

TITLE OF THE APPLICATION

REDUCTASE ENZYMES AND PROCESSES FOR MAKING AND USING REDUCTASE ENZYMES

5 FIELD OF THE INVENTION

The present invention relates to ketoreductase enzymes, useful in the biocatalytic and synthetic processes involving reduction of ketones to chiral alcohols. Such enzymes may be particularly useful in preparation of nucleosides and nucleotides, such as fluorinated nucleotides.

10 REFERENCE TO SEQUENCE LISTING SUBMITTED ELECTRONICALLY

The sequence listing of the present application is submitted electronically via EFS-Web as an ASCII-formatted sequence listing, with a file name of “24998WOPCT-SEQTEXT-14MAY2021.txt”, creation date of May 14, 2021, and a size of 24 KB. This sequence listing submitted via EFS-Web is part of the specification and is herein incorporated by reference in its
15 entirety.

BACKGROUND OF THE INVENTION

Enzymes are protein molecules that serve to accelerate the chemical reactions of living cells (often by several orders of magnitude). Without enzymes, most biochemical
20 reactions would be too slow to even carry out life processes. Enzymes display great specificity and are not permanently modified by their participation in reactions. Because they are not changed during the reactions, enzymes can be cost effectively used as catalysts for a desired chemical transformation.

Ketoreductases, also known as alcohol dehydrogenases, are enzymatic reducing
25 agents, a specific class of enzymes that catalyze the selective reduction of ketones to chiral alcohols. Enzymes belonging to the ketoreductase or carbonyl reductase class are useful for the synthesis of optically active alcohols. Ketoreductase enzymes may selectively convert a ketone or aldehyde substrate to the corresponding chiral alcohol product; these enzymes may also convert alcohols into the corresponding ketones or aldehydes, in a reverse reaction. Enzymatic
30 reduction of ketones and aldehydes requires participation of a co-factor that can act as an electron donor, while enzymatic oxidation of alcohols requires participation of a co-factor that can act as an electron acceptor.

Ketoreductase enzymes are well known in nature, and numerous genes that encode ketoreductase enzymes and ketoreductase enzyme sequences have been reported. *See, e.g., Candida magnoliae* (Genbank Acc. No. JC7338; GI:11360538) *Candida parapsilosis* (Genbank Acc. No. 10 BAA24528.1; GI:2815409), *Sporobolomyces salmonicolor* (Genbank
5 Acc. No. AF160799; GI:6539734), and *Rhodococcus erythropolis* (Genbank Acc. No. AAN73270.1; GI: 34776951).

Ketoreductase enzymes are being used with increasing frequency to provide alternative synthetic pathways to key compounds. When used, the ketoreductase enzymes may be provided as purified enzymes or as whole cells that express the desired ketoreductase. In
10 view of their promise for improved synthetic pathways, there remains a need to identify additional ketoreductase enzymes that can be used to carry out transformations to prepare specific chiral alcohols.

SUMMARY OF THE INVENTION

15 The present disclosure relates to ketoreductase enzymes capable of converting ketones to chiral alcohols, particularly as part of nucleotide synthesis. In embodiments, the subject ketoreductase enzymes described herein are capable of converting ketones to chiral alcohols in the synthesis of fluoronucleotides. In particular, the subject ketoreductase enzymes described herein may be useful in the preparation of fluorinated nucleosides, such as (trisodium
20 O-{(2*R*,3*S*,4*S*,5*R*)-5-(2-amino-6-oxido-9*H*-purin-9-yl)-3-fluoro-4-hydroxyoxolan-2yl}methyl} phosphorothioate hydrate (1:6)), also known as 3'-fluoro-thio-guanosine monophosphate or 3'-F-thio-GMP. Such fluorinated nucleosides may be useful as a biologically active compound and or as an intermediate for the synthesis of more complex biologically active compounds.

Additional embodiments describe processes for preparing the subject
25 ketoreductase enzymes and processes for using the subject ketoreductase enzymes.

Other embodiments, aspects and features of the present invention are either further described in or will be apparent from the ensuing description, examples and appended claims.

30 BRIEF DESCRIPTION OF THE DRAWINGS

Fig. 1 depicts SDS-PAGE gel shows the removal of IPA insoluble proteins from the lyophilized enzyme preparation.

DETAILED DESCRIPTION OF THE INVENTION***Definitions***

Certain technical and scientific terms are specifically defined below. Unless
5 specifically defined elsewhere in this document, all other technical and scientific terms used
herein have the meaning commonly understood by one of ordinary skill in the art to which this
disclosure relates. That notwithstanding and except where stated otherwise, the following
definitions apply throughout the specification and claims. Chemical names, common names, and
chemical structures may be used interchangeably to describe the same structure. If a chemical
10 compound is referred to using both a chemical structure and a chemical name, and an ambiguity
exists between the structure and the name, the structure predominates. These definitions apply
regardless of whether a term is used by itself or in combination with other terms, unless
otherwise indicated. Hence, the definition of “alkyl” applies to “alkyl” as well as the “alkyl”
portions of “hydroxyalkyl,” “haloalkyl,” “-O-alkyl,” etc.

15 As used herein, and throughout this disclosure, the following terms, unless
otherwise indicated, shall be understood to have the following meanings:

As used herein, including the appended claims, the singular forms of words, such
as “a,” “an,” and “the,” include their corresponding plural references unless the context clearly
dictates otherwise. In particular, “a,” “an,” and “the” item each include a single item selected
20 from a list as well as mixtures of two or more items selected from the list.

As used herein, the terms “at least one” item or “one or more” item each include a
single item selected from the list as well as mixtures of two or more items selected from the list.
For example, “at least one ketoreductase type enzyme” (alternatively referred to as
“ketoreductase type enzymes”) refers to a single ketoreductase type enzyme as well as to
25 mixtures of two or more different ketoreductase type enzymes. Similarly, the terms “at least
two” items and “two or more” items each include mixtures of two items selected from the list as
well as mixtures of three or more items selected from the list.

“Consists essentially of,” and variations such as “consist essentially of” or
“consisting essentially of,” as used throughout the specification and claims, indicate the inclusion
30 of any recited elements or group of elements, and the optional inclusion of other elements, of
similar or different nature than the recited elements, that do not materially change the basic or
novel properties of the specified dosage regimen, method, or composition.

Throughout this specification and claims, the word “comprise,” or variations such as “comprises” or “comprising” will be understood to imply the inclusion of a stated element or group of elements but not the exclusion of any other element or group of elements. In the case of integers, the word “comprise,” or variations such as “comprises” or “comprising” will be understood to imply the inclusion of a stated integer or group of integers but not the exclusion of any other integer or group of integers. Unless otherwise required by context, singular terms shall include pluralities and plural terms shall include the singular. Any example(s) following the term “e.g.” or “for example” is not meant to be exhaustive or limiting. It is understood that wherever embodiments are described herein with the language “comprising,” otherwise analogous embodiments described in terms of “consisting of” and/or “consisting essentially of” are also provided.

Unless expressly stated to the contrary, all ranges cited herein are inclusive; *i.e.*, the range includes the values for the upper and lower limits of the range as well as all values in between. All ranges also are intended to include all included sub-ranges, although not necessarily explicitly set forth. For example, temperature ranges, percentages, ranges of equivalents, and the like described herein include the upper and lower limits of the range and any value in the continuum there between. Numerical values provided herein, and the use of the term “about”, may include variations of $\pm 1\%$, $\pm 2\%$, $\pm 3\%$, $\pm 4\%$, $\pm 5\%$, and $\pm 10\%$ and their numerical equivalents. “About” when used to modify a numerically defined parameter means that the parameter may vary by as much as 10% below or above the stated numerical value; where appropriate, the stated parameter may be rounded to the nearest whole number. For example, an amount of about 5mg may vary between 4.5mg and 5.5mg. In addition, the term “or,” as used herein, denotes alternatives that may, where appropriate, be combined; that is, the term “or” includes each listed alternative separately as well as their combination.

The term “alkyl,” as used herein, refers to an aliphatic hydrocarbon group having one of its hydrogen atoms replaced with a bond having the specified number of carbon atoms. In different embodiments, an alkyl group contains from 1 to 6 carbon atoms (C₁-C₆ alkyl) or from 1 to 3 carbon atoms (C₁-C₃ alkyl). Non-limiting examples of alkyl groups include methyl, ethyl, n-propyl, isopropyl, n-butyl, *sec*-butyl, isobutyl, *tert*-butyl, n-pentyl, neopentyl, isopentyl, n-hexyl, isohexyl, and neohexyl. In one embodiment, an alkyl group is linear. In another embodiment, an alkyl group is branched.

The terms “halogen” and “halo,” as used herein, means -F (fluorine), -Cl

(chlorine), -Br (bromine) or -I (iodine).

The term "haloalkyl," as used herein, refers to an alkyl group as defined above, wherein one or more of the alkyl group's hydrogen atoms has been replaced with a halogen. In one embodiment, a haloalkyl group has from 1 to 6 carbon atoms. In another embodiment, a
5 haloalkyl group has from 1 to 3 carbon atoms. In another embodiment, a haloalkyl group is substituted with from 1 to 3 halogen atoms. Non-limiting examples of haloalkyl groups include -CH₂F, -CHF₂, and -CF₃. The term "C₁-C₄ haloalkyl" refers to a haloalkyl group having from 1 to 4 carbon atoms.

The term "alkoxy" as used herein, refers to an -O-alkyl group, wherein an alkyl
10 group is as defined above. Non-limiting examples of alkoxy groups include methoxy, ethoxy, n-propoxy, isopropoxy, n-butoxy, and *tert*-butoxy. An alkoxy group is bonded via its oxygen atom to the rest of the molecule.

The term "aryl," as used herein, refers to an aromatic monocyclic or multicyclic ring system comprising from about 6 to about 14 carbon atoms. In one embodiment, an aryl
15 group contains from about 6 to 10 carbon atoms (C₆-C₁₀ aryl). In another embodiment an aryl group is phenyl. Non-limiting examples of aryl groups include phenyl and naphthyl.

When a functional group in a compound is termed "protected," that functional group is in modified form to preclude undesired side reactions at the protected site when the compound is subjected to a reaction. The term "PG", as used herein, refers to a protecting group.
20 Those skilled in the art will readily envisage protecting groups suitable for use in compounds and processes according to the disclosure. Suitable protecting groups will be recognized by those of ordinary skill in the art as well as by reference to standard textbooks such as, for example, GREEN'S PROTECTIVE GROUPS IN ORGANIC SYNTHESIS (5th ed., Peter G.M. Wuts ed., 2014). Protecting groups suitable for use in the processes disclosed herein include acid-labile protecting
25 groups. Non-limiting examples of PG suitable for use herein include -S(O)₂R⁸, -C(O)OR⁸, -C(O)R⁸, -CH₂OCH₂CH₂SiR⁸, and -CH₂R⁸, wherein R⁸ is selected from the group consisting of -C₁₋₈ alkyl (straight or branched), -C₃₋₈ cycloalkyl, -CH₂(aryl), and -CH(aryl)₂, wherein each aryl is independently phenyl or naphthyl and each said aryl is optionally independently unsubstituted or substituted with one or more (*e.g.*, 1, 2, or 3) groups independently selected from -OCH₃, -Cl,
30 -Br, and -I.

The term "substituted" means that one or more hydrogens on the atoms of the designated moiety are replaced with a selection from the indicated group, provided that the

atoms' normal valencies under the existing circumstances are not exceeded, and that the substitution results in a stable compound. Combinations of substituents and/or variables are permissible only if such combinations result in stable compounds. By "stable compound" or "stable structure" is meant a compound that is sufficiently robust to survive isolation to a useful
5 degree of purity from a reaction mixture, and formulation into an efficacious therapeutic agent.

When any substituent or variable occurs more than one time in any compound, its definition on each occurrence is independent of its definition at every other occurrence, unless otherwise indicated. For example, description of radicals that include the expression "-N(C₁-C₃ alkyl)₂" means -N(CH₃)(CH₂CH₃), -N(CH₃)(CH₂CH₂CH₃), and -N(CH₂CH₃)(CH₂CH₂CH₃), as
10 well as -N(CH₃)₂, -N(CH₂CH₃)₂, and -N(CH₂CH₂CH₃)₂.

It should also be noted that any carbon or heteroatom with unsatisfied valences in the text, schemes, examples and tables herein is assumed to have sufficient hydrogen atom(s) to satisfy the valences. Any one or more of these hydrogen atoms can be deuterium.

The present disclosure also embraces isotopically-labelled compounds that are
15 identical to those recited herein, but for the fact that one or more atoms are replaced by an atom having an atomic mass or mass number different from the atomic mass or mass number usually found in nature. Examples of isotopes that can be incorporated into compounds of the invention include isotopes of hydrogen, carbon, nitrogen, oxygen, phosphorus, fluorine and chlorine and iodine, such as ²H, ³H, ¹¹C, ¹³C, ¹⁴C, ¹⁵N, ¹⁸O, ¹⁷O, ³¹P, ³²P, ³⁵S, ¹⁸F, ³⁶Cl, and ¹²³I, respectively.

20 Certain isotopically-labelled compounds (*e.g.*, those labeled with ³H and ¹⁴C) are useful in compound and/or substrate tissue distribution assays. Tritiated (*i.e.*, ³H) and carbon-14 (*i.e.*, ¹⁴C) isotopes are particularly preferred for their ease of preparation and detectability. Isotopic substitution at a site where epimerization occurs may slow or reduce the epimerization process and thereby retain the more active or efficacious form of the compound for a longer
25 period of time. Isotopically labeled compounds, in particular those containing isotopes with longer half-lives ($T_{1/2} > 1$ day), can generally be prepared by following procedures analogous to those disclosed in the Schemes and/or in the Examples herein below, by substituting an appropriate isotopically labeled reagent for a non-isotopically labeled reagent.

Compounds herein may contain one or more stereogenic centers and can occur as
30 racemates, racemic mixtures, single enantiomers, diastereomeric mixtures, and individual diastereomers. Additional asymmetric centers may be present depending upon the nature of the various substituents on the molecule. Each such asymmetric center will independently produce

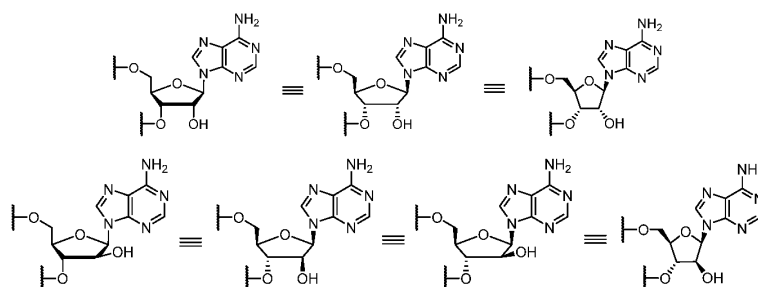
two optical isomers, and all possible optical isomers and diastereomers in mixtures and as pure or partially purified compounds are included within the disclosure. Any formulas, structures, or names of compounds described herein that do not specify a particular stereochemistry are meant to encompass any and all existing isomers as described above and mixtures thereof in any proportion. When stereochemistry is specified, the disclosure is meant to encompass that particular isomer in pure form or as part of a mixture with other isomers in any proportion.

Diastereomeric mixtures can be separated into their individual diastereomers on the basis of their physical chemical differences by methods well known to those skilled in the art, such as, for example, by chromatography and/or fractional crystallization. Enantiomers can be separated by converting the enantiomeric mixture into a diastereomeric mixture by reaction with an appropriate optically active compound (*e.g.*, chiral auxiliary such as a chiral alcohol or Mosher's acid chloride), separating the diastereomers and converting (*e.g.*, hydrolyzing) the individual diastereomers to the corresponding pure enantiomers. Enantiomers can also be separated by use of chiral HPLC column.

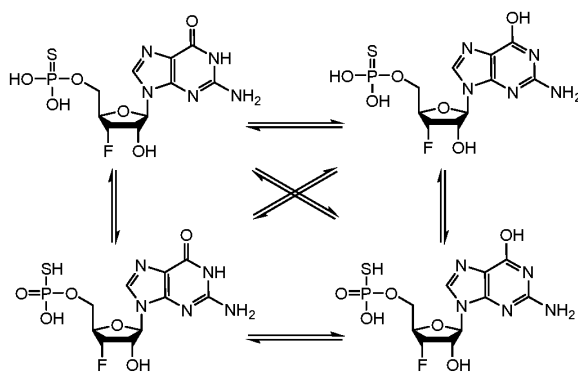
All stereoisomers (for example, geometric isomers, optical isomers, and the like) of disclosed compounds (including those of the salts and solvates of compounds as well as the salts, solvates, and esters of prodrugs), such as those that may exist due to asymmetric carbons on various substituents, including enantiomeric forms (which may exist even in the absence of asymmetric carbons), rotameric forms, atropisomers, and diastereomeric forms, are contemplated within the scope of this disclosure. Individual stereoisomers of compounds may, for example, be substantially free of other isomers, or may be admixed, for example, as racemates or with all other, or other selected, stereoisomers. The chiral centers can have the S or R configuration as defined by the *IUPAC* 1974 Recommendations.

The present disclosure further includes compounds and synthetic intermediates in all their isolated forms. For example, the above-identified compounds are intended to encompass all forms of the compounds such as, any solvates, hydrates, stereoisomers, and tautomers thereof.

Those skilled in the art will recognize that chiral compounds, and in particular, in sugars, can be drawn in a number of different ways that are equivalent. Those skilled in the art will further recognize that the identity and regiochemical position of the substituents on ribose can vary widely and that the same principles of stereochemical equivalence apply regardless of substituent. Non-limiting examples of such equivalence include those exemplified below.



Similarly, those skilled in the art will recognize that certain compounds, and in particular compounds containing certain heteroatoms and double or triple bonds, can be tautomers, structural isomers that readily interconvert. Thus, tautomeric compounds can be drawn in a number of different ways that are equivalent. Non-limiting examples of such tautomers include those exemplified below.



Compounds can form salts that are also within the scope of this disclosure.

