¹⁷ Docking calculations were carried out using Rosetta Design in which both the identity and conformation of non-catalytic residues within the active site were allowed to change, and the backbone was allowed to move during minimization. The calculated interface energy after the dock and design simulations was used to evaluate each GEO's ability to accommodate C8.

[0090] A phylogenetic tree based on sequence similarity was built for the 239 GEOs in which the lowest protein-ligand interface energy from the Rosetta Design simulation for each GEO are depicted. (Figure 2) The bar height above each GEO leaf is proportional to the lowest interface energy from all 239 GEOs, the higher the bar the lower the energy. Based on the calculated energies it is immediately apparent that a cluster of GEOs distant in sequence to KIVD (~15% sequence identity) is predicted to be capable of utilizing C8. The enzyme with the lowest energy, GEO 175, is a predicted protein in the genome database with no known function. The closest enzyme to GEO 175 (~50% sequence identity) for which significant biophysical characterization has been performed is benzoylformate decarboxylase from bacteria *Pseudomonas putida* (ppBFD).¹⁸ A structural analysis of the active site of ppBFD and GEO 175 reveals that the active sites are only 50% identical in sequence (Figure 10), and the activity of ppBFD on C8 has not been previously studied.

[0091] From the design simulations, ten GEOs were chosen for experimental characterization based on the predicted protein-ligand interface energy as well as sequence diversity (Figure 2, Figure 16). Before ordering these GEOs, each protein was manually analyzed and mutations introduced through the Rosetta Design simulations were reverted if they were not predicted to significantly enhance the calculated Rosetta interface energy. Synthetic genes encoding each

enzyme was obtained, expressed in *E. coli* and purified for *in vitro* kinetic characterization using a panel of 2-ketoacid substrates of different chain lengths including C8, C5, C3 and the natural substrate for KIVD ketoisovaleric acid (Figure 3). Out of the 10 selected GEOs, six express and are able to be purified in a soluble form using our standardized method. Of these six, three exhibit detectable activities on at least one of the four ketoacids used for kinetic characterization.

[0092] Among the active GEOs, the enzyme with the highest efficiency on C8 is the computationally top ranked GEO 175, with a k_{cat}/K_M of 17,000 $M^{-1}s^{-1}$. (Table 1, Figure 4) This is only 2-fold lower than the efficiency of native KIVD on C8. However, GEO 175's catalytic efficiency on C8 is 33,000-fold higher than on C3 and 354-fold higher than on C5. When assayed under the same conditions, native KIVD's catalytic efficiency on C8 is 762-fold higher than on C3 and 3.4-fold higher than on C5. This corresponds to over a 100-fold improvement in specificity (C8 versus C5) relative to the native KIVD.

Table 1: Kinetic Characteristics of 2-ketoacid Decarboxylase Enzymes

k_{cat}/K_M ($M^{-1} s^{-1}$)	C3	C5	C8	Iso C5
GEO179	0.90 ± 0.02	100 ± 15	1200 ± 130	41 ± 1.2
GEO195	2.4 ± 0.1	200 ± 14	1400 ± 180	8.0 ± 0.3
GEO175	0.51 ± 0.02	48.0 ± 3.7	17000 ± 2700	3.3 ± 0.2
GEO175 L376T_T240S	0.27 ± 0.02	30.0 ± 7.2	1100 ± 160	1.6 ± 0.1
1OVM	32.0 ± 0.5	2100 ± 420	80000 ± 4300	1300 ± 130
2VBI	5700 ± 1400	52.0 ± 1.8	14.0 ± 0.8	1.8 ± 0.1
3FZN	5.4 ± 0.1	1700 ± 110	350 ± 30	110 ± 10
1ZPD	8200 ± 545	140.0 ± 5.6	19.0 ± 2.0	0.33 ± 0.01
1OZF	n.d.	0.53 ± 0.04	n.d.	17.0 ± 0.8
Native KIVD	42.0 ± 0.6	9500 ± 470	32000 ± 5500	14000 ± 1100
KIVD_VLV	0.71 ± 0.07	1.3 ± 0.2	2800 ± 860	0.24 ± 0.02
k_{cat} (s^{-1})				
GEO179	n.d.	0.47 ± 0.03	0.32 ± 0.01	0.39 ± 0.01
GEO195	n.d.	0.49 ± 0.01	0.56 ± 0.02	0.072 ± 0.001
GEO175	n.d.	0.97 ± 0.04	10.1 ± 0.6	n.d.
GEO175 L376T_T240S	n.d.	0.28 ± 0.04	3.4 ± 0.2	n.d.
1OVM	0.2 ± 0.01	1.0 ± 0.1	1.7 ± 0.1	7.8 ± 0.4
2VBI	25 ± 2.6	n.d.	0.051 ± 0.01	0.025 ± 0.001
3FZN	n.d.	4.7 ± 0.1	1.3 ± 0.1	0.52 ± 0.02
1ZPD	46.7 ± 1.4	0.95 ± 0.02	0.021 ± 0.001	n.d.
1OZF	n.d.	n.d.	n.d.	0.030 ± 0.001
Native KIVD	n.d.	14.3 ± 0.2	7.0 ± 0.3	61.0 ± 2.1
KIVD_VLV	n.d.	0.013 ± 0.001	0.50 ± 0.03	n.d.
K_M (mM)				
GEO179	n.d.	4.6 ± 0.6	0.27 ± 0.03	9.4 ± 0.2
GEO195	n.d.	2.5 ± 0.2	0.40 ± 0.04	9.1 ± 0.3
GEO175	n.d.	20.0 ± 1.3	0.58 ± 0.09	n.d.
GEO175 L376T_T240S	n.d.	10 ± 2	3.1 ± 0.4	n.d.
1OVM	7.5 ± 0.1	0.5 ± 0.1	0.021 ± 0.001	6.1 ± 0.6
2VBI	4.3 ± 1.0	n.d.	3.6 ± 0.2	14.0 ± 0.5
3FZN	n.d.	2.7 ± 0.2	3.8 ± 0.3	4.9 ± 0.3
1ZPD	5.7 ± 0.3	7.1 ± 0.3	1.2 ± 0.1	n.d.
1OZF	n.d.	n.d.	n.d.	1.8 ± 0.1
Native KIVD	n.d.	1.5 ± 0.1	0.21 ± 0.04	4.5 ± 0.3
KIVD_VLV	n.d.	10.0 ± 1.0	0.18 ± 0.05	n.d.

[0093] To evaluate if the integrative genomic mining approach enriches the mining for C8 activity, an additional set of previously characterized ketoacid decarboxylases diverse in sequence were experimentally characterized in an equivalent manner. Five enzymes were selected (PDB: 1OVM, 2VBI, 3FZN, 1ZPD, 1OZF) and all were found to have detectable activity on at least one of the four ketoacids used for GEO characterization. The median C8 activity from the GEOs selected using the integrative genomic mining approach is 75-fold higher than the naively picked decarboxylases (Table 1, Figure 4). This indicates the integrative genomic mining approach is effective in identifying functional enzymes that perform the desired reaction on C8.

[0094] The protein sequence for GEO 175 is not predicted to require any mutations to be capable of utilizing the C8 substrate, and therefore represents the native enzyme from

Streptomyces sp. C. Comparing the active site pocket of GEO 175 with native KIVD (16.8% identity to each other) reveal significant differences which gave implications to the high specificity of this newly identified enzyme. (Figure 5) Based on molecular modeling, GEO 175 is predicted to have an extended active site that is significantly more open to solvent relative to native KIVD. This results in the predicted binding mode for C8 to be extended in GEO 175 as opposed to a “wrapped” conformation in KIVD. In GEO175, the third through fifth carbons of the C8 ketoacid ligand are predicted to occupy a solvent-exposed, open pocket with limited molecular interactions. However the sixth through eighth carbons are predicted to interact with a narrow and hydrophobic pocket. Contrary to GEO 175, KIVD has an enclosed hydrophobic binding pocket and is predicted to make molecular interactions throughout the ketoacid alkyl chain (Figure 5).

[0095] From these observations it was hypothesized that the specificity of GEO 175 could be attributed to a combination of hydrophobic interactions with the sixth through eighth carbons of the alkyl chain, while lacking interactions with the third through fifth carbons. (Figure 6) Therefore, an engineered a mutant of GEO 175 containing two amino acid changes (L376T, T240S) predicted to remove the direct molecular interactions with seventh and eighth carbons of the ketoacid alkyl chain, while maintaining overall protein stability was generated. Upon kinetic characterization, it was found that the mutant protein’s catalytic efficiency on C8 is 15-fold lower than GEO 175, but the decrease in efficiency for C5, C3, and isoC5 assayed under the same conditions is less than 2-fold. The ability to rationally modulate activity based on structure supports the accuracy of the GEO 175 molecular model.

Computationally Directed KIVD Active Site Redesign

[0096] In the second effort to reprogram the specificity of KIVD, Rosetta Design methods were used to explore the potential active site sequence space that would accommodate 2-ketooctanoate as a substrate.¹¹ Design simulations were carried out with the same reaction intermediate and functional constraints utilized in the genomic mining pipeline. Ten residues in the active site were targeted for mutagenesis. At each site, one of eleven relatively hydrophobic amino acids (Val, Leu, Ile, Met, Phe, His, Gly, Ala, Thr, Tyr, Trp) was systematically introduced, and the remaining nine sites were allowed to be redesigned. The identities of amino acids at all other residues were kept constant. (Figure 7) The sequence profile from the 50 lowest

energy designs was used to guide construction of a small library of roughly 400 KIVD mutants from the original combinatorial space of over 10^{11} possible active site mutations.

[0097] Since KIVD can already efficiently utilize C8 as a substrate, enzyme specificity is of primary importance. Therefore each mutant was screened against a panel of 2-ketoacid substrates in order to assess the specificity and activity of each mutant. (Figure 11) This assay was performed using a high-throughput pH-dependent colorimetric assay of enzyme activity in crude cell extracts. Mutations that increased specificity for C8 were then combined to produce combinatorial mutants and screened for specificity with the colorimetric assays. The mutant that exhibits the highest specificity and activity is G402V/M538L/F542V (KIVD_VLV). The engineered enzyme KIVD_VLV has a 600-fold improvement in specificity between C8 and C5, in terms of catalytic efficiency, relative to the native KIVD.

In Vivo Testing of Engineered and Genomic Enzymes for Long-chain Alcohol Production

[0098] The measured specificity and activity for each enzyme was used to select two candidates for further characterization of their ability to modulate alcohol production *in vivo*. The two enzymes with the highest activity and specificity (Figure 4), KIVD_VLV and GEO 175, were tested for their ability to reprogram the synthetic recursive +1 pathway to produce long-chain alcohol products *in vivo*. The alcohol-producing biochemical pathway was adapted from a previous study used to synthesize alcohols of various lengths.^{3, 4} Starting from glucose, the aforementioned +1 pathway, involving LeuABCD, recursively elongates 2-ketoacids, starting with 2-ketobutyrate, into longer chained 2-ketoacids. These 2-ketoacids are then converted to aldehydes by ketoacid decarboxylases (native KIVD, KIVD_VLV, and GEO 175) and subsequently to alcohols by an alcohol dehydrogenase. Carbon flux through this alcohol synthesizing biochemical pathway is enhanced by overexpressing thrABC, ilvA, LeuABCD, and ADH6 on inducible plasmids transformed into *E. coli*.³ The effects of overexpression were furthermore supported and maintained by knocking out the threonine exporter enzyme RhtA and the primary DNA recombination degradation enzyme RecA. The enzyme AdhE was also knocked out, which helps raise long chain alcohol production by eliminating a competing pathway where acetyl-CoA is directly converted to ethanol.¹⁹

[0099] Consistent with previous reports, the primary products (82% of total alcohol production) with native KIVD are short alcohols (C₂ – C₄) (Figure 8, Table 2). The KIVD_VLV

triple mutant completely switches the product profile of this pathway to predominantly long chain ($\geq C_5$) alcohols at 728 mg l⁻¹. Hexanol (C₆) is the major alcohol detected at a titer of 341 mg l⁻¹ (47%), with titers of 269 mg l⁻¹ (37%) heptanol (C₇), and 118 mg l⁻¹ pentanol (16%) also observed. GEO 175 also switches the product profile of the pathway so that long chain alcohols are the primary products and produced at a level of 522 mg l⁻¹. For GEO175 the product profile is further shifted, resulting in heptanol (C₇) as the major product at a titer of 274 mg l⁻¹ (50%), with hexanol production at 160 mg l⁻¹ (29%), and 88 mg l⁻¹ pentanol (16%). Similar to the results for KIVD_VLV, no ethanol or propanol production is observed, and only 19 mg l⁻¹ butanol is produced. In addition, 10 mg l⁻¹ of octanol was also observed, a product not observed when either native KIVD, or KIVD_VLV were tested under equivalent conditions. The *in vivo* butanol and octanol titers for GEO175 compared to KIVD_VLV are consistent with the *in vitro* enzyme kinetics in which GEO175 has a significantly higher activity on C₈, but lower specificity relative to C₅, than KIVD_VLV. Total long chain alcohol titers for both GEO 175 and KIVD_VLV are increased significantly in comparison to native KIVD, with a >10-fold increase in heptanol production in both cases, and a >95% yield of long chain alcohols.

Table 2: *In vivo* Alcohol Production

	Native KIVD (mg/L)	KIVD_VLV (mg/L)	GEO 175 (mg/L)
1-Ethanol	652 ± 32	n.d.	n.d.
1-Propanol	693 ± 4	n.d.	n.d.
1-Butanol	1964 ± 24	n.d.	19 ± 2
1-Pentanol	594 ± 20	118 ± 29	88 ± 12
1-Hexanol	75 ± 6	341 ± 63	160 ± 18
1-Heptanol	20 ± 3	269 ± 23	274 ± 19
1-Octanol	n.d.	n.d.	10 ± 1

Cells were incubated for forty hours in microaerobic conditions in a defined media. The best enzyme in shifting the product profile was GEO175 with 1-heptanol being the major product.

[0100] Replacing KIVD with GEO 175 or KIVD_VLV results in a significant increase of long chain alcohol production, however the overall alcohol production drops ~10-fold from the 4.0 g/L obtained with KIVD. We hypothesized that a potential reason for this could be due to the toxicity of long chain alcohols. To explore the potential of toxicity being a limiting factor in alcohol production, *E. coli* growth was monitored over a 7-hour incubation with 10-1000 mg l⁻¹ C₅-C₈ alcohols supplemented at the beginning of growth. (Figure 9) Long chain alcohols are found to be toxic as the supplemented concentration of C₆-C₈ alcohols reach 250 mg l⁻¹. The final OD of *E. coli* cultures drops from 1.9 to approximately 1.6 for C₆ and C₇ alcohols, and 1.0 for C₈. At 500 mg l⁻¹ the OD drops to 1.0 and 0.1 for C₆ and C₇ alcohols, respectively. No significant growth is observed for C₈. At 1000 mg l⁻¹ no significant growth is observed for heptanol, and only an OD of 0.2 is achieved in the presence of hexanol. However, growth up to an OD of 1.5 is observed for pentanol. Significant toxicity in the range of 250 to 1000 mg l⁻¹ for long chain alcohols is consistent with the level of long chain alcohol titer, roughly 500mg/L, produced in the engineered *E. coli* strains. This result indicates that in order to achieve higher total alcohol titer with this pathway, a strain of *E. coli* with high tolerance toward long chain alcohols is likely necessary.^{20, 21} However, similar issues with product toxicity have been addressed through the design of continuous extraction methods and could be applied to the system here to increase levels of long chain alcohols titers.²²

Discussion

[0101] In this study an integrative genomic mining approach was introduced to enable discovery of enzymes for a targeted function from sequence databases. The function of interest here was an enzyme with high activity for the decarboxylation of long chain ketoacids in order to modulate the carbon flux of a synthetic pathway and increase titers of long-chain alcohols. The integrative genomic mining approach led to the discovery of a previously uncharacterized protein that we demonstrate has the targeted functional properties, and only required experimental characterization of ten new proteins. In parallel, a more traditional computationally directed library screening approach was utilized to reengineer the activity of a well-established enzyme. This required two successive rounds of screening over four hundred mutants against a panel of substrates in order to identify an enzyme with the desired functional properties. When evaluated for their ability to modulate carbon flux *in vivo* both are able to completely shift the product profile towards long chain alcohols.

[0102] The enzyme GEO 175 represents the product of a general and rapid approach for obtaining enzymes with a desired function from the rapidly growing sequence databases. This approach provides a viable alternative to often expensive and laborious enzyme engineering efforts that require screening libraries of mutants. By combining bioinformatics and molecular modeling this approach enables the identification of proteins likely to carry out a targeted function regardless of their native or putatively annotated activity. This overcomes issues with misannotation of protein function²³ or biased assumptions based on the closest, but often distal in sequence, characterized protein. The enzyme identified in this study, GEO175, is a clear illustration of this as it is roughly 15% identical to KIVD and its closest characterized sequence homolog (ppBFD, roughly 50% sequence identity) has a significantly different functional profile.

[0103] While GEO175 and KIVD_VLV both increased long chain alcohol titer >10-fold, we identified that toxicity needs to be addressed in order to further increase total long-chain alcohol titer. The titer of long chain alcohols produced is on the order of 0.5 g/L, which we then demonstrate is highly toxic for the cell lines used in this study. For industrial applications, bio-alcohol production often requires yields on the order of grams per liter.²⁴ To avoid the need of a continuous extraction fermentation system, future efforts for increasing long chain alcohol production should focus on engineering or finding strains that are tolerant to this level of product formation. Efforts to integrate pumps and reengineer the cell wall to be resistant to alcohols could potentially mitigate toxicity effects and concurrently enable higher titers of long-chain alcohols to be produced through this pathway.^{20,21}

[0104] In summary, a new genomic mining approach and computationally directed library design efforts were both successfully implemented to obtain enzymes that enabled specific production of long chain alcohols *in vivo*. These results demonstrate the modularity of the synthetic +1 recursive pathway and provided a clear path forward to engineer industrial level production of long-chain alcohols. Finally, the integrative genomic mining approach introduced here is highly general, and with the rapid growth of sequence databases it has the potential to revolutionize the development and discovery of enzyme catalysts.

Methods

Integrative Genomic Mining

[0105] To obtain the GEO sequences, the native KIVD sequence was input for a homologous sequence search using HMMER3's* online server.¹³ The resulting sequences were filtered using the CD-HIT* online server with a 90% identity cutoff.^{15, 25} A homology model of each sequence was made using Rosetta Comparative Modeling.¹⁴ At this point, the intermediate was placed into the active site and 1000 simulations were ran to relax the intermediate according to the constraints. For each model, the lowest 100 in overall protein energy models were selected and then from that subset, the lowest protein-ligand interface was chosen as the energy for the GEO.

[0106] An intermediate of the C8 product-yielding reaction was modeled using Spartan*.²⁶ Different conformations of the alkyl chain were included in the modeling and a conformational library was made using OpenEye Omega*.²⁷ For the enzyme design of KIVD, this intermediate was placed into the active site using distance and angle constraints. Rosetta Enzyme Design was run with default settings, an example of which is provided in the Rosetta Molecular Modeling Suite demos.

[0107] From the design simulations for each GEO the lowest protein-ligand interface Rosetta energy was used to select a tractable number of GEOs for experimental characterization. During these simulations, any amino acids with a C α within 8 Å of the active site could be mutated to any of the 20 amino acids. An exemplary mapping of GEO175 active site residues to 2VBG active site residues is provided as Table 3. Ten GEOs of significant interest were chosen with the following criteria: Five GEOs were chosen because they had the lowest predicted energies; the second five were chosen with the purpose of maximizing sequence space diversity. The 234 GEOs (less the five lowest in energy) were filtered with a sequence identity cutoff of 40% and the five sequences from this filtered list with the lowest energy were picked as the second five GEOs in our final list. Each model was evaluated in the Foldit interface and mutations made during the design simulations were reverted to the native amino acid if not predicted to improve the interface energy by more than 2 Rosetta energy units.

Table 3: Mapping of GEO175 active site residues to active site residues in 2VBG crystal structure

GEO175 Active Site Amino Acid Position	Corresponding Amino Acid position in 2VBG Crystal Structure
23	23
24	24
25	25
26	26
27	27
48	50
49	51
54	56
71	gap before 73
74	75
75	76
78	79
81	82
112	113
113	114
286	287
356	352
376	373
377	374

378	375
379	376
380	377
382	379
399	394
401	396
402	397
403	398
404	399
427	gap before 424
428	424
429	425
430	426
431	427
432	428
434	430
454	450
455	451
456	452
457	453
458	454
459	455

460	456
461	457
462	458
465	461
535	532

[0108] Phylogenetic tree of GEOs was generated using Geneious software* using a Muscle sequence alignment.^{28*} The resulting tree was visualized using iTOL online tool in circular tree mode and rooted at native KIVD.^{29,30}

[0109] Synthetic genes coding for each GEO were synthesized as a DNA String by Life Technologies or the Joint Genome Institute. Genes were codon optimized for *Escherichia coli* and the amino acid sequences are provided herein. See, Figure 17 and Sequence Listing. The string was cloned into the pET-29b(+) plasmid vector using the Gibson assembly between the NdeI and XhoI restriction sites which added a C-terminal 6x-His tag in-frame to the gene.

[0110] All programs referenced in this section and denoted with an asterisk (*) were run using the default settings unless otherwise specified.

KIVD Active Site Redesign

[0111] In the design simulations, ten residues in the proposed active site pocket were allowed to either remain native or sample any of eleven relatively hydrophobic amino acids (Val, Leu, Ile, Met, Phe, His, Gly, Ala, Thr, Tyr, Trp). The identities of amino acids at all other positions were kept constant. Residues within 12 Å of the ligand were allowed to undergo conformational sampling during simulation. A total of 10,000 design simulations were ran, from which the 50 designs lowest in ligand-protein interface energy, and non-redundant in terms of sequence, were selected to represent the potential sequence space predicted to accommodate the C8 substrate. The profile was used to construct a small library of ~400 KIVD mutants. Each amino acid in the library was sampled as a single mutation, with the exception of residues that were within five residues from one another. These were sampled in a combinatorial fashion as both single and

double mutants in order to evaluate synergistic effects given their proximity to one another. . A complete list of the amino acids allowed for each site is provided in Figure 18.

Construction and Selection of KIVD Libraries

[0112] Partially degenerate oligonucleotides were ordered from Integrated DNA Technologies (San Diego, CA) and were used to generate libraries of *kivd* mutants (with N-terminal his tags) by PCR. Library DNA was purified, inserted into the PCR-amplified pQE9 vector by isothermal Gibson assembly,³¹ and transformed into XL1-Blue cells. Individual ampicillin-resistant colonies were picked and grown in 96-well blocks at 37 °C overnight. The number of colonies picked was three times the theoretical library size to ensure approximately 95% probability of all possible mutation combinations occurring. Glycerol was added to 25% (w/v) and libraries were stored at -80 °C until enzyme activity and specificity were assayed. KIVD mutant libraries were screened using a pH-monitored enzyme assay (method explained below) to measure the rate of H⁺ consumption resulting from the decarboxylation of 2-ketoacids.

[0113] KIVD mutant libraries were grown overnight in culture blocks, diluted 1:100 into fresh media (Luria Broth), grown at 37 °C for 3 hours to OD₆₀₀ ~0.6, induced with 0.1mM IPTG (GoldBio, Saint Louis, MO), and grown for 2 additional hours at 37 °C. Cells from the well blocks were transferred to 96-well assay plates (Costar, Corning, NY) and cell density was measured at 600nm. An equal volume of permeabilization solution (8.7mM potassium phosphate, 43.4mM KCl, 0.87mM MgSO₄, pH 7.1±0.1, 8.7% (v/v) chloroform, 0.0043% (w/v) SDS, 0.26% (v/v) 2-mercaptoethanol) was then added to break the cell membranes. Bromothymol blue, TPP (Sigma-Aldrich, St. Louis, MO), and substrate were added to 0.008% (w/v), 0.5mM, and 10mM, respectively. All substrates (2-ketobutyrate, 2-ketovalerate, 2-ketoisovalerate, 2-ketocaproate, and 2-ketooctanoate; Sigma-Aldrich, St. Louis, MO) were dissolved in MilliQ dH₂O and pH was adjusted to 7.1±0.1. Absorbance was measured at 615nm in a spectrophotometer (TEK Powerwave XS, BioTek, Winooski, VT). Data were acquired for 15 min at 30 °C. Cells containing pQE_hiskivd_wt and pQE9 were used as positive and negative controls, respectively. Enzyme activity data were corrected for cell density. Mutants chosen from this screening had a 25% higher activity than wild-type KIVD for any of the 2-ketoacids.

[0114] The library derived from the pH-coupled assay was further screened using a second enzyme kinetics assay. This assay involved a second enzyme, alcohol dehydrogenase from *S.*

Cerevisiae (ADH6), which is a key enzyme in the alcohol production pathway in reducing the aldehyde produced by KIVD into an alcohol. Alcohol dehydrogenase couples oxidation of NADPH to NADP⁺ to the reduction of aldehyde into an alcohol. Stoichiometrically, conversion of one mole of 2-ketoacid to n-alcohol depletes exactly one mole of NADPH. Therefore, activity of KIVD could be measured directly by measuring depletion of NADPH through ultra-violet spectrophotometry.

[0115] Overnight cultures of XL1B with pQE9 containing KIVD mutations were grown at 37 °C in a 96 well block. Cultures were diluted 1:100 and then grown 3 hours at 37 °C to OD₆₀₀ of 0.6. Cultures were induced with 0.1 mM IPTG (GoldBio, Saint Louis, MO) for 3 hours at 37 °C. Cultures were then centrifuged, and pellets were lysed with BugBuster (Novagen, Madison, WI). 5 µL of cell lysate from each culture and 175 µL of a buffer mix containing coenzyme 1.5 mM TPP (Sigma), 0.2 mM NADPH (Fisher, Waltham, MA), 0.045 U ADH6, 100 mM NaPO₄, 100 mM NaCl, and 10 mM MgCl₂, pH of 7 were added together into a 96 well plate. 20 µL of 100 mM substrate was added to dilute to make a final concentration of 10 mM. The substrates tested were IsoC5, C4, C5, C6, C8 and H₂O. (Figure 11) Absorbances were measured with a plate reader (TEK Powerwave XS, BioTek, Winooski, VT) at 340 nm at 30 °C for 15 minutes. The protein concentration in each sample was determined with a BCA assay (Thermo Scientific, Waltham, MA) and used to normalize slope values. Candidate KIVD mutants were chosen for greater activity on longer chained 2-ketoacids and decreased activity on shorter 2-ketoacids. This was determined from graphing absorbance vs. time for each substrate and qualitatively comparing slope values vs. wild-type (Figure 11).

