

naturally-occurring protein or polypeptide fragment thereof, which has an amino acid sequence that is at least 60%, e.g., at least 65%, 70%, 75%, 80%, 85%, 90%, 95%, 96%, 97%, 98%, or 99%, but less than 100%, amino acid sequence identity to the amino acid sequence of the naturally-occurring (native or wild-type) protein from which it is derived. The term "fusion polypeptide" or "fusion protein" refers to a chimeric protein containing a reference protein (e.g., luciferase) joined at the N- and/or C-terminus to one or more heterologous sequences (e.g., a non-luciferase polypeptide).

[0082] Protein primary structure (primary sequence, peptide sequence, protein sequence) is the sequence of amino acids. It is generally reported starting from the amino-terminal (N) end to the carboxyl-terminal (C) end. Protein secondary structure can be described as the local conformation of the peptide chain, independent of the rest of the protein. There are 'regular' secondary structure elements (e.g., helices, sheets or strands) that are generally stabilized by hydrogen bond interactions between the backbone atoms of the participating residues, and 'irregular' secondary structure elements (e.g., turns, bends, loops, coils, disordered or unstructured segments). Protein secondary structure can be predicted with different methods/programs, e.g., PSIPRED (McGuffin et al., Bioinformatics, 16:404 (2000)), PORTER (Pollastri et al., Bioinformatics, 21:1719 (2005)), DSC (King and Sternberg, Protein Sci., 5:2298 (1996)), see www.expasy.org/tools/#secondary for a list. Protein tertiary structure is the global three-dimensional (3D) structure of the peptide chain. It is described by atomic positions in three-dimensional space, and it may involve interactions between groups that are distant in primary structure. Protein tertiary structures are classified into folds, which are specific three-dimensional arrangements of secondary structure elements. Sometimes there is no discernable sequence similarity between proteins that have the same fold.

[0083] The term "wild-type" or "native" as used herein, refers to a gene or gene product that has the characteristics of that gene or gene product isolated from a naturally occurring source. A wild-type gene is that which is most frequently observed in a population and is thus arbitrarily designated the "wild-type" form of the gene. In contrast, the term "mutant" refers to a gene or gene product that displays modifications in sequence and/or functional properties (i.e., altered characteristics) when compared to the wild-type gene or gene product. It is noted that naturally occurring mutants can be isolated; these are identified by the fact that they have altered characteristics when compared to the wild-type gene or gene product.

[0088] Methods of alignment of sequence for comparison are well-known in the art. Optimal alignment of sequences for comparison can be conducted by the local homology algorithm of Smith et al. (1981), by the homology alignment algorithm of Needleman et al. (J. Mol. Biol., 48:443 (1970)), by the search for similarity method of Person et al. (Proc. Natl. Acad. Sci. USA, 85, 2444 (1988)), by computerized implementations of these algorithms (GAP, BESTFIT, FASTA, and TFASTA in the Wisconsin Genetics Software Package, Genetics Computer Group, 575 Science Dr., Madison, WI), or by manual alignment and visual inspection.

[0089] Computer implementations of these mathematical algorithms can be utilized for comparison of sequences to determine sequence identity. Such implementations include, but are not limited to: CLUSTAL in the PC/Gene program (available from Intelligenetics, Mountain View, California); the ALIGN program (Version 2.0) and GAP, BESTFIT, BLAST, FASTA, and TFASTA in the Wisconsin Genetics Software Package, Version 8 (available from Genetics Computer Group (GCG), 575 Science Drive, Madison, Wisconsin, USA). Alignments using these programs can be performed using the default parameters. The CLUSTAL program is well described by Higgins et al., Gene, 73:237 (1988); Higgins et al., CABIOS, 5:157 (1989); Corpet et al., Nucl. Acids Res., 16:1088 (1988); Huang et al., CABIOS, 8:155 (1992); and Pearson et al., Methods Mol. Biol., 24:307 (1994). The ALIGN program is based on the algorithm of Myers and Miller, LABIOS, 4:11 (1988). The BLAST programs of Altschul et al. (J. Mol. Biol., 215:403 (1990)) are based on the algorithm of Karlin and Altschul (PNAS USA, 90:5873 (1993)).

[0090] Software for performing BLAST analyses is publicly available through the National Center for Biotechnology Information (www.ncbi.nlm.nih.gov/). This algorithm involves first identifying high scoring sequence pairs (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., J. Mol. Biol., 215:403 (1990)). 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 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).

structure of OgLuc. There is no published experimental tertiary structure or tertiary structure model of OgLuc. Homology modeling was used to generate a tertiary structure model of OgLuc. Building a homology model comprises several steps including identification of 3D structural template(s), alignment of target sequence (e.g., OgLuc) and template structure(s), model building, and model quality evaluation. Identification of one or more 3D structural templates for OgLuc was not intuitive because standard sequence search methods did not identify significant overall similarity to proteins with known tertiary structure. To overcome this problem, two approaches were employed to identify remote OgLuc homologs with known tertiary structure.