Reference to a compound herein is understood to include reference to salts thereof, unless otherwise indicated. The term “salt(s),” as employed herein, denotes acidic salts formed with inorganic and/or organic acids, as well as basic salts formed with inorganic and/or organic bases. In addition, when a compound contains both a basic moiety, such as, but not limited to a pyridine or imidazole, and an acidic moiety, such as, but not limited to a carboxylic acid, zwitterions (“inner salts”) may be formed and are included within the term “salt(s)” as used herein. Pharmaceutically acceptable (*i.e.*, non-toxic, physiologically acceptable) salts are preferred, although other salts are also useful. Salts of the compounds may be formed, for example, by reacting a compound with an amount of acid or base, such as an equivalent amount, in a medium such as one in which the salt precipitates or in an aqueous medium followed by lyophilization.

Exemplary acid addition salts include acetates, ascorbates, benzoates,

benzenesulfonates, bisulfates, borates, butyrates, citrates, camphorates, camphorsulfonates, fumarates, hydrochlorides, hydrobromides, hydroiodides, lactates, maleates, methanesulfonates, naphthalenesulfonates, nitrates, oxalates, phosphates, propionates, salicylates, succinates, sulfates, tartarates, thiocyanates, toluenesulfonates (also known as tosylates,), and the like.

- 5 Additionally, acids that are generally considered suitable for the formation of pharmaceutically useful salts from basic pharmaceutical compounds are discussed, for example, by P. Stahl *et al.*, Camille G. (eds.) HANDBOOK OF PHARMACEUTICAL SALTS: PROPERTIES, SELECTION AND USE (2002) Zurich: Wiley-VCH; S. Berge *et al.*, J. Pharm. Sci. (1977) 66(1) 1-19; P. Gould, *International J. of Pharmaceutics* (1986) 33 201-217; Anderson *et al.*, THE PRACTICE OF
10 MEDICINAL CHEMISTRY (1996), Academic Press, New York; and in THE ORANGE BOOK (Food & Drug Administration, Washington, D.C.). These disclosures are incorporated herein by reference thereto.

- Exemplary basic salts include ammonium salts, alkali metal salts such as sodium, lithium, and potassium salts, alkaline earth metal salts such as calcium and magnesium salts,
15 salts with organic bases (for example, organic amines) such as dicyclohexylamines, t-butyl amines, and salts with amino acids such as arginine, lysine, and the like. Basic nitrogen-containing groups may be quarternized with agents such as lower alkyl halides (*e.g.*, methyl, ethyl, and butyl chlorides, bromides and iodides), dialkyl sulfates (*e.g.*, dimethyl, diethyl, and dibutyl sulfates), long chain halides (*e.g.*, decyl, lauryl, and stearyl chlorides, bromides, and
20 iodides), aralkyl halides (*e.g.*, benzyl and phenethyl bromides), and others.

All such acid salts and base salts are intended to be pharmaceutically acceptable salts within the scope of the invention and all acid and base salts are considered equivalent to the free forms of the corresponding compounds for purposes of the invention.

- One or more compounds herein may exist in unsolvated as well as solvated forms
25 with pharmaceutically acceptable solvents, such as water, ethanol, and the like, and this disclosure is intended to embrace both solvated and unsolvated forms. "Solvate" means a physical association of a compound with one or more solvent molecules. This physical association involves varying degrees of ionic and covalent bonding, including hydrogen bonding. In certain instances, the solvate will be capable of isolation, for example when one or more
30 solvent molecules are incorporated in the crystal lattice of a crystalline solid. "Solvate" encompasses both solution-phase and isolatable solvates. Non-limiting examples of suitable solvates include ethanulates, methanulates, and the like. "Hydrate" is a solvate in which the

solvent molecule is H₂O.

“Protein,” “polypeptide,” and “peptide” are used interchangeably herein to denote a polymer of at least two amino acids covalently linked by an amide bond, regardless of length or post-translational modification (*e.g.*, glycosylation or phosphorylation, lipidation, myristoylation, ubiquitination, etc.). Included within this definition are D- and L-amino acids, and mixtures of D- and L-amino acids, as well as polymers comprising D- and L-amino acids, and mixtures of D- and L-amino acids. Proteins, polypeptides, and peptides may include a tag, such as a histidine tag, which should not be included when determining percentage of sequence identity.

“Amino acid” or “residue” as used in context of the polypeptides disclosed herein refers to the specific monomer at a sequence position. Amino acids are referred to herein by either their commonly known three-letter symbols or by the one-letter symbols recommended by IUPAC-IUB Biochemical Nomenclature Commission. Nucleotides, likewise, may be referred to by their commonly accepted single letter codes.

The abbreviations used for the genetically encoded amino acids are conventional and are as follows: alanine (Ala or A), arginine (Arg or R), asparagine (Asn or N), aspartate (Asp or D), cysteine (Cys or C), glutamate (Glu or E), glutamine (Gln or Q), histidine (His or H), isoleucine (Ile or I), leucine (Leu or L), lysine (Lys or K), methionine (Met or M), phenylalanine (Phe or F), proline (Pro or P), serine (Ser or S), threonine (Thr or T), tryptophan (Trp or W), tyrosine (Tyr or Y), and valine (Val or V).

The abbreviations used for the genetically encoding nucleosides are conventional and are as follows: adenosine (A); guanosine (G); cytidine (C); thymidine (T); and uridine (U). Unless specifically delineated, the abbreviated nucleosides may be either ribonucleosides or 2'-deoxyribonucleosides. The nucleosides may be specified as being either ribonucleosides or 2'-deoxyribonucleosides on an individual basis or on an aggregate basis. When nucleic acid sequences are presented as a string of one-letter abbreviations, the sequences are presented in the 5' to 3' direction in accordance with common convention, and the phosphates are not indicated.

“Derived from” as used herein in the context of enzymes, identifies the originating enzyme, and/or the gene encoding such enzyme, upon which the enzyme was based. For example, the ketoreductase enzyme of SEQ ID NO: 7 was obtained by artificially evolving, over multiple generations the gene encoding the ketoreductase enzyme of SEQ ID NO: 1. Thus, this evolved ketoreductase enzyme is “derived from” the ketoreductase of SEQ ID NO: 1.

“Hydrophilic amino acid or residue” refers to an amino acid or residue having a

side chain exhibiting a hydrophobicity of less than zero according to the normalized consensus hydrophobicity scale of Eisenberg *et al.*, 1984, J. Mol. Biol. 179:125-142. Genetically encoded hydrophilic amino acids include L-Thr (T), L-Ser (S), L-His (H), L-Glu (E), L-Asn (N), L-Gln (Q), L-Asp (D), L-Lys (K), and L-Arg (R).

5 “Acidic amino acid or residue” refers to a hydrophilic amino acid or residue having a side chain exhibiting a pK value of less than about 6 when the amino acid is included in a peptide or polypeptide. Acidic amino acids typically have negatively charged side chains at physiological pH due to loss of a hydrogen ion. Genetically encoded acidic amino acids include L-Glu (E) and L-Asp (D).

10 “Basic amino acid or residue” refers to a hydrophilic amino acid or residue having a side chain exhibiting a pKa value of greater than about 6 when the amino acid is included in a peptide or polypeptide. Basic amino acids typically have positively charged side chains at physiological pH due to association with hydronium ion. Genetically encoded basic amino acids include L-Arg (R) and L-Lys (K).

15 “Polar amino acid or residue” refers to a hydrophilic amino acid or residue having a side chain that is uncharged at physiological pH, but which has at least one bond in which the pair of electrons shared in common by two atoms is held more closely by one of the atoms. Genetically encoded polar amino acids include L-Asn (N), L-Gln (Q), L-Ser (S), and L-Thr (T).

20 “Hydrophobic amino acid or residue” refers to an amino acid or residue having a side chain exhibiting a hydrophobicity of greater than zero according to the normalized consensus hydrophobicity scale of Eisenberg *et al.*, 1984, J. Mol. Biol. 179:125-142. Genetically encoded hydrophobic amino acids include L-Pro (P), L-Ile (I), L-Phe (F), L-Val (V), L-Leu (L), L-Trp (W), L-Met (M), L-Ala (A), and L-Tyr (Y).

25 “Aromatic amino acid or residue” refers to a hydrophilic or hydrophobic amino acid or residue having a side chain that includes at least one aromatic or heteroaromatic ring. Genetically encoded aromatic amino acids include L-Phe (F), L-Tyr (Y), L-His (H), and L-Trp (W). L-His (H) histidine is also classified herein as a hydrophilic residue or as a constrained residue.

30 As used herein, “constrained amino acid or residue” refers to an amino acid or residue that has a constrained geometry. Herein, constrained residues include L-Pro (P) and L-His (H). Histidine has a constrained geometry because it has a relatively small imidazole ring. Proline has a constrained geometry because it also has a five-membered ring.

“Non-polar amino acid or residue” refers to a hydrophobic amino acid or residue that has a side chain that is uncharged at physiological pH and that has bonds in which the pair of electrons shared in common by two atoms is generally held equally by each of the two atoms (*i.e.*, the side chain is not polar). Genetically encoded non-polar amino acids include L-Gly (G),
5 L-Leu (L), L-Val (V), L-Ile (I), L-Met (M), and L-Ala (A).

As used herein, “aliphatic amino acid or residue” refers to a hydrophobic amino acid or residue having an aliphatic hydrocarbon side chain. Genetically encoded aliphatic amino acids include L-Ala (A), L-Val (V), L-Leu (L), and L-Ile (I).

The ability of L-Cys (C) (and other amino acids with -SH containing side chains) to exist in a peptide in either the reduced free -SH or oxidized disulfide-bridged form affects whether L-Cys (C) contributes net hydrophobic or hydrophilic character to a peptide. While L-Cys (C) exhibits a hydrophobicity of 0.29 according to the normalized consensus scale of Eisenberg (Eisenberg *et al.*, 1984, *supra*), it is to be understood that for purposes of the present disclosure, L-Cys (C) is categorized into its own unique group. It is noted that cysteine (or “L-Cys” or “[C]”) is unusual in that it can form disulfide bridges with other L-Cys (C) amino acids
10 to exist in a peptide in either the reduced free -SH or oxidized disulfide-bridged form affects whether L-Cys (C) contributes net hydrophobic or hydrophilic character to a peptide. While L-Cys (C) exhibits a hydrophobicity of 0.29 according to the normalized consensus scale of Eisenberg (Eisenberg *et al.*, 1984, *supra*), it is to be understood that for purposes of the present disclosure, L-Cys (C) is categorized into its own unique group. It is noted that cysteine (or “L-Cys” or “[C]”) is unusual in that it can form disulfide bridges with other L-Cys (C) amino acids
15 Cys” or “[C]”) is unusual in that it can form disulfide bridges with other L-Cys (C) amino acids or other sulfanyl- or sulfhydryl-containing amino acids. The “cysteine-like residues” include cysteine and other amino acids that contain sulfhydryl moieties that are available for formation of disulfide bridges.

As used herein, “small amino acid or residue” refers to an amino acid or residue having a side chain that is composed of a total three or fewer carbon and/or heteroatoms (excluding the α -carbon and hydrogens). The small amino acids or residues may be further categorized as aliphatic, non-polar, polar or acidic small amino acids or residues, in accordance with the above definitions. Genetically-encoded small amino acids include L-Ala (A), L-Val (V), L-Cys (C), L-Asn (N), L-Ser (S), L-Thr (T), and L-Asp (D).
20 having a side chain that is composed of a total three or fewer carbon and/or heteroatoms (excluding the α -carbon and hydrogens). The small amino acids or residues may be further categorized as aliphatic, non-polar, polar or acidic small amino acids or residues, in accordance with the above definitions. Genetically-encoded small amino acids include L-Ala (A), L-Val (V), L-Cys (C), L-Asn (N), L-Ser (S), L-Thr (T), and L-Asp (D).

“Hydroxyl-containing amino acid or residue” refers to an amino acid containing a hydroxyl (-OH) moiety. Genetically-encoded hydroxyl-containing amino acids include L-Ser (S) L-Thr (T), and L-Tyr (Y).
25 “Hydroxyl-containing amino acid or residue” refers to an amino acid containing a hydroxyl (-OH) moiety. Genetically-encoded hydroxyl-containing amino acids include L-Ser (S) L-Thr (T), and L-Tyr (Y).

As used herein, “polynucleotide” and “nucleic acid” refer to two or more nucleotides that are covalently linked together. The polynucleotide may be wholly comprised of ribonucleotides (*i.e.*, RNA), wholly comprised of 2' deoxyribonucleotides (*i.e.*, DNA), or
30 ribonucleotides (*i.e.*, RNA), wholly comprised of 2' deoxyribonucleotides (*i.e.*, DNA), or comprised of mixtures of ribo- and 2' deoxyribonucleotides. While the nucleosides will typically be linked together via standard phosphodiester linkages, the polynucleotides may

include one or more non-standard linkages. The polynucleotide may be single-stranded or double-stranded, or the polynucleotide may include both single-stranded regions and double-stranded regions. Moreover, while a polynucleotide will typically be composed of the naturally occurring encoding nucleobases (*i.e.*, adenine, guanine, uracil, thymine, and cytosine), it may
5 include one or more modified and/or synthetic nucleobases, such as, for example, inosine, xanthine, hypoxanthine, etc. In some embodiments, such modified or synthetic nucleobases are nucleobases encoding amino acid sequences.

As used herein, “nucleoside” refers to glycosylamines comprising a nucleobase (*i.e.*, a nitrogenous base), and a 5-carbon sugar (*e.g.*, ribose or deoxyribose). Non-limiting
10 examples of nucleosides include cytidine, uridine, adenosine, guanosine, thymidine, and inosine. In contrast, the term “nucleotide” refers to the glycosylamines comprising a nucleobase, a 5-carbon sugar, and one or more phosphate groups. In some embodiments, nucleosides can be phosphorylated by kinases to produce nucleotides.

As used herein, “nucleoside diphosphate” refers to glycosylamines comprising a
15 nucleobase (*i.e.*, a nitrogenous base), a 5-carbon sugar (*e.g.*, ribose or deoxyribose), and a diphosphate (*i.e.*, pyrophosphate) moiety. In some embodiments herein, “nucleoside diphosphate” is abbreviated as “NDP.” Non-limiting examples of nucleoside diphosphates include cytidine diphosphate (CDP), uridine diphosphate (UDP), adenosine diphosphate (ADP), guanosine diphosphate (GDP), thymidine diphosphate (TDP), and inosine diphosphate (IDP).
20 The terms “nucleoside” and “nucleotide” may be used interchangeably in some contexts.

As used herein, “nucleoside triphosphate” refers to glycosylamines comprising a nucleobase (*i.e.*, a nitrogenous base), a 5-carbon sugar (*e.g.*, ribose or deoxyribose), and a triphosphate moiety. In some embodiments herein, “nucleoside triphosphate” is abbreviated as “NTP.” Non-limiting examples of nucleoside triphosphates include cytidine triphosphate (CTP),
25 uridine triphosphate (UTP), adenosine triphosphate (ATP), guanosine triphosphate (GTP), thymidine triphosphate (TTP), and inosine triphosphate (ITP). The terms “nucleoside” and “nucleotide” may be used interchangeably in some contexts.

As used herein, “conservative amino acid substitution” refers to a substitution of a residue with a different residue having a similar side chain, and thus typically involves
30 substitution of the amino acid in the polypeptide with amino acids within the same or similar defined class of amino acids. By way of example and not limitation, in some embodiments, an amino acid with an aliphatic side chain is substituted with another aliphatic amino acid (*e.g.*,

alanine, valine, leucine, and isoleucine); an amino acid with an hydroxyl side chain is substituted with another amino acid with an hydroxyl side chain (*e.g.*, serine and threonine); an amino acids having aromatic side chains is substituted with another amino acid having an aromatic side chain (*e.g.*, phenylalanine, tyrosine, tryptophan, and histidine); an amino acid with a basic side chain is substituted with another amino acid with a basic side chain (*e.g.*, lysine and arginine); an amino acid with an acidic side chain is substituted with another amino acid with an acidic side chain (*e.g.*, aspartic acid and glutamic acid); and/or a hydrophobic or hydrophilic amino acid is replaced with another hydrophobic or hydrophilic amino acid, respectively.

As used herein, “non-conservative substitution” refers to substitution of an amino acid in the polypeptide with an amino acid with significantly differing side chain properties. Non-conservative substitutions may use amino acids between, rather than within, the defined groups and affects (a) the structure of the peptide backbone in the area of the substitution (*e.g.*, proline for glycine) (b) the charge or hydrophobicity, or (c) the bulk of the side chain. By way of example and not limitation, an exemplary non-conservative substitution can be an acidic amino acid substituted with a basic or aliphatic amino acid; an aromatic amino acid substituted with a small amino acid; and a hydrophilic amino acid substituted with a hydrophobic amino acid.

As used herein, “deletion” refers to modification to the polypeptide by removal of one or more amino acids from the reference polypeptide. Deletions can comprise removal of 1 or more amino acids, 2 or more amino acids, 5 or more amino acids, 10 or more amino acids, 15 or more amino acids, or 20 or more amino acids, up to 10% of the total number of amino acids, or up to 20% of the total number of amino acids making up the reference enzyme while retaining enzymatic activity and/or retaining the improved properties of an evolved enzyme. Deletions can be directed to the internal portions and/or terminal portions of the polypeptide. In various embodiments, the deletion can comprise a continuous segment or can be discontinuous. Deletions are typically indicated by “-” in amino acid sequences.

As used herein, “insertion” refers to modification to the polypeptide by addition of one or more amino acids from the reference polypeptide. Insertions can be in the internal portions of the polypeptide, or to the carboxy or amino terminus. Insertions as used herein include fusion proteins as is known in the art. The insertion can be a contiguous segment of amino acids or separated by one or more of the amino acids in the naturally occurring polypeptide.

The term “amino acid substitution set” or “substitution set” refers to a group of

amino acid substitutions in a polypeptide sequence, as compared to a reference sequence. A substitution set can have 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, or more amino acid substitutions.

A “functional fragment” and “biologically active fragment” are used interchangeably herein to refer to a polypeptide that has an amino-terminal and/or carboxy-terminal deletion(s) and/or internal deletions, but where the remaining amino acid sequence is identical to the corresponding positions in the sequence to which it is being compared and that retains substantially all of the activity of the full-length polypeptide.