[0116] **Site-directed mutagenesis.** Oligonucleotides encoding specific *kivd* mutations were ordered from Integrated DNA Technologies (Coralville, Iowa) and were used to mutate *kivd* in pZE_LeuA*BCDKA6 and pQE_hiskivd_wt by PCR. Amplified DNA fragments were purified, inserted into either the PCR-amplified pZE vector containing *leuA*BCD* and *adh6* or the pQE vector by isothermal Gibson assembly, and transformed into XL1-Blue cells. Plasmid DNA was purified (Qiagen, Hilden, Germany) from overnight cultures of antibiotic-resistant colonies and the plasmid sequences were verified (Laragen, Culver City, CA).

[0117] **Chromosomal gene knockout.** Genes were removed from the ATCC 98082 *ΔrhtA* strain genome using P1 transduction from the Keio collection as previously.³² The aldehyde-

alcohol dehydrogenase gene (*adhE*) was knocked out to eliminate ethanol production from acetyl-coA. In all strains, *recA* was knocked out to prevent recombination between the genome and plasmids, thereby stabilizing the transformants. Primers used to target *recA* for knockout were designed based on the Keio collection (Genobase, ecoli.aist-nara.ac.jp) and ordered from Integrated DNA Technologies.

[0118] Fermentation procedure and analysis. For *n*-alcohol production, strains of ST128 were transformed with pZS_thrO, pZAlac_ilvA_{BS}leuA, and pZE_LeuA*BCDK*A6 containing various *kivd* mutations. Fermentation conditions were adapted from those in previous works,^{3,4} with the following changes: 20 mL of medium was used, with 100 µg/mL ampicillin, 50 µg/uL kanamycin, and 100 µg/mL spectinomycin added. Cells were grown to an optical density at 600 nm of ~0.6, followed by induction with 0.1 mM isopropyl-b-D-thiogalactoside (IPTG). After fermentation, cells were centrifuged for 15 minutes at 4000 x g and 4 °C. The supernatant was split into two fractions for analysis, 5 mL for short chain alcohols (ethanol, 1-propanol, and 1-butanol), and 15 mL for long chain alcohols (1-pentanol, 1-hexanol, 1-heptanol, and 1-octanol). Long chain alcohols were extracted from the 15 mL fraction by 3 mL *n*-hexane prior to analysis. GC-FID analysis was performed as previously described.⁴

[0119] Protein expression, purification, and enzymatic assay of KIVD. For the native KIVD and KIVD_VLV mutant, 2 ml overnight cultures of XL1B cells were transformed with pQE9 containing N-terminal his-tagged KIVD enzymes and grown in Terrific Broth (BP biomedical, Cat# 3046-042) with 50 µg/mL of carbenicillin. (Fisher scientific, Cat# BP2648-5) For the GEOs, 2 ml overnight cultures of BLR cells were transformed with pet29b+ plasmid containing N-terminal his-tagged GEOs, and grown in Terrific Broth with 50 µg/mL of kanamycin. (Fisher scientific, Cat# BP906-5) These cultures were diluted 1:1000 in 500 mL of Terrific Broth with 1 mM MgSO₄, 1% glucose, and 50ug/mL of corresponding antibiotics, then grown at 37 °C for 24 hours. Cultures were pelleted down at 4700 rpm for 10 mins and resuspended in auto-induction media (LB broth, 1 mM MgSO₄, 0.1 mM TPP, 1X NPS, and 1X 5052) for induction at 18 °C for 34 hours. At the end of induction, cells were centrifuged (4700 rpm, 4 °C, 20 minutes), supernatant was removed, and cells were resuspended in 40 mL lysis buffer (100 mM Hepes pH 7.5, 100 mM NaCl, 10% glycerol, 0.1 mM TPP, 1 mM MgSO₄, 10 mM Imidazole, 1 mM TCEP) and 1 mM PMSF) and sonicated for 2 minutes. Lysed cells were

centrifuged at 4700 rpm at 4 °C for 60 mins to remove cell debris. Supernatant was loaded on gravity flow column with 700 μ l of cobalt slurry (Fisher scientific, CAT# PI-90091) washed with 10 ml of wash buffer (100 mM Hepes pH 7.5, 100 mM NaCl, 10% glycerol, 0.1 mM TPP, 1 mM MgSO₄, 10 mM Imidazole, and 1 mM TCEP). Cobalt bead bed was washed with 15 ml of wash buffer 5 times and proteins were eluted with 1000 μ l of elution buffer (100 mM Hepes pH 7.5, 100 mM NaCl, 10% glycerol, 0.1 mM TPP, 1 mM MgSO₄, 200 mM Imidazole, and 1 mM TCEP). Protein samples were immediately buffer exchanged with spin concentrators (Satorius, CAT# VS0112) into storage buffer (100 mM Hepes pH 7.5, 100 mM NaCl, 10% glycerol, 0.1 mM TPP, 1 mM MgSO₄, and 1 mM TCEP) and stored at 4 °C until kinetics characterization. Protein concentrations were determined using a Synergy H1 spectrophotometer (Biotek) by measuring absorbance at 280 nm using their calculated extinction coefficients with ExPASy ProtParam Tool.³³ All other buffers and salts were purchased from Fisher Scientific unless otherwise specified.

[0120] The k_{cat} and K_M values of selective KIVD mutants were measured for the substrates: C₃, C₅, isoC₅, and C₈. All substrates were dissolved in MilliQ H₂O and pH was adjusted to 7.5 as necessary. Activity was measured at 0.005 mM to 10 mM substrates. The assay was performed in a 96-well half-area plate. Each reaction contains a final concentration of 0.5 mM NADH, 1 mM DTT, 0.1 mM TPP, 1 mM MgSO₄, reaction buffer (100 mM Hepes, 100 mM NaCl, 10% glycerol, pH 7.5) and ADH (Sigma Aldrich, A7011, 10 U/ml for C₃, C₅ and C₈ reactions, 500 U/ml for isoC₅ reactions). A wide range of ketoacid decarboxylase concentrations, 4.5 nM – 15 μ M, were used according to the activity of each enzyme toward different substrates in order to perform steady state kinetics measurement over a period of an hour. Absorbance readings were taken every 1-minute at OD₃₄₀ at 21 °C for 60 minutes using the Synergy H1. Kinetic parameters (k_{cat} and K_M) were determined by fitting initial velocity versus substrate concentration data to the Michaelis–Menten equation.

[0121] **Alcohol Toxicity.** Alcohol tolerance of the *in vivo* alcohol production strain was evaluated by supplementing the fermentation media with specified quantities of long chain alcohols, and measuring growth over 7 hours. Both the *E. coli* strain (ATCC 98082 pZS_thrO, pZAlac_ilvA_LeuA, pZE12LeuA*BCDKA6_KIVD_wt) and fermentation media (1X M9 metals + 1X trace metal mix + 0.5% yeast extract + 2% glucose + antibiotics) and conditions are the

same as described above (Fermentation procedure and analysis). Cells were grown in media without IPTG-induction to a starting OD₆₀₀ ~ 0.02 and then supplemented with either 1-pentanol, 1-hexanol, 1-heptanol, or 1-octanol at specified concentrations of 10 mg/L, 50 mg/L, 250 mg/L, 500 mg/L, or 1000 mg/L. OD₆₀₀ readings were taken hourly up to 7 hours. Figure 9 shows the final OD₆₀₀ recorded at 7 hours.

REFERENCES

1. House, US White. "National bioeconomy blueprint." *Washington DC, The White House*, April (2012).
2. USDA, US. "Bio-based products: market potential and projections through 2025." *US Department of Agriculture* (2008).
3. Zhang, K., Sawaya, M. R., Eisenberg, D. S., & Liao, J. C. "Expanding metabolism for biosynthesis of non-natural alcohols." *Proceedings of the National Academy of Sciences*, 20653-20658 (2008).
4. Marcheschi, R.J., *et al.* "A synthetic recursive "+ 1" pathway for carbon chain elongation." *ACS chemical biology* 7.4, 689-697 (2012).
5. Erickson, B., and Winters P. "Perspective on opportunities in industrial biotechnology in renewable chemicals." *Biotechnology journal* 7.2, 176-185 (2012).
6. Plaza, M., *et al.* "Biochemical and molecular characterization of α -ketoisovalerate decarboxylase, an enzyme involved in the formation of aldehydes from amino acids by *Lactococcus lactis*." *FEMS microbiology letters* 238.2, 367-374 (2004).
7. Felnagle, E.A., *et al.* "Engineering synthetic recursive pathways to generate non-natural small molecules." *Nature chemical biology* 8.6, 518-526 (2012).
8. Bayer, Travis S., *et al.* "Synthesis of methyl halides from biomass using engineered microbes." *Journal of the American Chemical Society*. 6508-6515 (2009).

9. Heins, Richard A., et al. "Phylogenomically guided identification of industrially relevant GH1 β -glucosidases through DNA synthesis and nanostructure-initiator mass spectrometry." *ACS chemical biology*. 2082-2091 (2014).
10. The National Academies. "Industrialization of Biology: A Roadmap to Accelerate Advanced Manufacturing of Chemicals." *The National Academies Press*. (2015)
11. Richter, F., et al. "De novo enzyme design using Rosetta3." *PLoS One* 6.5, e19230 (2011).
12. Finn, Robert D., et al. "Pfam: the protein families database." *Nucleic acids research* (2013).
13. Finn, R.D., Clements J., and Eddy S.R. "HMMER web server: interactive sequence similarity searching." *Nucleic acids research*, gkr367 (2011).
14. Song, Yifan, et al. "High-resolution comparative modeling with RosettaCM." *Structure* 21.10 (2013): 1735-1742.
15. Huang, Y., et al. "CD-HIT Suite: a web server for clustering and comparing biological sequences." *Bioinformatics* 26.5, 680-682 (2010).
16. Zhang, Y., and Skolnick J. "TM-align: a protein structure alignment algorithm based on the TM-score." *Nucleic acids research* 33.7, 2302-2309 (2005).
17. Berthold, Catrine L., et al. "Structure of the branched-chain keto acid decarboxylase (KdcA) from *Lactococcus lactis* provides insights into the structural basis for the chemoselective and enantioselective carbonylation reaction." *Acta Crystallographica Section D: Biological Crystallography*. 1217-1224 (2007).
18. Bruning, M., et al. "Structural and Kinetic Studies on Native Intermediates and an Intermediate Analogue in Benzoylformate Decarboxylase Reveal a Least Motion Mechanism with an Unprecedented Short-Lived Predecarboxylation Intermediate[‡]." *Biochemistry* 48.15, 3258-3268 (2009).

19. Shen, C.R., and Liao J.C. "Metabolic engineering of *Escherichia coli* for 1-butanol and 1-propanol production via the ketoacid pathways." *Metabolic engineering* 10.6, 312-320 (2008).
20. Kell, Douglas B., et al. "Membrane transporter engineering in industrial biotechnology and whole cell biocatalysis." *Trends in biotechnology* (2015).
21. Foo, Jee Loon, et al. "Improving Microbial Biogasoline Production in *Escherichia coli* Using Tolerance Engineering." *mBio* 5.6, e01932-14 (2014).
22. Brennan, Timothy CR, et al. "Alleviating monoterpene toxicity using a two-phase extractive fermentation for the bioproduction of jet fuel mixtures in *Saccharomyces cerevisiae*." *Biotechnology and bioengineering* 2513-2522 (2012).
23. Schnoes, Alexandra M., et al. "Annotation error in public databases: misannotation of molecular function in enzyme superfamilies." *PLoS computational biology* 5.12 (2009).
24. Parekh S. Strain improvement. *Desk encyclopedia of microbiology*. Academic Press, Waltham, (2010).
25. Kearse, M., et al. "Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data." *Bioinformatics* 28.12, 1647-1649 (2012).
26. Spartan'08, Wavefunction, Inc. Irvine, CA. Y. Shao et al. *Phys. Chem. Chem. Phys.*, 3172-3191 (2006).
27. OEChem, version 1.7.7, OpenEye Scientific Software, Inc., Santa Fe, NM, USA, www.eyesopen.com, (2010).
28. Edgar, Robert C. "MUSCLE: multiple sequence alignment with high accuracy and high throughput." *Nucleic acids research* 32.5, 1792-1797 (2004).
29. Letunic, Ivica, and Peer Bork. "Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation.", 127-128 *Bioinformatics* 23.1 (2007).
30. Letunic, Ivica, and Peer Bork. "Interactive Tree Of Life v2: online annotation and display of phylogenetic trees made easy." *Nucleic acids research* (2011).

31. Gibson, G.D., *et al.* "Enzymatic assembly of DNA molecules up to several hundred kilobases." *Nature methods* 6.5, 343-345 (2009).
32. Datsenko, A.K., and Wanner, B.L. "One-step inactivation of chromosomal genes in *Escherichia coli* K-12 using PCR products." *Proceedings of the National Academy of Sciences* 97.12, 6640-6645 (2000).
33. Gasteiger E., *et al.* *ExPASy*: the proteomics server for in-depth protein knowledge and analysis. *Nucleic Acids Res.* 31:3784-3788 (2003).
34. The PyMOL Molecular Graphics System, Version 1.7.4 Schrödinger, LLC.

INFORMAL SEQUENCE LISTING

>SEQ ID NO:1 KIVD

MYTVGDYLLDRLHELGIIEIFGVPGDYNLQFLDQIISRKDMKWVGNANELNASYMADGYA
RTKKAAAFLLTTFGVGELSAVNGLAGSYAENLPVVEIVGSPTSQVQNEGKFVHHTLADGDF
KHFMKMHEPVTAARTLLTAENATVEIDRVLSALLKERKPVYINLPVDVAAAKAEKPSLPL
KKENSTSNSTSDQEILNKIQESLKNACKPIVITGHEIISFGLEKTVSQFISKTKLPITTLN
FGKSSVDEALPSFLGIYNGKLSEPNLKEFVESADFILMLGVKLTDSSTGAFTHHLNENKM
ISLNIDEKGIFNESIQNFDFESLISSLLDLSEIEYKGYIDKKQEDFVPSNALLSQDRLW
QAVENLTQSNETIVAEQGTSTFFGASSIFLKPKSHFIGQPLWGSIGYTFPAALGSQIADKE
SRHLLFIGDGSLLQTLVQELGLAIREKINPFCFIINNDGYTVEREIHGPNQSYNDIPMWNV
SKLPESFGATEERVVSKIVRTENEFVSVMEKAQADPNRMWIELILAKEDAPKVLKKMGK
LF AEQNK

>SEQ ID NO:2 GEO175

MRTVRESALDVLRRGMTTVFGNPGSTELPMLKQFPDDFRYVLGLQEAVVVGMDGFFALASGTTGLVNLH
TGPGTGNAMGAILNARANRTPMVVTAGQQVRAMLTMEALLTNPQSTLLPQPAVKWAYEPPRAADVAPALA
RAVQVAETPPQGPVFVSLPMDDFDVVLGEDEDRAAQRAARTVTTHAAAPSAEVVRRLLAARLSGARSALV
AGNDVDASGAWDAVVELAERTGLPVWSAPTEGRVAFPKSHPOYRGMLPPAIAPLSRCLEGHDLVLVIGAP
VFCYYPYVPGAHLPELTELHTRDADEAARAPVGDAVVADLALTVRALLAELPAREAAAPAARTARAES
TAEVDGVLTPLAAMTAIAQGAPANTLWVNESPSNLGQFHDATRIDTPGSFLFTAGGGLGFGLAAGVGAQL
GAPDRPVVCVIGDGSTHYAVQALWTAAYKVPVTFVVLNQRYSILQWFAQVEGAQAGPLDIPGLDIAA
VATGYGVRAHRATGFGELSKLVRESALQDGPVLIDVPVTTELPTL

>SEQ ID NO:3 KIVD_VLV

MYTVGDYLLDRLHELGIIEIFGVPGDYNLQFLDQIISRKDMKWVGNANELNASYMADGYARTKKAAAFLLT
TFGVGELSAVNGLAGSYAENLPVVEIVGSPTSQVQNEGKFVHHTLADGDFKHFMKMHEPVTAARTLLTAE
NATVEIDRVLSALLKERKPVYINLPVDVAAAKAEKPSLPLKKENSTSNSTSDQEILNKIQESLKNACKPIV
ITGHEIISFGLEKTVSQFISKTKLPITTLNFGKSSVDEALPSFLGIYNGKLSEPNLKEFVESADFILMLG
VKLTDSSTGAFTHHLNENKMISLNIDEKGIFNESIQNFDFESLISSLLDLSEIEYKGYIDKKQEDFVPS
NALLSQDRLWQAVENLTQSNETIVAEQGTSTFFGASSIFLKPKSHFIGQPLWVSIGYTFPAALGSQIADKE
SRHLLFIGDGSLLQTLVQELGLAIREKINPFCFIINNDGYTVEREIHGPNQSYNDIPMWNVSKLPESFGAT
EERVVSKIVRTENEFVSVMEKAQADPNRMWIELILAKEDAPKVLKKLGLVAEQNK

>GEO249, NP_594083

MSSEKVLVGEYLFTRLLQLGIKSILGVPGDFNLALLDLIEKVGDDETRWVGNENELNGAYAADAYARVKG
ISAIVTTFGVGELSALNGFAGAYSERIPVVHIVGVPNTKAQATRPLLHHTLGNGDFKVFQRMSSSELSADV
AFLDSGDSAGRLIDNLLCVRVTSRPVYLAVPSDAGYFYTDASPLKTPLVFPVPENNKEIEHEVVSEILE
LIEKSKNPSILVDACVSRFHIQQETQDFIDATHFPTYVTPMGKTAINESSPYFDGVYIGSLTEPSIKERA
ESTDLLLIIGGLRSDFNSTFTYATPASQTIEFHSDYTKIRSGVYEGISMKHLLPKLTAADKKSQVQAKA
RPVHFEPKAVAAEGYAEGTITHKFWPPTFASFLRESVVTETGTSTNFGILDCIFPKGCQNLQVWLWGS
IGWSVGAMFGATLGIKDSDAPHRRSILIVGDGSLHLTVQEISATIRNGLTPPIIFVINNKGYTIERLIHGL
HAVYNDINTEWDYQNLKGYGAKNSRSYNIHSEKELLDLFKDEEFGKADVIQLVEVHMPVLDAPRVLIBQ
AKLTASLNKQ

>GEO243, WP_002115026

MELLSGGEMLVRLADEGVEHVFVGYPGGAVLHIYDALFQQDKINHLYVRHEQAAGHMADAYS RATGKTGV
VLVTSGPGATNTVTPIATAYMDSIPMVILSGQVASHLIGEDAFQETDMVGISRPVVKHSFQVRHASEIPA
I IKKAFYIAASGRPGPVVVDIPKDATNPAEKFAEYEPKVKMRSYQPPSRGHSGQIRKAIDELLSAKRPV
IYTGCGVQGNASALLTELAHLLGYPVTNTLMGLGGFPGDDPQFVGMLGMHGTYEANMAMHNADVILAIG
ARFDDRVTNNPAKFCVNAKVIHIDIDPASISKTIMAHIPIVGAVEPVLQEMLTQLKQLNVSKPNPEATAA
WWDQINERKVKHGLKFETPTDGTMPKQQVVEALYKATNGDAIITSVVGQHQMFALYKYKRPRQWINS
GLGTMGVGLPYMAAKLAFDPQVVCITGEASIQMCIQELSTCKQYGMNVKILCLNNRALGMVKQWQDMN
YEGRHSSSYVESLPDFGKLMEAYGHVGIQIDHADELESKLAEAMAINDKCVFINVMVDRTEHVYPMLIAG
QSMKDMWLKGKERT

>GEO250, WP_003075272

MPANTAPNAQAAEVFTVRHAVINMLRELGMTRIFGNPGSTELPLFRDYPEDFSYILGLQETVVVGMA
AQATRNASFVNLHSAAGVGHAMANIFTAFKNRTPMVITAGQOTRSLQFDFLHNSQAAELPKPYVKWSC
EPARAEDVPQALARAYYIAMQEPRGPVFVSI PADDWDVPCPITLRKVG FETRPDPRLDSIGQALEGAR
APAFVVGAAVDRSQAFEAVQALAEHQARVYVAPMSGRCGFPEDHALFGGFLPAMRERIVDRLSGHDVVF
VIGAPAFYTHVEGHGPFIAEGTQLFQLIEDPAIAAWAPVGDAAVGNIRMGVQELLARPLTHPRPALQPRP
AIPAPAAPEPGRIMTDAFLMHTLAQVRSRDSIIVEEAPGSRSI IQAHLPIYAAETFFTMCSGGLGHS
LPA SVGIALARPDKKVGIVIGDGSAMYAIQALWSAAHLKLPVTYIIVKNRRYAALQDFSRVFGYREGEK
VEGT DLPDIDFVALAKGQCGDGVVRTDAAQLSQVLRDALRS PRATLVEVEVA

>GEO242, WP_010764607

MYTVADYLLDRLKELGIDEVFGVPGDYNLQFLDHITARKDLEWIGNANELNAAYMADGYARTKGISALVT
TFGVGELSAINGLAGSYAESIPVIEIVGSPTTTVQONKKLVHHTLGDGDFLRFERIHEEVSAIAHLSTE
NAPSEIDRVLTVMTEKRPVYINLPIDIAEMKASAPTTPLNHTTDQLTTVETAILTKVEDALKQSKNPVV
IAGHEILSYHIENQLEQFIQKFNLPITVLPFGKGAFNEEDAHYLGTYTGSTTDESMKNRVDHADLVLLLG
AKLTDSATSGFSFGFTEKQMISIGSTEVLIFYGEKQETVQLDRFVSALSTLSFSRFTDEMPSVKRLATPKV
RDEKLTQKQFWMVESFLLQGD TVVGEQGSFFGLTNVPLKKDMHFIGQPLWGSIGYTFPSALGSQIANK
ESRHLFIGDGSLLTVQELGTAI REKLTPIVFVINNNGYTVEREIHGATEQYNDIPMWDYQKLPFVFGG
TDQTVATYKVVSTEIELDNAMTRARTDVRDLQWIEVMDQNDAPVLLKKLAKIFAKQNS

>GEO245, WP_018535238

MVSRPARVAILEQLRADGVRYMFGNPGTVEQGFLELNRNFPDIEYILALQEAGVVGLADGYARATRTPAV
LQLHTGVGVGNAVGMLYQAKRGHAPLVAIAGEAGLRYDAMEAQMAVDLVAMAEPVTKWATRVDPESTLR
VLRRAMKVAATPPYGPVLVVLPAVMDRDTSEAAVPTS YVDFAATPDPQVLDRAAELLAGAERP IVIAGD
GVHFAQAQEELGRLAQTWGAEVWGADWAEVNLSVEHPAYAGQLGHMFGDSSRRVTGAADAVLLVGTYALP
EVYPALDGVFADGAPVVHIDLDTDAIAKNFPVDLGLAADPRRALDGLARALERRMSPESRARAGEWFTGR
SAQRSYEIAAAREQDEAALAPDALPVTAFLQELARQLPEDAVVFDEALTASPDVTRHLPPTRPGHWHQTR
GGSLGVGIPGAIAAQLAHPDRTTVVGFTGDGGSLYTIQALWTAARYDIGATFVICNNSSYKLELNIEEY
W KSVDVAAHEQPEMFDLARPAIDFVALSRSLGVPVRVEKPDQAKAAVEQALGTPGPFLLIDLVTGRGRED

>GEO251, WP_020634527

MNV AELVGR TLAELGVGA AFGVVGSGNFVVTNGLRAGGVRFV AARHEGGAAS MADAYARMSGRVSVLSLH
QGCGLTNALTGITEAAKSRTPMIVLTGDTAASAVLSNFRIGQDALATAVGAVPERVHSAPTAVADTVRAY
RTAVQQRRRTVLLNLPLDVQAQE APEAVEI PKVRGPAPIRPDAGMVAKLADLLAEARRPVFIAGR GARASA
VPLRELAEISGALLATS AVAHGLFHDDPFS LGISGGFSSPRTADLIVDADLVIGWCALNMWTTTRHGTL
GPAARLVQVDVEQAALGAHRPIDLG VVG DVAGTAVDVHAELDKRGHQRSREAPTGTRWNDVPYNDLSGDG
RIDPRTL SRRLDEILPAERMVSIDSGNFMGYPSAYLSVPDENGFCFTQAFQSIGLGLGTAIGAALARPDR
LPVLGVGDGGFHMVASELETAVRLRIPLVIVVYNDAAYGAEIHHFGDADMTTVRFPD TDIAAIGRGFGCD
GVTVR SVGDLAAVKEWLG GPRDAPLVIDAKIADDGGSWWLAEAFRH

>GEO247, YP_005461458

MIDLDGTVTVAEYLGRLRLRHAGVEHLFGVPGDFNLNLLDGLAFVEGLRWVGS PNELGAGYAADAYARRRG
LSALFTTYGVGELSAINAVAGSAAEDSPVVHVVGSPRTTTVAGGALVHHTIADGDFRHFARAYAEVTV AQ
AMVTATDAGA QIDRVLLAALTHRKPVYLSIPQDLALHRIPAAPLREPLTPASDPAAVERFRTAVRDL LTP
AVRPIMLVGQLVSR YGLSTLVTDMTTRSGIPVAAQLSAKGVIDESVEGNLGLYAGSMLDGPAA SLIDSAD
VVLHLGTALTAE LTGFFTHRRPDARTVQLLSTAALVGTTFRDNVLF PDAMTTLAEVLTTFPAPARLAAPT
TRA EPTGLAASITPPAPSAVDLTASTATDLTAPTAGDISEMSRVLTQDAFWAGMQAWLPAGHALVADTGT
SYWGALALRLPGD TVFLGQPIWNSIGWALPAVLGQGLADPDRRPVLVIGDGAAQMTIQELSTIVAAGLRP
IILLNNRGYTIERALQSPNAGYNDVADWNWRAVVA AFAGPD TDYHHAATGT ELAKALTAASESNRPVFI
EVELDAFDTPPLLRRLAERATAPS