[00116] Approach 1:

[00117] An Hidden Markov Model (HMM) based template library search (Karplus et al., Bioinformatics, 14:846 (1998)) was used to detect distantly related template structures using the SWISS-MODEL Template Identification Tool at swissmodel.expasy.org/SWISS-MODEL.html (Arnold et al., Bioinformatics, 22:195 (2006)).

[00118] The best (highest E-value score) 3D structure template identified for OgLuc using this approach was a fatty acid binding protein (FABP) (Protein Data Bank (PDB) accession number 1VYF) (Angelucci et al., Biochemistry, 43:13000 (2004)). Additional FABPs with lower scores were also identified, including PDB accession numbers 1PMP and 1CRB.

[00119] Exemplary alignments of the target sequence (OgLuc, residues 1-2 and 168-169 omitted) and the sequences of the identified 3D structure templates (1VYF, 1PMP, 1CRB) are shown below. Note that due to the low sequence similarity, the placement of gaps in the alignment can vary.

1vyf 1 GSMSSFLGKWKLSSEHNFDVMSKLGVSWATRQIGNTVTPTVTFTMDGDK... 50
F G W N D V G S G VTP G
Target 3 --LADFVGDWQQTAGYNQDQVLEQGGLSSLFQALGVSVTPIQKVVLSGENgl 52

1vyf 51MTMLTESTFKN..LSCTFKF.....GEEF 72
S F FK G
Target 53 kadihviIPYEGLSGFQMgliEMIFKVVypvddhhfkiihygtlvidGVTP 104

1vyf 73 DEKTS DGRNVKSVVEKNSKLTQTQVDPKNTTVIVREV.DGDTMKTTVTVG 123
GR N R VT
Target 105 NMIDYFGRPYPGIAVFDGKQITVTGTLWNGNKIYDERLInPDGSLLFRVTIN 156

1vyf 124 DVTAIRNYKRLS 135 (SEQ ID NO:5)
VT R
Target 157 GVTGWRLCENI 167 (SEQ ID NO:7)

1pmp 3 SNKFLGTWKLVSSENFDEYMKALGVGLATRKLGNLAKPRVIISKKGDI.... 48
F G W N D G LG P G
Target 3 LADFVGDWQQTAGYNQDQVLEQGGLSSLFQALGVSVTPIQKVVLSGENglka 54

1pmp 49ITIRTESPFKNTESISFKL.....GQEFEE 72
P L G
Target 55 dihviiipyeglsqfmgliEMIFKVVYPVDDHHFKIILhygtlvidGVTPNM 106

1pmp 73 TTADNRKTKSTVTLARGSLNQVQK.WNGNETTIKRKL.VDGKMVVECKMKDV 122
R WNGN R DG V
Target 107 IDYFGRPYPGIAVFDGKQITVTGTLWNGNKIYDERLInPDGSLLFRVTINGV 158

1pmp 123 VCTRIYEKV 131 (SEQ ID NO:3)
R E
Target 159 TGWRLCENI 167 (SEQ ID NO:7)

1crb 1 PVDFNGYWKMLSNENFEEYLRALDVNVALRKIANLLKPDKEIVQDGDH.... 48
DF G W N L P V G
Target 3 LADFVGDWQQTAGYNQDQVLEQGGLSSLFQALGVSVTPIQKVVLSGENglka 54

1crb 49MIIRTLSTFRNYIMDFQV.....GKEFEE 72
MI G
Target 55 dihviiipyeglsqfmgliEMIFKVVYPVDDHHFKIILhygtlvidGVTPNM 106

1crb 73 DLTGIDDRKCMTTVSWDGDQLQCVQK.GEKEGRGWTQWI.EGDELHLEMRAE 122
R DG I L
Target 107 --IDYFGRPYPGIAVFDGKQITVTGTLWNGNKIYDERLInPDGSLLFRVTIN 156

1crb 123 GVTCKQVFKKVH 134 (SEQ ID NO:4)
GVT
Target 157 GVTGWRLCENI-- 165 (SEQ ID NO:7)

[00120] Approach 2:

[00121] A fold recognition method using the “GeneSilico meta-server” at genesilico.pl/meta2 (Kurowski et al., Nucl. Acids Res., 31:3305 (2003)) was also used to identify remote OgLuc homologs with known tertiary structure.