As used herein, “isolated polypeptide” refers to a polypeptide that is substantially separated from other contaminants that naturally accompany it (*e.g.*, protein, lipids, and polynucleotides). The term embraces polypeptides that have been removed or purified from their naturally occurring environment or expression system (*e.g.*, within a host cell or via *in vitro* synthesis). The recombinant polypeptides may be present within a cell, present in the cellular medium, or prepared in various forms, such as lysates or isolated preparations. As such, in some embodiments, the recombinant polypeptides can be an isolated polypeptide.

As used herein, “substantially pure polypeptide” or “purified protein” refers to a composition in which the polypeptide species is the predominant species present (*i.e.*, on a molar or weight basis it is more abundant than any other individual macromolecular species in the composition), and is generally a substantially purified composition when the object species comprises at least about 50 percent of the macromolecular species present by mole or % weight. However, in some embodiments, an enzyme comprising composition comprises enzymes that are less than 50% pure (*e.g.*, about 10%, about 20%, about 30%, about 40%, or about 50%). Generally, a substantially pure enzyme or polypeptide composition comprises about 60% or more, about 70% or more, about 80% or more, about 90% or more, about 95% or more, and about 98% or more of all macromolecular species by mole or % weight present in the composition. In some embodiments, the object species is purified to essential homogeneity (*i.e.*, contaminant species cannot be detected in the composition by conventional detection methods) wherein the composition consists essentially of a single macromolecular species. Solvent species, small molecules (<500 Daltons), and elemental ion species are not considered macromolecular species. In some embodiments, the isolated recombinant polypeptides are substantially pure polypeptide compositions.

“Improved enzyme property” refers to an enzyme that exhibits an improvement in

any enzyme property as compared to a reference enzyme. For the enzymes described herein, the comparison is generally made to the wild-type enzyme, although in some embodiments, the reference enzyme can be another improved enzyme. Enzyme properties for which improvement is desirable include, but are not limited to, enzymatic activity (which can be expressed in terms of percent conversion of the substrate), thermal stability, pH activity profile, cofactor requirements, refractoriness to inhibitors (*e.g.*, product inhibition), stereospecificity, and stereoselectivity (including enantioselectivity).

“Increased enzymatic activity” refers to an improved property of the enzymes, which can be represented by an increase in specific activity (*e.g.*, product produced/time/weight protein) or an increase in percent conversion of the substrate to the product (*e.g.*, percent conversion of starting amount of substrate to product in a specified time period using a specified amount of enzyme) as compared to the reference enzyme. Exemplary methods to determine enzyme activity are provided in the Examples. Any property relating to enzyme activity may be affected, including the classical enzyme properties of K_m , V_{max} , or k_{cat} , changes of which can lead to increased enzymatic activity. Improvements in enzyme activity can be from about 1.5 times the enzymatic activity of the corresponding wild-type enzyme, to as much as 2 times. 5 times, 10 times, 20 times, 25 times, 50 times, 75 times, 100 times, 150 times, 200 times, 500 times, 1000 times, 3000 times, 5000 times, 7000 times or more enzymatic activity than the naturally occurring enzyme or another enzyme from which the polypeptides were derived. In specific embodiments, the enzyme exhibits improved enzymatic activity in the range of 150 to 3000 times, 3000 to 7000 times, or more than 7000 times greater than that of the parent enzyme. It is understood by the skilled artisan that the activity of any enzyme is diffusion limited such that the catalytic turnover rate cannot exceed the diffusion rate of the substrate, including any required cofactors. The theoretical maximum of the diffusion limit, or k_{cat}/K_m , is generally about 10^8 to 10^9 ($M^{-1}s^{-1}$). Hence, any improvements in the enzyme activity will have an upper limit related to the diffusion rate of the substrates acted on by the enzyme. Enzyme activity can be measured by any one of standard assays used for measuring kinase activity, or via a coupled assay with an nucleoside phosphorylase enzyme which is capable of catalyzing reaction between the polypeptide product and a nucleoside base to afford a nucleoside, or by any of the traditional methods for assaying chemical reactions, including but not limited to HPLC, HPLC-MS, UPLC, UPLC-MS, TLC, and NMR. Comparisons of enzyme activities are made using a defined preparation of enzyme, a defined assay under a set condition, and one or more defined substrates,

as further described in detail herein. Generally, when lysates are compared, the numbers of cells and the amount of protein assayed are determined as well as use of identical expression systems and identical host cells to minimize variations in amount of enzyme produced by the host cells and present in the lysates.

5 As used herein, a “vector” is a DNA construct for introducing a DNA sequence into a cell. In some embodiments, the vector is an expression vector that is operably linked to a suitable control sequence capable of effecting the expression in a suitable host of the polypeptide encoded in the DNA sequence. In some embodiments, an “expression vector” has a promoter sequence operably linked to the DNA sequence (*e.g.*, transgene) to drive expression in a host
10 cell, and in some embodiments, also comprises a transcription terminator sequence.

 As used herein, the term “expression” includes any step involved in the production of the polypeptide including, but not limited to, transcription, post-transcriptional modification, translation, and post-translational modification. In some embodiments, the term also encompasses secretion of the polypeptide from a cell.

15 As used herein, the term “produces” refers to the production of proteins and/or other compounds by cells. It is intended that the term encompass any step involved in the production of polypeptides including, but not limited to, transcription, post-transcriptional modification, translation, and post-translational modification. In some embodiments, the term also encompasses secretion of the polypeptide from a cell.

20 As used herein, an amino acid or nucleotide sequence (*e.g.*, a promoter sequence, signal peptide, terminator sequence, etc.) is “heterologous” to another sequence with which it is operably linked if the two sequences are not associated in nature. For example, a “heterologous polynucleotide” is any polynucleotide that is introduced into a host cell by laboratory techniques, and the term includes polynucleotides that are removed from a host cell, subjected to laboratory
25 manipulation, and then reintroduced into a host cell.

 As used herein, the terms “host cell” and “host strain” refer to suitable hosts for expression vectors comprising DNA provided herein (*e.g.*, the polynucleotides encoding the variants). In some embodiments, the host cells are prokaryotic or eukaryotic cells that have been transformed or transfected with vectors constructed using recombinant DNA techniques as
30 known in the art.

 The term “analogue” means a polypeptide having more than 70% sequence identity but less than 100% sequence identity (*e.g.*, more than 75%, 78%, 80%, 83%, 85%, 88%,

90%, 91%, 92%, 93%, 94%, 95%, 96%, 97%, 98%, or 99% sequence identity) with a reference polypeptide. In some embodiments, “analogues” means polypeptides that contain one or more non-naturally occurring amino acid residues including, but not limited, to homoarginine, ornithine and norvaline, as well as naturally occurring amino acids. In some embodiments, analogues also include one or more D-amino acid residues and non-peptide linkages between two or more amino acid residues.

As used herein, “EC” number refers to the Enzyme Nomenclature of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (NC-IUBMB). The IUBMB biochemical classification is a numerical classification system for enzymes based on the chemical reactions they catalyze.

As used herein, “ATCC” refers to the American Type Culture Collection whose biorepository collection includes genes and strains.

As used herein, “NCBI” refers to National Center for Biological Information and the sequence databases provided therein.

“Coding sequence” refers to that portion of a nucleic acid (*e.g.*, a gene) that encodes an amino acid sequence of a protein.

“Naturally occurring” or “wild-type” refers to a form found in nature. For example, a naturally occurring or wild-type polypeptide or polynucleotide sequence is a sequence present in an organism that can be isolated from a source in nature and that has not been intentionally modified by human manipulation, with the sole exception that wild-type polypeptide or polynucleotide sequences as identified herein may include a tag, such as a histidine tag, which should not be included when determining percentage of sequence identity. Herein, “wild-type” polypeptide or polynucleotide sequences may be denoted “WT”.

“Recombinant” when used with reference to, *e.g.*, a cell, nucleic acid, or polypeptide, refers to a material, or a material corresponding to the natural or native form of the material, that has been modified in a manner that would not otherwise exist in nature, or is identical thereto but produced or derived from synthetic materials and/or by manipulation using recombinant techniques. Non-limiting examples include, among others, recombinant cells expressing genes that are not found within the native (non-recombinant) form of the cell or express native genes that are otherwise expressed at a different level.

“Percentage of sequence identity,” “percent identity,” and “percent identical” are used herein to refer to comparisons between polynucleotide sequences or polypeptide sequences,

and are determined by comparing two optimally aligned sequences over a comparison window, wherein the portion of the polynucleotide or polypeptide sequence in the comparison window may comprise additions or deletions (*i.e.*, gaps) as compared to the reference sequence for optimal alignment of the two sequences. The percentage is calculated by determining the
5 number of positions at which either the identical nucleic acid base or amino acid residue occurs in both sequences or a nucleic acid base or amino acid residue is aligned with a gap to yield the number of matched positions, dividing the number of matched positions by the total number of positions in the window of comparison and multiplying the result by 100 to yield the percentage of sequence identity. Determination of optimal alignment and percent sequence identity is
10 performed using the BLAST and BLAST 2.0 algorithms (*see e.g.*, Altschul *et al.*, 1990, J. Mol. Biol. 215: 403-410; and Altschul *et al.*, 1977, Nucleic Acids Res. 3389-3402). Software for performing BLAST analyses is publicly available through the National Center for Biotechnology Information website.

Briefly, the BLAST analyses involve first identifying high scoring sequence pairs
15 (HSPs) by identifying short words of length W in the query sequence, which either match or satisfy some positive-valued threshold score T when aligned with a word of the same length in a database sequence. T is referred to as, the neighborhood word score threshold (Altschul *et al.*, *supra*). These initial neighborhood word hits act as seeds for initiating searches to find longer HSPs containing them. The word hits are then extended in both directions along each sequence
20 for as far as the cumulative alignment score can be increased. Cumulative scores are calculated using, for nucleotide sequences, the parameters M (reward score for a pair of matching residues; always >0) and N (penalty score for mismatching residues; always <0). For amino acid sequences, a scoring matrix is used to calculate the cumulative score. Extension of the word hits in each direction are halted when: the cumulative alignment score falls off by the quantity X
25 from its maximum achieved value; the cumulative score goes to zero or below, due to the accumulation of one or more negative-scoring residue alignments; or the end of either sequence is reached. The BLAST algorithm parameters W, T, and X determine the sensitivity and speed of the alignment. The BLASTN program (for nucleotide sequences) uses as defaults a word length (W) of 11, an expectation (E) of 10, M = 5, N = -4, and a comparison of both strands. For
30 amino acid sequences, the BLASTP program uses as defaults a word length (W) of 3, an expectation (E) of 10, and the BLOSUM62 scoring matrix (*see* Henikoff and Henikoff, 1989, PROC. NATL. ACAD. SCI. USA 89:10915).

Numerous other algorithms are available that function similarly to BLAST in providing percent identity for two sequences. Optimal alignment of sequences for comparison can be conducted, *e.g.*, by the local homology algorithm of Smith and Waterman, 1981, *Adv. Appl. Math.* 2:482, by the homology alignment algorithm of Needleman and Wunsch, 1970, *J. Mol. Biol.* 48:443, by the search for similarity method of Pearson and Lipman, 1988, *Proc. Natl. Acad. Sci. USA* 85:2444, by computerized implementations of these algorithms (GAP, BESTFIT, FASTA, and TFASTA in the GCG Wisconsin Software Package), or by visual inspection (*see generally*, *Current Protocols in Molecular Biology*, F. M. Ausubel *et al.*, eds., *Current Protocols*, a joint venture between Greene Publishing Associates, Inc. and John Wiley & Sons, Inc., (1995 Supplement) (Ausubel)). Additionally, determination of sequence alignment and percent sequence identity can employ the BESTFIT or GAP programs in the GCG Wisconsin Software package (Accelrys, Madison WI), using default parameters provided.

“Substantial identity” refers to a polynucleotide or polypeptide sequence that has at least 80 percent sequence identity, preferably at least 85 percent sequence identity, more preferably at least 89 percent sequence identity, more preferably at least 95 percent sequence identity, and even more preferably at least 99 percent sequence identity as compared to a reference sequence over a comparison window of at least 20 residue positions, frequently over a window of at least 30-50 residues, wherein the percentage of sequence identity is calculated by comparing the reference sequence to a sequence that includes deletions or additions which total 20 percent or less of the reference sequence over the window of comparison. In specific embodiments applied to polypeptides, the term “substantial identity” means that two polypeptide sequences, when optimally aligned, such as by the programs GAP or BESTFIT using default gap weights, share at least 80 percent sequence identity, preferably at least 89 percent sequence identity, more preferably at least 95 percent sequence identity or more (*e.g.*, 99 percent sequence identity). Preferably, residue positions which are not identical differ by conservative amino acid substitutions.

“Corresponding to”, “reference to” or “relative to” when used in the context of the numbering of a given amino acid or polynucleotide sequence refers to the numbering of the residues of a specified reference sequence when the given amino acid or polynucleotide sequence is compared to the reference sequence. In other words, the residue number or residue position of a given polymer is designated with respect to the reference sequence rather than by the actual numerical position of the residue within the given amino acid or polynucleotide

sequence. For example, a given amino acid sequence can be aligned to a reference sequence by introducing gaps to optimize residue matches between the two sequences. In these cases, although the gaps are present, the numbering of the residue in the given amino acid or polynucleotide sequence is made with respect to the reference sequence to which it has been aligned.

“Stereoselectivity” refers to the preferential formation in a chemical or enzymatic reaction of one stereoisomer over another. Stereoselectivity can be partial, where the formation of one stereoisomer is favored over the other, or it may be complete where only one stereoisomer is formed. When the stereoisomers are enantiomers, the stereoselectivity is referred to as enantioselectivity, the fraction (typically reported as a percentage) of one enantiomer in the sum of both. It is commonly alternatively reported in the art (typically as a percentage) as the enantiomeric excess (EE) calculated therefrom according to the formula $[\text{major enantiomer} - \text{minor enantiomer}] / [\text{major enantiomer} + \text{minor enantiomer}]$. Where the stereoisomers are diastereoisomers, the stereoselectivity is referred to as diastereoselectivity, the fraction (typically reported as a percentage) of one diastereomer in a mixture of two diastereomers, commonly alternatively reported as the diastereomeric excess (DE). Enantiomeric excess and diastereomeric excess are types of stereomeric excess.

“Highly stereoselective” refers to a chemical or enzymatic reaction that is capable of converting a substrate to its corresponding product with at least about 85% stereoisomeric excess.

“Chemoselectivity” refers to the preferential formation in a chemical or enzymatic reaction of one product over another.

“Conversion” refers to the enzymatic transformation of a substrate to the corresponding product. “Percent conversion” refers to the percent of the substrate that is converted to the product within a period of time under specified conditions. Thus, for example, the “enzymatic activity” or “activity” of a polypeptide can be expressed as “percent conversion” of the substrate to the product.

“Chiral alcohol” refers to amines of general formula $R^1\text{-CH(OH)-R}^2$ wherein R^1 and R^2 are nonidentical and is employed herein in its broadest sense, including a wide variety of aliphatic and alicyclic compounds of different, and mixed, functional types, characterized by the presence of a primary hydroxyl group bound to a secondary carbon atom which, in addition to a hydrogen atom, carries either (i) a divalent group forming a chiral cyclic structure, or (ii) two

substituents (other than hydrogen) differing from each other in structure or chirality. Divalent groups forming a chiral cyclic structure include, for example, 2-methylbutane-1,4-diyl, pentane-1,4-diyl, hexane-1,4-diyl, hexane-1,5-diyl, 2-methylpentane-1,5-diyl. The two different substituents on the secondary carbon atom (R^1 and R^2 above) also can vary widely and include

5 alkyl, aralkyl, aryl, halo, hydroxy, lower alkyl, lower alkoxy, lower alkylthio, cycloalkyl, carboxy, carboalkoxy, carbamoyl, mono- and di-(lower alkyl) substituted carbamoyl, trifluoromethyl, phenyl, nitro, amino, mono- and di-(lower alkyl) substituted amino, alkylsulfonyl, arylsulfonyl, alkylcarboxamido, arylcarboxamido, etc., as well as alkyl, aralkyl, or aryl substituted by the foregoing.

10 Immobilized enzyme preparations have a number of recognized advantages. They can confer shelf life to enzyme preparations, they can improve reaction stability, they can enable stability in organic solvents, they can aid in protein removal from reaction streams, as examples. "Stable" refers to the ability of the immobilized enzymes to retain their structural conformation and/or their activity in a solvent system that contains organic solvents. Stable

15 immobilized enzymes lose less than 10% activity per hour in a solvent system that contains organic solvents. Stable immobilized enzymes lose less than 9% activity per hour in a solvent system that contains organic solvents. Preferably, the stable immobilized enzymes lose less than 8% activity per hour in a solvent system that contains organic solvents. Preferably, the stable immobilized enzymes lose less than 7% activity per hour in a solvent system that contains

20 organic solvents. Preferably, the stable immobilized enzymes lose less than 6% activity per hour in a solvent system that contains organic solvents. Preferably, the stable immobilized enzymes lose less than 5% activity per hour in a solvent system that contains organic solvents. Preferably, the stable immobilized enzymes lose less than 4% activity per hour in a solvent system that contains organic solvents. Preferably, the stable immobilized enzymes lose less than 3% activity per hour

25 in a solvent system that contains organic solvents. Preferably, the stable immobilized enzymes lose less than 2% activity per hour in a solvent system that contains organic solvents. Preferably, the stable immobilized enzymes lose less than 1% activity per hour in a solvent system that contains organic solvents.

"Thermostable" refers to a polypeptide that maintains similar activity (more than

30 60% to 80%, for example) after exposure to elevated temperatures (*e.g.*, 40°C to 80°C) for a period of time (*e.g.*, 0.5h to 24h) compared to the untreated enzyme.

"Solvent stable" refers to a polypeptide that maintains similar activity (more than

e.g., 60% to 80%) after exposure to varying concentrations (*e.g.*, 5% to 99%) of solvent (isopropyl alcohol, tetrahydrofuran, 2-methyltetrahydrofuran, acetone, toluene, butylacetate, methyl *tert*-butylether, etc.) for a period of time (*e.g.*, 0.5h to 24h) compared to the untreated enzyme.