>GEO244, YP_005756646

MKQRIGAYLIDAIHRAGVDKIFGVPGDFNLAFLDDIISNPNVDWVGNTNELNASYAADGYARLNGLAALV
TTFGVGELSAVNGIAGSYAERIPVIAITGAPTRAVEHAGKYVHHS LGEGTFDDYRKMF AHTVAQGYITP
ENATTEIPRLINTAIAERRPVHLHLPIDVAISEIEIPTPFEVTA AKD TDASTYIELLTSKLHQSKQPIII
TGHEINSFHLHQELED FVNQTQIPVAQLSLGKGAFNEENPYMGIYDGKIAEDKIRDYVDNSDLILNIGA
KLTD SATAGFSYQFNIDDVMLNHHNIKIDVD TND EISLPSLLKQLSNISHTNNATFPAYHRPTSPDYTV
GTEPLTQQTYFKMMQNFLKPNDVIIADQGTSFFGAYDLALYKNNTFIGQPLWGSIGYTLPATLGSQ LADK
DRRNL LLLIGDGS LQLTVQAISTMIRQH IKPVLFVINNDGYTVERLIHGMYEPYNEIHMWDYKALPAVFGG
KNVEIHDV ESSKDLQDTFNAINGHPDVMHFVEVKMSVEDAPKKLIDIAKAFSQQNK

>GEO246, YP_006485164

MKTVHSASYEILRRHGLTTVFGNPGSNELPFLKDFPEDFRYIILGLHEGAVVGMADGFALASGRPAFVNLH
AAAGTGNGMGALTNAWYSHSPLVITAGQQVRSMIGVEAMLANVDAGQLPKPLVKWSHEPACAQDVPRALS
QAIQTASLP RPAPVYLSIPYDDWAQ PAPAGVEHLAARQVSGAALPAPALLAELGERLSRSRNPVLVLGPD
VDGANANGLAVELAEKLRMPAWGAPSASRCFPFTRHACFRGVLPAAIAGISRLLDGHDLILVVGAPVFRY
HQFAPGDYLPAGAE LVQVTCDPGEAARAPMGDALVGDIALTLEALLEQVRPSARPLPEALPRPPALAE EG
GPLRPETVFDVIDALAPRDAIFVKESTSTVTAFWQVRVEMREPGSYFFPAAGGLGFGLPAAVGAQLAQPRR
QVIGIIGDGSANYGITALWSAAQYRVP AVFIILKNGTYGALRWFAGVLEVPDAPGLDVPGLD FCAIARGY
GVEALHAATREELEGALKHALAADRPVLI EVPTQTIEP

>GEO248, YP_006991301

MYTVGNYLLDRLTELGIRDFGVPGDYNLKF LDHVMTHKELNWNIGNANELNAAYAADGYARTKGIAALVT
TFGVGELSAANGTAGSYAEKVPVVQIVGTPTTAVQNSHKL VHHTLGDGRFDHFEKMQTEINGAIAHLTAD
NALAEIDRVLR IAVTERCPVYINLAIDVAEVVAEKPLKPLMEESKKVEEETTLVLNKIEKALQDSKNPVV
LIGNEIASFHLESALAD FVKKFNLPVTVLPFGKGGFDEEDAHFIGVYTGAPTAESIKERVEKADLILIG
AKLTD SATAGFSYDFEDRQVISVGSDEVSFYGEIMKPVAF AQFVNGLSNLNYLGYTGEIKQVERVADIEA
KASNLTONNFWK FVEKYLSNGDTLVAEQGTSFFGASLVPLKSKMKFIGQPLWGSIGYTFPAMLG SQIANP
ASRHLLFIGDGS LQLTIQELGMTFREKLTPIVFVINNDGYTVEREIHGPNELYNDIPMWDYQNL PYVFGG
NKG NVATYKVTT EELVAAMSQARQD TTRLQWIEVVMGKQDSPDLLVQLGKVFAKQNS

>GEO179, ZP_06547677

MKTIHSAAYALLRRHGMTTIFGNPGSNELPFLKSPFEDFQYVLGLHEGAVVGMADGYALASGKPAFVNLH
AAAGTGNGMGALTNSWYSHSPLVITAGQQVRPMIGVEAMLANVDATQLPKPLVKWSYEPANAQDVPRLS
QAIHYANTTPKAPVYLSIPYDDWDQPSGPGVEHLIERDVQTAGTPDARQLQVLVQQVQDARNPVLVLGPD
VDATLSNDHAVALADKLRMPVWIAPSASRCPPTRHPSFRGVLPAAIAGISKTLQGHDLIIVVGAPVFRY
HQFAPGDYLPVGAQLLHITSDPLEATRPMGHALVGDIRETLRVLAEEVVQQSRPYPEALAAPECVTDEP
HHLHPETLFDVLDVAVPHDAIYVKESTSTTTAFWQRMNLRHPSYFPAAGGLGFLPAAVGVQLAQQR
RVVALIGDGSANYGITALWTAAYRI PVVFIILKNGTYGALRWFAVLKAEDSPGLDVPGLDFCAIAKGY
GVKAVHTDTRDSFEAALRTALDANEPTVIEVPTLTIQPH

>GEO240 ZP_06549025

MSNAITKVQANANARRGGDVLLLEVLESEGEVYVFGNPGTTTELPMFDALLRKPSIQYVLALQEASAVAMADG
YAQAACKPGFLNLHTAGGLGHGMGNLLNAKCSQTPLVVTAGQQDSRHTTTDPLLLGLDVGGMGKTFKWSQ
EVTHVDQLPVLVRRAFHDSDAAPKGSVFLSLPMDVMEAMSAIGIGAPSTIDRNAVAGSLPLLASKLAAFT
PGNVALIAGDEIYQSEANEVVALAEMLAADVVGSTWPNRI PYPTAHPLWRGNLSTKATEINRALSQYDA
IFALGGKSLITILYTAGQAVPEQCKVFQLSADAGDLGRYSSSELSVVGDIKSSLKVLLEPELEKATANHRR
DYQRRFEKAINFEKLSKESLLGQVQEQQSATVITPLVAAFEAARAIGPDVAIVDEAIATSGSLRKSLNSH
RADQYAFRLGGGLGWGMPPAAVGYSLGLKAPVVCVFGDGAAMYS PQALWTAHEKLPVTFIVMNNTTEYNV
LKNFMRSQADYTSQTDRTFIAMD LVNPSVDYQALGASMGLETRKVI RAGDIAPAVEAALASGKPNVIEII
ISKS

>GEO195, ZP_06846103

MTSRSSFSPPSASEQRGADIFAEVLQCEGVRYIFGNPGTTTELPLLDALTDITGIHYVLGLHEASVVAMAD
GYAASGKPGFVNLHTAGGLGNAMGAILNAKMANTPLVVTAGQQDTRHGVTDP LLHGDLTG IARPNVKWA
EEIHHPEHIPMLLRALQDCRTGPAGPVFLSLPIDTMRCTSVGAGEASRIERASVANMLHALATALAEV
TAGHIALVAGEEVFTANASVEAVALAEALGAPVFGASWPGHIPFPTAHPQWQGTLP PKASDIRETLGPF
AVLILGGHSLISYPYSEGPAIPPHCRFLQLTGDGHQIGRVHETTLGLVGDQLSLRALLPLLARKLQPN
GAVARLRQVATLKRDRARTEAAERSAREFDASATTPFVAAFETIRAIGPDVPIVDEAPVTIPHVRACLDS
ASARQYLFTRSAILGWGMPPAAVGVSLGLDRSPVCLVGDGSAMYS PQALWTAHERLPVTFVVFNNGEYN
ILKNYARAQTNYRSARANRFIGLDISDPAIDFPALASSLGVPARRVERAGDIAI AVEDGIRSGRPNLIDV
LISSSS

>GEO241, ZP_07033476

MNIAYETRENVASGRECLLEILRDEGVTHVFGNPGTTTELALIDALAGDDDFHFILGLQEAAVVGMDGY
AQATGRPSFVNLHTTAGLGNMGNLTNAFATNPMVVTAGQQDIRHLAYDPLLSGDLVGLARATVKWAHE
VRSLOELPIILRRAFRDANTEPRGPVVFVSLPMNIIDEIGTVSIPPRSTIVQAESGDISQLVRLLVESAGN
LCLVVGDEVGRYGATEAAVRVAELLGAPVYGSPFHSNVPEPTDHPLWRFTLPNTGEMRKVLGGYDRILL
IGDRAFMSTYTSDELPLSPKTQLLQIAVDRHSLGRCHAVELGLYGDPLSLLAAVGDALSQERALAPSRDS
RLAIARDWRASWEQDLKDECERLAPSRPLYPLVAADAVLRGVPPGTIVIVDECLATNKYVRQLYPVRKPGE
YYYFRGAGLWGMPPAAVGVSLGLERQQRVCLLGDGAAMYS PQALWSAAHESLPITFVVFNNSEYNILKN
FMRSRPGYNAQSGRFVGMIEINQPSIDFCALARSMGVDVRLTEPDDITAYMIAAGDREGPSLLEIPIAAT
AS

PROGRAM CODE AND INPUT

[0122] Described herein are various program codes and input that can be utilized in the methods of the present invention for identifying, designing, scoring, and/or characterizing 2-ketoacid decarboxylases of the present invention.

Ligand.pdb

[0123] The following exemplary ligand.pdb file contents can be used for calculating Rosetta energy units for ligand:enzyme interactions. The ligand.pdb file can also be used to calculate a ligand binding pocket (*e.g.*, active site) solvent accessible surface area.

```

HETATM17213 C11 X00 X 1 20.201 113.106 7.257 1.00 0.00
HETATM17214 N4 X00 X 1 19.791 111.778 7.282 1.00 0.00
HETATM17215 C6 X00 X 1 20.602 110.614 6.899 1.00 0.00
HETATM17216 C8 X00 X 1 20.906 109.687 8.079 1.00 0.00
HETATM17217 C7 X00 X 1 19.941 108.811 8.567 1.00 0.00
HETATM17218 N3 X00 X 1 20.229 107.980 9.634 1.00 0.00
HETATM17219 C10 X00 X 1 21.458 108.021 10.258 1.00 0.00
HETATM17220 N2 X00 X 1 22.411 108.875 9.787 1.00 0.00
HETATM17221 C9 X00 X 1 22.160 109.704 8.707 1.00 0.00
HETATM17222 N1 X00 X 1 23.154 110.476 8.289 1.00 0.00
HETATM17223 C2 X00 X 1 21.726 107.127 11.428 1.00 0.00
HETATM17224 C12 X00 X 1 18.500 111.725 7.696 1.00 0.00
HETATM17225 C4 X00 X 1 17.701 110.455 7.838 1.00 0.00
HETATM17226 C13 X00 X 1 18.045 113.015 7.978 1.00 0.00
HETATM17227 C14 X00 X 1 16.677 113.426 8.453 1.00 0.00
HETATM17228 C15 X00 X 1 16.683 114.488 9.549 1.00 0.00
HETATM17229 O7 X00 X 1 15.322 114.843 9.716 1.00 0.00
HETATM17230 P1 X00 X 1 14.842 115.976 10.759 1.00 0.00
HETATM17231 O1 X00 X 1 15.535 115.665 12.077 1.00 0.00
HETATM17232 O3 X00 X 1 13.339 115.886 10.849 1.00 0.00
HETATM17233 MG1 X00 X 1 11.877 117.206 10.165 1.00 0.00
HETATM17234 O6 X00 X 1 13.320 117.904 8.896 1.00 0.00
HETATM17235 P2 X00 X 1 14.706 118.411 9.193 1.00 0.00
HETATM17236 O2 X00 X 1 15.588 118.497 7.985 1.00 0.00
HETATM17237 O4 X00 X 1 14.642 119.707 9.942 1.00 0.00
HETATM17238 O5 X00 X 1 15.431 117.343 10.179 1.00 0.00
HETATM17239 S1 X00 X 1 19.109 114.205 7.732 1.00 0.00
HETATM17240 C17 X00 X 1 21.225 113.527 6.136 1.00 0.00
HETATM17241 C16 X00 X 1 20.888 112.379 5.024 1.00 0.00
HETATM17242 O8 X00 X 1 21.804 111.560 4.740 1.00 0.00
HETATM17243 O9 X00 X 1 19.772 112.545 4.414 1.00 0.00
HETATM17244 C18 X00 X 1 20.969 114.995 5.653 1.00 0.00
HETATM17245 C1 X00 X 1 22.079 116.060 5.757 1.00 0.00
HETATM17246 C19 X00 X 1 23.348 115.563 5.034 1.00 0.00
HETATM17247 C20 X00 X 1 24.602 116.384 5.305 1.00 0.00
HETATM17248 C3 X00 X 1 24.784 116.577 6.823 1.00 0.00
HETATM17249 C5 X00 X 1 24.614 115.223 7.536 1.00 0.00
HETATM17250 O10 X00 X 1 22.455 113.469 6.632 1.00 0.00
HETATM17251 H10 X00 X 1 20.062 110.047 6.141 1.00 0.00
HETATM17252 H24 X00 X 1 21.553 110.989 6.523 1.00 0.00

```

HETATM17253	H11	X00	X	1	18.954	108.779	8.106	1.00	0.00
HETATM17254	H8	X00	X	1	23.019	111.090	7.498	1.00	0.00
HETATM17255	H22	X00	X	1	24.047	110.452	8.762	1.00	0.00
HETATM17256	H7	X00	X	1	21.522	106.093	11.149	1.00	0.00
HETATM17257	H21	X00	X	1	22.769	107.223	11.729	1.00	0.00
HETATM17258	H28	X00	X	1	21.081	107.412	12.258	1.00	0.00
HETATM17259	H9	X00	X	1	17.647	110.175	8.891	1.00	0.00
HETATM17260	H23	X00	X	1	16.694	110.613	7.452	1.00	0.00
HETATM17261	H27	X00	X	1	18.183	109.657	7.274	1.00	0.00
HETATM17262	H12	X00	X	1	16.170	112.540	8.838	1.00	0.00
HETATM17263	H25	X00	X	1	16.170	113.867	7.596	1.00	0.00
HETATM17264	H13	X00	X	1	17.272	115.352	9.244	1.00	0.00
HETATM17265	H26	X00	X	1	17.130	114.121	10.474	1.00	0.00
HETATM17266	H1	X00	X	1	20.115	115.397	6.223	1.00	0.00
HETATM17267	H2	X00	X	1	20.635	114.977	4.607	1.00	0.00
HETATM17268	H6	X00	X	1	22.309	116.243	6.807	1.00	0.00
HETATM17269	H14	X00	X	1	21.738	116.983	5.289	1.00	0.00
HETATM17270	H4	X00	X	1	23.163	115.581	3.953	1.00	0.00
HETATM17271	H5	X00	X	1	23.539	114.517	5.305	1.00	0.00
HETATM17272	H3	X00	X	1	24.539	117.378	4.856	1.00	0.00
HETATM17273	H17	X00	X	1	25.493	115.880	4.920	1.00	0.00
HETATM17274	H15	X00	X	1	24.036	117.279	7.193	1.00	0.00
HETATM17275	H16	X00	X	1	25.781	116.969	7.022	1.00	0.00
HETATM17276	H18	X00	X	1	24.733	115.360	8.611	1.00	0.00
HETATM17277	H19	X00	X	1	25.366	114.523	7.173	1.00	0.00
HETATM17278	H20	X00	X	1	23.620	114.826	7.329	1.00	0.00

Ligand.params

[0124] The following exemplary ligand.params file contents can be used for calculating Rosetta energy units for ligand:enzyme interactions. The ligand.pdb file can also be used to calculate a ligand binding pocket (*e.g.*, active site) solvent accessible surface area.

```

NAME X00
IO_STRING X00 Z
TYPE LIGAND
AA UNK
ATOM C11 CNH2 X 0.54
ATOM N4 Npro X -0.38
ATOM C6 CH2 X -0.19
ATOM C8 aroC X -0.13
ATOM C7 aroC X -0.13
ATOM N3 Nhis X -0.54
ATOM C10 aroC X -0.13
ATOM N2 Nhis X -0.54

```

ATOM C9 aroC X -0.13
ATOM N1 NH2O X -0.48
ATOM H8 Hpol X 0.42
ATOM H22 Hpol X 0.42
ATOM C2 CH3 X -0.28
ATOM H7 Hapo X 0.08
ATOM H21 Hapo X 0.08
ATOM H28 Hapo X 0.08
ATOM H11 Haro X 0.10
ATOM H10 Hapo X 0.08
ATOM H24 Hapo X 0.08
ATOM C12 COO X 0.61
ATOM C4 CH3 X -0.28
ATOM H9 Hapo X 0.08
ATOM H23 Hapo X 0.08
ATOM H27 Hapo X 0.08
ATOM C13 COO X 0.61
ATOM C14 CH2 X -0.19
ATOM C15 CH2 X -0.19
ATOM O7 OH X -0.67
ATOM P1 Phos X 1.49
ATOM O1 OOC X -0.77
ATOM O3 Oaro X -0.67
ATOM MG1 Mg2p X 1.99
ATOM O6 Oaro X -0.67
ATOM P2 Phos X 1.49
ATOM O2 OOC X -0.77
ATOM O4 OOC X -0.77
ATOM O5 Oaro X -0.67
ATOM H13 Hapo X 0.08
ATOM H26 Hapo X 0.08
ATOM H12 Hapo X 0.08
ATOM H25 Hapo X 0.08
ATOM S1 S X -0.17
ATOM C17 CH1 X -0.10
ATOM C16 COO X 0.61
ATOM O8 OOC X -0.77
ATOM O9 OOC X -0.77
ATOM C18 CH2 X -0.19
ATOM C1 CH2 X -0.19
ATOM C19 CH2 X -0.19
ATOM C20 CH2 X -0.19
ATOM C3 CH2 X -0.19
ATOM C5 CH3 X -0.28
ATOM H18 Hapo X 0.08
ATOM H19 Hapo X 0.08

ATOM H20 Hapo X 0.08
ATOM H15 Hapo X 0.08
ATOM H16 Hapo X 0.08
ATOM H3 Hapo X 0.08
ATOM H17 Hapo X 0.08
ATOM H4 Hapo X 0.08
ATOM H5 Hapo X 0.08
ATOM H6 Hapo X 0.08
ATOM H14 Hapo X 0.08
ATOM H1 Hapo X 0.08
ATOM H2 Hapo X 0.08
ATOM O10 OOC X -0.77
BOND N1 H8
BOND N1 H22
BOND N2 C9
BOND N3 C10
BOND N4 C6
BOND N4 C11
BOND C1 C19
BOND C1 H6
BOND C1 H14
BOND C2 H7
BOND C2 H21
BOND C2 H28
BOND C3 C5
BOND C3 H15
BOND C3 H16
BOND C4 H9
BOND C4 H23
BOND C4 H27
BOND C5 H18
BOND C5 H19
BOND C5 H20
BOND C6 C8
BOND C6 H10
BOND C6 H24
BOND C7 N3
BOND C7 H11
BOND C8 C7
BOND C8 C9
BOND C9 N1
BOND C10 N2
BOND C10 C2
BOND C11 C17
BOND C11 S1
BOND C12 N4

BOND C12 C4
BOND C13 C12
BOND C13 C14
BOND C13 S1
BOND C14 C15
BOND C14 H12
BOND C14 H25
BOND C15 O7
BOND C15 H13
BOND C15 H26
BOND C16 O8
BOND C16 O9
BOND C17 C16
BOND C17 C18
BOND C17 O10
BOND C18 C1
BOND C18 H1
BOND C18 H2
BOND C19 C20
BOND C19 H4
BOND C19 H5
BOND C20 C3
BOND C20 H3
BOND C20 H17
BOND O3 MG1
BOND O6 P2
BOND O6 MG1
BOND O7 P1
BOND P1 O1
BOND P1 O3
BOND P1 O5
BOND P2 O2
BOND P2 O4
BOND P2 O5
CHI 1 C11 N4 C6 C8
CHI 2 C18 C1 C19 C20
CHI 3 N4 C6 C8 C7
CHI 4 N4 C11 C17 C16
CHI 5 C12 C13 C14 C15
CHI 6 C13 C14 C15 O7
CHI 7 C14 C15 O7 P1
CHI 8 C11 C17 C16 O8
CHI 9 C11 C17 C18 C1
CHI 10 C17 C18 C1 C19
CHI 11 C1 C19 C20 C3
CHI 12 C19 C20 C3 C5

CHI 13 C15 O7 P1 O1

NBR_ATOM C11

NBR_RADIUS 12.161170

ICOOR_INTERNAL	C11	0.000000	0.000000	0.000000	C11	N4	C6
ICOOR_INTERNAL	N4	0.000000	180.000000	1.391812	C11	N4	C6
ICOOR_INTERNAL	C6	0.000000	53.916539	1.468135	N4	C11	C6
ICOOR_INTERNAL	C8	96.000197	67.116300	1.531771	C6	N4	C11
ICOOR_INTERNAL	C7	111.300524	59.137106	1.391982	C8	C6	N4
ICOOR_INTERNAL	N3	-179.917537	59.736377	1.382921	C7	C8	C6
ICOOR_INTERNAL	C10	1.618450	58.841090	1.378822	N3	C7	C8
ICOOR_INTERNAL	N2	-1.665903	61.021480	1.364563	C10	N3	C7
ICOOR_INTERNAL	C9	0.593863	58.813216	1.384115	N2	C10	N3
ICOOR_INTERNAL	N1	-178.163678	62.733766	1.326764	C9	N2	C10
ICOOR_INTERNAL	H8	179.365571	60.001398	1.010016	N1	C9	N2
ICOOR_INTERNAL	H22	179.989152	59.997284	1.009957	N1	C9	H8
ICOOR_INTERNAL	C2	-179.539104	60.287405	1.495790	C10	N3	N2
ICOOR_INTERNAL	H7	54.439218	70.499071	1.089994	C2	C10	N3
ICOOR_INTERNAL	H21	120.082237	70.494917	1.090006	C2	C10	H7
ICOOR_INTERNAL	H28	119.957562	70.562820	1.089996	C2	C10	H21
ICOOR_INTERNAL	H11	179.993306	60.133454	1.089916	C7	C8	N3
ICOOR_INTERNAL	H10	120.583822	71.350394	1.089994	C6	N4	C8
ICOOR_INTERNAL	H24	121.121616	72.544242	1.090009	C6	N4	H10
ICOOR_INTERNAL	C12	-178.860538	71.160846	1.356234	N4	C11	C6
ICOOR_INTERNAL	C4	179.432481	55.532238	1.507704	C12	N4	C11
ICOOR_INTERNAL	H9	102.069181	70.493434	1.089997	C4	C12	N4
ICOOR_INTERNAL	H23	120.080362	70.495628	1.090075	C4	C12	H9
ICOOR_INTERNAL	H27	119.957609	70.556165	1.090015	C4	C12	H23
ICOOR_INTERNAL	C13	179.339904	70.347924	1.397384	C12	N4	C4
ICOOR_INTERNAL	C14	179.451180	52.280217	1.504307	C13	C12	N4
ICOOR_INTERNAL	C15	117.854883	65.606771	1.526234	C14	C13	C12
ICOOR_INTERNAL	O7	-175.865383	75.197950	1.417354	C15	C14	C13
ICOOR_INTERNAL	P1	-164.747946	57.441505	1.612898	O7	C15	C14
ICOOR_INTERNAL	O1	-43.185207	73.683638	1.520783	P1	O7	C15
ICOOR_INTERNAL	O3	-121.006572	72.974924	1.508629	P1	O7	O1
ICOOR_INTERNAL	MG1	-110.188590	50.123302	2.084810	O3	P1	O7
ICOOR_INTERNAL	O6	21.371051	94.278939	2.043304	MG1	O3	P1
ICOOR_INTERNAL	P2	-54.013740	50.009149	1.505128	O6	MG1	O3
ICOOR_INTERNAL	O2	164.799552	66.240624	1.498838	P2	O6	MG1
ICOOR_INTERNAL	O4	127.318605	69.476366	1.498457	P2	O6	O2
ICOOR_INTERNAL	O5	116.143547	71.979188	1.624755	P2	O6	O4
ICOOR_INTERNAL	H13	-119.330756	69.335680	1.090084	C15	C14	O7
ICOOR_INTERNAL	H26	-118.726623	67.835020	1.089992	C15	C14	H13
ICOOR_INTERNAL	H12	-120.796185	71.737001	1.089999	C14	C13	C15
ICOOR_INTERNAL	H25	-121.711373	73.431344	1.089949	C14	C13	H12
ICOOR_INTERNAL	S1	-178.555063	64.242905	1.615075	C13	C12	C14
ICOOR_INTERNAL	C17	32.779163	62.660465	1.575235	C11	N4	C6

ICOOR_INTERNAL	C16	73.771627	80.628634	1.632675	C17	C11	N4
ICOOR_INTERNAL	O8	-125.893768	62.597308	1.261160	C16	C17	C11
ICOOR_INTERNAL	O9	-170.771560	65.639200	1.283554	C16	C17	O8
ICOOR_INTERNAL	C18	121.091764	68.746111	1.567025	C17	C11	C16
ICOOR_INTERNAL	C1	159.715274	59.432661	1.541992	C18	C17	C11
ICOOR_INTERNAL	C19	-179.885210	70.242962	1.542018	C1	C18	C17
ICOOR_INTERNAL	C20	179.671288	65.188952	1.523201	C19	C1	C18
ICOOR_INTERNAL	C3	-179.189402	70.082610	1.540108	C20	C19	C1
ICOOR_INTERNAL	C5	-179.841491	70.561809	1.540037	C3	C20	C19
ICOOR_INTERNAL	H18	-179.464013	70.502707	1.090032	C5	C3	C20
ICOOR_INTERNAL	H19	-120.076286	70.498233	1.089998	C5	C3	H18
ICOOR_INTERNAL	H20	-119.965258	70.557224	1.089971	C5	C3	H19
ICOOR_INTERNAL	H15	-119.950407	70.509198	1.090018	C3	C20	C5
ICOOR_INTERNAL	H16	-120.079178	70.498264	1.089841	C3	C20	H15
ICOOR_INTERNAL	H3	-119.086173	68.309294	1.093502	C20	C19	C3
ICOOR_INTERNAL	H17	-121.483665	68.933404	1.094687	C20	C19	H3
ICOOR_INTERNAL	H4	-120.658650	71.441727	1.096942	C19	C1	C20
ICOOR_INTERNAL	H5	-116.222819	70.419964	1.097278	C19	C1	H4
ICOOR_INTERNAL	H6	-120.110603	70.565925	1.090080	C1	C18	C19
ICOOR_INTERNAL	H14	-120.044272	70.701604	1.089923	C1	C18	H6
ICOOR_INTERNAL	H1	-121.585119	71.959730	1.101515	C18	C17	C1
ICOOR_INTERNAL	H2	-113.963932	70.907736	1.098815	C18	C17	H1
ICOOR_INTERNAL	O10	118.803514	71.041960	1.327227	C17	C11	C18

SASA.xml

[0125] The following exemplary SASA.xml file contents can be used to calculate solvent accessible surface area (SASA) of a 2-ketoacid decarboxylase active site.

```

<ROSETTASCRIPTS>
  <SCOREFXNS>
  </SCOREFXNS>
  <FILTERS>
    <Sasa name=sasa_filter jump=1/>
  </FILTERS>
  <MOVERS>
  </MOVERS>
  <APPLY_TO_POSE>
  </APPLY_TO_POSE>
  <PROTOCOLS>

    <Add filter_name=sasa_filter/>

  </PROTOCOLS>
</ROSETTASCRIPTS>

```

Ligand Energies

[0126] Ligand binding energies described herein refer to the transition state substrate that is docked into the active site of a 2-ketoacid decarboxylase in a catalytically relevant conformation and calculated using Rosetta Enzyme Design style constraints. The base energy function is the `talaris2013_csts.wts` file which contains the following terms: `fa_atr fa_rep fa_sol fa_intra_rep fa_elec pro_close hbond_sr_bb hbond_lr_bb hbond_bb_sc hbond_sc dslf_fa13 rama omega fa_dun p_aa_pp ref`

[0127] This base score function can be invoked with the flag: `-score:weights Rosetta/main/database/scoring/weights/talaris2013_cst.wts`

[0128] The flag `-ligand::old_estat` turns the `fa_elec` term into the `hack_elec` term.