CLAIMS

What is claimed is:

1. A polypeptide comprising an amino acid sequence with at least 70% sequence identity with SEQ ID NO:1 and comprising at least one amino acid substitution relative to SEQ ID NO: 1 at a position corresponding to position 2, 4, 11, 20, 23, 28, 33, 34, 44, 45, 51, 54, 68, 72, 75, 76, 77, 89, 90, 92, 99, 104, 115, 124, 135, 138, 139, 143, 144, 166, 167, or 169 of SEQ ID NO:1, wherein the polypeptide is a luciferase that is capable of utilizing coelenterazine as a substrate to generate luminescence, and wherein the polypeptide has at least one of enhanced luminescence, enhanced signal stability, and enhanced protein stability relative to a luciferase comprising the amino acid sequence of SEQ ID NO: 1, wherein the modified luciferase polypeptide has luciferase activity.
2. The polypeptide of claim 1, wherein the at least one amino acid substitution relative to SEQ ID NO: 1 is at a position corresponding to at least one of positions 4, 11, 20, 28, 33, 54, 68, 75, 115, 124, 138, 143, or 166 of SEQ ID NO:1.
3. The polypeptide of claim 1, comprising amino acid substitutions relative to SEQ ID NO: 1 at positions corresponding to positions 11, 33, 44, 54, 115, 124, 138, and 166 of SEQ ID NO:1.
4. The polypeptide of claim 3, wherein the amino acid at position 11 is arginine, the amino acid at position 33 is lysine, the amino acid at position 44 is isoleucine, the amino acid at position 54 is phenylalanine, the amino acid at position 115 is glutamic acid, the amino acid at position 124 is lysine, the amino acid at position 138 is isoleucine, and the amino acid at position 166 is arginine.
5. The polypeptide of claim 3, further comprising amino acid substitutions relative to SEQ ID NO: 1 at positions corresponding to positions 45, 104, 135, 139, and 167 of SEQ ID NO:1.

6. The polypeptide of claim 5, wherein the amino acid at position 45 is glutamic acid, the amino acid at position 104 is leucine, the amino acid at position 135 is lysine, the amino acid at position 139 is glutamic acid, and the amino acid at position 167 is valine.
7. The polypeptide of claim 3, further comprising amino acid substitutions relative to SEQ ID NO: 1 at positions corresponding to positions 28, 34, 51, 99 and 143 of SEQ ID NO:1.
8. The polypeptide of claim 7, wherein the amino acid at position 28 is proline, the amino acid substitution made at position 34 is methionine, the amino acid substitution made at position 51 is valine, the amino acid substitution made at position 99 is valine, and the amino acid substitution made at position 143 is leucine.
9. The polypeptide of claim 3, further comprising an amino acid substitution relative to SEQ ID NO: 1 at one or more positions corresponding to positions 20, 54, 68, 72, 75, 77, 79, 89, 90, 92, and 164 of SEQ ID NO:1.
10. The polypeptide of claim 9, further comprising one or more of an arginine at position 20, an isoleucine at position 54, a serine at position 68, a glutamine at position 72, a lysine at position 75, a tryptophan at position 77, an isoleucine at position 79, a glutamic acid at position 89, either a threonine or valine at position 90, an amino acid selected from the group consisting of glycine, glutamine, serine and alanine at position 92, and a serine at position 164.
11. The polypeptide of claim 3, further comprising amino acid substitutions relative to SEQ ID NO: 1 at positions corresponding to positions 23, 28, 143, and 166 of SEQ ID NO:1.
12. The polypeptide of claim 11, wherein the amino acid at position 23 is valine, the amino acid at position 28 is proline, the amino acid at position 143 is valine, and the amino acid at position 166 is arginine.

13. The polypeptide of claim 1, wherein the amino acid substitutions relative to SEQ ID NO: 1 are at positions corresponding to positions 4, 34, 76, and 166 of SEQ ID NO:1.
14. The polypeptide of claim 13, wherein the amino acid at position 4 is serine, the amino acid at position 35 is methionine, the amino acid at position 76 is valine, and the amino acid at position 166 is arginine.
15. The polypeptide of claim 1, wherein the amino acid substitutions relative to SEQ ID NO: 1 are at positions corresponding to positions 51, 99, and 166 of SEQ ID NO:1.
16. The polypeptide of claim 15, wherein the amino acid at position 51 is valine, the amino acid at position 99 is valine, and the amino acid at position 177 is arginine.
17. A fusion protein comprising the polypeptide of claim 1 linked to a polypeptide of interest.