- 5 “pH stable” refers to a polypeptide that maintains similar activity (more than *e.g.*, 60% to 80%) after exposure to high or low pH (*e.g.*, 4.5 to 6 or 8 to 12) for a period of time (*e.g.*, 0.5h to 24h) compared to the untreated enzyme.

 “Thermo- and solvent stable” refers to a polypeptide that is both thermostable and solvent stable.

- 10 As used herein, the terms “biocatalysis,” “biocatalytic,” “biotransformation,” and “biosynthesis” refer to the use of enzymes to perform chemical reactions on organic compounds.

 The term “effective amount” means an amount sufficient to produce the desired result. One of general skill in the art may determine what the effective amount by using routine experimentation.

- 15 The terms “isolated” and “purified” are used to refer to a molecule (*e.g.*, an isolated nucleic acid, polypeptide, etc.) or other component that is removed from at least one other component with which it is naturally associated. The term “purified” does not require absolute purity, rather it is intended as a relative definition.

- Exemplary methods and materials are described herein, although methods and
20 materials similar or equivalent to those described herein can also be used in the practice or testing of the present disclosure. The materials, methods, and examples are illustrative only and not intended to be limiting.

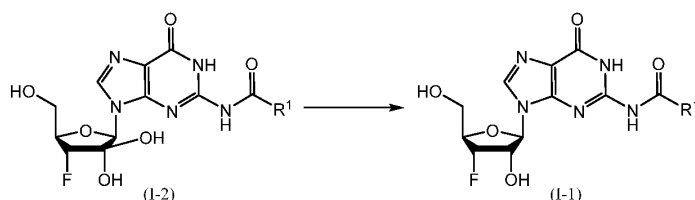
ABBREVIATIONS

ACN, MeCN	Acetonitrile
DMSO	Dimethyl sulfoxide
g	Grams
h	Hour
Hz	Hertz
IPA	Isopropyl alcohol
IPTG	Isopropyl β -D-1-thiogalactopyranoside
<i>J</i>	NMR Coupling constant

L	Liter
M	Molar, moles per liter
mg	Milligrams
min	Minute(s)
mL, ml	Milliliters
mm	Millimeter
mM	Millimole per liter
nm	Nanometer
NaCl	Sodium chloride
NADP	Nicotinamide adenine dinucleotide phosphate
OD ₆₀₀	Optical density at 600nm wavelength
PTFE	Polytetrafluoroethylene
RH	Relative humidity
RPM, rpm	Revolutions per minute
RT	Room temperature, approximately 25°C
TFA	Trifluoroacetic acid
THF	Tetrahydrofuran
UPLC	Ultra-Performance Liquid Chromatography
vol	Volumes
v/v	Volume per volume
μg, ug	Micrograms
μL, μl, uL, ul	Microliters

Ketoreductases

- This disclosure relates to ketoreductase enzymes capable of reducing ketones to chiral alcohols in the synthesis of nucleosides, particularly fluoronucleosides. In embodiments,
- 5 the ketoreductase enzymes are capable of the following conversion:



In particular embodiments, the ketoreductase enzymes are capable of the following conversion:



In embodiments, the ketoreductase enzymes described herein have an amino acid sequence that has one or more amino acid differences as compared to a reference amino acid sequence of a wild-type ketoreductase or a modified ketoreductase that result in an improved property of the enzyme for the defined keto substrate.

The ketoreductase enzymes described herein are the product of directed evolution from a commercially available, wild-type ketoreductase, which was identified by screening a Prozomix metagenomic panel and which has the amino acid sequence as set forth below in SEQ ID NO: 1.

MHHHHHHNKTIVVTGGTKGIGRAIVEKFAKEGFTVLTCARTKGDNFPENVHFFK
ADLSKKVEVLAFADFIKQTVNQVDILVNNTGFFLPGEINNEAEGTLEAMIETNLY
SAYYLTRALVGDMITKKEGHIFNICSTASITAYTNGGSYCISKFALLGMSKVLREE
LKPHHVRVTSILPGATLTDSWAGVELPAERFIASEDIAQIVWTAHCLPSTTVLEEIL
IRPQLGDL (SEQ ID NO: 1)

In embodiments, the ketoreductase enzymes of the disclosure may demonstrate improvements relative the ketoreductase enzymes of SEQ ID NO: 1, such as increases in enzyme activity, stereoselectivity, stereospecificity, thermostability, solvent stability, or reduced product inhibition.

In some embodiments, the ketoreductase enzymes of the disclosure may demonstrate improvements in the rate of enzymatic activity, i.e., the rate of converting the substrate to the product. In some embodiments, the ketoreductase polypeptides are capable of converting the substrate to the product at a rate that is at least 1.5-times, 2-times, 3-times, 4-times, 5-times, 10-times, 25-times, 50-times, 100-times, 150-times, 200-times, 400-times, 1000-times, 3000-times, 5000-times, 7000-times or more than 7000-times the rate exhibited by the enzyme of SEQ ID NO: 1.

In some embodiments, such ketoreductase polypeptides are also capable of converting the substrate to the product with a percent stereometric excess of at least about 80%. In some embodiments, such ketoreductase polypeptides are also capable of converting the

substrate to the product with a percent stereometric excess of at least about 90%. In some embodiments, such ketoreductase polypeptides are also capable of converting the substrate to the product with a percent stereometric excess of at least about 99%.

In some embodiments, the ketoreductase polypeptide is highly stereoselective, wherein the polypeptide can reduce the substrate to the product in greater than about 99%, 99.1%, 99.2%, 99.3%, 99.4%, 99.5%, 99.6%, 99.7%, 99.8%, or 99.9% stereometric excess.

In embodiments, the ketoreductase enzymes described herein include ketoreductase enzymes having the amino acid sequence as set forth below in SEQ ID NO: 2.

MHHHHHHNKTIVVTGGTKGIGRAIVEKFAKEGFTVLT CARTKGDNFPENVHFFK
 10 ADLSKKVEVLAFADFIKQTVNQVDILVNNTGFFLPGEINNEAEGTLEAMIETNLY
 SAYYLTRALVGDMITKKEGHIFNICSTASITAYTNGGSYCISKFALLGMSKVLREE
 LKPHHVRVTSILPGATLND SWAGVELPAERFIASEDIAQIVWTAHCLPSTTVLEEI
 LIRPQLGDL (SEQ ID NO: 2)

In additional embodiments, the ketoreductase enzymes described herein include ketoreductase enzymes having the amino acid sequence as set forth below in SEQ ID NO: 3.

MHHHHHHNKTIVVTGGTKGIGRAIVEKFAKEGFTVLT CARTKGDNFPENVHFFK
 ADLSKKVEVLAFADFIKQTVNQVDILVNNTGFFLPGEINNEAEGTLEAMIETNLY
 SAYYLTRALVGDMITKKEGHIFNICSYASITAYTSGGSYCISKFALLGMSKVLREE
 LKPHHVRVTSILPGATLND SWAKVELPAERFIASEDIAQIVWTAHCLPSTTVLEEI
 20 LIRPQLGDL (SEQ ID NO: 3)

In additional embodiments, the ketoreductase enzymes described herein include ketoreductase enzymes having the amino acid sequence as set forth below in SEQ ID NO: 4.

MHHHHHPATIVVTGGTKGIGRAIVEKFAKEGFTVLT CARTKGDNFPENVHFFKA
 DLSKKVEVLAFADFIKQTVNQVDILVNNTGFFLPGEINNEAEGTLEAMIETNLYS
 25 AYYLTRALVGDMITKKEGHIFNICSYASITAYTSGGSYCISK TALLGMSKVLREEL
 KPHHVRVTSILPGATLND SWAKVELPAEWFIASEDIAQIVWTAHCLPSTTVLEEIL
 IRPQLGDL (SEQ ID NO: 4)

In additional embodiments, the ketoreductase enzymes described herein include ketoreductase enzymes having the amino acid sequence as set forth below in SEQ ID NO: 5.

MHHHHHPATIVVTGGTKGIGRAIVEKFAKEGFTVLT CARTKGDNFPENVHFFKA
 30 DLSKKVEVLAFADFIKQTVNQVDILVNNTGWFLPGEINNEAEGTLEAMIETNLYS
 AYYLTRALVGDMITKKEGHIFNICSYASIVPYTSGGSYCISKTAQLGMSKVLREEL

KPHHVRVTSILPGAVLNDWSAKVELPAEWFIASEDIAQIVWTAHCLPSTTVLEEIL
IRPQLGDL (SEQ ID NO: 5)

In additional embodiments, the ketoreductase enzymes described herein include ketoreductase enzymes having the amino acid sequence as set forth below in SEQ ID NO: 6.

5 MHHHHHPATIVVTGGTKGIGRAIVEKFAKEGFTVLT CARTKGDNFPENVHFFKA
DLSKKVEVLAFADFIKQTVNQVDILVNNTGWFLPGEINNEAEGTLEAMIETNLYS
AAYLTRALVGDMITKKEGHIFNICSYASIVPYTSGGSYCISKTAQLGMSKVLREEL
KPHHVRVTSILPGAVLNDWSAKVELPAELFIAPEDIAQIVWTAHCLPSTTVLEEILI
RPQTGDL (SEQ ID NO: 6)

10 In additional embodiments, the ketoreductase enzymes described herein include ketoreductase enzymes having the amino acid sequence as set forth below in SEQ ID NO: 7.

MHHHHHPATIVVTGGTKGIGRAIVEKFAKEGFTVLT CARTAGDNFPENVHFFKA
DLSKKVEVLAFADFIKQTVNQVDILVNNTGWFLPGEINNEEEGTLEAMIETNLYS
AAYLTRALVGDMITKKEGHIFNICSYASIVPYTSGGSYCISKTAQLGMSKVLREEL
15 KPHHVRVTSILPGAVLNDWSAKVELPAELFIAPEDIAQIVWTAHCLPSTTVLEEILI
RPQEGDL (SEQ ID NO: 7)

In additional embodiments, the ketoreductase enzymes described herein include ketoreductase enzymes having the amino acid sequence as set forth below in SEQ ID NO: 8.

MHHHHHPATIVVTGGTKGIGRAIVEKFAKEGFTVLT CARTAGDNFPENVHFFKA
20 DLSKKVEVLAFADFIKQTVNQVDILVNNTGHFLPGEINNEEEGTLEAMIETNLYS
AAYLTRALVGDMITKKEGHIFNICSYASIVPYTSGGSYCISKTAELGMSKVLREEL
KPHHVRVTSILPGAVLNDWSAKAELPAELFIAPEDIAQIVWTAHCLPSTTVLEEILI
RPQEGDL (SEQ ID NO: 8)

In additional embodiments, the ketoreductase enzymes described herein include ketoreductase enzymes having the amino acid sequence as set forth below in SEQ ID NO: 9.

MHHHHHPATIVVTGGTKGIGRAIVEKFAKEGFTVLT CARTAGDNFPENVHFFKA
DLSKKVEVLAFADFIKQTVNQVDILVNNTGHFLPGEINNEEEGTLEAMIETNLYS
AAYLTRALVGDMITKKEGHIFNICSYASIVPYTSGGSYCISKTAELAMSRVLREEL
KPHHVRVTSILPGAVLNDNWAKAELPAELFIAPEDIAQIVWTAHCLPSTTVLEEIL
30 IRPTEGDL (SEQ ID NO: 9)

In additional embodiments, the ketoreductase enzymes described herein include ketoreductase enzymes having the amino acid sequence as set forth below in SEQ ID NO: 10.

MHHHHHPATIVVTGGTKGIGRAIVEKFAKEGFTVLT CARTAGDNFPENVHFFKA
 DLSKKVEVLAFADFIKATVNQVDILVNNTGHFLPGEINNEEEGTLEAMIETNLYS
 AYYLTRALVGDMITKKEGHIFNICSYASIVPYTPGGSYCISKTAELAMSRVLREEL
 KPHHVVRTSILPGAVLNDNWAKAELPAELFIAPEDIAQIVWTAHCLPSTTVLEEIL
 5 IRPTEGDL (SEQ ID NO: 10)

In additional embodiments, the ketoreductase enzymes described herein include ketoreductase enzymes having the amino acid sequence as set forth below in SEQ ID NO: 11.

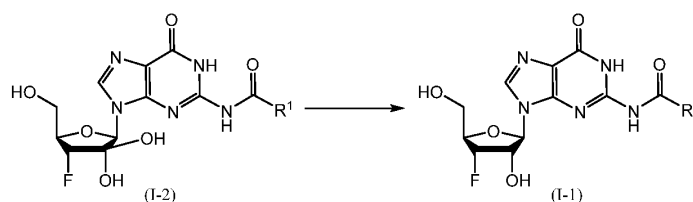
MHHHHHPATIVVTGGTKGIGRAIVEKFAKEGFTVLT CARTAGDNFPENVHFFKA
 DLSKKVEVLAFADFIKATVNQVDILVNNTGHFLPGEINNEEEGTLEAMIETNLYS
 10 AYYLTRALVGDMITKKEGHIFNICSYASIVPYTSGGSYCISKTAELAMSRVLREE
 LKPHHVVRTSILPGAVLNDNWAKAELPAELFIAPEDIAQIVWTAHCLPSTTVLEEIL
 LIRPTEGDL (SEQ ID NO: 11)

In some embodiments, an improved ketoreductase enzyme of the disclosure is based on the sequence formulas of SEQ ID NO: 2, 3, 4, 5, 6, 7, 8, 9, 10, or 11 and can comprise an amino acid sequence that is at least about 85%, 86%, 87%, 88%, 89%, 90%, 91%, 92%, 93%, 94%, 95%, 96%, 97%, 98%, or 99% identical to the reference sequence of SEQ ID NO: 2, 3, 4, 5, 6, 7, 8, 9, 10, or 11. These differences can be amino acid insertions, deletions, substitutions, or any combinations of such changes. In some embodiments, the amino acid sequence differences can comprise non-conservative, conservative, as well as a combination of non-conservative and conservative amino acid substitutions.

Additional embodiments provide host cells comprising the polynucleotides and/or expression vectors described herein. The host cells may be *E. coli*, or they may be a different organism, such as *L. brevis*. The host cells can be used for the expression and isolation of the ketoreductase enzymes described herein, or, alternatively, they can be used directly for the conversion of the substrate to the stereoisomeric product.

Whether carrying out the method with whole cells, cell extracts or purified ketoreductase enzymes, a single ketoreductase enzyme may be used or, alternatively, mixtures of two or more ketoreductase enzymes may be used.

Embodiments relate to ketoreductase enzymes capable of selectively preparing chiral alcohols in the synthesis of nucleosides, particularly fluoronucleosides. In embodiments, the ketoreductase enzymes are capable of the following conversion:



In particular embodiments, the ketoreductase enzymes are capable of the following conversion:



Enzyme properties for which improvement is desirable include, but are not

- 5 limited to, enzymatic activity, thermal stability, pH activity profile, cofactor requirements, refractoriness to inhibitors (*e.g.*, product inhibition), stereospecificity, stereoselectivity, and solvent stability. The improvements can relate to a single enzyme property, such as enzymatic activity, or a combination of different enzyme properties, such as enzymatic activity and stereoselectivity.

- 10 Table 1 below provide a list of the SEQ ID NOs disclosed herein with associated activities. The sequences below are based on the ketoreductase sequence of SEQ ID NO: 1, unless otherwise specified. In table below, each row lists a SEQ ID NO. The column listing the number of mutations (*i.e.*, residue changes) refers to the number of amino acid substitutions as compared to the ketoreductase sequence of SEQ ID NO: 1. In the Conversion column of the
- 15 Table, + indicates <10% conversion of substrate to product, ++ indicates 10-60% conversion, +++ indicates >60% conversion. In the %DE column of the Table, – indicates R selectivity, + indicates <50% S, S-diastereomeric product, ++ indicates >50% S, S-diastereomeric product.

Table 1: Activity of Ketoreductase Enzymes				
SEQ ID NO:	Residue Changes Relative to SEQ ID NO: 1	Number of Changes Relative to SEQ ID NO: 1	Conversion	% DE
1	-	0	+	++
2	T183N	1	+	++
3	T136Y N144S T183N G188K	4	++	++

Table 1: Activity of Ketoreductase Enzymes				
SEQ ID NO:	Residue Changes Relative to SEQ ID NO: 1	Number of Changes Relative to SEQ ID NO: 1	Conversion	% DE
4	H7- N8P K9A T136Y N144S F153T T183N G188K R195W	9	++	++
5	H7- N8P K9A F86W T136Y T140V A141P N144S F153T L155Q T181V T183N G188K R195W	14	+++	++
6	H7- N8P K9A F86W T136Y T140V A141P N144S F153T L155Q T181V T183N G188K R195L S199P L227T	16	+++	++
7	H7- N8P K9A K42A F86W A96E T136Y T140V A141P N144S F153T L155Q T181V T183N G188K R195L S199P L227E	18	+++	++
8	H7- N8P K9A K42A F86H A96E T136Y T140V A141P N144S F153T L155E T181V T183N G188K V189A R195L S199P L227E	19	+++	++
9	H7- N8P K9A K42A F86H A96E T136Y T140V A141P N144S F153T L155E G156A K159R T181V T183N S184N G188K V189A R195L S199P Q225T L227E	23	+++	++
10	H7- N8P K9A K42A Q76A F86H A96E T136Y T140V A141P N144S S148P F153T L155E G156A K159R T181V T183N S184N G188K V189A R195L S199P Q225T L227E	25	+++	++

Table 1: Activity of Ketoreductase Enzymes				
SEQ ID NO:	Residue Changes Relative to SEQ ID NO: 1	Number of Changes Relative to SEQ ID NO: 1	Conversion	% DE
11	H7- N8P K9A K42A Q76A F86H A96E T136Y T140V A141P N144S F153T L155E G156A K159M T181V T183N S184N G188K V189A R195L S199P Q225T L227E	24	+++	++

Polynucleotides Encoding Ketoreductases

In another aspect, the present disclosure provides polynucleotides encoding the ketoreductase enzymes disclosed herein. The polynucleotides may be operatively linked to one or more heterologous regulatory sequences that control gene expression to create a recombinant polynucleotide capable of expressing the polypeptide. Expression constructs containing a heterologous polynucleotide encoding the ketoreductase can be introduced into appropriate host cells to express the corresponding ketoreductase polypeptide.