[0129] The flag `-enzdes::favor_native_res 2` turns on the `res_type_constraint` term with a weight of 2.

[0130] This results in the final score function of: `fa_atr fa_rep fa_sol fa_intra_rep hack_elec pro_close hbond_sr_bb hbond_lr_bb hbond_bb_sc hbond_sc dslf_fa13 rama omega fa_dun p_aa_pp ref res_type_constraint` with weights of `.8 .44 .75 .004 .7 1 1.17 1.17 1.17 1.1 1 .2 .5 .56 .32 1 2`

[0131] The ligand score is weighted sum of this modified `talaris 2013` score function.

Solvent Accessible Surface Area (SASA) Calculation

[0132] The SASA can be calculated using Rosetta with an input PDB file (the protein structure to score that has a ligand bound in it). The ligand and associated params file in the Rosetta format can be used to instruct the Rosetta program in how to treat and score the ligand. The binding pocket is auto detected. RosettaScripts can be used to run the SASA filter on the input PDB. The score is output in the log file/trace.

[0133] In an exemplary embodiment, SASA values are calculated using Rosetta version `c2bf8f674e7b416b5b756630d7ccd5d64c57512c 2015-05-04 14:47:46 -0700` from `git@github.com:RosettaCommons/main.git`. This version can be used for any of the Rosetta

functions described herein (*e.g.*, enzyme design, homology modeling, docking, energy function evaluation, *etc.*).

[0134] The SASA score can be calculated using the following command line input:

```
~/Rosetta/main/source/bin/rosetta_scripts.default.macosclangrelease -database
```

```
~/Rosetta/main/database -parser:protocol sasa.xml -s input.pdb -extra_res_fa ligand.params .
```

WHAT IS CLAIMED IS:

1 1. An expression cassette comprising a heterologous promoter operably
2 linked to a nucleic acid encoding an enzyme comprising 2-ketoacid decarboxylase activity,
3 wherein the enzyme:

4 comprises a 3-layer alpha-beta-alpha sandwich;
5 comprises a thiamine pyrophosphate (TPP) binding pocket;
6 has a TM align score of at least 0.5 as compared against PDB ID:2VBG; and
7 comprises at least a 100-fold greater catalytic efficiency for C8 2-ketoacid
8 substrates as compared to C3 or isoC5 2-ketoacid substrates, wherein the catalytic efficiency is
9 measured by kcat/Km.

1 2. The expression cassette of claim 1, wherein the enzyme comprising 2-
2 ketoacid decarboxylase activity has a TM align score of at least 0.7 as compared against PDB
3 ID:2VBG.

1 3. The expression cassette of claim 1, wherein the substrate binding pocket
2 of the enzyme has a solvent accessible surface area of at least about 7500 Å².

1 4. The expression cassette of claim 1, wherein the enzyme comprising 2-
2 ketoacid decarboxylase activity comprises an amino acid sequence that is at least 14% identical
3 to SEQ ID NO:1.

1 5. The expression cassette of claim 1, wherein the enzyme comprising 2-
2 ketoacid decarboxylase activity comprises an amino acid sequence that contains a mutation of a
3 native sequence wherein
4 the mutation is in an amino acid residue having a Cα within 8 Å of a 2-ketoacid
5 decarboxylase active site, wherein
6 the 2-ketoacid decarboxylase active site comprises the amino acids corresponding
7 to amino acid positions 23-27, 50, 51, 56, 75, 76, 79, 82, 113, 114, 287, 352, 373-377, 379, 394,
8 396-399, 424-428, 430, 450-458, 461, and 532 of PDB ID:2VBG.

1 6. The expression cassette of claim 1, wherein the enzyme comprising 2-
2 ketoacid decarboxylase activity comprises an amino acid sequence that contains a mutation of a
3 native sequence wherein

4 the mutation is in an amino acid residue within 8 Å of a 2-ketoacid decarboxylase
5 active site, wherein

6 the 2-ketoacid decarboxylase active site comprises the amino acids corresponding
7 to amino acid positions 23-27, 50, 51, 56, 71, 75, 76, 79, 82, 112, 114, 287, 356, 373-377, 379,
8 394, 396-399, 423-428, 430, 450-458, 461, and 535 of SEQ ID NO:1.

1 7. The expression cassette of claim 1, wherein the enzyme comprising 2-
2 ketoacid decarboxylase activity comprises an amino acid sequence that is at least 90% or 99%
3 identical to SEQ ID NO:1.

1 8. The expression cassette of claim 7, wherein the enzyme comprising 2-
2 ketoacid decarboxylase activity comprises an amino acid sequence that contains a mutation in an
3 amino acid residue corresponding to a residue selected from the group consisting of G402,
4 M538, and F542 of SEQ ID NO:1.

1 9. The expression cassette of claim 8, wherein the mutation in an amino acid
2 residue corresponding to a residue selected from the group consisting of G402, M538, and F542
3 of SEQ ID NO:1 is to a residue selected from the group consisting of Val, Leu, Ile, Met, Phe,
4 His, Gly, Ala, Thr, Tyr, and Trp.

1 10. The expression cassette of claim 8, wherein the mutation corresponds to a
2 mutation selected from the group consisting of G402V, M538L, and F542V of SEQ ID NO:1.

1 11. The expression cassette of claim 7, wherein the enzyme comprising 2-
2 ketoacid decarboxylase activity comprises two mutations in amino acid residues corresponding
3 to residues selected from the group consisting of G402, M538, and F542 of SEQ ID NO:1.

1 12. The expression cassette of claim 7, wherein the enzyme comprising 2-
2 ketoacid decarboxylase activity comprises mutations corresponding to G402V, M538L, and
3 F542V of SEQ ID NO:1.

1 13. The expression cassette of claim 1, wherein the enzyme comprising 2-
2 ketoacid decarboxylase activity comprises an amino acid sequence at least 90% or 99% identical,
3 or identical, to SEQ ID NO:2.

1 14. The expression cassette of claim 13, wherein the enzyme comprising 2-
2 ketoacid decarboxylase activity comprises an amino acid sequence that contains a mutation,
3 wherein
4 the mutation is in an amino acid residue of SEQ ID NO:2 having a C α within 8 Å
5 of a 2-ketoacid decarboxylase active site, wherein
6 the 2-ketoacid decarboxylase active site comprises the amino acids corresponding
7 to amino acid positions 23-27, 48, 49, 54, 71, 74, 75, 78, 81, 112, 113, 286, 356, 376-380, 382,
8 399, 401-404, 427-432, 434, 454-462, 465, and 535 of SEQ ID NO:2.

1 15. The expression cassette of claim 1, wherein the enzyme comprising 2-
2 ketoacid decarboxylase activity interacts with a C8 2-ketoacid substrate with a calculated
3 interaction energy of less than -10.4 rosetta energy units.

1 16. A modified enzyme comprising 2-ketoacid decarboxylase activity,
2 wherein the enzyme comprises:
3 a 3-layer alpha-beta-alpha sandwich;
4 thiamine pyrophosphate; and
5 at least a 100-fold greater catalytic efficiency for C8 2-ketoacid substrates as
6 compared to C3 or isoC5 2-ketoacid substrates, wherein the catalytic efficiency is measured by
7 kcat/Km,
8 wherein the modification comprises a mutation in the primary amino acid sequence of SEQ ID
9 NO:1 or SEQ ID NO:2.

1 17. The modified enzyme of claim 16, wherein the enzyme has a TM align
2 score against PDB ID:2VBG of at least 0.5.

1 18. The modified enzyme of claim 16, wherein the enzyme has a C α root
2 mean squared deviation of less than 1 Å over at least 100 structurally aligned residues as
3 compared against PDB ID:2VBG.

1 19. The modified enzyme of claim 16, wherein the modification comprises a
2 mutation in an amino acid residue, wherein

3 the amino acid residue is within 8 Å of a 2-ketoacid decarboxylase active site,
4 wherein

5 the 2-ketoacid decarboxylase active site comprises the amino acids corresponding
6 to amino acid positions 23-27, 50, 51, 56, 71, 75, 76, 79, 82, 112, 114, 287, 356, 373-377, 379,
7 394, 396-399, 423-428, 430, 450-458, 461, and 535 of SEQ ID NO:1.

1 20. The modified enzyme of claim 16, wherein the modification comprises a
2 mutation in an amino acid residue corresponding to a residue selected from the group consisting
3 of G402, M538, and F542 of SEQ ID NO:1.

1 21. The modified enzyme of claim 20, wherein the mutation in an amino acid
2 residue corresponding to a residue selected from the group consisting of G402, M538, and F542
3 of SEQ ID NO:1, and wherein the mutation is to a residue selected from the group consisting of
4 Val, Leu, Ile, Met, Phe, His, Gly, Ala, Thr, Tyr, and Trp.

1 22. The modified enzyme of claim 20, wherein the modification comprises
2 two mutations in amino acid residues corresponding to residues selected from the group
3 consisting of G402, M538, and F542 of SEQ ID NO:1.

1 23. The modified enzyme of claim 22, wherein the modification comprises
2 one or two mutations corresponding to the mutations selected from the group consisting of
3 G402V, M538L, and F542V of SEQ ID NO:1.

1 24. The modified enzyme of claim 23, wherein the modification comprises
2 mutations corresponding to G402V, M538L, and F542V of SEQ ID NO:1.

1 25. The modified enzyme of claim 16, wherein the ratio of k_{cat}/K_m against
2 C8 2-ketoacid substrates over the k_{cat}/K_m against isoC5 2-ketoacid substrates is at least 50-fold
3 greater than the native enzyme.

1 26. The modified enzyme of claim 16, wherein the solvent accessible surface
2 area of the substrate binding pocket is at least about 7500 Å².

1 27. A host cell comprising the expression cassette of claim 1, or the modified
2 enzyme of claim 16, wherein the host cell produces long chain ($\geq C5$) alcohols at a concentration
3 of at least 500 mg/L.

1 28. The host cell of claim 27, wherein the host cell produces heptanol as the
2 major alcohol product.

1 29. The host cell of claim 27, wherein the host cell produces octanol.

1 30. The host cell of claim 27, wherein the host cell does not produce ethanol
2 or propanol.

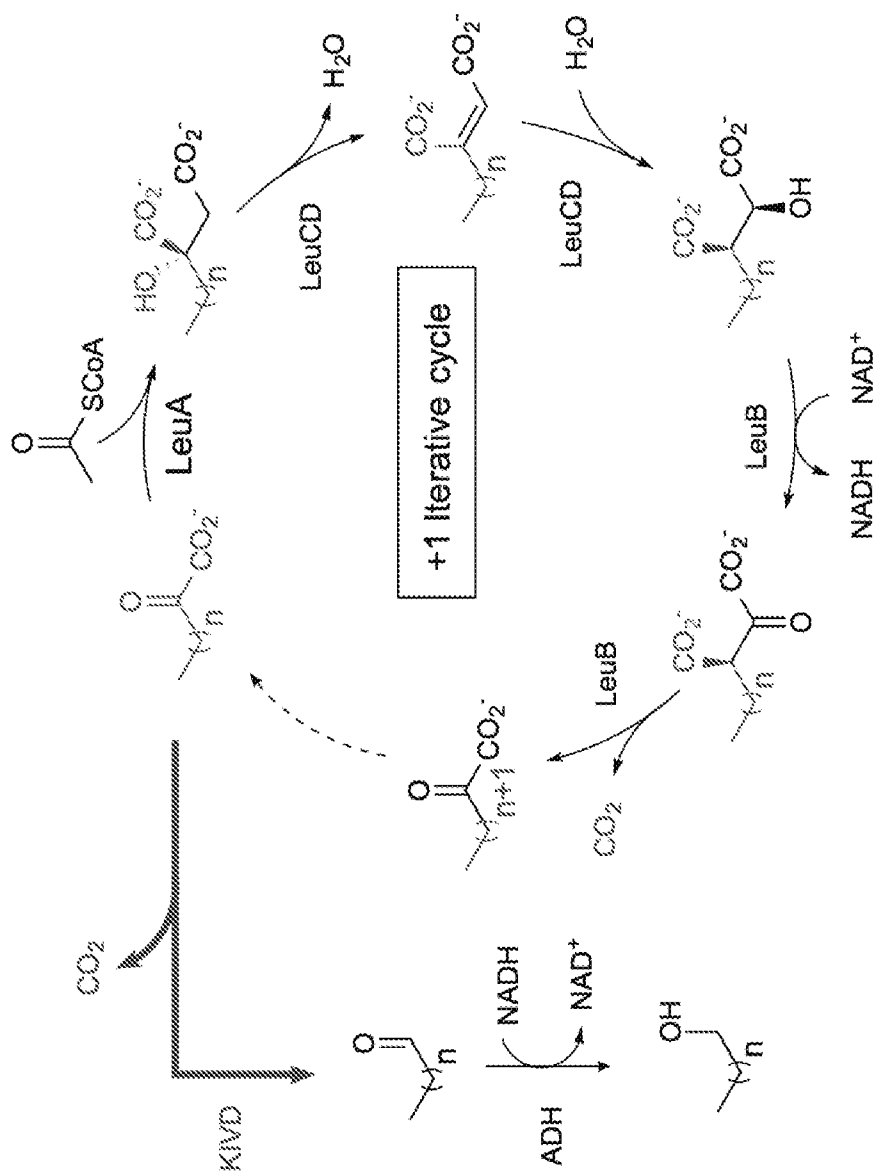


Figure 1 The synthetic recursive +1 pathway. This pathway employs enzymes LeuABCD from leucine biosynthesis for carbon chain elongation of 2-ketoacids. The enzyme ketoisovalerate decarboxylase (KIVD) performs the decarboxylation of 2-ketoacids and diverts carbon out from the +1 iterative cycle.

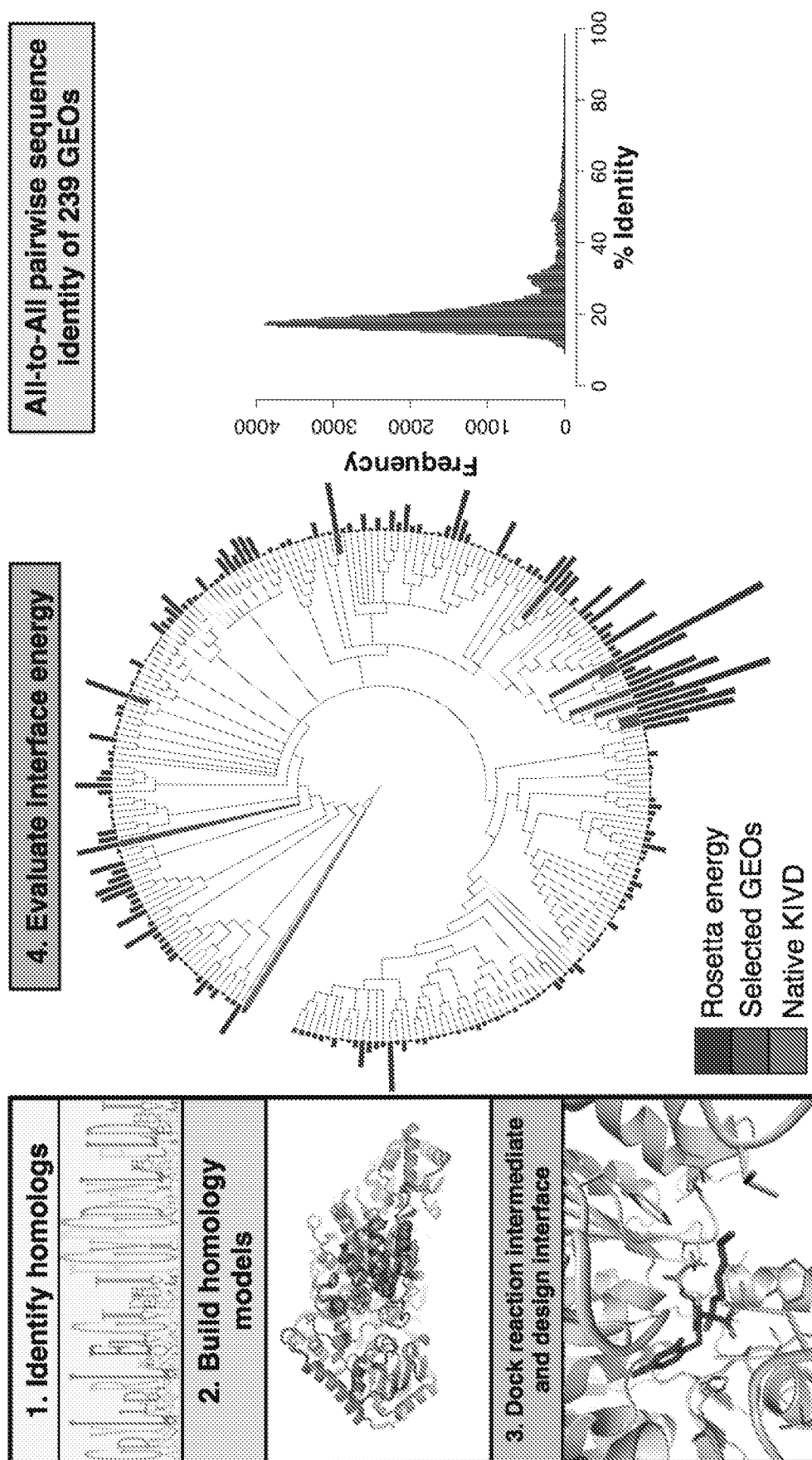


Figure 2. The computationally directed genomic enzyme-mining pipeline. (Left) GEOs were identified based on sequence homology to KIVD. Bioinformatics filters were used to identify a set of protein sequences from genomic databases that are predicted to be KIVD homologs and likely to be decarboxylases. Once identified, homology models were built to obtain a predicted ternary structure of each GEO. Ligand docking and design simulations were subsequently run in the presence of our target ligand (e.g. C8) to evaluate the potential protein-ligand interface energy. **(Middle)** a phylogenetic tree for the 239 GEOs was depicted with a bar chart above each sequence. The bar height indicated the predicted protein-ligand interface energy, the higher the bar the lower the energy. Bar height was scaled linearly relative to the lowest protein-ligand interface energy. 10 GEOs (Brown) were selected for experimental characterization. **(Right)** The pairwise sequence identity of all 239 GEOs to each other has a mode of 20%, indicating how diverse this set of sequences is.

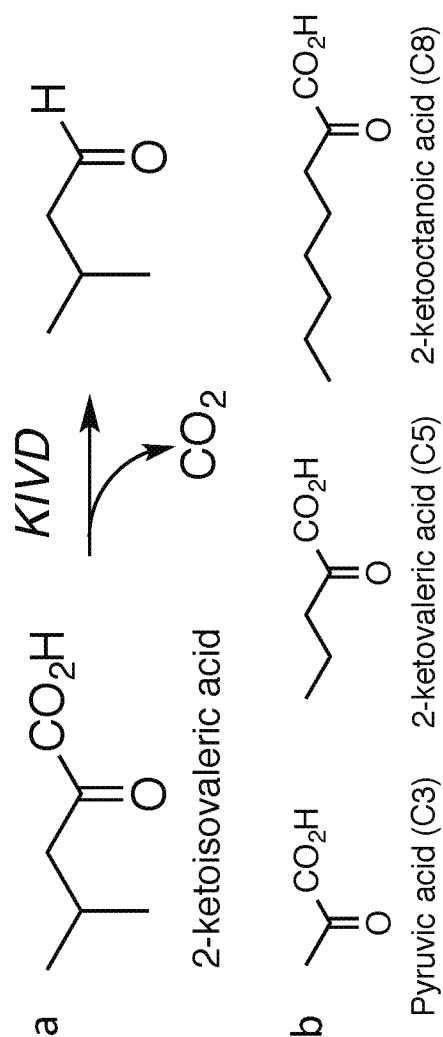


Figure 3. Characterized reaction specificities. (a) Ketoisovalerate decarboxylase (KIVD) utilizes various ketoacids as substrates to produce the corresponding aldehyde. (b) Ketoacids used for *in vitro* kinetic constants in addition to 2-ketoisovaleric acid.

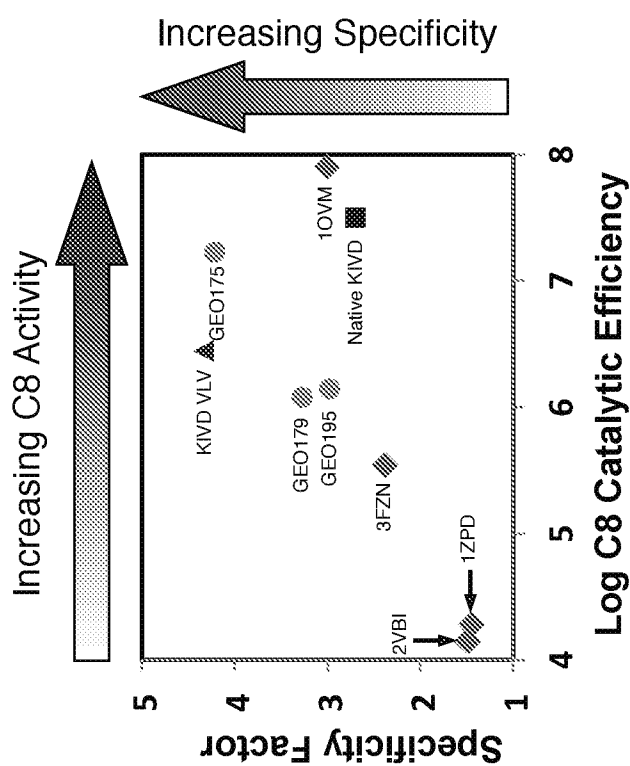


Figure 4. Catalytic efficiency and specificity of characterized ketoacid decarboxylases. Specificity factor is calculated as where the log scaled catalytic efficiencies for C8 relative to C5 and C3 are compared. The three active GEOs and the naively selected set of decarboxylases are represented in gold and green, respectively. Native KIVD and KIVD_VLV are depicted in black and blue, respectively. The genomic mining method and computational active site redesign approach both produced enzymes with enhanced specificity relative to native KIVD.

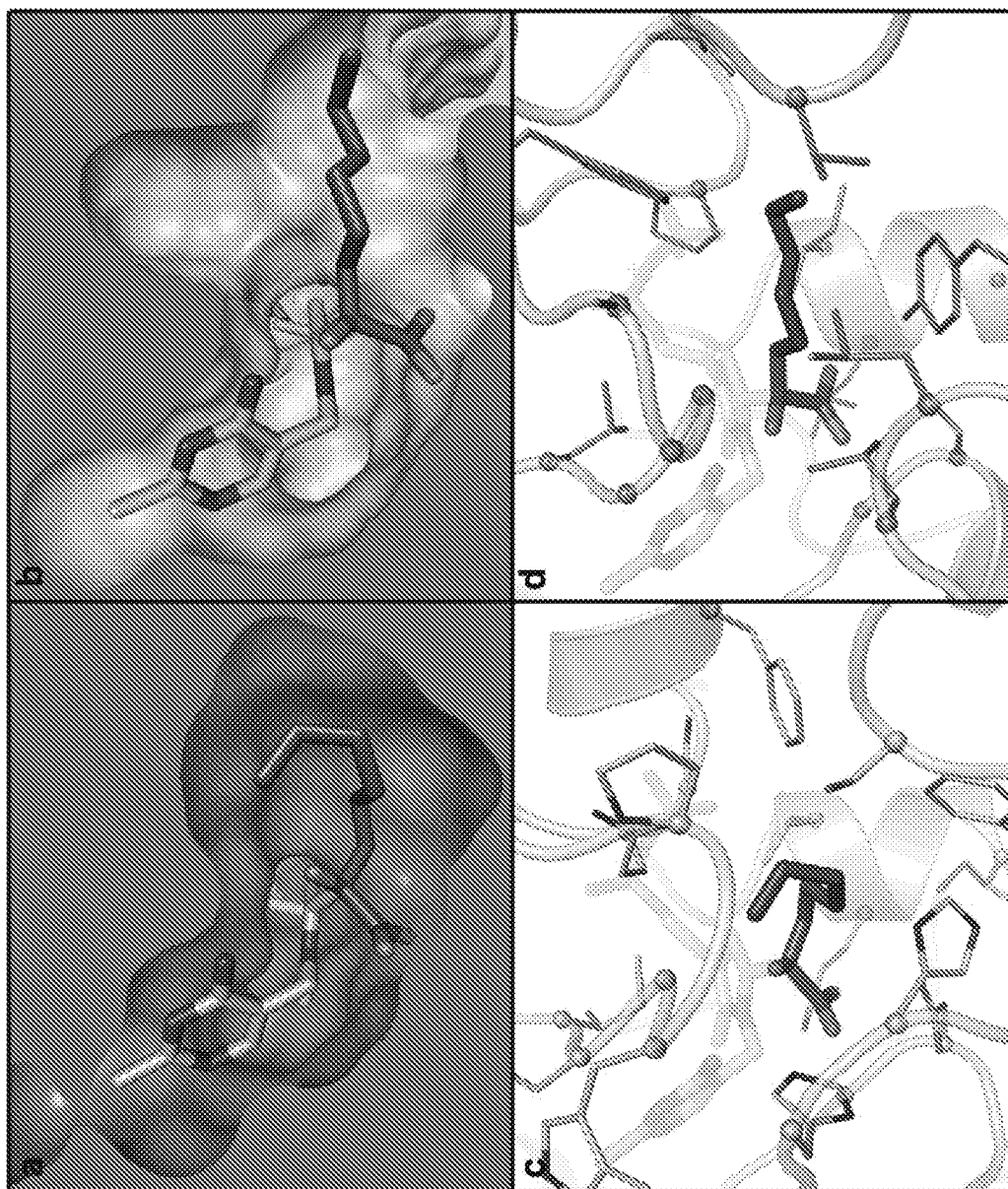


Figure 5. Molecular model of the C8 intermediate docked into KIVD (a, c) and GEO 175 (b, d). Cross sections of the space filled active site are represented in a and b where the C8 keto acid and TPP ligand are shown in brown and grey respectively. A detailed view of the amino acids within the active site are illustrated in c and d. Residues within 5 Å of C8 ligand are shown in sticks and their corresponding C-alpha are shown in spheres. Figure was generated using PyMol v1.7.4.³⁴

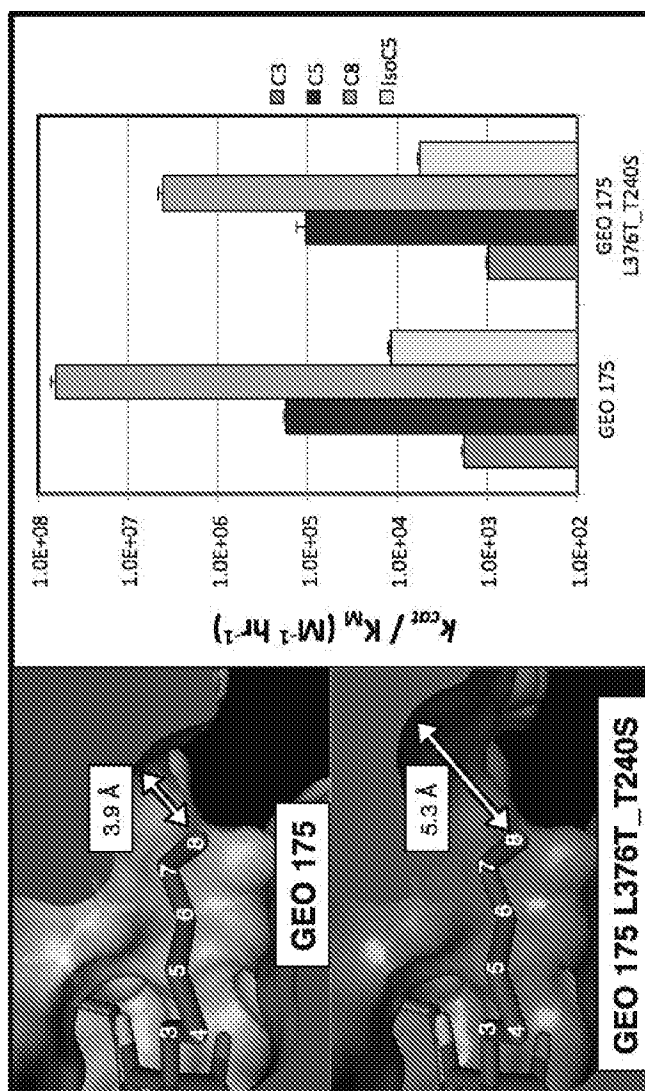


Figure 6. Kinetic characterization of GEO 175 and GEO 175 L376T_T240S. Kinetic constants were measured as described in materials and method. The substrate 2-ketooctanoate carbon chain numbering referred to in the text is numbered in white. According to the molecular model of GEO 175, the double mutation L376T_T240S is predicted to recede the pocket by 1.4 Å and remove interactions between the binding pocket and carbon 8 of the ketoacid alkyl chain. This mutant was observed to decrease catalytic efficiency on C8, but has a negligible effect on shorter chain substrates.

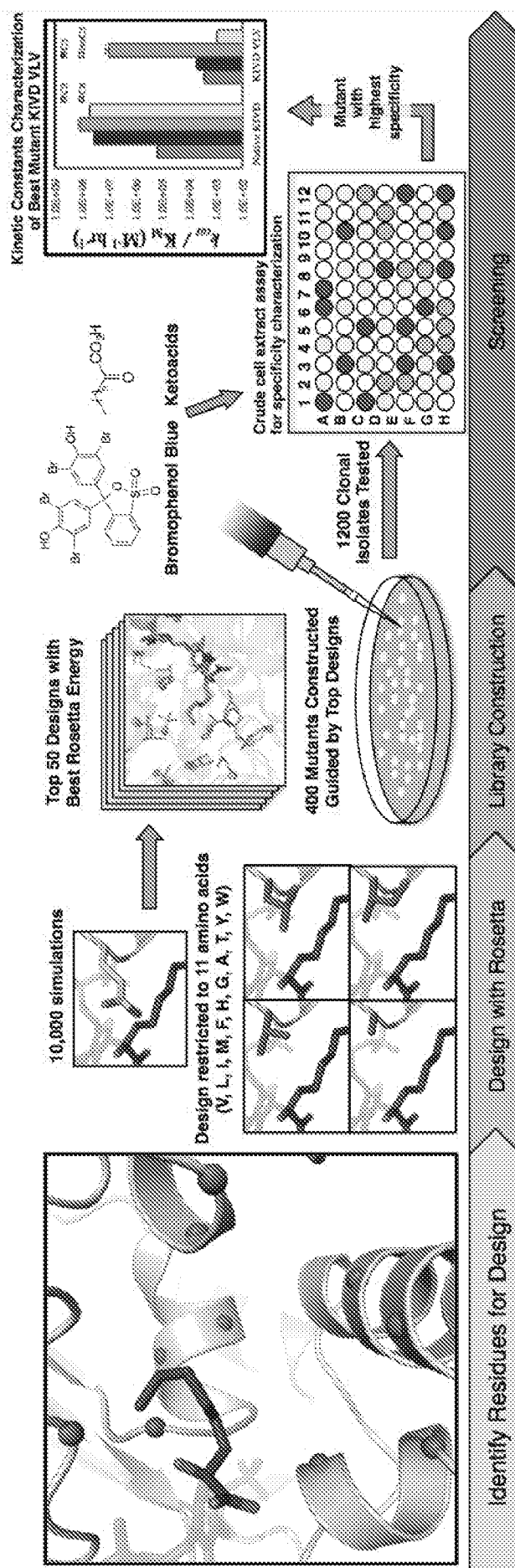


Figure 7. Screening process for identifying an engineered KIVD with altered substrate specificity. The design process started with identifying active site residues of KIVD (depicted in red). These amino acids were allowed to either remain native or sample any of eleven relatively hydrophobic amino acids. A total of 10,000 design simulations were ran and the amino acids identified in the 50 lowest energy designs were used to guide construction of a small library of roughly 400 KIVD mutants from the original combinatorial space of 10^{11} possible active site mutations. 1200 clonal isolates were screened for activity and specificity. The KIVD_VLV mutant was selected and subsequently purified for *in vitro* kinetic constant characterization.

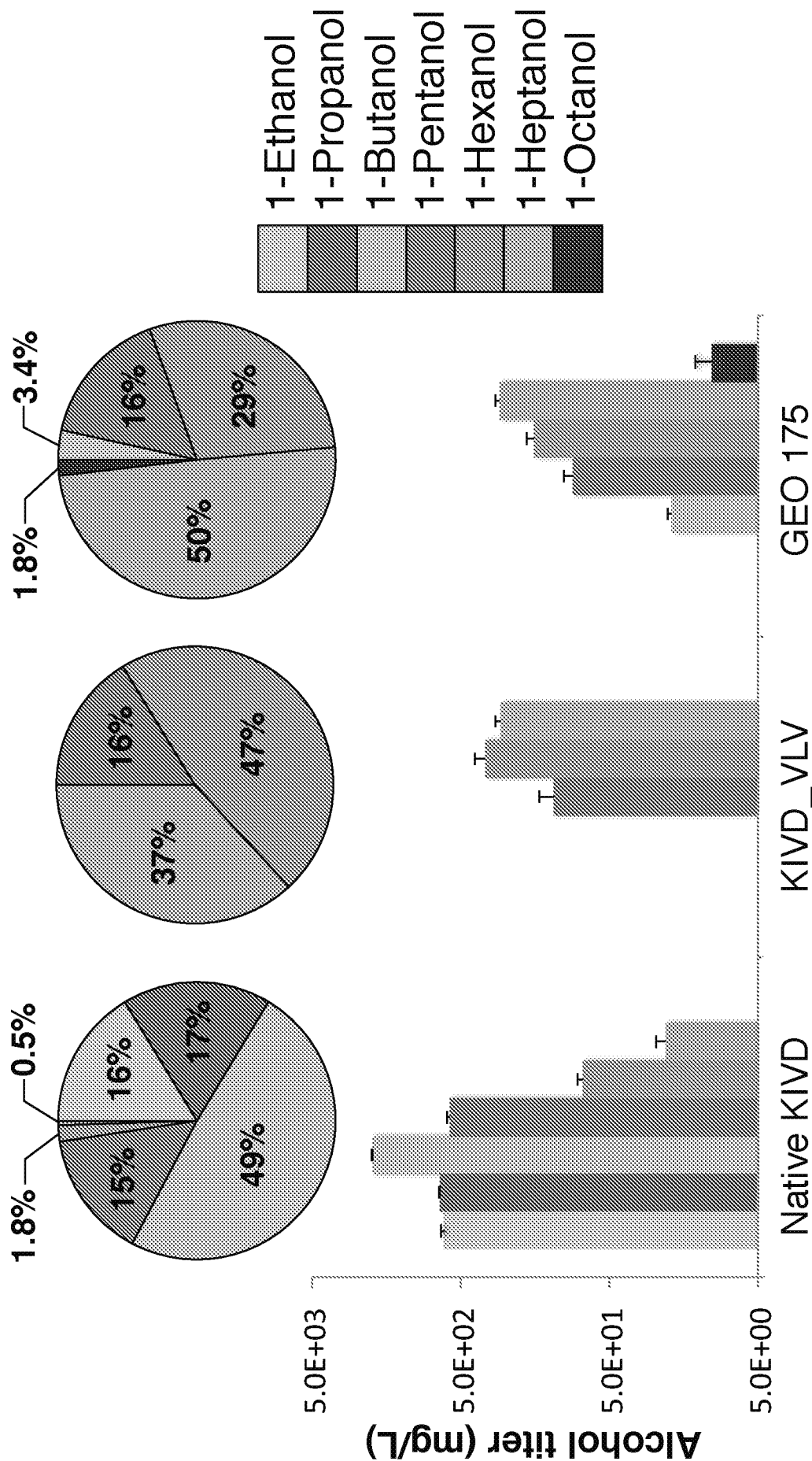


Figure 8. *In vivo* alcohol production of the synthetic recursive +1 pathway with native KIVD, KIVD_VLV, and GEO 175. Cells were incubated for forty hours in microaerobic conditions in a defined media as described in materials and methods. Each assay was performed in triplicate and titers reported if all three samples had observed product production above the limit of quantitation (5 mg/L).

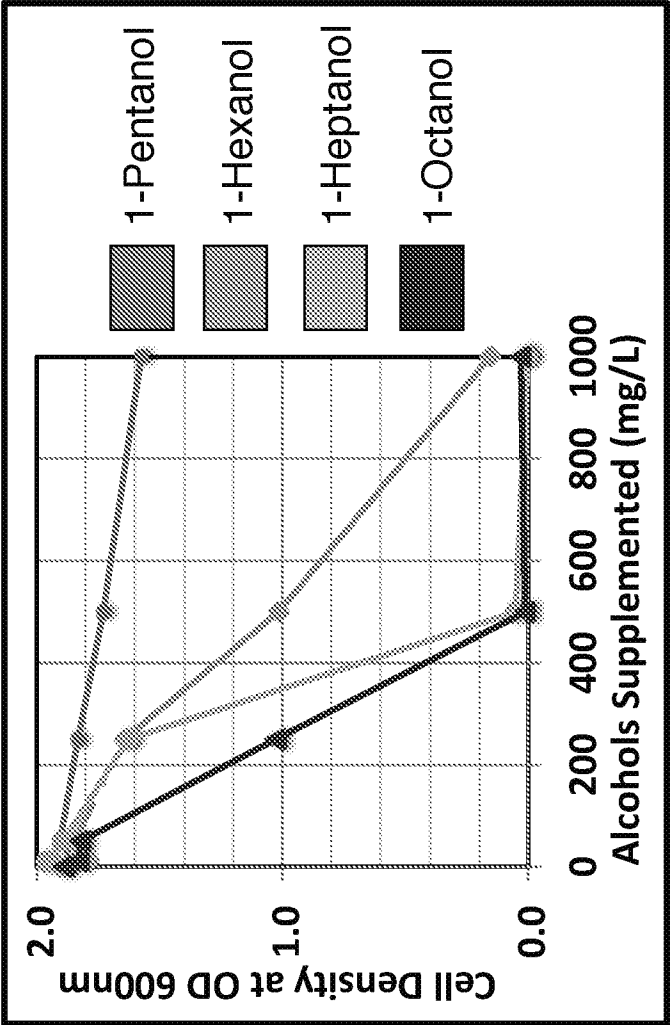


Figure 9. Toxicity as a function of alcohol chain length. The optical density of the engineered *E. coli* strain after a 7-hour growth was measured as a function of the alcohol concentration added to growth media at the beginning of incubation.

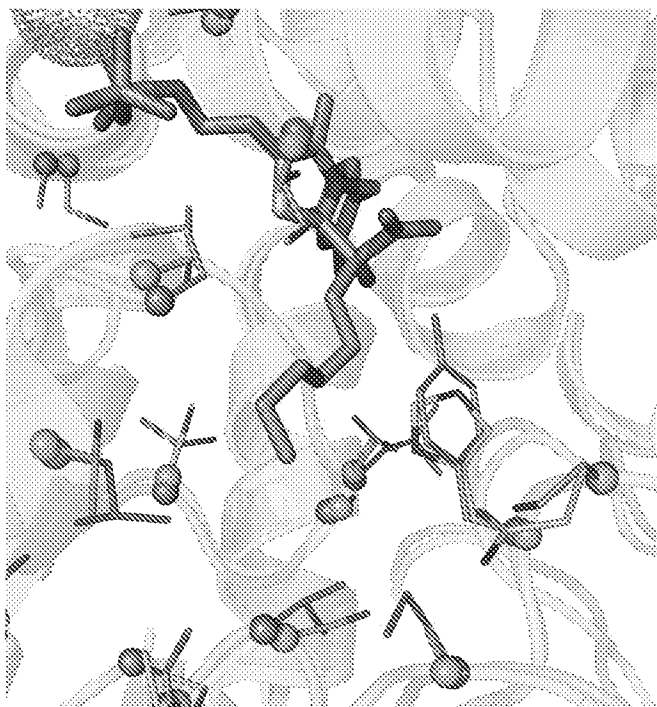


Figure 10. Overlay of active sites between GEO 175 (orange) and BFD (green). Residues that are different between the proteins are shown in sticks and their corresponding alpha carbons are shown in spheres.

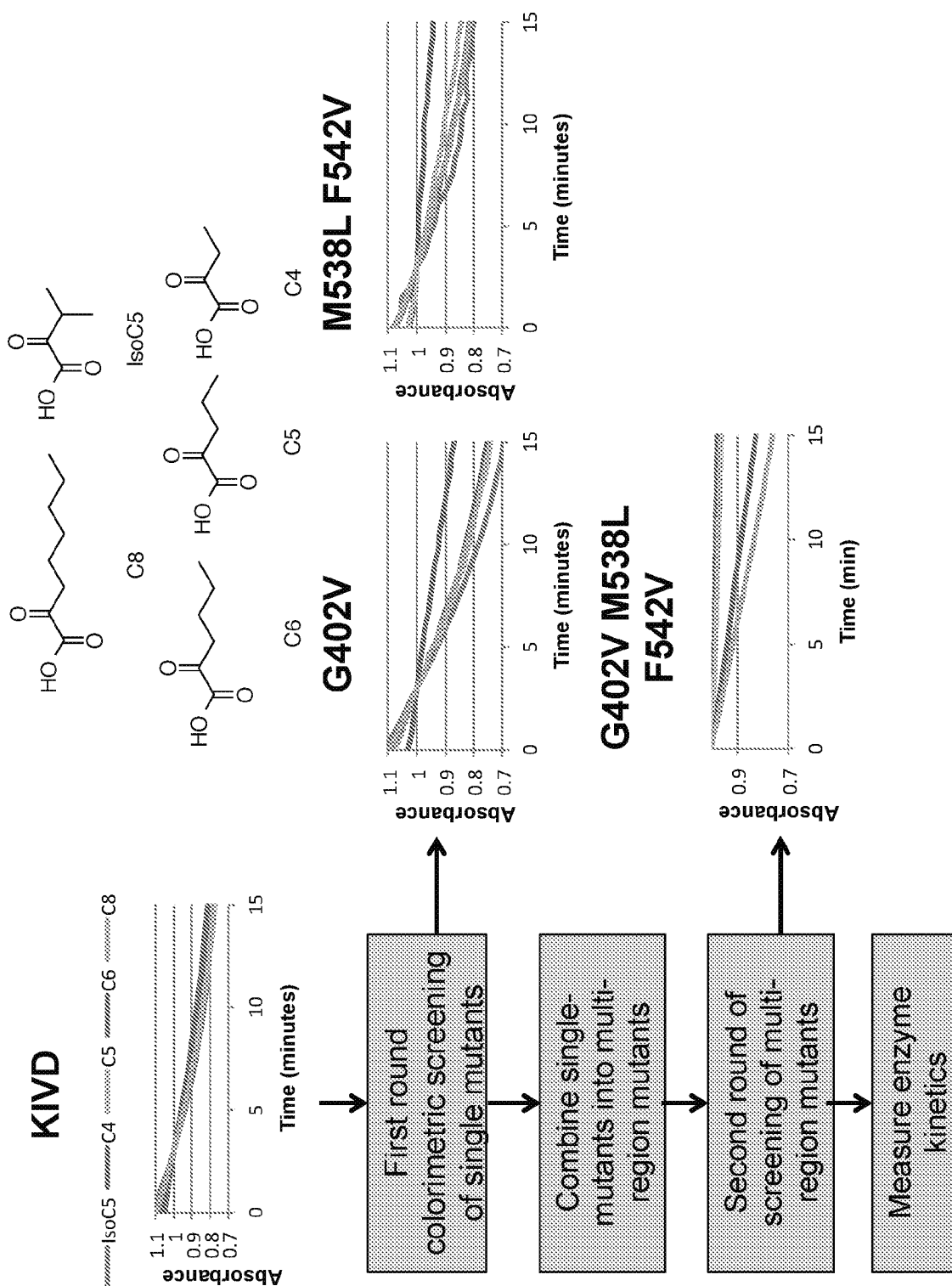


Figure 11. Screening process for identifying engineered KIVD with altered substrate specificity. Pertinent screening results/candidate mutants are shown (graphs) which contributed directly to finding KIVD_VLV. Colorimetric assays (graphs) measured approximate enzymatic activity of screened mutants for ketoisovalerate (IsoC5), 2-ketobutyrate (C4), 2-ketovalerate (C5), 2-ketohexanoate (C6), and 2-ketooctanoate (C8). Higher activity is shown by larger downwards slope of [Absorbance]/[Minute]; greater specificity, relative Kivd, was observed in mutants with reduced activity (lower downward slope) for IsoC5, C4, and C5, and only slightly affected or greater activity (similar or larger downward slope) for C6 and C8. First round high-throughput colorimetric screening of single mutants identified candidates such as G402V and M538L/F542V. Mutations from each region that showed potential in altering the specificity of KIVD were then combined to produce multi-region mutants through PCR-site directed mutagenesis using the plasmids of KIVD single region mutants as templates. These KIVD multi-region mutants, such as G402V M538L F542V (KIVD_VLV) were then screened a second time with the colorimetric assays. Enzyme kinetics was then measured for mutant KIVD_VLV.

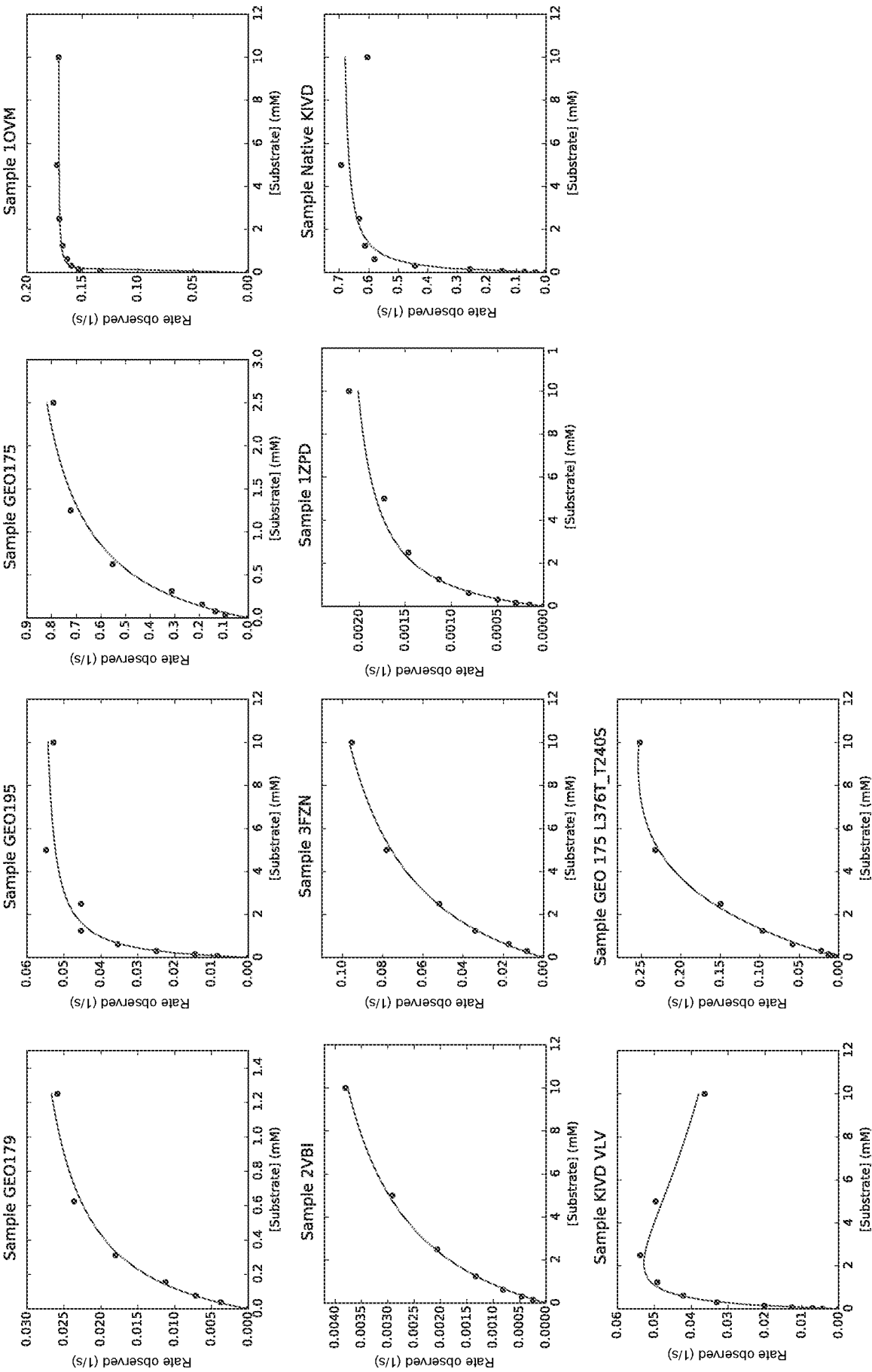


Figure 12. Michaelis-Menton curve fit of tested enzymes with C8 substrate.