Because of the knowledge of the codons corresponding to the various amino acids, availability of a protein sequence provides a description of all the polynucleotides capable of encoding the subject. The degeneracy of the genetic code, where the same amino acids are encoded by alternative or synonymous codons allows an extremely large number of nucleic acids to be made, all of which encode the improved ketoreductase enzymes disclosed herein. Thus, having identified a particular amino acid sequence, those skilled in the art could make any number of different nucleic acids by simply modifying the sequence of one or more codons in a way that does not change the amino acid sequence of the protein. In this regard, the present disclosure specifically contemplates each and every possible variation of polynucleotides that could be made by selecting combinations based on the possible codon choices, and all such variations are to be considered specifically disclosed for any polypeptide disclosed herein.

In various embodiments, the codons are preferably selected to fit the host cell in which the protein is being produced. For example, preferred codons used in bacteria are used to express the gene in bacteria; preferred codons used in yeast are used for expression in yeast; and preferred codons used in mammals are used for expression in mammalian cells. By way of

example, the polynucleotide of SEQ ID NO: 3 has been codon optimized for expression in *E. coli*.

In certain embodiments, all codons need not be replaced to optimize the codon usage of the ketoreductases since the natural sequence will comprise preferred codons and because use of preferred codons may not be required for all amino acid residues. Consequently, codon optimized polynucleotides encoding the ketoreductase enzymes may contain preferred codons at about 40%, 50%, 60%, 70%, 80%, or greater than 90% of codon positions of the full length coding region.

In various embodiments, an isolated polynucleotide encoding an improved ketoreductase polypeptide may be manipulated in a variety of ways to provide for expression of the polypeptide. Manipulation of the isolated polynucleotide prior to its insertion into a vector may be desirable or necessary depending on the expression vector. The techniques for modifying polynucleotides and nucleic acid sequences utilizing recombinant DNA methods are well known in the art. Guidance is provided in Sambrook *et al.*, 2001, Molecular Cloning: A Laboratory Manual, 3rd Ed., Cold Spring Harbor Laboratory Press; and Current Protocols in Molecular Biology, Ausubel, F. ed., Greene Pub. Associates, 1998, updates to 2006.

In some embodiments, an isolated polynucleotide encoding any of the ketoreductase polypeptides herein is manipulated in a variety of ways to facilitate expression of the ketoreductase polypeptide. In some embodiments, the polynucleotides encoding the ketoreductase polypeptides comprise expression vectors where one or more control sequences is present to regulate the expression of the ketoreductase polynucleotides and/or polypeptides. Manipulation of the isolated polynucleotide prior to its insertion into a vector may be desirable or necessary depending on the expression vector utilized. Techniques for modifying polynucleotides and nucleic acid sequences utilizing recombinant DNA methods are well known in the art. In some embodiments, the control sequences include among others, promoters, leader sequences, polyadenylation sequences, propeptide sequences, signal peptide sequences, and transcription terminators. In some embodiments, suitable promoters are selected based on the host cells selection. For bacterial host cells, suitable promoters for directing transcription of the nucleic acid constructs of the present disclosure, include, but are not limited to, promoters obtained from the *E. coli* lac operon, *Streptomyces coelicolor* agarase gene (dagA), *Bacillus subtilis* levansucrase gene (sacB), *Bacillus licheniformis* alpha-amylase gene (amyL), *Bacillus stearothermophilus* maltogenic amylase gene (amyM), *Bacillus amyloliquefaciens* alpha-amylase

gene (amyQ), *Bacillus licheniformis* penicillinase gene (penP), *Bacillus subtilis* xylA and xylB genes, and prokaryotic beta-lactamase gene (See e.g., Villa-Kamaroff *et al.*, Proc. Natl Acad. Sci. USA 75: 3727-3731 [1978]), as well as the *tac* promoter (See e.g., DeBoer *et al.*, Proc. Natl Acad. Sci. USA 80: 21-25 [1983]). Exemplary promoters for filamentous fungal host cells, include, but are not limited to, promoters obtained from the genes for *Aspergillus oryzae* TAKA amylase, *Rhizomucor miehei* aspartic proteinase, *Aspergillus niger* neutral alpha-amylase, *Aspergillus niger* acid stable alpha-amylase, *Aspergillus niger* or *Aspergillus awamori* glucoamylase (glaA), *Rhizomucor miehei* lipase, *Aspergillus oryzae* alkaline protease, *Aspergillus oryzae* triose phosphate isomerase, *Aspergillus nidulans* acetamidase, and *Fusarium oxysporum* trypsin-like protease (See e.g., WO 96/00787), as well as the NA2-tpi promoter (a hybrid of the promoters from the genes for *Aspergillus niger* neutral alpha-amylase and *Aspergillus oryzae* triose phosphate isomerase), and mutant, truncated, and hybrid promoters thereof. Exemplary yeast cell promoters can be from the genes for *Saccharomyces cerevisiae* enolase (ENO-1), *Saccharomyces cerevisiae* galactokinase (GAL1), *Saccharomyces cerevisiae* alcohol dehydrogenase/glyceraldehyde-3-phosphate dehydrogenase (ADH2/GAP), and *Saccharomyces cerevisiae* 3-phosphoglycerate kinase. Other useful promoters for yeast host cells are known in the art (See e.g., Romanos *et al.*, Yeast 8:423-488 [1992]).

In some embodiments, the control sequence is also a suitable transcription terminator sequence (i.e., a sequence recognized by a host cell to terminate transcription). In some embodiments, the terminator sequence is operably linked to the 3' terminus of the nucleic acid sequence encoding the enzyme polypeptide. Any suitable terminator that is functional in the host cell of choice finds use in the present invention. Exemplary transcription terminators for filamentous fungal host cells can be obtained from the genes for *Aspergillus oryzae* TAKA amylase, *Aspergillus niger* glucoamylase, *Aspergillus nidulans* anthranilate synthase, *Aspergillus niger* alpha-glucosidase, and *Fusarium oxysporum* trypsin-like protease. Exemplary terminators for yeast host cells can be obtained from the genes for *Saccharomyces cerevisiae* enolase, *Saccharomyces cerevisiae* cytochrome C (CYC1), and *Saccharomyces cerevisiae* glyceraldehyde-3-phosphate dehydrogenase. Other useful terminators for yeast host cells are known in the art (See e.g., Romanos *et al.*, supra).

In some embodiments, the control sequence is also a suitable leader sequence (i.e., a non-translated region of an mRNA that is important for translation by the host cell). In

some embodiments, the leader sequence is operably linked to the 5' terminus of the nucleic acid sequence encoding the ketoreductase. Any suitable leader sequence that is functional in the host cell of choice find use in the present invention. Exemplary leaders for filamentous fungal host cells are obtained from the genes for *Aspergillus oryzae* TAKA amylase, and *Aspergillus*
 5 *nidulans* triose phosphate isomerase. Suitable leaders for yeast host cells are obtained from the genes for *Saccharomyces cerevisiae* enolase (ENO-1), *Saccharomyces cerevisiae* 3-phosphoglycerate kinase, *Saccharomyces cerevisiae* alpha-factor, and *Saccharomyces cerevisiae* alcohol dehydrogenase/glyceraldehyde-3-phosphate dehydrogenase (ADH2/GAP).

In some embodiments, the control sequence is also a polyadenylation sequence
 10 (*i.e.*, a sequence operably linked to the 3' terminus of the nucleic acid sequence and which, when transcribed, is recognized by the host cell as a signal to add polyadenosine residues to transcribed mRNA). Any suitable polyadenylation sequence that is functional in the host cell of choice may be used in the present invention. Exemplary polyadenylation sequences for filamentous fungal host cells include, but are not limited to, the genes for *Aspergillus oryzae* TAKA amylase,
 15 *Aspergillus niger* glucoamylase, *Aspergillus nidulans* anthranilate synthase, *Fusarium oxysporum* trypsin-like protease, and *Aspergillus niger* alpha-glucosidase. Useful polyadenylation sequences for yeast host cells are known (*See e.g.*, Guo and Sherman, Mol. Cell. Biol., 15:5983-5990 [1995]).

In some embodiments, the control sequence is also a signal peptide (*i.e.*, a coding
 20 region that codes for an amino acid sequence linked to the amino terminus of a polypeptide and directs the encoded polypeptide into the cell's secretory pathway). In some embodiments, the 5' end of the coding sequence of the nucleic acid sequence inherently contains a signal peptide coding region naturally linked in translation reading frame with the segment of the coding region that encodes the secreted polypeptide. Alternatively, in some embodiments, the 5' end of the
 25 coding sequence contains a signal peptide coding region that is foreign to the coding sequence. Any suitable signal peptide coding region that directs the expressed polypeptide into the secretory pathway of a host cell of choice finds use for expression of the engineered polypeptide(s). Effective signal peptide coding regions for bacterial host cells are the signal peptide coding regions include, but are not limited to, those obtained from the genes for *Bacillus*
 30 NCIB 11837 maltogenic amylase, *Bacillus stearothermophilus* alpha-amylase, *Bacillus licheniformis* subtilisin, *Bacillus licheniformis* beta-lactamase, *Bacillus stearothermophilus* neutral proteases (nprT, nprS, nprM), and *Bacillus subtilis* prsA. Further signal peptides are

known in the art (*See e.g.*, Simonen and Palva, Microbiol. Rev., 57:109-137 [1993]). In some embodiments, effective signal peptide coding regions for filamentous fungal host cells include, but are not limited to, the signal peptide coding regions obtained from the genes for *Aspergillus oryzae* TAKA amylase, *Aspergillus niger* neutral amylase, *Aspergillus niger* glucoamylase, 5 *Rhizomucor miehei* aspartic proteinase, *Humicola insolens* cellulase, and *Humicola lanuginosa* lipase. Useful signal peptides for yeast host cells include, but are not limited to, those from the genes for *Saccharomyces cerevisiae* alpha-factor and *Saccharomyces cerevisiae* invertase.

In some embodiments, regulatory sequences are also utilized. These sequences facilitate the regulation of the expression of the polypeptide relative to the growth of the host 10 cell. Examples of regulatory systems are those that cause the expression of the gene to be turned on or off in response to a chemical or physical stimulus, including the presence of a regulatory compound. In prokaryotic host cells, suitable regulatory sequences include, but are not limited to, the *lac*, *tac*, and *trp* operator systems. In yeast host cells, suitable regulatory systems include, but are not limited to, the ADH2 system or GAL1 system. In filamentous fungi, suitable 15 regulatory sequences include, but are not limited to, the TAKA alpha-amylase promoter, *Aspergillus niger* glucoamylase promoter, and *Aspergillus oryzae* glucoamylase promoter.

In another aspect, the present invention is directed to a recombinant expression vector comprising a polynucleotide encoding ketoreductase polypeptide, and one or more expression regulating regions such as a promoter and a terminator, a replication origin, etc., 20 depending on the type of hosts into which they are to be introduced. In some embodiments, the various nucleic acid and control sequences described herein are joined together to produce recombinant expression vectors that include one or more convenient restriction sites to allow for insertion or substitution of the nucleic acid sequence encoding the enzyme polypeptide at such sites. Alternatively, in some embodiments, the nucleic acid sequence of the present invention is 25 expressed by inserting the nucleic acid sequence or a nucleic acid construct comprising the sequence into an appropriate vector for expression. In some embodiments involving the creation of the expression vector, the coding sequence is located in the vector so that the coding sequence is operably linked with the appropriate control sequences for expression.

The recombinant expression vector may be any suitable vector (*e.g.*, a plasmid or 30 virus), that can be conveniently subjected to recombinant DNA procedures and bring about the expression of the enzyme polynucleotide sequence. The choice of the vector typically depends

on the compatibility of the vector with the host cell into which the vector is to be introduced. The vectors may be linear or closed circular plasmids.

In some embodiments, the expression vector is an autonomously replicating vector (*i.e.*, a vector that exists as an extra-chromosomal entity, the replication of which is independent of chromosomal replication, such as a plasmid, an extra-chromosomal element, a minichromosome, or an artificial chromosome). The vector may contain any means for assuring self-replication. In some alternative embodiments, the vector is one in which, when introduced into the host cell, it is integrated into the genome and replicated together with the chromosome(s) into which it has been integrated. Furthermore, in some embodiments, a single vector or plasmid, or two or more vectors or plasmids which together contain the total DNA to be introduced into the genome of the host cell, and/or a transposon is utilized.

In some embodiments, the expression vector contains one or more selectable markers, which permit easy selection of transformed cells. A “selectable marker” is a gene, the product of which provides for biocide or viral resistance, resistance to heavy metals, prototrophy to auxotrophs, and the like. Examples of bacterial selectable markers include, but are not limited to, the *dal* genes from *Bacillus subtilis* or *Bacillus licheniformis*, or markers, which confer antibiotic resistance such as ampicillin, kanamycin, chloramphenicol or tetracycline resistance. Suitable markers for yeast host cells include, but are not limited to, ADE2, HIS3, LEU2, LYS2, MET3, TRP1, and URA3. Selectable markers for use in filamentous fungal host cells include, but are not limited to, *amdS* (acetamidase; *e.g.*, from *A. nidulans* or *A. oryzae*), *argB* (ornithine carbamoyltransferase), *bar* (phosphinothricin acetyltransferase; *e.g.*, from *S. hygrosopicus*), *hph* (hygromycin phosphotransferase), *niaD* (nitrate reductase), *pyrG* (orotidine-5'-phosphate decarboxylase; *e.g.*, from *A. nidulans* or *A. oryzae*), *sC* (sulfate adenylyltransferase), and *trpC* (anthranilate synthase), as well as equivalents thereof.

In another aspect, the present invention provides a host cell comprising at least one polynucleotide encoding at least one ketoreductase of the present invention, the polynucleotide(s) being operatively linked to one or more control sequences for expression of the at least one ketoreductase in the host cell. Host cells suitable for use in expressing the polypeptides encoded by the expression vectors of the present invention are well known in the art and include but are not limited to, bacterial cells, such as *E. coli*, *Vibrio fluvialis*, *Streptomyces* and *Salmonella typhimurium* cells; fungal cells, such as yeast cells (*e.g.*, *Saccharomyces cerevisiae* or *Pichia pastoris* (ATCC Accession No. 201178));. Exemplary host

cells also include various *Escherichia coli* strains (e.g., W3110 (ΔfhuA) and BL21). Examples of bacterial selectable markers include, but are not limited to, the *dal* genes from *Bacillus subtilis* or *Bacillus licheniformis*, or markers, which confer antibiotic resistance such as ampicillin, kanamycin, chloramphenicol, and or tetracycline resistance.

5 In some embodiments, the expression vectors of the present invention contain an element(s) that permits integration of the vector into the host cell's genome or autonomous replication of the vector in the cell independent of the genome. In some embodiments involving integration into the host cell genome, the vectors rely on the nucleic acid sequence encoding the polypeptide or any other element of the vector for integration of the vector into the genome by
10 homologous or nonhomologous recombination.

 In some alternative embodiments, the expression vectors contain additional nucleic acid sequences for directing integration by homologous recombination into the genome of the host cell. The additional nucleic acid sequences enable the vector to be integrated into the host cell genome at a precise location(s) in the chromosome(s). To increase the likelihood of
15 integration at a precise location, the integrational elements preferably contain a sufficient number of nucleotides, such as 100 to 10,000 base pairs, preferably 400 to 10,000 base pairs, and most preferably 800 to 10,000 base pairs, which are highly homologous with the corresponding target sequence to enhance the probability of homologous recombination. The integrational elements may be any sequence that is homologous with the target sequence in the genome of the host cell.
20 Furthermore, the integrational elements may be non-encoding or encoding nucleic acid sequences. On the other hand, the vector may be integrated into the genome of the host cell by non-homologous recombination.

 For autonomous replication, the vector may further comprise an origin of replication enabling the vector to replicate autonomously in the host cell in question. Examples
25 of bacterial origins of replication are P15A ori or the origins of replication of plasmids pBR322, pUC19, pACYC177 (which contains the P15A ori), or pACYC184 (which contains the P15A ori) permitting replication in *E. coli*, and pUB110, pE194, or pTA1060 permitting replication in *Bacillus*. Examples of origins of replication for use in a yeast host cell are the 2 micron origin of replication, ARS1, ARS4, the combination of ARS1 and CEN3, and the combination of ARS4
30 and CEN6. The origin of replication may be one having a mutation which makes its functioning temperature-sensitive in the host cell (See e.g., Ehrlich, Proc. Natl. Acad. Sci. USA 75:1433 [1978]).

In some embodiments, more than one copy of a nucleic acid sequence of the present invention is inserted into the host cell to increase production of the gene product. An increase in the copy number of the nucleic acid sequence can be obtained by integrating at least one additional copy of the sequence into the host cell genome or by including an amplifiable
5 selectable marker gene with the nucleic acid sequence where cells containing amplified copies of the selectable marker gene, and thereby additional copies of the nucleic acid sequence, can be selected for by cultivating the cells in the presence of the appropriate selectable agent.

Many of the expression vectors for use in the present invention are commercially available. Suitable commercial expression vectors include, but are not limited to, Novagen's®
10 pET *E. coli* T7 expression vectors (Millipore Sigma) and the p3xFLAG™ expression vectors (Sigma-Aldrich Chemicals). Other suitable expression vectors include, but are not limited to, pBluescriptII SK(-) and pBK-CMV (Stratagene), and plasmids derived from pBR322 (Gibco BRL), pUC (Gibco BRL), pREP4, pCEP4 (Invitrogen) or pPoly (*See e.g.*, Lathe *et al.*, Gene 57:193-201 [1987]).

15 Thus, in some embodiments, a vector comprising a sequence encoding at least one variant ketoreductase is transformed into a host cell in order to allow propagation of the vector and expression of the variant ketoreductase(s). In some embodiments, the transformed host cell described above is cultured in a suitable nutrient medium under conditions permitting the expression of the variant ketoreductase(s). Any suitable medium useful for culturing the host
20 cells finds use in the present invention, including, but not limited to minimal or complex media containing appropriate supplements. In some embodiments, host cells are grown in HTP media. Suitable media are available from various commercial suppliers or may be prepared according to published recipes (*e.g.*, in catalogues of the American Type Culture Collection).

25 ***Host Cells for Expression of Ketoreductases***

In another aspect, the present disclosure provides a host cell comprising a polynucleotide encoding an improved ketoreductase polypeptide of the present disclosure, the polynucleotide being operatively linked to one or more control sequences for expression of the ketoreductase enzyme in the host cell. Host cells for use in expressing the ketoreductase
30 polypeptides encoded by the expression vectors of the present invention are well known in the art and include but are not limited to, bacterial cells, such as *E. coli*, *B. subtilis*, *B. licheniformis*, *B. megaterium*, *B. stearothermophilus*, *B. amyloliquefaciens*, *Lactobacillus kefir*, *Lactobacillus*

brevis, *Lactobacillus minor*, *Streptomyces* and *Salmonella typhimurium* cells; fungal cells, such as yeast cells (e.g., *Saccharomyces cerevisiae* or *Pichia pastoris* (ATCC Accession No. 201178)). Appropriate culture mediums and growth conditions for the above-described host cells are well known in the art.

5 Polynucleotides for expression of the ketoreductases may be introduced into cells by various methods known in the art. Techniques include among others, electroporation, biolistic particle bombardment, liposome mediated transfection, calcium chloride transfection, and protoplast fusion. Various methods for introducing polynucleotides into cells will be apparent to the skilled artisan.

10 In some embodiments of the present invention, the filamentous fungal host cells are of any suitable genus and species, including, but not limited to *Achlya*, *Acremonium*, *Aspergillus*, *Aureobasidium*, *Bjerkandera*, *Ceriporiopsis*, *Cephalosporium*, *Chrysosporium*, *Cochliobolus*, *Corynascus*, *Cryphonectria*, *Cryptococcus*, *Coprinus*, *Coriolus*, *Diplodia*, *Endothia*, *Fusarium*, *Gibberella*, *Gliocladium*, *Humicola*, *Hypocrea*, *Myceliophthora*, *Mucor*,
15 *Neurospora*, *Penicillium*, *Podospira*, *Phlebia*, *Piromyces*, *Pyricularia*, *Rhizomucor*, *Rhizopus*, *Schizophyllum*, *Scytalidium*, *Sporotrichum*, *Talaromyces*, *Thermoascus*, *Thielavia*, *Trametes*, *Tolypocladium*, *Trichoderma*, *Verticillium*, and/or *Volvariella*, and/or teleomorphs, or anamorphs, and synonyms, basionyms, or taxonomic equivalents thereof.

In some embodiments of the present invention, the host cell is a yeast cell,
20 including but not limited to cells of *Candida*, *Hansenula*, *Saccharomyces*, *Schizosaccharomyces*, *Pichia*, *Kluyveromyces*, or *Yarrowia* species. In some embodiments of the present invention, the yeast cell is *Hansenula polymorpha*, *Saccharomyces cerevisiae*, *Saccharomyces carlsbergensis*, *Saccharomyces diastaticus*, *Saccharomyces norbensis*, *Saccharomyces kluyveri*, *Schizosaccharomyces pombe*, *Pichia pastoris*, *Pichia finlandica*, *Pichia trehalophila*, *Pichia*
25 *kodamae*, *Pichia membranaefaciens*, *Pichia opuntiae*, *Pichia thermotolerans*, *Pichia salictaria*, *Pichia quercuum*, *Pichia pijperi*, *Pichia stipitis*, *Pichia methanolica*, *Pichia angusta*, *Kluyveromyces lactis*, *Candida albicans*, or *Yarrowia lipolytica*.

In some other embodiments, the host cell is a prokaryotic cell. Suitable prokaryotic cells include, but are not limited to, Gram-positive, Gram-negative and Gram-
30 variable bacterial cells. Any suitable bacterial organism finds use in the present invention, including but not limited to *Agrobacterium*, *Alicyclobacillus*, *Anabaena*, *Anacystis*, *Acinetobacter*, *Acidothermus*, *Arthrobacter*, *Azobacter*, *Bacillus*, *Bifidobacterium*,

Brevibacterium, Butyrivibrio, Buchnera, Campestris, Campylobacter, Clostridium, Corynebacterium, Chromatium, Coprococcus, Escherichia, Enterococcus, Enterobacter, Erwinia, Fusobacterium, Faecalibacterium, Francisella, Flavobacterium, Geobacillus, Haemophilus, Helicobacter, Klebsiella, Lactobacillus, Lactococcus, Ilyobacter, Micrococcus,
5 *Microbacterium, Mesorhizobium, Methylobacterium, Methylobacterium, Mycobacterium, Neisseria, Pantoea, Pseudomonas, Prochlorococcus, Rhodobacter, Rhodopseudomonas, Rhodopseudomonas, Roseburia, Rhodospirillum, Rhodococcus, Scenedesmus, Streptomyces, Streptococcus, Synechococcus, Saccharomonospora, Staphylococcus, Serratia, Salmonella, Shigella, Thermoanaerobacterium, Tropheryma, Tularensis, Temecula, Thermosynechococcus,*
10 *Thermococcus, Ureaplasma, Xanthomonas, Xylella, Yersinia and Zymomonas.* In some embodiments, the host cell is a species of *Agrobacterium, Acinetobacter, Azobacter, Bacillus, Bifidobacterium, Buchnera, Geobacillus, Campylobacter, Clostridium, Corynebacterium, Escherichia, Enterococcus, Erwinia, Flavobacterium, Lactobacillus, Lactococcus, Pantoea, Pseudomonas, Staphylococcus, Salmonella, Streptococcus, Streptomyces, or Zymomonas.* In
15 some embodiments, the bacterial host strain is non-pathogenic to humans. In some embodiments the bacterial host strain is an industrial strain. Numerous bacterial industrial strains are known and suitable in the present invention. In some embodiments of the present invention, the bacterial host cell is an *Agrobacterium* species (e.g., *A. radiobacter, A. rhizogenes, and A. rubi*). In some embodiments of the present invention, the bacterial host cell is an *Arthrobacter* species
20 (e.g., *A. aureus, A. citreus, A. globiformis, A. hydrocarboglutamicus, A. mysores, A. nicotianae, A. paraffineus, A. protophominae, A. roseoparaffinus, A. sulfureus, and A. ureafaciens*). In some embodiments of the present invention, the bacterial host cell is a *Bacillus* species (e.g., *B. thuringensis, B. anthracis, B. megaterium, B. subtilis, B. lentus, B. circulans, B. pumilus, B. lautus, B. coagulans, B. brevis, B. firmus, B. alkaophilus, B. licheniformis, B. clausii,*
25 *B. stearothermophilus, B. halodurans, and B. amyloliquefaciens*). In some embodiments, the host cell is an industrial *Bacillus* strain including but not limited to *B. subtilis, B. pumilus, B. licheniformis, B. megaterium, B. clausii, B. stearothermophilus, or B. amyloliquefaciens.* In some embodiments, the *Bacillus* host cells are *B. subtilis, B. licheniformis, B. megaterium, B. stearothermophilus, and/or B. amyloliquefaciens.* In some embodiments, the bacterial host cell is
30 a *Clostridium* species (e.g., *C. acetobutylicum, C. tetani E88, C. lituseburens, C. saccharobutylicum, C. perfringens, and C. beijerinckii*). In some embodiments, the bacterial host cell is a *Corynebacterium* species (e.g., *C. glutamicum and C. acetoacidophilum*). In some

embodiments the bacterial host cell is an *Escherichia* species (e.g., *E. coli*). In some embodiments, the host cell is *Escherichia coli* W3110. In some embodiments the host is *Escherichia coli* BL21 or BL21(DE3). In some embodiments, the bacterial host cell is an *Erwinia* species (e.g., *E. uredovora*, *E. carotovora*, *E. ananas*, *E. herbicola*, *E. punctata*, and *E. terreus*). In some embodiments, the bacterial host cell is a *Pantoea* species (e.g., *P. citrea*, and *P. agglomerans*). In some embodiments the bacterial host cell is a *Pseudomonas* species (e.g., *P. putida*, *P. aeruginosa*, *P. mevalonii*, and *P. sp.* D-01 10). In some embodiments, the bacterial host cell is a *Streptococcus* species (e.g., *S. equisimiles*, *S. pyogenes*, and *S. uberis*). In some embodiments, the bacterial host cell is a *Streptomyces* species (e.g., *S. ambofaciens*, *S. achromogenes*, *S. avermitilis*, *S. coelicolor*, *S. aureofaciens*, *S. aureus*, *S. fungicidicus*, *S. griseus*, and *S. lividans*). In some embodiments, the bacterial host cell is a *Zymomonas* species (e.g., *Z. mobilis*, and *Z. lipolytica*).

Many prokaryotic and eukaryotic strains that find use in the present invention are readily available to the public from a number of culture collections such as American Type Culture Collection (ATCC), Deutsche Sammlung von Mikroorganismen und Zellkulturen GmbH (DSM), Centraalbureau Voor Schimmelcultures (CBS), and Agricultural Research Service Patent Culture Collection, Northern Regional Research Center (NRRL).

In some embodiments, host cells are genetically modified to have characteristics that improve protein secretion, protein stability and/or other properties desirable for expression and/or secretion of a protein. Genetic modification can be achieved by genetic engineering techniques and/or classical microbiological techniques (e.g., chemical or UV mutagenesis and subsequent selection). Indeed, in some embodiments, combinations of recombinant modification and classical selection techniques are used to produce the host cells. Using recombinant technology, nucleic acid molecules can be introduced, deleted, inhibited or modified, in a manner that results in increased yields of ketoreductase variant(s) within the host cell and/or in the culture medium. In one genetic engineering approach, homologous recombination is used to induce targeted gene modifications by specifically targeting a gene *in vivo* to suppress expression of the encoded protein. In alternative approaches, siRNA, antisense and/or ribozyme technology find use in inhibiting gene expression. A variety of methods are known in the art for reducing expression of protein in cells, including, but not limited to deletion of all or part of the gene encoding the protein and site-specific mutagenesis to disrupt expression or activity of the gene product. (See e.g., Chaverroche *et al.*, Nucl. Acids Res., 28:22 e97 [2000]; Cho *et al.*, Molec.

Plant Microbe Interact., 19:7-15 [2006]; Maruyama and Kitamoto, Biotechnol. Lett., 30:1811-1817 [2008]; Takahashi *et al.*, Mol. Gen. Genom., 272: 344-352 [2004]; and You *et al.*, Arch. Microbiol., 191:615-622 [2009], all of which are incorporated by reference herein). Random mutagenesis, followed by screening for desired mutations also finds use (*See e.g.*, Combier *et al.*, 5 *FEMS Microbiol. Lett.*, **220**:141-8 [2003]; and Firon *et al.*, *Eukary. Cell* **2**:247-55 [2003], both of which are incorporated by reference).

Introduction of a vector or DNA construct into a host cell can be accomplished using any suitable method known in the art, including but not limited to calcium phosphate transfection, DEAE-dextran mediated transfection, PEG-mediated transformation, 10 electroporation, or other common techniques known in the art.

In some embodiments, the engineered host cells (*i.e.*, “recombinant host cells”) of the present invention are cultured in conventional nutrient media modified as appropriate for activating promoters, selecting transformants, or amplifying the ketoreductase polynucleotide. Culture conditions, such as temperature, pH and the like, are those previously used with the host 15 cell selected for expression, and are well-known to those skilled in the art. As noted, many standard references and texts are available for the culture and production of many cells, including cells of bacterial, plant, animal (especially mammalian) and archebacterial origin.

In some embodiments, cells expressing the ketoreductase of the invention are grown under batch or continuous fermentations conditions. Classical “batch fermentation” is a 20 closed system, wherein the compositions of the medium are set at the beginning of the fermentation and is not subject to artificial alternations during the fermentation. A variation of the batch system is a “fed-batch fermentation” that also finds use in the present invention. In this variation, the substrate is added in increments as the fermentation progresses. Fed-batch systems are useful when catabolite repression is likely to inhibit the metabolism of the cells and where it 25 is desirable to have limited amounts of substrate in the medium. Batch and fed-batch fermentations are common and well known in the art. “Continuous fermentation” is an open system where a defined fermentation medium is added continuously to a bioreactor and an equal amount of conditioned medium is removed simultaneously for processing. Continuous fermentation generally maintains the cultures at a constant high density where cells are primarily 30 in log phase growth. Continuous fermentation systems strive to maintain steady state growth conditions. Methods for modulating nutrients and growth factors for continuous fermentation

processes as well as techniques for maximizing the rate of product formation are well known in the art of industrial microbiology.

More than one copy of a nucleic acid sequence of the present invention may be inserted into the host cell to increase production of the gene product. An increase in the copy
5 number of the nucleic acid sequence can be obtained by integrating at least one additional copy of the sequence into the host cell genome or by including an amplifiable selectable marker gene with the nucleic acid sequence where cells containing amplified copies of the selectable marker gene, and thereby additional copies of the nucleic acid sequence, can be selected for by cultivating the cells in the presence of the appropriate selectable agent.

10 In some embodiments of the present invention, cell-free transcription and translation systems find use in producing the ketoreductase(s). Several systems are commercially available, and the methods are well-known to those skilled in the art.

Methods of Evolving Ketoreductases

15 In some embodiments, to make the ketoreductases of the present disclosure, the ketoreductase enzyme that catalyzes the reduction reaction is obtained (or derived) from *E. coli*. In some embodiments, the parent polynucleotide sequence is codon optimized to enhance expression of the ketoreductase in a specified host cell. The parental polynucleotide sequence, designated as SEQ ID NO: 1, was codon optimized for expression in *E. coli* and the codon-
20 optimized polynucleotide cloned into an expression vector, placing the expression of the ketoreductase gene under the control of the T7 promoter. The T7 polymerase needed to express the gene of interest is under control of the lac promoter, and both the gene of interest and the T7 polymerase are subject to lacI repression. The presence of IPTG activates the T7 polymerase production and eliminates the repression, resulting in production of the ketoreductase gene.
25 Clones expressing the active ketoreductase in *E. coli* were identified and the genes sequenced to confirm their identity.

The ketoreductases of the disclosure may be obtained by subjecting the polynucleotide encoding the parent sequence to mutagenesis and/or directed evolution methods. An exemplary directed evolution technique is mutagenesis and/or DNA shuffling as described in
30 Stemmer, 1994, Proc. Natl. Acad. Sci. USA 91:10747-10751; WO 95/22625; WO 97/20078; WO 97/35966; WO 98/27230; WO 00/42651; WO 01/75767 and U.S. Pat. No. 6,537,746. Other directed evolution procedures that can be used include, among others, staggered extension

process (StEP), in vitro recombination (Zhao *et al.*, 1998, Nat. Biotechnol. 16:258-261), mutagenic PCR (Caldwell *et al.*, 1994, PCR Methods Appl. 3:S136-S140), and cassette mutagenesis (Black *et al.*, 1996, Proc. Natl. Acad. Sci. USA 93:3525-3529).

- The clones obtained following mutagenesis treatment are screened for
- 5 ketoreductases having a desired improved enzyme property. Measuring enzyme activity from the expression libraries can be performed using standard chemistry analytical techniques for measuring substrates and products such as UPLC-MS, as well as the standard biochemistry technique of monitoring the rate of decrease (via a decrease in absorbance or fluorescence) of NADH or NADPH concentration, as it is converted into NAD⁺ or NADP⁺. In this reaction, the
- 10 NADH or NADPH is consumed (oxidized) by the ketoreductase as the ketoreductase reduces a ketone substrate to the corresponding hydroxyl group. The rate of decrease of NADH or NADPH concentration, as measured by the decrease in absorbance or fluorescence, per unit time indicates the relative (enzymatic) activity of the ketoreductase polypeptide in a fixed amount of the lysate (or a lyophilized powder made therefrom). Where the improved enzyme property
- 15 desired is thermal stability, enzyme activity may be measured after subjecting the enzyme preparations to a defined temperature and measuring the amount of enzyme activity remaining after heat treatments. Clones containing a polynucleotide encoding a ketoreductase are then isolated, sequenced to identify the nucleotide sequence changes (if any), and used to express the enzyme in a host cell.
- 20 Where the sequence of the polypeptide is known, the polynucleotides encoding the enzyme can be prepared by standard solid-phase methods, according to known synthetic methods. In some embodiments, fragments of up to about 100 bases can be individually synthesized, then joined (*e.g.*, by enzymatic or chemical ligation methods, or polymerase mediated methods) to form any desired continuous sequence. For example, polynucleotides and
- 25 oligonucleotides of the invention can be prepared by chemical synthesis using, *e.g.*, the classical phosphoramidite method described by Beaucage *et al.*, 1981, Tet. Lett. 22:1859-69, or the method described by Matthes *et al.*, 1984, EMBO J. 3:801-05, *e.g.*, as it is typically practiced in automated synthetic methods. According to the phosphoramidite method, oligonucleotides are synthesized, *e.g.*, in an automatic DNA synthesizer, purified, annealed, ligated and cloned in
- 30 appropriate vectors. In addition, essentially any nucleic acid can be obtained from any of a variety of commercial sources, such as The Midland Certified Reagent Company, Midland, Tex.,

The Great American Gene Company, Ramona, Calif., ExpressGen Inc. Chicago, Ill., Operon Technologies Inc., Alameda, Calif., and many others.

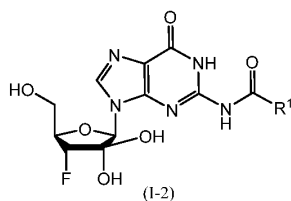
5 Ketoreductase enzymes expressed in a host cell can be recovered from the cells and or the culture medium using any one or more of the well-known techniques for protein purification, including, among others, lysozyme treatment, sonication, filtration, salting-out, ultra-centrifugation, and chromatography. Suitable solutions for lysing and the high efficiency extraction of proteins from bacteria, such as *E. coli*, are commercially available under the trade name CellLytic B® from Sigma-Aldrich of St. Louis Mo.

10 Chromatographic techniques for isolation of the ketoreductase polypeptide include, among others, reverse phase chromatography high performance liquid chromatography, ion exchange chromatography, gel electrophoresis, and affinity chromatography. Conditions for purifying a particular enzyme will depend, in part, on factors such as net charge, hydrophobicity, hydrophilicity, molecular weight, molecular shape, etc., and will be apparent to those having skill in the art.

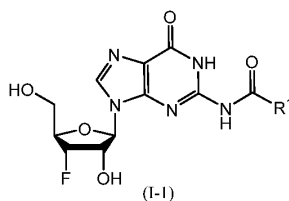
15 In some embodiments, affinity techniques may be used to isolate the improved ketoreductase enzymes. For affinity chromatography purification, the protein sequence can be tagged with a recognition sequence to enable purification. Common tags include cellulose-binding domains, poly His-tags, di-His chelates, FLAG-tags and many others that will be apparent to those having skill in the art. Antibodies can also be used as affinity purification reagents. Any antibody that specifically binds the ketoreductase polypeptide may be used.

Methods of Using the Ketoreductases

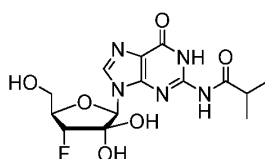
The ketoreductase enzymes described herein can catalyze the reduction of substrate compounds of Formula (I-2):



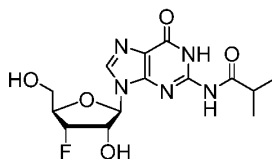
25 to the corresponding isomeric product of Formula (I-3):



In particular embodiments, the ketoreductase enzymes described herein can catalyze the reduction of the substrate compound, *N*-(7-((2*S*,4*R*,5*R*)-4-fluoro-3,3-dihydroxy-5-(hydroxymethyl)tetrahydrofuran-2-yl)-4-oxo-4,7-dihydro-3*H*-pyrrolo[3,2-*d*]pyrimidin-2-yl)isobutyramide (Compound A):



to the corresponding isomeric product, *N*-(9-((2*R*,3*S*,4*S*,5*R*)-4-fluoro-3-hydroxy-5-(hydroxymethyl)tetrahydrofuran-2-yl)-6-oxo-6,9-dihydro-1*H*-purin-2-yl)isobutyramide (Compound B):



10

In some embodiments, the method for reducing Compound A to Compound B comprises contacting or incubating the substrate with a ketoreductase disclosed herein under reaction conditions suitable for reducing or converting Compound A to Compound B. Compound B is an intermediate for the synthesis of (trisodium O-[(2*R*,3*S*,4*S*,5*R*)-5-(2-amino-6-oxido-9*H*-purin-9-yl)-3-fluoro-4-hydroxyoxolan-2-yl]methyl} phosphorothioate hydrate (1:6)), also known as 3'-fluoro-thio-guanosine monophosphate or 3'-F-thio-GMP. Thus, in a method for synthesizing 3'-F-thio-GMP, the method can comprise a step in which Compound A to Compound B using a ketoreductase disclosed herein. In some embodiments, the product in greater than about 99%, 99.1%, 99.2%, 99.3%, 99.4%, 99.5%, 99.6%, 99.7%, 99.8%, or 99.9% stereometric excess over the corresponding (R) alcohol product.

20

As noted herein, in some embodiments, the ketoreductases can comprise an amino acid sequence that is at least about 85%, 86%, 87%, 88%, 89%, 90%, 91 %, 92%, 93%, 94%, 95%, 96%, 97%, 98%, or 99% identical as compared a reference sequence comprising the

sequence of SEQ ID NO: 2, 3, 4, 5, 6, 7, 8, 9, 10, or 11. In some embodiments, these ketoreductase polypeptides can have one or more modifications to the amino acid sequence of SEQ ID NO: 2, 3, 4, 5, 6, 7, 8, 9, 10, or 11. The modifications can include substitutions, deletions, and insertions. The substitutions can be non-conservative substitutions, conservative substitutions, or a combination of nonconservative and conservative substitutions.

5 In some embodiments of the method for reducing the substrate to the product, the substrate is reduced to the product in greater than about 99% stereometric excess, wherein the ketoreductase polypeptide comprises a sequence that corresponds to SEQ ID NO: 2, 3, 4, 5, 6, 7, 8, 9, 10, or 11.

10 In another embodiment of this method for reducing the substrate to the product, at least about 95% of the substrate is converted to the product in less than about 24 hours when carried out with greater than about 100 g/L of substrate and less than about 5 g/L of the polypeptide, wherein the polypeptide comprises an amino acid sequence corresponding to SEQ ID NO: 2, 3, 4, 5, 6, 7, 8, 9, 10, or 11.

15 As is known by those of skill in the art, ketoreductase catalyzed reduction reactions typically require a cofactor. Reduction reactions catalyzed by the ketoreductase enzymes described herein also typically require a cofactor, although many embodiments of the engineered ketoreductases require far less cofactor than reactions catalyzed with wild-type ketoreductase enzymes. As used herein, the term “cofactor” refers to a non-protein compound
20 that operates in combination with a ketoreductase enzyme. Cofactors suitable for use with the engineered ketoreductase enzymes described herein include, but are not limited to, NADP⁺ (nicotinamide adenine dinucleotide phosphate), NADPH (the reduced form of NADP⁺), NAD⁺ (nicotinamide adenine dinucleotide) and NADH (the reduced form of NAD⁺). Generally, the reduced form of the cofactor is added to the reaction mixture. The reduced NAD(P)H form can
25 be optionally regenerated from the oxidized NAD(P)⁺ form using a cofactor regeneration system.

The term “cofactor regeneration system” refers to a set of reactants that participate in a reaction that reduces the oxidized form of the cofactor (*e.g.*, NADP⁺ to NADPH). Cofactors oxidized by the ketoreductase-catalyzed reduction of the keto substrate are
30 regenerated in reduced form by the cofactor regeneration system. Cofactor regeneration systems comprise a stoichiometric reductant that is a source of reducing hydrogen equivalents and that is capable of reducing the oxidized form of the cofactor. The cofactor regeneration system may

further comprise a catalyst, for example an enzyme catalyst that catalyzes the reduction of the oxidized form of the cofactor by the reductant. Cofactor regeneration systems to regenerate NADH or NADPH from NAD⁺ or NADP⁺, respectively, are known in the art and may be used in the methods described herein.

5 Suitable exemplary cofactor regeneration systems that may be employed include, but are not limited to, glucose and glucose dehydrogenase, formate and formate dehydrogenase, glucose-6-phosphate and glucose-6-phosphate dehydrogenase, a secondary (*e.g.*, isopropanol) alcohol and secondary alcohol dehydrogenase, phosphite and phosphite dehydrogenase, molecular hydrogen and hydrogenase, and the like. These systems may be used in combination
10 with either NADP⁺/NADPH or NAD⁺/NADH as the cofactor. Electrochemical regeneration using hydrogenase may also be used as a cofactor regeneration system. *See e.g.*, U.S. Pat. Nos. 5,538,867 and 6,495,023, both of which are incorporated herein by reference. Chemical cofactor regeneration systems comprising a metal catalyst and a reducing agent (for example, molecular hydrogen or formate) are also suitable. *See e.g.*, PCT publication WO 2000/053731, which is
15 incorporated herein by reference.

The terms “glucose dehydrogenase” and “GDH” are used interchangeably herein to refer to an NAD⁺ or NADP⁺-dependent enzyme that catalyzes the conversion of D-glucose and NAD⁺ or NADP⁺ to gluconic acid and NADH or NADPH, respectively.

Glucose dehydrogenases that are suitable for use in the practice of the methods
20 described herein include both naturally occurring glucose dehydrogenases, as well as nonnaturally occurring glucose dehydrogenases. Naturally occurring glucose dehydrogenase encoding genes have been reported in the literature. For example, the *Bacillus subtilis* 61297 GDH gene was expressed in *E. coli* and was reported to exhibit the same physicochemical properties as the enzyme produced in its native host (Vasantha *et al.*, 1983, Proc. Natl. Acad. Sci. USA 80:785). The gene sequence of the *B. subtilis* GDH gene, which corresponds to Genbank
25 Acc. No. M12276, was reported by Lampel *et al.*, 1986, J. Bacterial. 166:238-243, and in corrected form by Yamane *et al.*, 1996, Microbiology 142:3047-3056 as Genbank Acc. No. D50453. Naturally occurring GDH genes also include those that encode the GDH from *B. cereus* ATCC 14579 (Nature, 2003, 423:87-91; Genbank Acc. No. AE017013) and *B.*
30 *megaterium* (Eur. J. Biochem., 1988, 174:485-490, Genbank Acc. No. X12370; J. Ferment. Bioeng., 1990, 70:363-369, Genbank Acc. No. GI216270). Glucose dehydrogenases from

Bacillus sp. are provided in PCT publication WO 2005/018579, the disclosure of which is incorporated herein by reference.

Non-naturally occurring glucose dehydrogenases may be generated using known methods, such as, for example, mutagenesis, directed evolution, and the like. GDH enzymes
5 having suitable activity, whether naturally occurring or non-naturally occurring, may be readily identified by one of ordinary skill in the art, including using the assay described in Example 4 of PCT publication WO 2005/018579, the disclosure of which is incorporated herein by reference.

The ketoreductase-catalyzed reduction reactions described herein are generally carried out in a solvent. Suitable solvents include water, organic solvents, (*e.g.*, ethyl acetate,
10 butyl acetate, 1-octanol, heptane, octane, methyl t-butyl ether (MTBE), toluene, and the like), and ionic liquids (*e.g.*, 1-ethyl 4-methylimidazolium tetrafluoroborate, 1-butyl-3-methylimidazolium tetrafluoroborate, 1-butyl-3-methylimidazolium hexafluorophosphate, and the like). In some embodiments, aqueous solvents, including water and aqueous co-solvent systems, are used.

15 Exemplary aqueous co-solvent systems have water and one or more organic solvent. In general, an organic solvent component of an aqueous co-solvent system is selected such that it does not completely inactivate the ketoreductase enzyme. Appropriate co-solvent systems can be readily identified by measuring the enzymatic activity of the specified engineered ketoreductase enzyme with a defined substrate of interest in the candidate solvent system,
20 utilizing an enzyme activity assay, such as those described herein.

The organic solvent component of an aqueous co-solvent system may be miscible with the aqueous component, providing a single liquid phase, or may be partly miscible or immiscible with the aqueous component, providing two liquid phases. Generally, when an aqueous co-solvent system is employed, it is selected to be biphasic, with water dispersed in an
25 organic solvent, or vice-versa. Generally, when an aqueous co-solvent system is utilized, it is desirable to select an organic solvent that can be readily separated from the aqueous phase. In general, the ratio of water to organic solvent in the co-solvent system is typically in the range of from about 90:10 to about 10:90 (v/v) organic solvent to water, and between 80:20 and 20:80 (v/v) organic solvent to water. The co-solvent system may be pre-formed prior to addition to the
30 reaction mixture, or it may be formed in situ in the reaction vessel.

The aqueous solvent (water or aqueous co-solvent system) may be pH-buffered or unbuffered. Generally, the reduction can be carried out at a pH of about 10 or below, usually in

the range of from about 5 to about 10. In some embodiments, the reduction is carried out at a pH of about 9 or below, usually in the range of from about 5 to about 9. In some embodiments, the reduction is carried out at a pH of about 8 or below, often in the range of from about 5 to about 8, and usually in the range of from about 6 to about 8. The reduction may also be carried out at a pH of about 7.8 or below, or 7.5 or below. Alternatively, the reduction may be carried out at a neutral pH, *i.e.*, about 7.

During the course of the reduction reactions, the pH of the reaction mixture may change. The pH of the reaction mixture may be maintained at a desired pH or within a desired pH range by the addition of an acid or a base during the course of the reaction. Alternatively, the pH may be controlled by using an aqueous solvent that comprises a buffer. Suitable buffers to maintain desired pH ranges are known in the art and include, for example, phosphate buffer, triethanolamine buffer, and the like. Combinations of buffering and acid or base addition may also be used.

When the glucose/glucose dehydrogenase cofactor regeneration system is employed, the co-production of gluconic acid ($pK_a=3.6$), as represented in equation (1) causes the pH of the reaction mixture to drop if the resulting aqueous gluconic acid is not otherwise neutralized. The pH of the reaction mixture may be maintained at the desired level by standard buffering techniques, wherein the buffer neutralizes the gluconic acid up to the buffering capacity provided, or by the addition of a base concurrent with the course of the conversion. Combinations of buffering and base addition may also be used. Suitable buffers to maintain desired pH ranges are described above. Suitable bases for neutralization of gluconic acid are organic bases, for example amines, alkoxides and the like, and inorganic bases, for example, hydroxide salts (*e.g.*, NaOH), carbonate salts (*e.g.*, NaHCO_3), bicarbonate salts (*e.g.*, K_2CO_3), basic phosphate salts (*e.g.*, K_2HPO_4 , Na_3PO_4), and the like. The addition of a base concurrent with the course of the conversion may be done manually while monitoring the reaction mixture pH or, more conveniently, by using an automatic titrator as a pH stat. A combination of partial buffering capacity and base addition can also be used for process control.

When base addition is employed to neutralize gluconic acid released during a ketoreductase-catalyzed reduction reaction, the progress of the conversion may be monitored by the amount of base added to maintain the pH. Typically, bases added to unbuffered or partially buffered reaction mixtures over the course of the reduction are added in aqueous solutions.

In some embodiments, the co-factor regenerating system can comprises a formate dehydrogenase. The terms "formate dehydrogenase" and "FDH" are used interchangeably herein to refer to an NAD⁺ or NADP⁺-dependent enzyme that catalyzes the conversion of formate and NAD⁺ or NADP⁺ to carbon dioxide and NADH or NADPH, respectively. Formate dehydrogenases that are suitable for use as cofactor regenerating systems in the ketoreductase-catalyzed reduction reactions described herein include both naturally occurring formate dehydrogenases, as well as non-naturally occurring formate dehydrogenases. Formate dehydrogenases include those corresponding to SEQ ID NOS: 70 (*Pseudomonas sp.*) and 72 (*Candida boidinii*) of PCT publication WO 2005/018579, which are encoded by polynucleotide sequences corresponding to SEQ ID NOS: 69 and 71, respectively, of PCT publication WO 2005/018579, the disclosures of which are incorporated herein by reference. Formate dehydrogenases employed in the methods described herein, whether naturally occurring or non-naturally occurring, may exhibit an activity of at least about 1 $\mu\text{mol/min/mg}$, sometimes at least about 10 $\mu\text{mol/min/mg}$, or at least about $10^2 \mu\text{mol/min/mg}$, up to about $10^3 \mu\text{mol/min/mg}$ or higher, and can be readily screened for activity in the assay described in Example 4 of PCT publication WO 2005/018579.

As used herein, the term "formate" refers to formate anion (HCO_2^-), formic acid (HCO_2H), and mixtures thereof. Formate may be provided in the form of a salt, typically an alkali or ammonium salt (for example, HCO_2Na , KHCO_2NH_4 , and the like), in the form of formic acid, typically aqueous formic acid, or mixtures thereof. Formic acid is a moderate acid. In aqueous solutions within several pH units of its pKa (pKa=3.7 in water) formate is present as both HCO_2^- and HCO_2H in equilibrium concentrations. At pH values above about pH 4, formate is predominantly present as HCO_2^- . When formate is provided as formic acid, the reaction mixture is typically buffered or made less acidic by adding a base to provide the desired pH, typically of about pH 5 or above. Suitable bases for neutralization of formic acid include, but are not limited to, organic bases, for example amines, alkoxides and the like, and inorganic bases. When the glucose/glucose dehydrogenase cofactor regeneration system is employed, the co-production of gluconic acid (pKa=3.6), as represented in equation (1) causes the pH of the reaction mixture to drop if the resulting aqueous gluconic acid is not otherwise neutralized. The pH of the reaction mixture may be maintained at the desired level by standard buffering techniques, wherein the buffer neutralizes the gluconic acid up to the buffering capacity provided, or by the addition of a base concurrent with the course of the conversion.

Combinations of buffering and base addition may also be used. Suitable buffers to maintain desired pH ranges are described above. Suitable bases for neutralization for example, hydroxide salts (*e.g.*, NaOH), carbonate salts (*e.g.*, NaHCO₃), bicarbonate salts (*e.g.*, K₂CO₃), basic phosphate salts (*e.g.*, K₂HPO₄, Na₃PO₄), and the like.

5 When formate and formate dehydrogenase are employed as the cofactor regeneration system, the pH of the reaction mixture may be maintained at the desired level by standard buffering techniques, wherein the buffer releases protons up to the buffering capacity provided, or by the addition of an acid concurrent with the course of the conversion. Suitable acids to add during the course of the reaction to maintain the pH include organic acids, for
10 example carboxylic acids, sulfonic acids, phosphonic acids, and the like, mineral acids, for example hydrohalic acids (such as hydrochloric acid), sulfuric acid, phosphoric acid, and the like, acidic salts, for example dihydrogenphosphate salts (*e.g.*, KH₂PO₄), bisulfate salts (*e.g.*, NaHSO₄) and the like. Some embodiments utilize formic acid, whereby both the formate concentration and the pH of the solution are maintained.

15 When acid addition is employed to maintain the pH during a reduction reaction using the formate/formate dehydrogenase cofactor regeneration system, the progress of the conversion may be monitored by the amount of acid added to maintain the pH. Typically, acids added to unbuffered or partially buffered reaction mixtures over the course of conversion are added in aqueous solutions.

20 In carrying out embodiments of the ketoreductase-catalyzed reduction reactions described herein employing a cofactor regeneration system, either the oxidized or reduced form of the cofactor may be provided initially. As described above, the cofactor regeneration system converts oxidized cofactor to its reduced form, which is then utilized in the reduction of the ketoreductase substrate.

25 In some embodiments, cofactor regeneration systems are not used. For reduction reactions carried out without the use of a cofactor regenerating systems, the cofactor is added to the reaction mixture in reduced form.

 In some embodiments, when the process is carried out using whole cells of the host organism, the whole cell may natively provide the cofactor. Alternatively or in
30 combination, the cell may natively or recombinantly provide the glucose dehydrogenase.

 In carrying out the stereoselective reduction reactions described herein, the ketoreductase enzyme, and any enzymes comprising the optional cofactor regeneration system,

may be added to the reaction mixture in the form of the purified enzymes, whole cells transformed with gene(s) encoding the enzymes, and/or cell extracts and/or lysates of such cells. The gene(s) encoding the ketoreductase enzyme and the optional cofactor regeneration enzymes can be transformed into host cells separately or together into the same host cell. For example, in
5 some embodiments one set of host cells can be transformed with gene(s) encoding the ketoreductase enzyme and another set can be transformed with gene(s) encoding the cofactor regeneration enzymes. Both sets of transformed cells can be utilized together in the reaction mixture in the form of whole cells, or in the form of lysates or extracts derived therefrom. In other embodiments, a host cell can be transformed with gene(s) encoding both the ketoreductase
10 enzyme and the cofactor regeneration enzymes.

Whole cells transformed with gene(s) encoding the ketoreductase enzyme and/or the optional cofactor regeneration enzymes, or cell extracts and/or lysates thereof, may be employed in a variety of different forms, including solid (*e.g.*, lyophilized, spray-dried, and the like) or semisolid (*e.g.*, a crude paste).

15 The cell extracts or cell lysates may be partially purified by precipitation (ammonium sulfate, polyethyleneimine, heat treatment or the like), followed by a desalting procedure prior to lyophilization (*e.g.*, ultrafiltration, dialysis, and the like). Any of the cell preparations may be stabilized by crosslinking using known crosslinking agents.

In some embodiments, ketoreductase enzymes with improved purity may be
20 desired. In such embodiments, clarified cell lysates used to obtain the ketoreductase enzymes may be pre-treated with isopropanol, to bring the volume% of isopropanol to 25%. The isopropanol treated lysates may be incubated for between 1 hour and overnight at room temperature. The isopropanol-treated lysate can then be centrifuged, and the pellet may be removed. The supernatant may then be transferred to a petri dish and frozen at -80°C for a
25 minimum of 2 hours. Samples may then be lyophilized using a standard automated protocol. As showing in **Fig. 1**, gel electrophoresis, using sodium dodecyl sulfate (SDS, also known as sodium lauryl sulfate) and polyacrylamide gel (also known as SDS-PAGE), shows the removal of proteins that are insoluble in isopropyl alcohol (IPA), providing purified ketoreductase. In **Fig. 1**, T is total culture; L is lysate supernatant; S-IPA is the supernatant after 25% IPA
30 treatment; SFP-IPA is the shake flask powder of IPA treated supernatant; and P-IPA refers to the pellet removed by 25% IPA treatment.

Suitable conditions for carrying out the ketoreductase catalyzed reduction reactions described herein include a wide variety of conditions that can be readily optimized by routine experimentation that includes, but is not limited to, contacting the engineered ketoreductase enzyme and substrate at an experimental pH and temperature and detecting product, for example, using the methods described in the Examples provided herein.

The ketoreductase catalyzed reduction is typically carried out at a temperature in the range of from about 15°C to about 75°C. For some embodiments, the reaction is carried out at a temperature in the range of from about 20°C to about 55°C. In still other embodiments, it is carried out at a temperature in the range of from about 20°C to about 45°C. The reaction may also be carried out under ambient conditions.

The reduction reaction is generally allowed to proceed until essentially complete, or near complete, reduction of substrate is obtained. Reduction of substrate to product can be monitored using known methods by detecting substrate and/or product. Suitable methods include gas chromatography, HPLC, and the like. Conversion yields of the alcohol reduction product generated in the reaction mixture are generally greater than about 50%, may also be greater than about 60%, may also be greater than about 70%, may also be greater than about 80%, may also be greater than 90%, and are often greater than about 97%.

EXAMPLES

Example 1: Enzyme Preparation

E. coli cultures, each harboring a plasmid that encodes a ketoreductase enzyme that can be represented by amino acid sequence as set forth below in SEQ ID NO. 3-11, as set forth above, were diluted serially to dilutions of 10^{-4} , 10^{-5} , and 10^{-6} using Luria-Bretani Broth (culture media for cells) as a diluent. 100µl of the dilutions were each spread on a petri dish containing LB agar, supplemented with 50µg/ml of Kanamycin. The plates were placed in a 30°C incubator overnight.

200µl per well of Luria-Bretani Broth (culture media for cells) (500ml LB + 50µg/ml Kanamycin) was aliquoted into labeled 96-well shallow well plates. The shallow well plates were loaded into a plate stacker of a colony picker. Agar plates containing colonies sufficiently diluted so that the majority of colonies were isolated from one another (known as single colonies to those schooled in the art) were picked into unique wells of the shallow well plates. The colonies were allowed to grow overnight, at 200rpm, at 30°C, and 85%RH.

390µl of Terrific Broth (TB) growth media (commercially available from ThermoFisher Scientific as Catalog #A1374301) (TB + 50µg/ml Kanamycin) was aliquoted into labeled 96-well deep well subculture plates. 13µl of the overnight growth culture was transferred from each well of the master shallow well plate into the corresponding labeled deep well subculture plates. The plates were sealed with breathable film, and the plates were shaken for 2-2.5h at 250rpm, at 30°C and 85%RH. After shaking, optical density (OD₆₀₀, optical density at 600nm wavelength) of at least one plate was measured for growth. When the OD₆₀₀ of this plate was in the range of 0.4-0.8, the deep well plates were inducted with 40µl per well of induction media. The plates were resealed and incubated with shaking for 18-20h at 250rpm, at 30°C, and 85%RH.

After incubation, all deep well plates were centrifuged for 15min. at 4000rpm at 4°C. Following centrifuge, the supernatant was discarded. The plates were heat-sealed and stored at -80°C.

The cell plates were removed from -80°C storage and defrosted at RT. A lysis buffer of 100mM sodium phosphate pH 6.0, 1mg/mL lysozyme, 0.50mg/mL polymyxin B sulfate (PMBS), 3units/ml DNase I and 4mM MgCl₂ was prepared. 400µl of the lysis buffer was aliquoted into each well. The lysis mixture was shaken for 1.5-2h at 1000rpm on a plate shaker at RT. The lysis mixture was then centrifuged for 15min at 4000rpm at 4°C to make the enzyme containing lysate solution (which is in the supernatant). In some instances, the lysate solution was further incubated with isopropanol in solutions of 25-30% isopropanol for 1-5h. After incubation, the lysate was centrifuged as before.

Example 2: Ketoreductase Reaction in Well Plates

A 10x ketone stock solution was prepared by making a 100g/L of solution of substrate ketone in 100% DMSO and adding an equal volume of THF. A reaction buffer was prepared by mixing 5.8g/L of the commercially available ketoreductase KRED-P1 B10 (available from Codexis), 5.8g/L of NADP, and 11.5% isopropanol in 100mM sodium phosphate, to a pH of 6.0.

350µl of the reaction buffer was added to 1ml round bottom well plates. 40µl of the 10x ketone stock solution was added to each well, followed by 10µl of the enzyme-containing lysate solution of Example 1. The plate was heat sealed and shaken overnight for 30°C at 1000rpm.

After overnight shaking, 190µl of a mixture of 50%ACN/water + 0.05%TFA was aliquoted into filter stacks (filter plates on top of the round bottom plates, with 0.20µM hydrophilic PTFE, commercially available from Millipore MSRLN2250). The reaction plates were unsealed. 10µl of reaction mixture from the reaction plates was transferred into the
5 corresponding filter plates with a receiving plate underneath. The plates were then centrifuged for 3min at 4000rpm at 4°C. The filter plates were removed, and the clear solutions in the receiving plates were heat sealed.

The filtered solutions were analyzed by ultra-performance liquid chromatography (UPLC), using a high throughput screening method to monitor the substrate and degradation of
10 the substrates, and the product peak areas. UPLC was conducted on a Waters HSS T3 1.8µm, 2.1x75mm column using an isocratic method of 15% CH₃CN/H₂O + v 0.05% TFA over 1.5min; flow rate 1ml/min. The starting material eluted at 0.84min, the desired enantiomer at 1min and the undesired enantiomer at 1.09min.

15 **Example 3: Enzyme Preparation in Shaker Flasks**

5ml of Luria-Bretani Broth (culture media for cells) (250ml LB + 50µg/ml Kanamycin + 1% glucose) was inoculated with 10µl of *E. coli* cells, each harboring a plasmid that encodes a ketoreductase enzyme that can be represented by amino acid sequence as set forth below in SEQ ID NO. 3-11, as set forth above, that had been stored at -80°C in 20% glycerol and
20 aliquoted into labeled 15mL cell culture tubes. The cell culture tubes were sealed and incubated with shaking for 20-24h at 250rpm, at 30°C.

After overnight growth, the overnight growth cultures (2-5ml of cell culture (having a starting OD₆₀₀ of 0.2)) were added to 250ml of Terrific Broth (TB) growth media (commercially available from ThermoFisher Scientific as Catalog #A1374301) (TB + 50µg/ml
25 Kanamycin) to a final volume of 250mL. The flasks were shaken for 3-4h at 250rpm, at 30°C. After shaking, OD₆₀₀ was measured for growth until OD₆₀₀ reached 0.4-0.6. At this point, 1mM of IPTG (250µl of 1M IPTG) was added to the culture to induce expression, and the culture was allowed to grow 20-24h at 250rpm, at 30°C.

After the additional growth period, the cultures were transferred to a centrifuge
30 bottle of known weight and centrifuged for 20min. at 4000rpm at 4°C. Following centrifugation, the supernatant was discarded, and the remaining cell pellet in the bottle was weighed. The weight of the cell pellet was calculated by subtracting the known weight of the bottle, and the

cell pellet was resuspended in 5 volumes of 50mM sodium phosphate buffer (pH = 7), at a volume that is 5 times the weight of the cell pellet.

The cells from the resuspended cell pellets were lysed using a microfluidizer, and the cell lysate was collected and centrifuged for 60min. at 10000rpm at 4°C. The clarified lysate
5 was occasionally further treated with isopropanol in solutions of 25-30% isopropanol for 1-5h. After incubation, the lysate was centrifuged as before. The clarified supernatant was transferred to a petri dish and frozen at -80°C for approximately 2h. Samples were lyophilized using a standard automated protocol.

10 ***Example 4: Ketoreductase Reaction in Vials***

10µL each of ketoreductase enzymes that can be represented by amino acid sequence as set forth below in SEQ ID NO. 3-11 were inoculated into 5mL of Luria-Bretani Broth (culture media for cells) in labeled 15mL cell culture tubes, supplemented with 1% glucose and 50ug/ml of Kanamycin antibiotic and grown overnight for 20-24h at 30°C, 250 rpm,
15 in a shaking incubator.

5mL of the overnight cultures were used to subculture 250mL of Terrific Broth growth media (commercially available from ThermoFisher Scientific as Catalog #A1374301) in a 1L flask. The subcultures were allowed to grow at 30°C for 3h at 250rpm, in a shaking incubator. When the OD₆₀₀ measures between 0.4 and 0.6, the IPTG was introduced to an IPTG
20 final concentration of 1mM. The subcultures were allowed to grow overnight, for 18-20h.

After the growth period, the cultures were transferred to centrifuge bottles and centrifuged for 20min at 4000rpm at 4°C. Following centrifuge, the supernatants were discarded. The cell pellets were resuspended in 50mM sodium phosphate buffer (pH = 7).

The cells from the resuspended cell pellets were lysed using a microfluidizer, and
25 the cell lysate was collected and centrifuged for 60min at 10000rpm at 4°C. In some instances, the lysate solution was further incubated with isopropanol in solutions of 25-30% isopropanol for 1-5 h. After incubation, the lysate was centrifuged as before. The supernatant was transferred to a petri dish and frozen at -80°C for a minimum of 2h. Samples were optionally lyophilized using a standard automated protocol.

30 The ketoreductase enzyme (20mg, harvested from the subculture) and Compound A (20mg) were added to each of six 4mL vials.

To a separate vial was added 1mL of 1mg/mL solution of a commercially available ketoreductase enzyme (KRED-P1 B10, available from CODEXIS) along with NADPH (32mg), followed by 3mL of 100mM phosphate buffer (pH = 6.0).

To each of the vials containing ketoreductase enzyme and Compound A was added 600µL of the buffered solution, followed by 120µL of IPA. The contents of the vials were combined in a shaker and held at 30°C for 20h.

Example 5: Ketoreductase Reaction

10µL each of ketoreductase enzymes that can be represented by amino acid sequence as set forth below in SEQ ID NO. 3-11 were inoculated into 5mL of Luria-Bretani Broth (culture media for cells), supplemented with 1% glucose and 50ug/ml of Kanamycin antibiotic and grown overnight for 20-24h at 30°C, 250 rpm, in a shaking incubator.

5mL of the overnight cultures were used to subculture 250mL of Terrific Broth growth media (commercially available from ThermoFisher Scientific as Catalog #A1374301) in a 1L flask. The subcultures were allowed to grow at 30°C for 3h at 250rpm, in a shaking incubator. When the OD₆₀₀ measures between 0.4 and 0.6, the IPTG was introduced to an IPTG final concentration of 1mM). The subcultures were allowed to grow overnight, for 18-20h.

After the growth period, the cultures were transferred to centrifuge bottles and centrifuged for 20min at 4000rpm at 4°C. Following centrifuge, the supernatants were discarded. The cell pellets were resuspended in 50mM sodium phosphate buffer (pH = 7).

The cells from the resuspended cell pellets were lysed using a microfluidizer, and the cell lysate was collected and centrifuged for 60min at 10000rpm at 4°C. In some instances, the lysate solution was further incubated with isopropanol in solutions of 25-30% isopropanol for 1-5h. After incubation, the lysate was centrifuged as before. The supernatant was transferred to a petri dish and frozen at -80°C for a minimum of 2h. Samples were optionally lyophilized using a standard automated protocol.

20mg of a commercially available ketoreductase enzyme (KRED-P1 B10, available from CODEXIS) added to a reaction vessel, along with NADPH (20mg), ketoreductase enzyme (250mg, harvested from the subculture) and Compound A (250mg). 10mL of phosphate buffer (0.1M, pH = 6.0) and 1mL IPA were added to the reaction vessel. The temperature was set at 30°C, and the reaction mixture was stirred at 350rpm. After 20h, the mixture was cooled to 15°C. NaCl (2g) was added to the reaction vessel, and the reaction mixture was allowed to de-

supersaturate overnight. The solids were filtered and washed with 2.5 mL (10 vol) of water. The wet cake was placed into a 50°C vacuum oven to dry overnight.

¹H NMR (500M Hz, DMSO-d6) δ 11.68 (s, 2H), 8.27 (s, 1H), 5.96 (d, J = 5.4 Hz, 1H), 5.83 (d, J = 8.1 Hz, 1H), 5.22 (t, J = 5.4 Hz, 1H), 5.07 (dd, J = 54.3, 4.1 Hz, 1H), 4.77
5 (dddd, 27.3, 8.1, 4.1, 4.1 Hz, 1H), 4.25 (dddd, J = 27.2, 8.1, 4.1 Hz, 4.1 Hz, 1H), 3.61 (m, 2H), 2.75 (sept, J = 6.8 Hz, 1H), 1.12 (d, J = 6.8 Hz, 6H).

¹³C NMR (125 MHz, DMSO-d6) δ 180.2, 154.8, 149.3, 148.4, 137.4, 120.1, 92.8 (d, J= 183 Hz), 85.0, 83.6 (d, J = 21.1 Hz), 83.5, 72.6 (d, J = 16.1 Hz), 60.6 (d, J = 11.2 Hz), 34.8, 18.8, 18.8.

10 **¹⁹F NMR** (500 MHz, DMSO-d6) δ -197.5.

It will be appreciated that various of the above-discussed and other features and functions, or alternatives thereof, may be desirably combined into many other different systems or applications. Also that various presently unforeseen or unanticipated alternatives,
15 modifications, variations, or improvements therein may be subsequently made by those skilled in the art which are also intended to be encompassed by the following claims.

WHAT IS CLAIMED IS:

1. A polypeptide, wherein said polypeptide comprises an amino acid
sequence having at least a 90% sequence identity to any one of SEQ ID NO: 2, 3, 4, 5, 6, 7, 8, 9,
5 10, or 11.
2. The polypeptide according to claim 1, having at least a 90% sequence
identity to SEQ ID NO: 2.
- 10 3. The polypeptide according to claim 1, having at least a 90% sequence
identity to SEQ ID NO: 3.
4. The polypeptide according to claim 1, having at least a 90% sequence
identity to SEQ ID NO: 4.
15
5. The polypeptide according to claim 1, having at least a 90% sequence
identity to SEQ ID NO: 5.
6. The polypeptide according to claim 1, having at least a 90% sequence
20 identity to SEQ ID NO: 6.
7. The polypeptide according to claim 1, having at least a 90% sequence
identity to SEQ ID NO: 7.
- 25 8. The polypeptide according to claim 1, having at least a 90% sequence
identity to SEQ ID NO: 8.
9. The polypeptide according to claim 1, having at least a 90% sequence
identity to SEQ ID NO: 9.
30
10. The polypeptide according to claim 1, having at least a 90% sequence
identity to SEQ ID NO: 10.

11. The polypeptide according to claim 1, having at least a 90% sequence identity to SEQ ID NO: 11.
- 5 12. An expression vector comprising the polypeptide according to any one of claim 1 through claim 11, operably linked to one or more control sequences suitable for directing expression of the encoded polypeptide in a host cell.
- 10 13. The expression vector of claim 11, wherein the control sequence comprises a promoter.
14. The expression vector of claim 12, wherein the promoter comprises an *E. coli* promoter.
- 15 15. A host cell comprising the expression vector of claim 11.
16. The host cell of claim 15, wherein said host cell is *E. coli*.
- 20 17. A process for precipitating proteins from a cell lysate containing ketoreductase enzymes, comprising treating the cell lysate containing ketoreductase enzymes with >20% isopropanol.
- 25 18. The process of claim 17, wherein the volume% of isopropanol is about 25%.