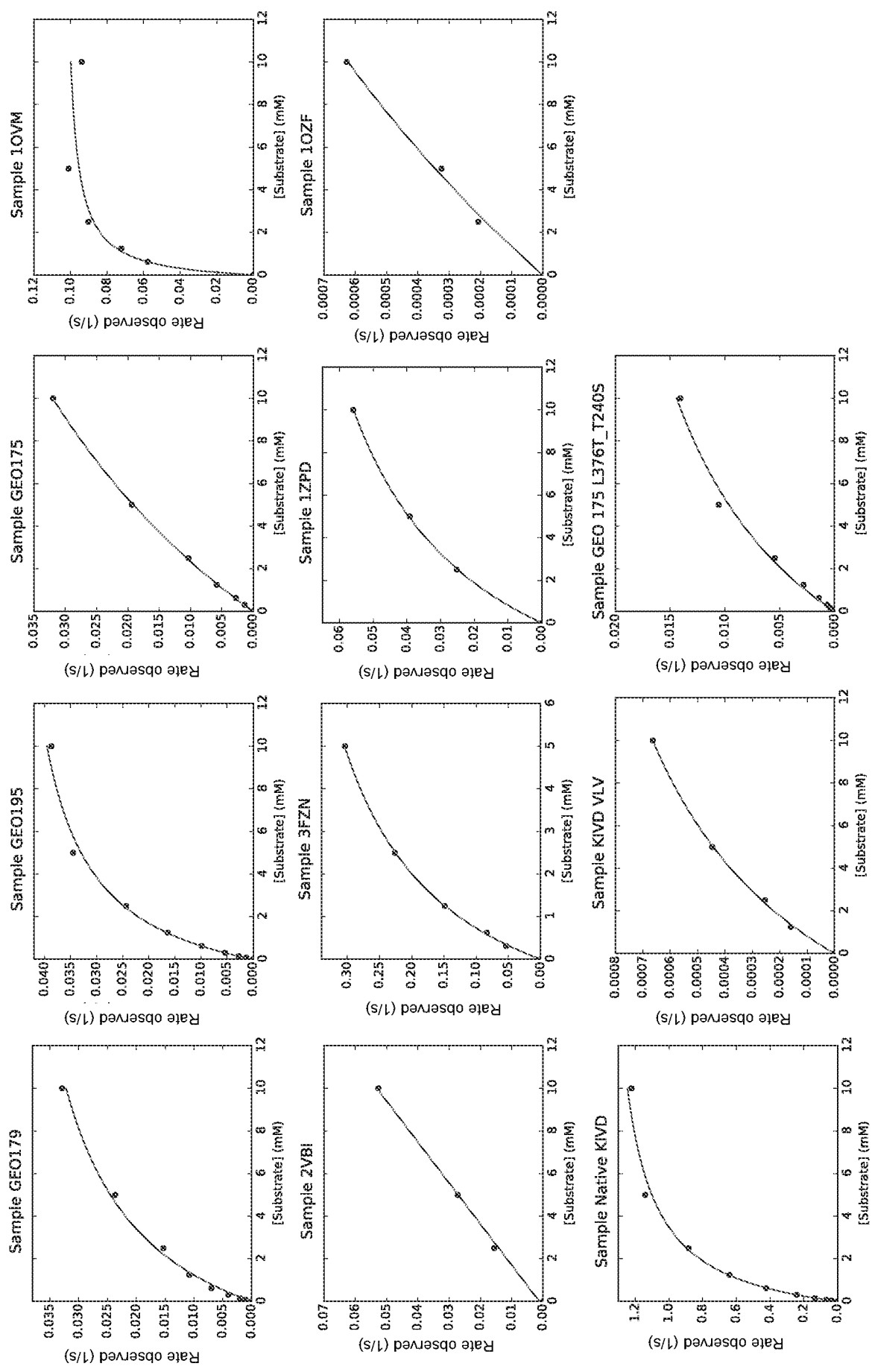


Figure 13. Michaelis-Menton curve fit of tested enzymes with C5 substrate.

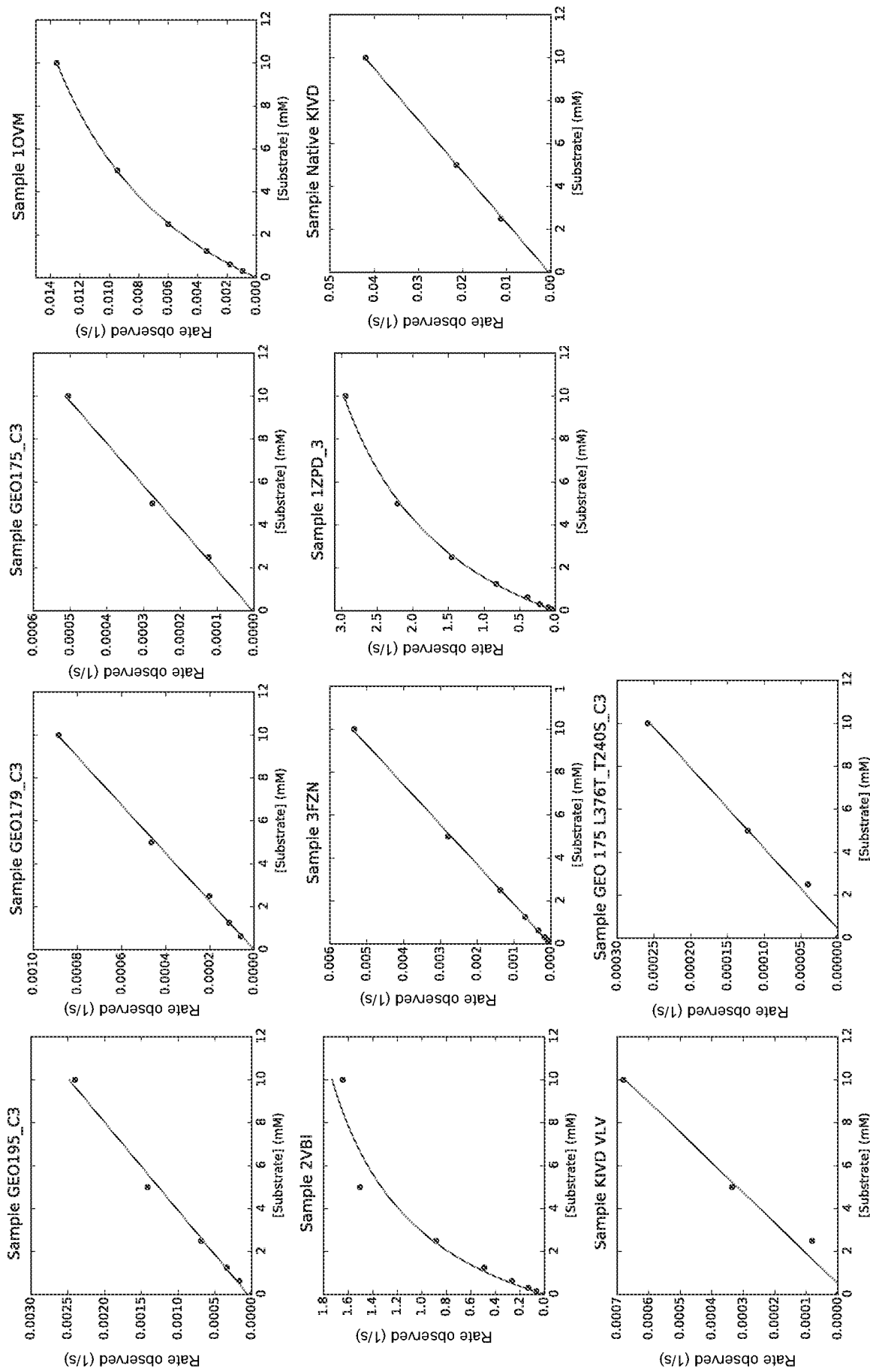
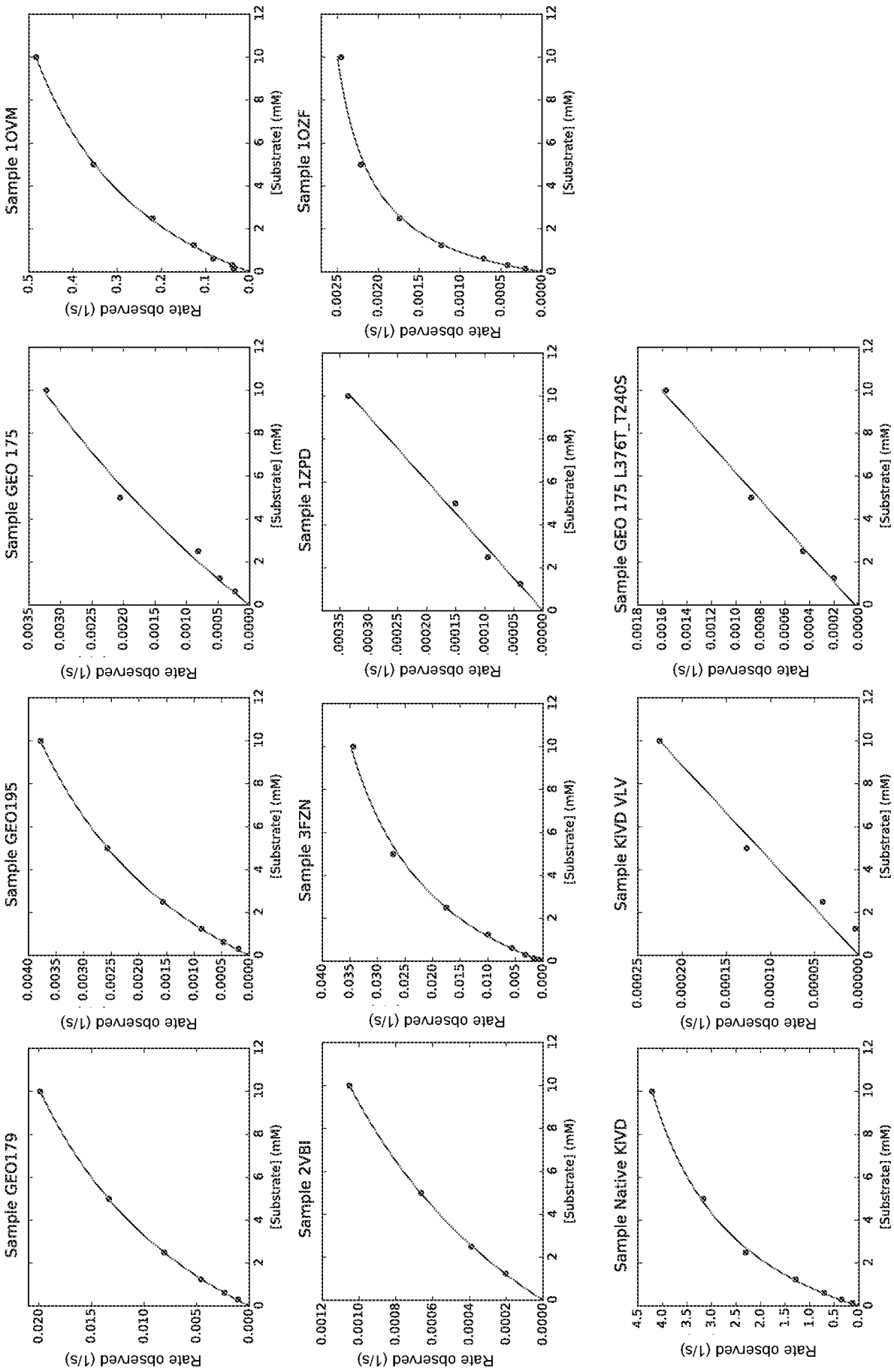


Figure 14. Michaelis-Menton curve fit of tested enzymes with C3 substrate.



Michaelis-Menton curve fit of tested enzymes with isoC5 substrate.

Figure 15.

1	2VBG	41	ZP_08570611	81	YP_001208174	121	ADP12107	161	O53554	201	YP_001070609
2	YP_004211875	42	YP_001236700	82	YP_830466	122	YP_004115182	162	YP_003909136	202	YP_004418205
3	YP_004506975	43	YP_001208817	83	YP_675840	123	ZP_08465980	163	ZP_08198305	203	YP_675476
4	ZP_07949728	44	YP_002353797	84	ZP_06842871	124	ZP_06711201	164	ZP_07042487	204	YP_675461
5	ZP_07380174	45	2Q5Q	85	ZP_05126709	125	ZP_08142152	165	YP_004210411	205	ZP_05074445
6	YP_004116674	46	YP_002379932	86	ZP_05094980	126	YP_511062	166	YP_925359	206	YP_004524050
7	ZP_08307744	47	O33112	87	ZP_07031890	127	ZP_07657182	167	CAE38656	207	ZP_06822743
8	ZP_08461032	48	POA623	88	YP_004416763	128	ZP_07356713	168	ZP_01901192	208	ZP_01044868
9	1OVM	49	ZP_04998899	89	EGD05600	129	ZP_05100555	169	ZP_07973795	209	ZP_05844198
10	ADP13207	50	YP_004291227	90	ZP_08543603	130	YP_001240610	170	CAM60234	210	ZP_07985939
11	YP_001281252	51	YP_001353840	91	ZP_06271645	131	YP_001237056	171	ZP_06711620	211	BAJ49000
12	YP_001177638	52	3D7K	92	ZP_06273171	132	AAM12352	172	YP_00420384	212	ZP_01626788
13	O53865	53	YP_004519441	93	ABI47993	133	YP_001768986	173	YP_002133221	213	ZP_06710664
14	ZP_04562872	54	CAP40738	94	ZP_07282849	134	YP_004216045	174	YP_001378281	214	YP_509362
15	AAL18557	55	ZP_08275845	95	YP_001208015	135	YP_001280485	175	ZP_07290467	215	YP_004076523
16	Q9CDB6	56	YP_004523976	96	ZP_06370153	136	ZP_00956146	176	BAD32743	216	ZP_08631155
17	YP_003731250	57	ZP_07706802	97	ZP_05125236	137	ZP_050333599	177	YP_004522883	217	AAR37826
18	AEG26729	58	YP_004081235	98	ZP_06757044	138	ZP_08591801	178	YP_004083736	218	ZP_01302864
19	ZP_01623010	59	ZP_08124815	99	YP_830089	139	ZP_06412949	179	ZP_06547677	219	ZP_05001859
20	YP_003137934	60	ZP_07949857	100	YP_004230473	140	CAM62618	180	ZP_04607738	220	CAP43929
21	ZP_06368393	61	ZP_08303118	101	YP_002536327	141	CAE37944	181	ZP_05135002	221	ZP_03267724
22	YP_002380104	62	YP_987677	102	ZP_01130961	142	CAP42719	182	ZP_08119462	222	ZP_06418208
23	YP_004674227	63	YP_686086	103	YP_002377567	143	YP_01999289	183	ZP_07280171	223	ZP_07216338
24	YP_004646634	64	ZP_05068817	104	YP_001242472	144	ZP_08124133	184	ZP_07380799	224	ZP_01886096
25	ZP_08317680	65	ZP_06549025	105	YP_004075455	145	ZP_01902746	185	YP_641501	225	YP_004695810
26	AAM49566	66	YP_003810578	106	YP_004505036	146	YP_003607818	186	YP_004085144	226	ZP_04997132
27	YP_004017390	67	YP_004524326	107	ZP_07378926	147	ZP_07280108	187	YP_830320	227	AD117252
28	YP_001224592	68	ZP_07282256	108	ZP_07284792	148	ZP_08317504	188	ZP_04996569	228	ZP_05780727
29	2VBI	69	BAF52673	109	ZP_07278109	149	YP_001769586	189	CAQ52617	229	CAE50484
30	ZP_01125207	70	YP_003909474	110	CAK95977	150	ZP_08272344	190	YP_001684628	230	CAC11911
31	YP_004210504	71	YP_004212185	111	YP_004490314	151	ZP_05102608	191	AAA17301	231	ZP_06273867
32	1ZPD	72	YP_001279645	112	CAK95978	152	ZP_01306645	192	Q9HUR2	232	YP_001275313
33	YP_001806325	73	CAE45665	113	BAC79260	153	YP_001505772	193	YP_004080965	233	YP_001240047
34	AAA27685	74	YP_831380	114	ZP_08387358	154	YP_001203581	194	AEQ30045	234	ZP_06712511
35	YP_730696	75	YP_370152	115	ZP_07278993	155	ZP_07703403	195	ZP_06846103	235	ZP_07033476
36	YP_003797329	76	ZP_05101111	116	YP_004230546	156	YP_004416777	196	ZP_07281735	236	YP_004081191
37	ZP_08315793	77	AAR05436	117	ZP_06841705	157	ZP_07282273	197	YP_047867	237	ZP_08333578
38	ZP_01079084	78	YP_001203568	118	ACR33042	158	ZP_07674873	198	YP_004317430	238	ZP_01725425
39	ZP_08074854	79	YP_381143	119	ZP_01749177	159	YP_481591	199	YP_004534047	239	ZP_05285963
40	CBE69591	80	CAC11945	120	ZP_08137686	160	NP_484471	200	YP_004077246		

Figure 16. PDB code and accession numbers of GEOs modeled as listed on the phylogenetic tree depicted in Figure 2. Sequences were obtained from homologous sequence search in the Non-Redundant database using HMMER3.

Protein Sequences	
3FZN	MASVHGTYTLLRRQGDITVFGNPGSNELPFKDPEDFRYLALQEAQVVGADGYAQSRKPAFINIHSAGTGNAMGALSNAWNSHSPITVAGQQTIRAMIGVEALLTNVDAANIPLPLVKWSEYEPASAAEVPHAMISRAIHMASMAPQGPPVLSVPYDDW DKDAPQSHHFLDRHVSRLINDODLILVKALNSASNPAILGPDVDAANANADCVMLAERLKAPVWVAPSAKPCPEPTRHPCFRGLIMPAGIAASQLLEGHDVVLVIGAPVFRYHQYDPGQYLKPGTRLSVTCDDPLEAARAPMGDAIVADIGAMASALNV ESSRQLPTAAPAPAKVDDQDAGRLHPETVFDTLNDMAPENAYLNESTSTTAQMWQRLNMRNPGSYFFCAAGGLGFALPAAGVQLAEPERQVIAVIGDGSANYSALWTAQAQNYPIPIFVIMNNGTYGALRWFAGVLEAENVPLGLDVP/GIDFRALAKGYGVQA LKADNLEQLKSLQEAIAKGPVLEIVSTSPVKGSTENLYFQSGALE
ZVBI	MTYIVGMVLAERLYOIGLHFFHFAVAGDYNVLIDQLINKDMQKQYCCNEINCGFSAEYARSNGAAAVVTSVGAISAMNALGGAYAEINPVILISGAPNSNDQGTGHILHHTGKTQSYQLEMARQVTCAESITDAHSAKIDHIVRTALRERKPAYIDAC NIASEPCVRPGVPSLLSATEDEHTSLTAAVADATVALLKESASPMVLLGSKLRAANALAAATETLADKLQCAVTIMAAAKGFFPEDHAGFRGLYWGVEVSNPGVQELVETSDALLCIAPVFNDSYTVGWSAWPKGNVILAEPRVTV/DGRAYDGTFLRAFLQALAEKAP ARPASAKSSVPTCSLTSDEAGLTNDEIVRHINALLTSNTTLVAETGDSWFNAMRMTLPRGARVELEMQWGHIGWSPVSAFGNAMSGSDRQHVVMVVDGSGFQLTAQEAQAMVRYELPVIIILINNRGVVIEIAIHDPVNYIKNWWDYAGLMEVFNAGEGHG LGIKATTTPLEAARAKANTGPTLIECQDRTDCTDMLVQWGRKVASTNARKTLTAGSTENLYFQSGALE
ZPD	MSYIVGTIYLAERLYOIGLHFFHFAVAGDYNVLIDQLINKDMQKQYCCNEINCGFSAEYARSNGAAAVVTSVGAISAMNALGGAYAEINPVILISGAPNSNDQGTGHILHHTGKTQSYQLEMARQVTCAESITDAHSAKIDHIVRTALRERKPAYIDAC NIASEPCVRPGVPSLLSATEDEHTSLTAAVADATVALLKESASPMVLLGSKLRAANALAAATETLADKLQCAVTIMAAAKGFFPEDHAGFRGLYWGVEVSNPGVQELVETSDALLCIAPVFNDSYTVGWSAWPKGNVILAEPRVTV/DGRAYDGTFLRAFLQALAEKAP ARPASAKSSVPTCSLTSDEAGLTNDEIVRHINALLTSNTTLVAETGDSWFNAMRMTLPRGARVELEMQWGHIGWSPVSAFGNAMSGSDRQHVVMVVDGSGFQLTAQEAQAMVRYELPVIIILINNRGVVIEIAIHDPVNYIKNWWDYAGLMEVFNAGEGHG GYDGAAGLAKKTGGELAEIAIKVALANTDGTPTLIECFIGREDCTEELVKWGRKVAASNRKPNKVVGSTENLYFQSGALE
ZP_08570611	MISSINFTVADYLLTRLHQLGKRVQVPGDYVWVNFMDALEQFNGIEAVGDTLTELGAAGYAADGYARLTGIGAVSQVQGVGTFHSLNAJAGSVERNPVVWITASPSTGNRKTIKETGVLFHHTGDLADSKVFNVTAAEVLSDPSDARQKIDKALTCAITFRPIYL EAWQDVWGLACEKPEGELKALPLISEGALKAMLADSKLLNSARQPLVLGVEINRFLQDQAVLIDLKASGLPYSTTSKAKTVISENEGIFVGTYADGASFPATVEYIEKADCVLALGVFTDDYLTMLSKQFDQMIVVNNDETSRLGHAYYHQLYLADFILQLTDEIKKS SLYPRONSALPLIPQOQTPALLQQQLSYQNFFDLFYGLLQHLQDNLISILGESSLYMSARLYGLPQDSFIADAAGWGLSILGHETGCVTGIAYSADKRAMAAGDGGFMMMCQCLSTISRHLQNSWFEVSNKYAIEQSFVDICAFKGGHFAPFDLLPTWDYLSL AKAFSEGVYRVQNGEELLQALEHIMTKQKDPALVEVVIQSDQLAPAMAGLVKSITGHTVEQCAIPTGSTENLYFQSGALE
ZP_831380	MTTVHAAAYELLRSNRLTIIFGNPGDNEPFLDAIMPADFRYLIGLHGGVVGWADGFAQAGQAAAFVNLHAAAGTGNAMIGALTNAWYSHTPLVTAGQQVPRPMIGLMEILSNVDAASIPRLVKWSAEPAQAPDVPRALSQAIHTATSDPKGPVYLSIPYDDW NQDTGNISEHLSSRSVRAGNPSAEQLDDLSALREANPALVFGPDVDAARANHHAVRLAEKLAAPVWIAAPAPRCPPEPTRHPNFRGVLPASIASIGALLNGHDLVIGAPVFRYHQYQPSYLPENSRLIHITCDAGEAARAPMGDALVADIGQTLRALADIPOQ KRPLRPVIPPVPSQSDILLAPDAVFEVMNEVAPEDVWVWVNESVSTVTALWERVELKHPGSYFPASGGGLGFGMPAAVGVQLANDRRRRIAVIIGDGSANYGITALTWTAQAEKIPVVFIIILNNGTYGALRAFAKLLNAENAAGLDVP/GICFCAIEGYGVEAHITSL ENFKDKLSAALQSDIPTLLLEVPSTSTSPFGSTENLYFQSGALE
ZP_06547677	MKTHSAAYALLRRHGMITIFGNPGSNELPFKSPEDFQYVLGLHGGVVGWADGYALASGRPATVNLHAAAGTGNMGALTNSWYSHSPVLTAGQQVPRPMIGLMEILSNVDAASIPRLVKWSAEPAQAPDVPRALSQAIHTATSDPKGPVYLSIPYDDW DQPSGPGVEHLIERDVTAGTPDARQLQVLVQQVQDARNPVVLVIGPDVDTLSNDHVALADKLMPVWIAAPASRCPPEPTRHPSFRGVLPAAIAGISKTLOGHDLIIVVGAPVFRYLQFAPGDYLPVGAQLLHITSDDPLEATRAPMGHALVGDIRETLRLVLAEEV QQSRPYPEALAAPECVTDEPHHLHPETLFDVLDVAPHDAIWKESTSTVTAFWQRMNLRHPGSYFPAAAGGLGFGLPAAVGVQLAQPPRRVVALIGDGSANYGITALTWTAQAEKIPVVFIIILNNGTYGALRWFAVGLKAEDSPGLDVPGLDFCAIAKGYGVKAVH TIDTRDSFEALRTALDANEPTVEVPTLTIQPHGSTENLYFQSGALE
TOVM_P23234	MIRTPYCVADYLLDRLTDCGADHFLGVPVGYNLQFDHVIDSPDICWYGCANELNASYAADGYARCKGFAALLTIFGVGELSAMNGIAGSYAEHPVPHVYVGTAAQQRGELHHTLGDGERHHYHMSPTVAAQVLTETQACYEIDRVLTMLERRPGYL MLPADVAKKAATPPVNALTHKQAHADSACLKAFRAAENKLAAMSKRTALLADFLVLRHGLKHALQKWWKEVPMHAHTMLMGKIFDERQAGFYGYSASTGAVKEAIEGADTVLCVGTFRFTDLTAGFTHQLTPAQITVQPPHAARVGDVWFTGPMNQAI ETLVELCKQHVHAGLMSSSSGAIPFPQDGSLTQENFWRTLQTFIRPGDILLADQGTSAFGAIDLRIPADVNFVQPLWGSIGYTLAAAFGAQACPNRRRVNLTGDDAAQLTIQELGSMILRDQKPHIILVLINEGYTVVERAIGHAEQRYNDIALWNWTHIPQALSLDP QSECWVSEAEQLADVLEKVAHHERLSLEVMLPKADIPILLGALKALEACNAGSTENLYFQSGALE

Protein sequences for all genes used in this study.

Figure 17.

Protein Sequences

CP_06418208

CP_07290467

CP_07282849

CP_06846103

CP_04996569

CP_381143

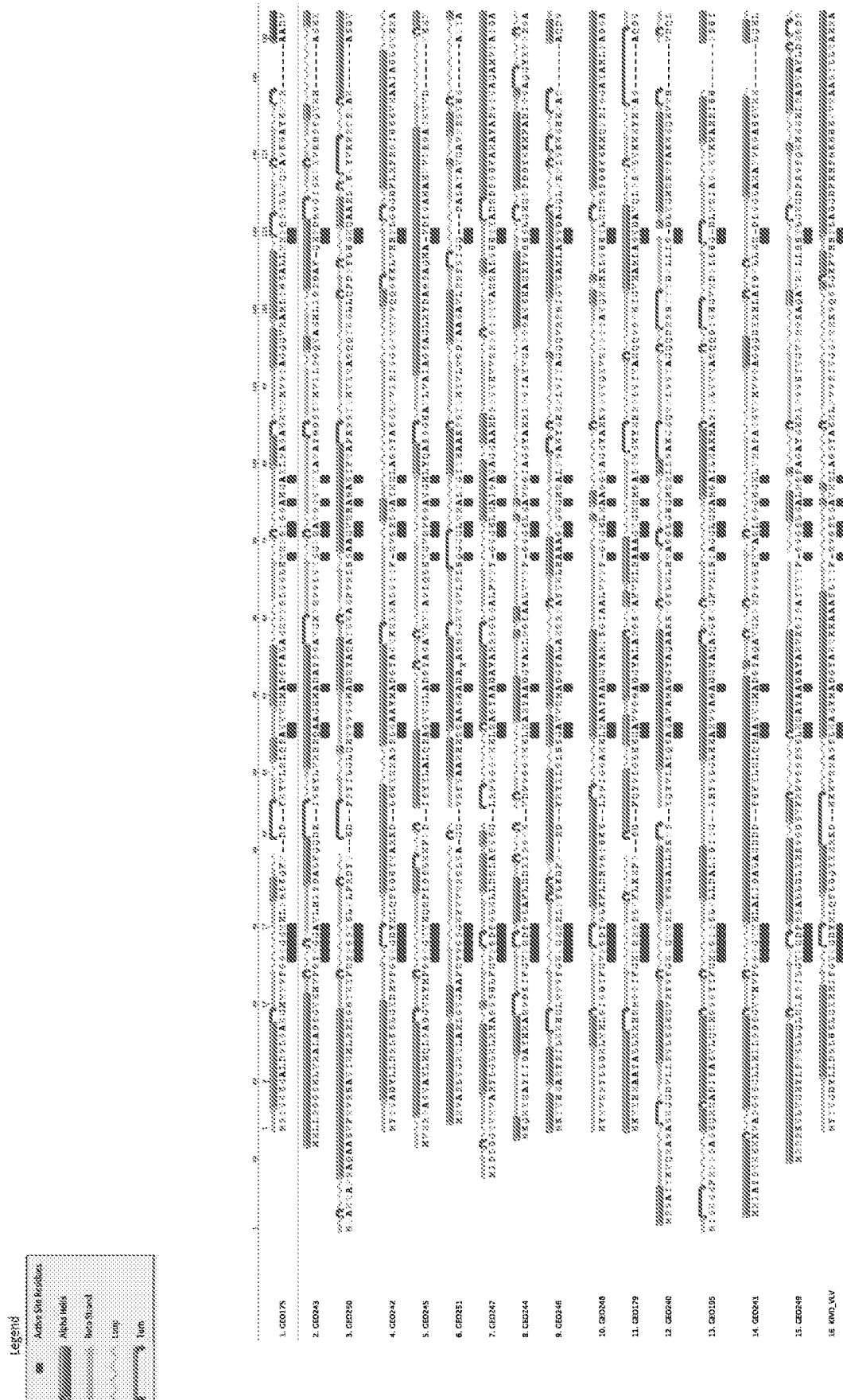
Figure 17 (continued). Protein sequences for all genes used in this study.

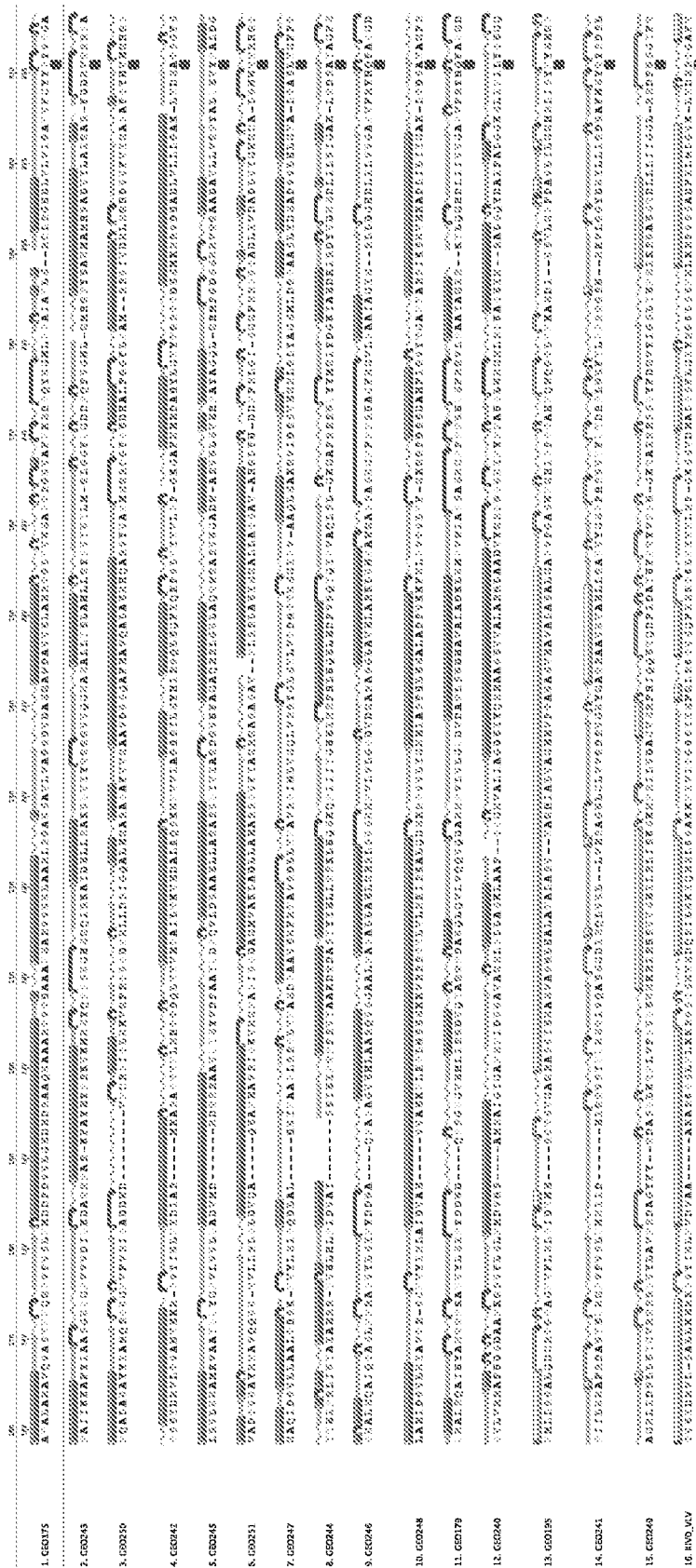
Library	Native	Mutations
1	D285 S286	ADLF FMSTWY
2	Q377 F381 F382	AFMWQ(ILR*) ^a ALF(SV) ^a AFMW(IL) ^a
3	G402	AILMTV(S) ^a
4	V461 I465	AMWTV IV
5	M538 F542	AHLMV(T) ^a AFHIMTVWY(L) ^a

Figure 18. KIVD Mutants with Altered Specificity for Long-chain 2-Ketoacids from the Results of High-throughput Kinetic Assays. *(a) Potential mutations in parenthesis were included due to degenerate codon used for mutagenesis.

Enzyme ID	k_{cat}/K_M ($M^{-1} s^{-1}$)				Rosetta Energy (Ligand: TPP-C8)
	Pyruvate (C3)	2-keto Valerate (C5)	2-keto octanoate (C8)	2-keto isovalerate (isoC5)	
GEO175	0.5	48.0	17000.0	3.3	-14.0
GEO179	0.9	100.0	1200.0	41.0	-13.44
GEO195	2.4	200.0	1400.0	8.0	-12.5
GEO240	<0.2	not tested	67.0	1500.0	-12.1
GEO241	5.7	not tested	320.0	9.0	-12.4
GEO242	8.2	not tested	8300.0	680.0	not tested
GEO243	<0.2	not tested	<0.2	<0.2	not tested
GEO244	78.0	not tested	10000.0	8500.0	not tested
GEO245	<0.2	not tested	6.6	1.0	not tested
GEO246	0.6	not tested	61.0	2.6	not tested
GEO247	1.8	not tested	260.0	0.8	not tested
GEO248	10.0	not tested	12000.0	2100.0	not tested
GEO249	41.0	not tested	28.0	<0.2	not tested
GEO250	0.7	not tested	43.0	6.7	not tested
GEO251	<0.2	not tested	1.1	<0.2	not tested
KIVD_VLV	0.71	1.3	2800	0.24	-11.08

Figure 19





1. GR0175
2. GR0180
3. GR0185
4. GR0190
5. GR0195
6. GR0200
7. GR0205
8. GR0210
9. GR0215
10. GR0220
11. GR0225
12. GR0230
13. GR0235
14. GR0240
15. GR0245
16. GR0250
17. GR0255
18. GR0260
19. GR0265
20. GR0270
21. GR0275
22. GR0280
23. GR0285
24. GR0290
25. GR0295
26. GR0300
27. GR0305
28. GR0310
29. GR0315
30. GR0320
31. GR0325
32. GR0330
33. GR0335
34. GR0340
35. GR0345
36. GR0350
37. GR0355
38. GR0360
39. GR0365
40. GR0370
41. GR0375
42. GR0380
43. GR0385
44. GR0390
45. GR0395
46. GR0400
47. GR0405
48. GR0410
49. GR0415
50. GR0420
51. GR0425
52. GR0430
53. GR0435
54. GR0440
55. GR0445
56. GR0450
57. GR0455
58. GR0460
59. GR0465
60. GR0470
61. GR0475
62. GR0480
63. GR0485
64. GR0490
65. GR0495
66. GR0500
67. GR0505
68. GR0510
69. GR0515
70. GR0520
71. GR0525
72. GR0530
73. GR0535
74. GR0540
75. GR0545
76. GR0550
77. GR0555
78. GR0560
79. GR0565
80. GR0570
81. GR0575
82. GR0580
83. GR0585
84. GR0590
85. GR0595
86. GR0600
87. GR0605
88. GR0610
89. GR0615
90. GR0620
91. GR0625
92. GR0630
93. GR0635
94. GR0640
95. GR0645
96. GR0650
97. GR0655
98. GR0660
99. GR0665
100. GR0670

Figure 20 (Continued)



Alignment Position	SEQ ID Position	Amino Side-Amino Acids	Predicted Secondary Structure Profile
36	23	['Y', 'V', 'N']	['Loop', 'beta-strand']
37	24	['P', 'V']	['Loop', 'beta-strand']
38	25	['G']	['Loop', 'turn']
39	26	['S', 'D', 'G', 'T']	['Loop', 'turn']
40	27	['A', 'G', 'F', 'N', 'T', 'V', 'Y']	['Loop', 'beta-strand', 'turn']
65	48	['A', 'Q', 'T', 'G', 'L']	['Helix', 'Loop', 'beta-strand', 'turn']
66	49	['A', 'S', 'N', 'G', 'V']	['Helix', 'Loop', 'beta-strand', 'turn']
71	54	['A']	['Helix', 'beta-strand']
88	71	['Q', 'A', 'S', '-', 'T']	['Helix', 'Loop', 'beta-strand', 'turn']
91	74	['G']	['Helix', 'Loop', 'turn']
92	75	['A', 'E', 'T', 'L', 'V']	['Helix', 'Loop']
95	78	['A', 'T', 'G']	['Helix', 'Loop', 'beta-strand']
98	81	['A', 'P', 'M', 'G', 'N']	['Helix', 'Loop', 'beta-strand']
129	112	['E', 'G', 'H', 'L', 'N', 'Q', 'S']	['Helix', 'Loop', 'beta-strand', 'turn']
130	113	['-', 'N', 'P', 'S', 'T', 'V']	['Helix', 'Loop', 'beta-strand', 'turn']
311	286	['A', 'L', 'N', 'Q', 'P', 'S', 'T', 'V', 'Y']	['Helix', 'Loop', 'beta-strand', 'turn']
401	356	['A', 'E', 'D', 'G', 'H', 'L', 'P', 'R']	['Helix', 'Loop', 'beta-strand', 'turn']
433	376	['A', 'I', 'M', 'L', 'P', 'V']	['Helix', 'beta-strand']
434	377	['F', 'I', 'T', 'W', 'V', 'Y']	['Helix', 'beta-strand']
435	378	['A', 'G', 'F', 'S', 'T', 'V']	['Helix', 'beta-strand']
436	379	['E', 'D', 'I', 'K', 'N', 'S']	['Helix', 'beta-strand']
437	380	['Q', 'E', 'D', 'T']	['Helix', 'Loop', 'beta-strand', 'turn']
439	382	['I', 'P', 'T', 'G', 'L']	['Helix', 'Loop', 'turn']
456	399	['E', 'G', 'H', 'K', 'Q', 'S', 'T', 'V']	['Helix', 'Loop', 'beta-strand', 'turn']
458	401	['A', 'C', 'F', 'I', 'H', 'L', 'Y']	['beta-strand', 'turn']
459	402	['G', 'F', 'N', 'Q', 'S', 'T', 'Y']	['Helix', 'Loop', 'beta-strand']
460	403	['F', 'M', 'L', 'Q', 'P', 'S', 'T']	['Helix', 'Loop', 'beta-strand', 'turn']
461	404	['A', 'C', 'G', 'Q', 'P', 'R', 'V']	['Helix', 'Loop', 'beta-strand', 'turn']
490	427	['H', 'S', 'N', 'P', 'V']	['Helix', 'beta-strand']
491	428	['I', 'L', 'V']	['beta-strand']
492	429	['A', 'C', 'L', 'G']	['beta-strand']
493	430	['I', 'F', 'L', 'G', 'V']	['beta-strand']
494	431	['I', 'T', 'L', 'V']	['Loop', 'beta-strand', 'turn']
495	432	['G']	['Loop', 'beta-strand', 'turn']
497	434	['A', 'G']	['Helix', 'Loop', 'turn']
517	454	['I', 'K', 'T', 'V']	['beta-strand']
518	455	['C', 'F', 'I', 'L', 'V', 'Y']	['beta-strand']
519	456	['I', 'F', 'L', 'V']	['beta-strand']
520	457	['I', 'C', 'L', 'V']	['beta-strand']
521	458	['C', 'F', 'I', 'M', 'L', 'V', 'Y']	['Turn', 'beta-strand']
522	459	['S', 'K', 'N']	['Helix', 'Loop', 'beta-strand', 'turn']
523	460	['D', 'N']	['Helix', 'Loop', 'turn']
524	461	['A', 'D', 'G', 'K', 'N', 'Q', 'S', 'R', 'T']	['Helix', 'Loop', 'turn']
525	462	['A', 'E', 'G', 'S', 'R', 'T']	['Helix', 'Loop', 'turn']
528	465	['I', 'A', 'M', 'L', 'V']	['Helix', 'beta-strand']
608	535	['D', 'H', 'L', 'P', 'S', 'T', 'V']	['Helix', 'Loop', 'turn']

Figure 21

Alignment Position	GEO175 Position	Active Site Amino Acids	Predicted Secondary Structure Profile
35	22	[G]	['Loop', 'beta-strand']
36	23	[Y, V, N]	['Loop', 'beta-strand']
37	24	[P, V]	['Loop', 'beta-strand']
38	25	[G]	['Loop', 'turn']
39	26	[S, D, G, T]	['Loop', 'turn']
40	27	[A, G, F, N, T, V, Y]	['Loop', 'beta-strand', 'turn']
41	28	[N, E, V]	['Helix', 'Loop', 'beta-strand']
42	29	[Q, L, F]	['Helix', 'beta-strand']
45	32	[F, I, M, L, T, Y]	['Helix', 'Loop', 'beta-strand']
60	43	[A, L, G]	['Helix', 'Loop', 'beta-strand']
63	46	[Q, H, N]	['Helix', 'Loop']
64	47	[E]	['Helix', 'Loop', 'beta-strand']
65	48	[A, Q, T, G, L]	['Helix', 'Loop', 'beta-strand', 'turn']
66	49	[A, S, N, G, V]	['Helix', 'Loop', 'beta-strand', 'turn']
67	50	[A, G, V]	['Helix', 'beta-strand']
71	54	[A]	['Helix', 'beta-strand']
86	69	[L, T, V]	['Helix', 'beta-strand']
87	70	[Y, H, T, F]	['Helix', 'Loop', 'beta-strand', 'turn']
88	71	[A, G, V, L, T, P, F, Y, W, S, T, C, M, N, Q, K, R, H, D, E, -]	['Helix', 'Loop', 'beta-strand', 'turn']
89	72	[A, T, G]	['Helix', 'Loop', 'beta-strand', 'turn']
90	73	[A, P, C, G, V]	['Helix', 'Loop', 'turn']
91	74	[G]	['Helix', 'Loop', 'turn']
92	75	[A, E, T, L, V]	['Helix', 'Loop']
93	76	[T, G, L]	['Helix', 'Loop', 'beta-strand']
94	77	[H, S, N]	['Helix', 'Loop', 'beta-strand']
95	78	[A, T, G]	['Helix', 'Loop', 'beta-strand']
96	79	[I, A, M, L, V]	['Helix', 'Loop', 'beta-strand']
98	81	[A, P, M, G, N]	['Helix', 'Loop', 'beta-strand']
99	82	[I, F, L, T, V]	['Helix', 'Loop', 'beta-strand']
128	111	[A, G, H, L, Q, S, T]	['Helix', 'Loop', 'beta-strand']
129	112	[A, G, V, L, T, P, F, Y, W, S, T, C, M, N, Q, K, R, H, D, E, -]	['Helix', 'Loop', 'beta-strand', 'turn']
130	113	[A, G, V, L, T, P, F, Y, W, S, T, C, M, N, Q, K, R, H, D, E, -]	['Helix', 'Loop', 'beta-strand', 'turn']
263	240	[F, M, L, S, T, W, V]	['Helix', 'Loop', 'beta-strand', 'turn']
310	285	[E, F, I, H, S, R, W, V, Y]	['Helix', 'Loop', 'beta-strand', 'turn']
311	286	[A, L, N, Q, P, S, T, V, Y]	['Helix', 'Loop', 'beta-strand', 'turn']
312	287	[E, F, P, S, T, Y]	['Helix', 'Loop', 'beta-strand', 'turn']
315	290	[P, E, D, G, F]	['Helix', 'Loop', 'beta-strand', 'turn']
401	356	[A, G, V, L, T, P, F, Y, W, S, T, C, M, N, Q, K, R, H, D, E, -]	['Helix', 'Loop', 'beta-strand', 'turn']
416	360	[Q, P, T, H, V]	['Helix', 'Loop', 'beta-strand', 'turn']
433	376	[A, I, M, L, P, V]	['Helix', 'beta-strand']
434	377	[F, I, T, W, V, Y]	['Helix', 'beta-strand']
435	378	[A, G, F, S, T, V]	['Helix', 'beta-strand']
436	379	[E, D, I, K, N, S]	['Helix', 'beta-strand']
437	380	[Q, E, D, T]	['Helix', 'Loop', 'beta-strand', 'turn']
438	381	[A, S, C, G, V]	['Helix', 'Loop', 'beta-strand', 'turn']
439	382	[I, P, T, G, L]	['Helix', 'Loop', 'turn']

FIG. 22

440	383	[A, G, N, Q, S, T, V]	[Helix, Loop, turn]
455	398	[C, E, D, G, M, N, S, R, T]	[Helix, Loop, beta-strand, turn]
456	399	[E, G, H, K, Q, S, T, V]	[Helix, Loop, beta-strand, turn]
457	400	[Y, N, W, F]	[Helix, Loop, beta-strand, turn]
458	401	[A, C, F, I, H, L, Y]	[beta-strand, turn]
459	402	[G, F, N, Q, S, T, Y]	[Helix, Loop, beta-strand]
460	403	[F, M, L, Q, P, S, T]	[Helix, Loop, beta-strand, turn]
461	404	[A, C, G, Q, P, R, V]	[Helix, Loop, beta-strand, turn]
462	405	[A, I, S, L, G]	[Helix, Loop, beta-strand, turn]
464	406	[Q, A, G, N]	[Loop, beta-strand, turn]
465	407	[I, S, T, G]	[Loop, beta-strand, turn]
489	426	[A, G, V, L, I, P, F, Y, W, S, T, C, M, N, Q, K, R, H, D, E, -]	[Helix, Loop, beta-strand, turn]
490	427	[H, S, N, P, V]	[Helix, beta-strand]
491	428	[I, L, V]	[beta-strand]
492	429	[A, C, L, G]	[beta-strand]
493	430	[I, F, L, G, V]	[beta-strand]
494	431	[I, T, L, V]	[Loop, beta-strand, turn]
495	432	[G]	[Loop, beta-strand, turn]
496	433	[E, D]	[Loop, turn]
497	434	[A, G]	[Helix, Loop, turn]
498	435	[A, S, G]	[Helix, Loop]
516	453	[A, P, I, L, V]	[Helix, beta-strand, turn]
517	454	[I, K, T, V]	[beta-strand]
518	455	[C, F, I, L, V, Y]	[beta-strand]
519	456	[I, F, L, V]	[beta-strand]
520	457	[I, C, L, V]	[beta-strand]
521	458	[C, F, I, M, L, V, Y]	[Turn, beta-strand]
522	459	[S, K, N]	[Helix, Loop, beta-strand, turn]
523	460	[D, N]	[Helix, Loop, turn]
524	461	[A, D, G, K, N, Q, S, R, T]	[Helix, Loop, turn]
525	462	[A, E, G, S, R, T]	[Helix, Loop, turn]
526	463	[Y, L]	[Helix, Loop, turn]
527	464	[A, K, T, G, N]	[Helix, Loop, beta-strand, turn]
528	465	[I, A, M, L, V]	[Helix, beta-strand]
529	466	[E, L, V]	[Helix, beta-strand]
535	472	[A, G, V, L, I, P, F, Y, W, S, T, C, M, N, Q, K, R, H, D, E, -]	[Helix, Loop, beta-strand, turn]
536	473	[A, G, V, L, I, P, F, Y, W, S, T, C, M, N, Q, K, R, H, D, E, -]	[Helix, Loop, beta-strand, turn]
545	475	[A, G, V, L, I, P, F, Y, W, S, T, C, M, N, Q, K, R, H, D, E, -]	[Helix, Loop, beta-strand, turn]
547	477	[Q, R, D, G, N]	[Helix, Loop, beta-strand, turn]
599	526	[E, D, N]	[Helix, beta-strand]
600	527	[A, I, L, V]	[Helix, beta-strand]
601	528	[E, I, H, K, M, L, P, V]	[Helix, Loop, beta-strand]
603	530	[A, D, G, L, Q, P, S, T]	[Helix, Loop, beta-strand, turn]
604	531	[A, G, V, L, I, P, F, Y, W, S, T, C, M, N, Q, K, R, H, D, E, -]	[Helix, Loop, beta-strand, turn]
605	532	[A, G, V, L, I, P, F, Y, W, S, T, C, M, N, Q, K, R, H, D, E, -]	[Helix, Loop, beta-strand, turn]
607	534	[A, G, V, L, I, P, F, Y, W, S, T, C, M, N, Q, K, R, H, D, E, -]	[Helix, Loop, beta-strand, turn]
608	535	[A, G, V, L, I, P, F, Y, W, S, T, C, M, N, Q, K, R, H, D, E, -]	[Helix, Loop, beta-strand, turn]

FIG. 22 (Cont.)

	GEO175	GEO243	GEO250	GEO242	GEO245	GEO251	GEO247	GEO244	GEO246	GEO248	GEO179	GEO240	GEO195	GEO241	GEO249	KVD_VLV
GEO175	20.9%	20.9%	40.0%	17.6%	29.9%	24.0%	20.3%	17.8%	48.5%	17.8%	46.1%	29.5%	35.9%	33.7%	17.3%	16.3%
GEO243	20.9%	20.9%	21.2%	16.1%	20.7%	19.5%	19.1%	16.3%	22.4%	18.0%	21.9%	19.3%	20.4%	22.5%	14.1%	17.0%
GEO250	40.0%	21.2%	21.2%	18.1%	26.4%	20.7%	18.7%	17.4%	40.4%	20.0%	38.7%	27.1%	29.8%	31.3%	17.2%	17.4%
GEO242	17.6%	16.1%	18.1%	18.6%	18.6%	18.1%	33.8%	43.3%	18.6%	64.8%	17.4%	16.8%	16.8%	17.1%	36.4%	54.9%
GEO245	29.9%	20.7%	26.4%	18.6%	22.3%	22.3%	18.2%	17.9%	27.9%	17.9%	27.4%	28.8%	31.6%	31.6%	12.1%	17.2%
GEO251	24.0%	19.5%	20.7%	18.1%	22.3%	21.8%	21.8%	18.9%	21.5%	16.7%	20.2%	18.7%	20.0%	19.4%	15.1%	18.3%
GEO247	20.3%	19.1%	18.7%	33.8%	18.2%	21.8%	30.9%	18.8%	18.8%	31.7%	18.1%	15.7%	18.3%	16.9%	28.4%	31.1%
GEO244	17.8%	16.3%	17.4%	43.3%	17.9%	18.9%	30.9%	18.0%	18.0%	43.0%	17.6%	16.0%	17.4%	14.1%	35.8%	42.6%
GEO246	48.5%	22.4%	40.4%	18.6%	27.9%	21.5%	18.8%	18.0%	19.5%	19.5%	74.3%	28.0%	32.2%	32.1%	16.7%	17.1%
GEO248	17.8%	18.0%	20.0%	64.8%	17.9%	16.7%	31.7%	43.0%	19.5%	18.9%	18.9%	16.3%	15.7%	16.3%	35.9%	53.6%
GEO179	46.1%	21.9%	38.7%	17.4%	27.4%	20.2%	18.1%	17.6%	74.3%	18.9%	18.9%	28.5%	31.7%	34.1%	14.7%	18.3%
GEO240	29.5%	19.3%	27.1%	16.8%	28.8%	18.7%	15.7%	16.0%	28.0%	16.3%	28.5%	52.4%	40.7%	40.7%	12.5%	18.7%
GEO195	35.9%	20.4%	29.8%	16.8%	31.6%	20.0%	18.3%	17.4%	32.2%	15.7%	31.7%	52.4%	43.3%	43.3%	15.6%	14.8%
GEO241	33.7%	22.5%	31.3%	17.1%	31.6%	19.4%	16.9%	14.1%	32.1%	16.3%	34.1%	40.7%	43.3%	43.3%	13.7%	15.7%
GEO249	17.3%	14.1%	17.2%	36.4%	12.1%	15.1%	28.4%	35.8%	16.7%	35.9%	14.7%	12.5%	15.6%	13.7%	32.9%	32.9%
KVD_VLV	16.3%	17.0%	17.4%	54.9%	17.2%	18.3%	31.1%	42.6%	17.1%	53.6%	18.3%	18.7%	14.8%	15.7%	32.9%	32.9%

FIG. 23

>GEO175

-----MRTVRESALDVLRRARGMTTVFGNPGSTELPMLKQFP--DD--
 FRYVLGLQEAVVGMADGFALASGTTGLVNLHTGPGTGNAMGAILNARANRTPMVVTTAGQQVRAMLTEA
 LLTNPQSTLLPQPAVKWAYEPPR-----
 AADVAPALARAVQVAETPPQGPVFSVSLPMDDFDVLGEDEDRAAQRAAARTVTHAAAPSAEVRRLAARLS
 GARS AVL VAGNDVDASGAWDAVVELAERTGLPVWSAPTEGRVAFPKSHPOYRGMLPPAIAPLS--
 RCLEGHDLVLVIGAPVFCYYPYVPGAHLPELTELHTRDADEAARAPVGD-AVVADLALTVRALLAELP-----
 -----A-----REAAAPAARTARAESTAEDVGV-----LTPLAAMTAIAQGA-
 PANTLWVNESPSNLGQFHDATRIDTPGSFLFTAG-GGLGFGLAAGVGAQLGAPDR-----
 PVVCVIGDGSTHYAVQALWTAAYKVPVTFVVLNQRYSAILQWFAQVE-----GAQGAPGLDI--
 PGLDIAAVATGYGVRAHRATGFGELSKLVRESA-LQQDGPVLIDVPVTTTELPTL-----

>GEO243

-----MELLSGGEMLVRLADEGVEHVFGYPGGAVLHIYDALFQQDK--
 INHYLVRHEQAAGHMADAYSRTATGKTGVVLTSGPGATNTVTPIATAYMDSIPMVILSGQVASHLIGEDAF-
 QETDMVGISRPVKSFSQVRH-----ASEIPAIKKAFYIAASGRPGPVVDIPKDATNPAAE-
 KFAYEYPEKVKMRSYQPPSRGHSGQIRKAIDELLSAKRPVIYTGCGVVGQGNASALLTELAHLLGYPVTNTLM-
 GLGGFPGDDPQFVGML-GMHGTYEANMAMHNADVILAIGAR-
 FDDRVTNNPAKFCVNAKVIHIDIDPASISKTIMAHPIVGAVEPVLQEMLTQLKQL-----
 NVSKPNPEAIAAWWDQINEWRKVHGLKFETPTDGT-----MKPQQVVEALYKAT-
 NGDAIITSDVGQHQMFALYYKYKRPRQWINSGLGTMGVGLPYAMAAKLAFPDQ-----
 QVVCITGEASIQMCIQELSTCKQYGMNVKILCLNNRAGLMVKQWQDMNY-----
 EGRHSSSYVESLPDFGKLMEAYGHVGIQIDHADELESKLAEM-
 AINDKCVFINVMVDRTEHVYPMLIAGQSMKDMWLKKGERT

>GEO250

MPANTAPNAQAAEVFTVRHAVINMLRELGMTRIFGNPGSTELPLFRDYP--ED--
 FSYILGLQETVVGMADGYAATRNASFVNLHSAAGVGHAMANIFTAFKNRTPMVITAGQQTRSLLOQDFPFL
 HSNQAAELPKPYVKWSCEPAR-----AEDVPQALARAYYIAMQEPRGPVFSIPADDWD-----
 VPCEPITLRKVGFEETRPDRLLDSIGQALEGARAPAFVVGAAVDRSQAFEAVQALAEHRHQARVYVAPMSGRCG
 FPDHALFGGFLPAM--
 RERIVDRLSGHDVVFVIGAPAFYTHVEGHGPFIAEGTQLFQLIEDPAIAAWAPVGDAAV-GNIRMGVQELLAR--
 -----P-----LTHPRPALQPRPAIPAPAAPEPGR-----
 LMTDAFLMHTLAQVRSRDSIIVEEAPGSRSHQAHLPYAAETFFTMCS-GGLGHSLPASVGIALARPDK-----
 KVIGVIGDGSAMYAIALWSAAHLKLPVTYIIVKNRRYAALQDFSRVF-----
 GYREGKEKVEGTDLPDIDFVALAKGQCGDGVVTDAAQLSQVLRDA--LRSRATLVEVEVA-----

>GEO242

-----MYTVADYLLDRLKELGIDEVFGVPGDYNLQFLDHITARKD--
 LEWIGNANELNAAYMADGYARTKGISALVTTF-
 GVGELSAINGLAGSYAESIPVIEIVGSPTTTVQQNKKLVHHTLGDGDFLRFERIHEEVSAIAHLSTENAPSEIDR
 VLTVMATEKR-PVYINLPIDIAE-----
 MKASAPTTPLNHTTDQLTTVETAILTKVEDALKQSKNPVVIAGHEILSYHIENQLEQFIQKFNLPITVLPF-
 GKGAFFNEEDAHYLGTYTGSTTDESMKNRVDHADLVLLGAK-
 LTDSATSGFSFGFTEKQMSISGSTVLFYGEKQETVQL-DRFVSALSTLSFS-----R-----
 FTDEMPSVKRLATPKVRDEK-----LTQKQFWQMVESFLLQGDVTVGEQGTSTFFGLT--
 NVPLKKDMHFIGQPLWGSIGYTFPSALGSQIANKES-----
 RHLLFIGDGSGLQTLVQELGTAIREKLTPIVFINNNGYTVREIHGAT-----EQYNDIPM-
 WDYQKLPFVFGGTDQTVATYKVSTEILDNAMTRAR-TDVDRLQWIEVMDQNDAPVLLKKLAKIFAKQNS-

>GEO245

-----MVSRPARVAILEQLRADGVRYMFGNPGTVEQGFLDELNFPD--
 IEYILALQEAGVGLADGYARATRPVAVLQLHTGVGVGNVGMILYQAKRGHAPLVAIAGEAGLRYDAMEAQ
 MA-VDLVAMAEPVTKWATRVVD-----PESTLRVLRAMKVAATPPYGPVLVLPADVMD-----
 RDTSEAAVPTSyvDFAATPDPQVLDRAAELLAGAERPVIAGDGVHFAGAQEELGRLAQTWGAEVWGADW-

FIG. 24

AEVNLSVEHPAYAGQL-
 GHMFGDSSRRVTGAADAVLLVGTYALPEVYPALDGVFADGAPVVHIDLDTDAIAKNFPVDLGLAADPRRALD
 GLARALERRMSPESTRARAGEWFTGRSAQRSYEIAAAAREQDEAALAPDA-----LPVTAFLQELARQL-
 PEDAVWFDEALTASPDVTRHLPPTRPBGHWHQTRG-GSLGVGIPGAIAAQLAHPDR-----
 TVVGFTGDGGSlyTIQALWTAARYDIGATFVICNNSSYKLELNEEYWKSVDVAAHEQPEMFDLARPAIDFV
 ALSRSLGVPVAVRVEKPDQAKAAVEQA--LGTPGPFLIDLVTGRGRED-----
 >GEO251
 -----MNVAELVGRTLAELGVGAAGFVVGSGNFVVTNGLRA-GG--
 VRFVAARHEGGAASMA DAYARMSGRVSVLSLHQGCGLTNALTGITEAAKSRTPMIVLTGDTAASAVLSNFRI
 GQ--DALATAVGAVPERVHS-----APTAVADTVRAYRTAVQQR--TVLLNLPLDVQA-----
 QEAPAEVEIPKVRGPAPIRDPAGMVAKLADLLAEARRPVFIAGR GARASAV--PLRELAEISGALLATSAV-
 AHGLFH-DDPFSLGI-SGGFSSPRTADLIVDADLVIGWGCA-
 LNMWTTTRHGTLLGPAARLVQVDVEQAALGAHRPIDLGVVGDVAGTAVDVHAEIDKR-----GH-----
 QRSREAPTGTRWNDVPYNDLSGDGR-----IDPRTLSRRLDEIL-
 PAERMVSIIDSGNFMGYPSAYLSVPDENGFCFTQAFQSIGLGLGTAIGAALARPDR-----
 LPVLGVGDGGFHMMAVSELETAVRLRIPLVIVVYNDAAAYGAEIHHFG-----
 DADMTTVRFPDTDIAAIGRGFGCDGVTVRVSGDL-AAVKEWLGGPRDAPLVIDAKIADDGGSWWLAFAFRH--

 >GEO247
 -----MIDL DGTVTVAEYLGRLRHAGVEHLFGVPGDFNLNLLDGLAFVEG--
 LRWVGSPNELGAGYAADAYARRRGLSALFTTY-
 GVGELSAINAVAGSAAEDSPVVHVVGSPRTTAVAGGALVHHTIADGDFRHFARAYAEVTVAQAMVTATDAG
 AQIDRVLLAALTHRKP-VYLSIPQDLAL-----
 HRIPAAPLREPLTPASDPAAVERFRTAVRDLLTPAVRPIMLVGQLVSRYGLSTLVTDMTTRSGIPV-
 AAQLSAKGVIDESVEGNLGLYAGSMLDGPAA SLIDSADVVLHLGTA-
 LTAELTGFFTHRRPDARTVQLLSTAALVGTRFDNV-
 LFPDAMTTLAEVLTFFPAPARLAAPTTRAETGLAASITPPAPSAVDLTASTATDLTAPTAGDISEMSRVLTQDA
 FWAGMQAWLPAGHALVADTGTSYWGAL--ALRLPGDTVFLGQPIWNSIGWALPAVLGQGLADPDR----R-
 PVLVIGDGAAQMTIQELSTIVAAGLRPIILLNNRGYTIERALQSPN-----AGYNDVA-DWNWRVAV--
 VAAFAGPDTDYHHAATGT ELAKALTA--SESNRPVFIEVELDAFDTPPLLRLRAERATAPS-----
 >GEO244
 -----MKQRIGAYLIDAIHRAGVDKIFGVPGDFNLAFLLDDIISNP--
 VDWWGNTNELNASYAADGYARLNGLAALVTTF-
 GVGELSAVNGIAGSYAERIPVIAITGAPTRAVEHAGKYVHSLGEGTFDDYRKMF AHITVAQGYITPENATTEIP
 RLINTAIAERR-PVHLHLPIDVAI-----
 SEIEIPTFEVTA AKD TDASTYIELLTSKLHQSKQPIITGHEINSFHLHQELEDVFNQTQIPVAQLSL-
 GKGA FNEENPYMGIYDGKIAEDKIRDYVDNSDLILNIGAK-
 LTDSATAGFSYQFNIDDVMLNHHNIKIDDVTNDEISL-PSLLKQLSNISHT-----N-----
 NATFPAYHRPTSPDYTVGTEP-----LTQQT YFKMMQNF LKPN DVIIADQGTSFFGAY--
 DLALYKNNTFIGQPLWGSIGYTLPATLGSQ LADKDR-----
 RNLLLIGDGLQLTVQAISTMIRQHIKPVLFVINNDGYTVERLIHGM Y-----EPYNEIHM-WDYKALP--
 AVFGGKNVEIHDVSSKDLQDTFNAIN-GHPDVMHFVEVKMSVEDAPKKLIDIAKA FSQQNK-----
 >GEO246
 -----MKTVHSASYEILRRHGLTTVFGNPGSNELPFLKDFP--ED--
 FRYILGLHEGAVVGMADGFALASGRPAFVNLAHAAAGTGNGMGALTNAWYSHSPLVITAGQQVRSMIGVEAM
 LANVDAGQLPKPLVKWSHEPAC-----AQDVPRALSQAIQTASLPPRAPVYLSIPYDDWA----
 QPAPAGVEHLAARQVSGAALPAPALLAELGERLSRSRNPVLVLGPDVDGANANGLAELAEKLMPAWGAP
 SASRCFPFTRHACFRGVLPAAIAGIS--
 RLLDGHDLILVVGAPVFRYHQFAPGDYLPAGAELVQVTCDPGEAARAPMGD-ALVGDIALTLEALLEQVR-----
 -----P-----SARPLPEALPRPPALAEEGGP-----LRPETVF-
 DVIDALAPRDAIFVKESTSTVAFWQVRVEMREPGSYFFPAA-GGLGFGLPAAVGAQLAQPRR-----
 QVIGIIGDGSANYGITALWSAAQYRVPVFIILKNGTYGALRWFAVL-----EVPDAPGLDV--
 PGLDFCAIARGYGVEALHAATREELEGALKHA--LAADRPVLIEVPTQTIEP-----
 >GEO248

FIG. 24 (Cont. 1)

-----MYTVGNLYLLDRLTELGIRDIFGVPGDYNLKF LDHVMTHKE--
 LNWIGNANELNAAYAADGYARTKGIAALVTF-
 GVGELSAANGTAGSYAEKVPVQIVGTPTTAVQNSHKL VHHTLGDGRFDHFEKMQTEINGAIAHLTADNALA
 EIDRVLRIAVTE-RCPVYINLAIDVAE-----
 VVAEKPLKPLMEESKKVEEETTLVLNKIEKALQDSKNPVVLIGNEIASFHLESALAD FVKKFNL PVTVL PF-
 GKGGFDEEDAHFIGVYTGAETAESIKERVEKADLILIGAK-
 LTDSATAGFSYDFEDRQVISVGSDEVSYFGEIMKPVAF-AQFVNGLNLSLNYL-----G-----
 YTGEIKQVERVADIEAKASN-----LTQNNFWKFVEKYLSNGDTLVAEQGTSFFGAS--
 LVPLKSKMKFIGQPLWGSIGYTFFPAMLGSIQIANPAS-----
 RHLLFIGDGSQSLTIQELGMTFREKLTPIVFVINNDGYTVEREIHGPN-----ELYNDIPM-
 WDYQNLPPYVFGGNKGNVATYKVTTEEELVAAMSQAR-
 QDTRLQWIEVVMGKQDSPDLLVQLGKVF AKQNS-----
 >GEO179
 -----MKTIHSAAYALLRRHGM TTFGNPGSNELPFLKSFP--ED--
 FQYVLGLHEGAVVGMADGYALASGKPAFVNLHAAAGTGNMGALTNSWYSHSPLVITAGQQVRPMIGVEA
 MLANVDATQLPKPLVKWSYEPAN-----AQDVPRALSQAIHYANTTPKAPVYLSIPYDDWD---
 QPSGPGVEHLIERDVQTAGTPDARQLQVLVQQVQDARNPVLVLGPDVDATLSNDHAVALADKL RMPVWIAPS
 ASRCPFPTRHPSFRGVLPAIAGIS--
 KTLOGHDLIIVVGAPVFRYHQFAPGDYLPVGAQLLHITS DPLEATRAPMGH-ALVGDIRETLRVLAEEVVQ-----
 -----Q-----SRPYPEALAAPECVTDEPHH-----LHPETLF-
 DVLDAVAPHDAIYVKESTSTTTAFWQRMNLRHPGSYFPAAGGLGFLPAAVGVQLAQPQR-----
 RVVALIGDGSANYGITALWTAQYRIPVVFILKNGTYGALRWFAGVL-----KAEDSPGLDV--
 PGLDFCAIAKGYGVKAVHTDTRDSFEALRTA--LDANEPTVIEVPTLTIQPH-----
 >GEO240
 -MSNAITKVQANANARRGGDVLLLEVLESEGVEYVFGNPGTTTELPMFDALLRKPS--
 IQYVLALQEASAVAMADGYAQA AKKPGFLNLHTAGGLGHGMGNLLNAKCSQTPLVVTAGQQDSRHTTTDPL
 LLG-DLVGMGKTFAKWSQEVTH-----VDQLPVLVRRAFHDSDAAPKGSVFLSLPMDVME---
 AMSAIGIGAPSTIDRNAVAGSLPLLASKLAAF--
 TPGNVALIAGDEIYQSEAA NEVVALAEMLAADVYGSTWPNRIPYPTAHPLWRGNLSTKATEIN--
 RALSQYDAIFALGGKSLITILYTEGQAVPEQCKVFLSADAGDLGRYSSSELSVWGDIKSSLKVLLEPEKATAN
 HRRDYQRRFEKAINFEKLSKESLLGQVQEQQSATV-----ITPLVAAFEAARAIGP-
 DVAIVDEAIATSGSLRKSLSNSHRADQYAFLRG-GGLGWGMPAAVGYSLGLGKA-----
 PVVCFVGDGAAMYSPOALWTAHEKLPVTFIVMNNTEYNVLKNFMRSQ-
 ADYTSAQTDRIAMD LVNPSVDYQALGASMGLETRKVIRAGDIAPAVEAA--LASGKPNVIEIIISKS-----

 >GEO195
 MTSRSSFSPPSASEQRGADIFAEVLQCEGVRYIFGNPGTTTELPLLDALTDITG--
 IHYVLGLHEASVAMADGYAQASGKPGFVNLHTAGGLGNAMGAILNAKMANTPLVVTAGQQDTRHGVTDP
 LLHG-DLTGIARNPVKWAEEIHH-----PEHIPMLLRALQDCRTGPAGPVFLSLPIDTME---
 RCTSVGAGEASRIERASVANMLHALATALAEV--
 TAGHIALVAGEEVFTANASVEAVALAEALGAPVFGASWPGHIPPTAHPQWQGTLP PKASDI--
 RETLGPFDVAVLILGGHSLISYPYSEGPAIPPHCRFLQLTGDGHQIGRVHETTLGLVGDQLSLRALLPLLARKLQP
 QNGAVARLRQVATLKR DARRTEAAERSAREFDASA-----TTPFVAAFETIRAIGP-
 DVPIVDEAPVTIPHVRACLDASARQYLFTRS-AILGWGMPAAVGVSLGLDRS-----
 PVVCLVGDGSAMYSPOALWTAHERLPVTFVFNNGEYNILKNYARAQ-
 TNYRSARANRFIGLDISDPAIDFPALASSLGVPARRVERAGDIAIAVEDG--IRSGRPNLIDVLISSSS-----

 >GEO241
 --MNIAYETRENKVASGRECLLEILRDEGVTHVFGNPGTTTELALIDALAGDDD--
 FHFILGLQEA AVVGMADGYAQATGRPSFVNLHTTAGLGNMGMLTNAFATNPMVVTAGQQDIRHLAYDPL
 LSG-DLVGLARATVKWAHEVRS-----LQELPIILRRAFRDANTEPRGPV FVSLPMNIID-----
 EIGTVSIPPRSTIVQAESGDISQLVRL--
 LVESAGNLCLVVGDEVGRYGATEAAVRVAELLGAPVYGSPFHSNVPFPTDHPLWRFTLPNTGEM--
 RKLVLGGYDRILLIGDRAFMSTYSDELPLSPKTQLLQIAVDRHSLGRCHAVELGLYGDPLSLLAAVGDALSQER

FIG. 24 (Cont. 2)

AL---APSRDSRLAIARDWRASWEQDLKDECERLAPSRP-----LYPLVAADAVLRGVPPG-
 TVIVDECLATNKYVRQLYPVRKPGEYYYFRG-AGLGWGMPPAAVGVSLGLERQ----
 QRVVCLLDGGAAMYSPQALWSAAHESLPITFVVFNNSEYNILKNFMRSR--
 PGYNAQSGRFVGMIEINQPSIDFCALARSMGVDAVRLTEPDDITAYMIAA--GDREGPSLLEIPAAATAS-----

 >GEO249

 MSSEKVLVGEYLFTRLLQLGIKSILGVPGDFNLALLDLIEKVGDFTFRWWGNENELNGAYAADAYARVKGISAI
 VTTF-
 GVGELSALNGFAGAYSERIPVHVIVGVPNTKAQATRPLLHHTLGNGDFKVFQRMSSSELSADVAFLDSGDSAGR
 LIDNLLCTCVRTSRPVYLA VPSDAGYFY--
 TDASPLKTPLVFPVPENNKIEHEVVSILELIEKSKNPSILVDACVSRFHQQETQDFIDATHFPTYVTPM-
 GKTAINESSPYFDGVYIGSLTEPSIKERAESTDLLLIIGGL-
 RSDFNSTFTYATPASQTIEFHSDYTKIRSGVYEGISMKHLPLKLTAAIDKKS-----Q-----
 AKARPVHFEPPKAVAAEGYAGT-----ITHKWFWPTFASFLRESDVTTETGTSNFGIL--
 DCIFPKGCQNLSQVLWGSIGWSVGAMFGATLGKDSAPHRRSILIVGDGSLHLTVQEISATIRNGLTPIIFVINN
 KGYTIERLIHGLH-----AVYNDINTEWDYQNL--
 LKGYGAKNSRSYNIHSEKELDLDFKDEEFGKADVIQLVEVHMPVLDAPRVLIEQAKLTASLNKQ-----
 >KIVD_VLV
 -----MYTVGDYLLDRLHELIGIEIFGVPGDYNLQFLDQIISRKD--
 MKWVGNANELNASYMADGYARTKKAAAFLTTF-
 GVGELSAVNLGAGSYAENLPVVEIVGSPTSQVQNEGKFVHHTLADGDFKHFMMKHEPVTAARTLLTAENATV
 EIDRVL-SALLKERKPVYINLPVDVAA-----
 AKAEKPSLPLKKENSTSDQEIILNKIQESLKNAPKPIVITGHEIISFGLEKTVSQFISKTKLPITTLNF-
 GKSSVDEALPSFLGIYNGKLSEPNLKEFVESADFILMLGVK-
 LTDSSTGAFTHHLNENKMISLNIDEKGIFNESIQNFDFESLISSLLDLSEIE-----Y-----
 KGKYIDKKQEDFVPSNAL-----LSQDRLWQAVENLTQSNETIVAEQGSFFGAS--
 SIFLKPKSHFIGQPLWWSIGYTFPAALGSQIADKES-----
 RHLLFIGDGLQLTVQELGLAIREKINPICFIINNDGYTVEREIHGPN-----QSYNDIPM-
 WNYSKLPESFGATEERVVSKIVRTENEFVSVMEKAQ-ADPNRMYWIELILAKEDAPKVLKKGKLVAEQNKS-

FIG. 24 (Cont. 3)

INTERNATIONAL SEARCH REPORT

International application No.

PCT/US 16/49210

A. CLASSIFICATION OF SUBJECT MATTER

IPC(8) - C12N 15/52, C12N 9/88, C12N 1/15, C12N 1/21, C12P 7/04 (2016.01)

CPC - C12N 9/1022, C12N 9/1025, C12P 7/04, C12N 9/88

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

IPC(8) - C12N 15/52, C12N 9/88, C12N 1/15, C12N 1/21, C12P 7/04 (2016.01)

CPC - C12N 9/1022, C12N 9/1025, C12P 7/04, C12N 9/88

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

CPC - C07H 21/00, Y02E 50/10

(keyword limited; terms below)

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

PatBase, Google Patents, Google Scholar, Google Web

Search terms: decarboxylase, 2-ketoacid, 2-keto acid, alpha-ketoacid, alpha-keto acid, 2-oxoacid, 2-oxo acid, layer, sandwich, thiamine pyrophosphate, thiamine diphosphate, cocarboxylase, C8, 2-ketoacid, kcat/Km, 2VBG, KIVD, TM, align, score, heptanol, octanol

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y --- A	WO 2013/016724 A2 (GEVO INC. et al.) 31 January 2013 (31.01.2013) para [0072], [00127], [00148], [00154], [00157], [00197], [00290]; SEQ ID NO: 197	1-12, 15-30 ----- 13-14
Y --- A	ODOKONYERO et al., Loss of quaternary structure is associated with rapid sequence divergence in the OSBS family. Proc Natl Acad Sci U S A. 10 June 2014, Vol 111, No 23, pp 8535-40. Especially p 8536, Table 1; p 8537, col 1, para 1; p 8539, col 1, para 3	1-12, 15-30 ----- 13-14
Y --- A	US 2011/0201083 A1 (LIAO et al.) 18 August 2011 (18.08.2011) para [0005], [0006], [0017]	1-12, 15-30 ----- 13-14
A	D9VSN5, UniProtKB submission no. D9VSN5_9ACTN, 1 October 2014 [online]. [Retrieved on 23 November 2016]. Retrieved from the internet <URL: http://www.uniprot.org/uniprot/D9VSN5.txt?version=14 > Entire document	13-14
A	A0A0F4JHH3, UniProtKB submission no. A0A0F4JHH3_9ACTN, 24 June 2015 [online]. [Retrieved on 23 November 2016]. Retrieved from the internet <URL: http://www.uniprot.org/uniprot/A0A0F4JHH3.txt?version=1 > Entire document	13-14
A	A0A0F4K2I7, UniProtKB submission no. A0A0F4K2I7_9ACTN, 24 June 2015 [online]. [Retrieved on 23 November 2016]. Retrieved from the internet <URL: http://www.uniprot.org/uniprot/A0A0F4K2I7.txt?version=1 > Entire document	13-14

☒ Further documents are listed in the continuation of Box C.

* Special categories of cited documents:

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search

23 November 2016

Date of mailing of the international search report

13 DEC 2016

Name and mailing address of the ISA/US

Mail Stop PCT, Attn: ISA/US, Commissioner for Patents
P.O. Box 1450, Alexandria, Virginia 22313-1450

Facsimile No. 571-273-8300

Authorized officer:

Lee W. Young

PCT Helpdesk: 571-272-4300

PCT OSP: 571-272-7774

INTERNATIONAL SEARCH REPORT

International application No.

PCT/US 16/49210

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X, P	MAK et al., Integrative genomic mining for enzyme function to enable engineering of a non-natural biosynthetic pathway. Nat Commun. 24 November 2015, Vol 6, page 10005.	1-30

INTERNATIONAL SEARCH REPORT

International application No.

PCT/US 16/49210

Box No. 1 Nucleotide and/or amino acid sequence(s) (Continuation of item 1.c of the first sheet)

1. With regard to any nucleotide and/or amino acid sequence disclosed in the international application, the international search was carried out on the basis of a sequence listing:
 - a. ☐ forming part of the international application as filed:
 - ☐ in the form of an Annex C/ST.25 text file.
 - ☐ on paper or in the form of an image file.
 - b. ☐ furnished together with the international application under PCT Rule 13*ter*.1(a) for the purposes of international search only in the form of an Annex C/ST.25 text file.
 - c. ☒ furnished subsequent to the international filing date for the purposes of international search only:
 - ☒ in the form of an Annex C/ST.25 text file (Rule 13*ter*.1(a)).
 - ☐ on paper or in the form of an image file (Rule 13*ter*.1(b) and Administrative Instructions, Section 713).
2. ☒ In addition, in the case that more than one version or copy of a sequence listing has been filed or furnished, the required statements that the information in the subsequent or additional copies is identical to that forming part of the application as filed or does not go beyond the application as filed, as appropriate, were furnished.
3. Additional comments: