(12) STANDARD PATENT (11) Application No. **AU 2007256780 B2**
(19) AUSTRALIAN PATENT OFFICE

(54) Title
Protein surface remodeling

(51) International Patent Classification(s)
G01N 33/50 (2006.01)

(21) Application No: **2007256780** (22) Date of Filing: **2007.06.01**

(87) WIPO No: **WO07/143574**

(30) Priority Data

(31) Number	(32) Date	(33) Country
60/836,607	**2006.08.09**	**US**
60/810,364	**2006.06.02**	**US**

(43) Publication Date: **2007.12.13**
(44) Accepted Journal Date: **2013.08.29**

(71) Applicant(s)
President and Fellows of Harvard College

(72) Inventor(s)
Phillips, Kevin John;Liu, David R.;Lawrence, Michael S.

(74) Agent / Attorney
Davies Collison Cave, Level 15 1 Nicholson Street, MELBOURNE, VIC, 3000

(56) Related Art
Strickler et al., Biochemistry, 2006
WO 91/00345

(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

CORRECTED VERSION

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
13 December 2007 (13.12.2007)

PCT

(10) International Publication Number
WO 2007/143574 A1

(51) International Patent Classification:
G01N 33/50 (2006.01)

(21) International Application Number:
PCT/US2007/070254

(22) International Filing Date: 1 June 2007 (01.06.2007)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
60/810,364 2 June 2006 (02.06.2006) US
60/836,607 9 August 2006 (09.08.2006) US

(71) Applicant *(for all designated States except US)*: **THE PRESIDENT AND FELLOWS OF HARVARD COLLEGE** [US/US]; 17 Quincy Avenue, Cambridge, MA 02139 (US).

(72) Inventors; and

(75) Inventors/Applicants *(for US only)*: **LIU, David, R.** [US/US]; 1 Fox Run Lane, Lexington, MA 02420 (US). **PHILLIPS, Kevin, John** [US/US]; 119 College Avenue, Apt 32, Somerville, MA 02144 (US). **LAWRENCE, Michael, S.** [US/US]; 11 Brookside Terrace, Atkinson, NH 03811 (US).

(74) Agent: **BAKER, C., Hunter**; Choate, Hall & Stewart LLP, Two International Place, Boston, MA 02110 (US).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BH, BR, BW, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PG, PH, PL, PT, RO, RS, RU, SC, SD, SE, SG, SK, SL, SM, SV, SY, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

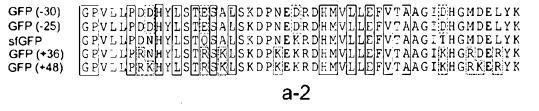
(84) Designated States *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH,

[Continued on next page]

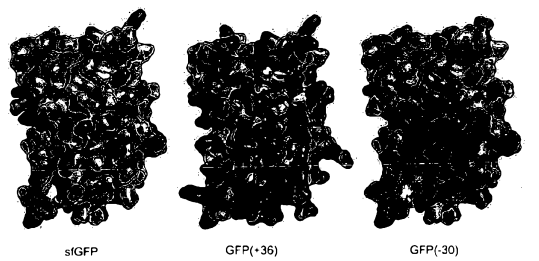
(54) Title: PROTEIN SURFACE REMODELING

a

a-1



a-2



b

(57) Abstract: Aggregation is a major cause of the misbehavior of proteins. A system for modifying a protein to create a more stable variant is provided. The method involves identifying non- conserved hydrophobic amino acid residues on the surface of a protein, suitable for mutating to more hydrophilic residues (e.g., charged amino acids). Any number of residues on the surface may be changed to create a variant that is more soluble, resistant to aggregation, has a greater ability to re-fold, and/or is more stable under a variety of conditions. The invention also provides GFP, streptavidin, and GST variants with an increased theoretical net charge created by the inventive technology. Kits are also provided for carrying out such modifications on any protein of interest.

WO 2007/143574 A1

GM, KE, LS, MW, MZ, NA, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HU, IE, IS, IT, LT, LU, LV, MC, MT, NL, PL, PT, RO, SE, SI, SK, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Published:

— *with international search report*

(48) Date of publication of this corrected version:
16 October 2008

(15) Information about Correction:
see Notice of 16 October 2008

PROTEIN SURFACE REMODELING

Related Applications

[0001] The present application claims priority under 35 U.S.C. § 119(e) to U.S. provisional patent application, USSN 60/810,364, filed June 2, 2006; the entire contents of which is incorporated herein by reference.

Government Support

[0002] The work described herein was supported, in part, by grants from the National Institutes of Health (GM065400). The United States government may have certain rights in the invention.

Background of the Invention

[0003] Proteins are the workhorses of the cell. Proteins catalyze chemical reactions, transduce signals in biological systems, provide structural elements in cells and the extracellular matrix, act as messengers, *etc.* One of the major causes of misbehavior of proteins is aggregation. This is not only a problem in the laboratory but also a problem in many diseases such as Alzheimer's disease. Aggregation is a particularly vexing problem when it comes to computationally designed proteins. For example, TOP7 is a computationally designed protein with a novel fold. A longer version of TOP7, TOP7 extended, is very prone to aggregation. TOP7ex is expressed predominantly as insoluble aggregates.

[0004] As more proteins are either designed or modified to be used a tools to study biological systems or as more proteins—wild type or modified—are used as therapeutic agents, there needs to be a system for routinely modifying these proteins to be more stable and/or to prevent aggregation.

Summary of the Invention

[0005] The present invention provides a system for modifying proteins to make them more stable. The invention stems from the recognition that modifying the hydrophobic areas on the surface of a protein can improve the extrathermodynamic properties of the protein. The inventive system is particularly useful in improving the solubility of a protein of interest,

improving the protein's resistance to aggregation, and/or improving the protein's ability to renature. All of these properties are particularly useful in protein production, protein purification, and the use of proteins as therapeutic agents and research tools.

[0006] In one aspect, the invention provides a method of altering the primary sequence of a protein in order to increase the protein's resistance to aggregation, solubility, ability to refold, and/or general stability under a wide range of conditions. The activity of the modified protein is preferably approximately or substantially the same as the protein without modification. In certain embodiments, the modified protein retains at least 50%, 75%, 90%, or 95% of the wild type protein's activity. In one embodiments, the method includes the steps of (a) identifying the surface residues of a protein of interest; (b) identifying the particular surface residues that are not highly conserved among other proteins related to the protein of interest (*i.e.*, determining which amino acids are not essential for the activity or function of the protein); (c) determining the hydrophobicity of the identified non-conserved surface residues; and (e) replacing at least one or more of the identified hydrophobic, non-conserved residues with an amino acid that is more polar or is charged at physiological pH. Each of the above steps may be carried out using any technique, computer software, algorithm, paradigm, *etc.* known in the art. After the modified protein is created, it may be tested for its activity and/or the desired property being sought. In certain embodiments, the modified protein is more stable. In certain embodiments, the modified protein is less susceptible to aggregation. The inventive method typically increases the net charge (positive or negative) on the protein at physiological pH.

[0007] In another aspect, the invention provides a method of altering the primary sequence of a protein in order to increase the protein's resistance to aggregation, solubility, ability to refold, and/or general stability under a wide range of conditions by "supercharging" the protein. That is, the overall net charge on the modified protein is increased (either positive charge or negative charge) compared to the wild type protein. Preferably, the activity of the modified protein is approximately or substantially the same as the protein without modification. In certain embodiments, the method includes the steps of (a) identifying the surface residues of a protein of interest; (b) identifying the particular surface residues that are not highly conserved among other proteins related to the protein of interest (*i.e.*, determining which amino acids are not essential for the activity or function of the protein); (c) determining the hydrophilicity of the identified non-conserved surface residues;

and (e) replacing at least one or more of the identified charged or polar, solvent-exposed, non-conserved residues with a charged amino acid that is charged at physiological pH. In certain embodiments, to make a negatively charged "supercharged" protein, the residues identified for modification are mutated either to aspartate (Asp) or glutamate (Glu) residues. In certain other embodiments, to make a positively charged "supercharged" protein, the residues identified for modification are mutated either to lysine (Lys) or arginine (Arg) residues. Each of the above steps may be carried out using any technique, computer software, algorithm, paradigm, *etc.* known in the art. After the modified protein is created, it may be tested for its activity and/or the desired property being sought. In certain embodiments, the modified protein ("supercharged protein") is more stable. In certain embodiments, the modified protein is less susceptible to aggregation. The inventive method typically increases the net charge (positive or negative) on the protein at physiological pH.

[0008] The theoretical net charge on over 80% of the proteins catalogued in the Protein Data Bank (PDB) fall within ±10. The modified protein created by the present invention typically have a net charge less than -10 or greater than +10. In certain embodiments, the modified protein has a net charge less than -20 or greater than +20. In certain embodiments, the modified protein has a net charge less than -30 or greater than +30. In certain embodiments, the modified protein has a net charge less than -40 or greater than +40. In certain embodiments, the modified protein has a net charge less than -50 or greater than +50. The modified proteins are able to fold correctly and retain their biological activity.

[0009] Any protein may be modified using the inventive system, and protein variants created by the inventive system are considered to be part of the present invention, as well as polynucleotides or vectors encoding the variant protein and cells expressing the variant protein. The inventive system has been used to create several new variants of green fluorescent protein (GFP). These variants retain their fluorescence; however, they are more stable than current versions of GFP under a wide range of environments. The inventive GFPs are immune to aggregation even over long periods of time and in environments that induce aggregation and are capable of refolding into a fluorescent protein even after being denatured by boiling. The inventive system has also been used to create new variants of streptavidin and glutathione-*S*-transferase (GST). These variants retain their biological activity and remain soluble when heated. The invention also includes polynucleotide sequences encoding the inventive GFP, streptavidin, and GST protein sequences, vectors including any of these

nucleotide sequences, and cells that include such a polynucleotide sequence or vector, or express the inventive variants. In certain embodiments, the invention includes bacteria or other cells that overexpress an inventive variant. The inventive variants may be used in a variety of biological assays known in the art. For example, supercharged GFPs may be used in any assay that currently uses GFP as a reporter protein.

[0010] In another aspect, the invention provides other proteins that have been modified by the inventive system. These modified proteins preferably retain a significant portion of their original activity. In certain embodiments, the modified protein retains at least 99%, 98%, 95%, or 90% of the activity of the unmodified version. The modified protein may be more soluble, resistant to aggregation, have a increased ability to refold, and/or have greater stability under a variety of conditions. The proteins modified by the inventive system include hydrophobic proteins, recombinant proteins, membrane proteins, structural proteins, enzymes, extracellular proteins, therapeutic proteins (*e.g.*, insulin, cytokines, immunoglobulins, fragments of immunoglobulins, *etc.*), receptors, cell signaling proteins, cytoplasmic proteins, nuclear proteins, transcription factors, *etc.* In certain specific embodiments, the proteins are therapeutic proteins for use in human or veterinary medicine. In certain embodiments, the proteins are unnatural proteins, for example, computationally designed proteins. In other embodiments, the proteins are hybrid proteins, fusion proteins, altered proteins, mutated proteins, genetically engineered proteins, or any other protein that has been altered by the hands of man.

[0011] Kits are also provided for the practice of the invention. The kits may include the reagents needed to modify a protein of interest to make it more resistant to aggregation, increase its ability to renature, or increase its stability overall. Such kits may include all or some of the following: polynucleotides, computer software, nucleotides, primers, vectors, cell lines, instructions, plates, media, buffers, enzymes, Eppendorf tubes, site-directed mutagenesis kits, *etc.* Preferably, the kit is conveniently packaged for use in a laboratory setting. The researcher typically provides the DNA coding sequence of the protein to be modified using the inventive technique.

Definitions

[0012] "Amino acid": The term "amino acid" refers to the basic structural subunits of proteins. An alpha-amino acid consists of an amino group, a carboxyl group, a hydrogen

atom, and a side chain (*i.e.*, R group) all bonded to a central carbon atom. This central carbon atom is referred to as the alpha carbon because it is adjacent to the carboxyl group. There are twenty natural amino acids including glycine, alanine, valine, leucine, isoleucine, phenylalanine, tyrosine, trypotphan, cysteine, methionine, serine, threonine, lysine, arginine, histidine, aspartate, glutamate, asparagine, glutamate, and proline. Hydrophobic amino acids include alanine, valine, leucine, isoleucine, and phenylalanine. Aromatic amino acids includes phenylalanine, tyrosine, tryptophan, and histine. Polar amino acids include tyrosine, cysteine, serine, threonine, lysine, arginine, histidine, aspartate, glutamate, asparagine, and glutamine. Sulfur-containing amino acids include cysteine and methionine. Basic amino acids include lysine, arginine, and histidine. Acidic amino acids include aspartate and glutamate. Unnatural amino acids have also been inserted into proteins. In certain embodiments, the twenty natural amino acids are referred to when the term "amino acid" is used.

[0013] "Antibody": The term "antibody" refers to an immunoglobulin, whether natural or wholly or partially synthetically produced. All derivatives thereof which maintain specific binding ability are also included in the term. The term also covers any protein having a binding domain which is homologous or largely homologous to an immunoglobulin binding domain. These proteins may be derived from natural sources, or partly or wholly synthetically produced. An antibody may be monoclonal or polyclonal. The antibody may be a member of any immunoglobulin class, including any of the human classes: IgG, IgM, IgA, IgD, and IgE.

[0014] "Conserved": The term "conserved" refers nucleotides or amino acid residues of a polynucleotide sequence or amino acid sequence, respectively, that are those that occur unaltered in the same position of two or more related sequences being compared. Nucleotides or amino acids that are relatively conserved are those that are conserved amongst more related sequences than nucleotides or amino acids appearing elsewhere in the sequences.

[0015] "Homologous": The term "homologous", as used herein is an art-understood term that refers to nucleic acids or proteins that are highly related at the level of nucleotide or amino acid sequence. Nucleic acids or proteins that are homologous to each other are termed homologues. Homologous may refer to the degree of sequence similarity between two sequences (*i.e.*, nucleotide sequence or amino acid). The homology percentage figures

referred to herein reflect the maximal homology possible between two sequences, *i.e.*, the percent homology when the two sequences are so aligned as to have the greatest number of matched (homologous) positions. Homology can be readily calculated by known methods such as those described in: Computational Molecular Biology, Lesk, A. M., ed., Oxford University Press, New York, 1988; Biocomputing: Informatics and Genome Projects, Smith, D. W., ed., Academic Press, New York, 1993; Sequence Analysis in Molecular Biology, von Heinje, G., Academic Press, 1987; Computer Analysis of Sequence Data, Part I, Griffin, A. M., and Griffin, H. G., eds., Humana Press, New Jersey, 1994; and Sequence Analysis Primer, Gribskov, M. and Devereux, J., eds., M Stockton Press, New York, 1991; each of which is incorporated herein by reference. Methods commonly employed to determinehomology between sequences include, but are not limited to those disclosed in Carillo, H., and Lipman, D., SIAM J Applied Math., 48:1073 (1988); incorporated herein by reference. Techniques for determining homology are codified in publicly available computer programs. Exemplary computer software to determine homology between two sequences include, but are not limited to, GCG program package, Devereux, J., et al., Nucleic Acids Research, 12(1), 387 (1984)), BLASTP, BLASTN, and FASTA Atschul, S. F. et al., J Molec. Biol., 215, 403 (1990)).

[0016] The term "homologous" necessarily refers to a comparison between at least two sequences (nucleotides sequences or amino acid sequences). In accordance with the invention, two nucleotide sequences are considered to be homologous if the polypeptides they encode are at least about 50-60% identical, preferably about 70% identical, for at least one stretch of at least 20 amino acids. Preferably, homologous nucleotide sequences are also characterized by the ability to encode a stretch of at least 4-5 uniquely specified amino acids. Both the identity and the approximate spacing of these amino acids relative to one another must be considered for nucleotide sequences to be considered homologous. For nucleotide sequences less than 60 nucleotides in length, homology is determined by the ability to encode a stretch of at least 4-5 uniquely specified amino acids.

[0017] "Peptide" or "protein": According to the present invention, a "peptide" or "protein" comprises a string of at least three amino acids linked together by peptide bonds. The terms "protein" and "peptide" may be used interchangeably. Inventive peptides preferably contain only natural amino acids, although non-natural amino acids (*i.e.*, compounds that do not occur in nature but that can be incorporated into a polypeptide chain)

and/or amino acid analogs as are known in the art may alternatively be employed. Also, one or more of the amino acids in an inventive peptide may be modified, for example, by the addition of a chemical entity such as a carbohydrate group, a phosphate group, a farnesyl group, an isofarnesyl group, a fatty acid group, a linker for conjugation, functionalization, or other modification (*e.g.*, alpha amindation), *etc.* In a preferred embodiment, the modifications of the peptide lead to a more stable peptide (*e.g.*, greater half-life *in vivo*). These modifications may include cyclization of the peptide, the incorporation of D-amino acids, *etc.* None of the modifications should substantially interfere with the desired biological activity of the peptide. In certain embodiments, the modifications of the peptide lead to a more biologically active peptide.

[0018] "Polynucleotide" or "oligonucleotide": Polynucleotide or oligonucleotide refers to a polymer of nucleotides. Typically, a polynucleotide comprises at least three nucleotides. The polymer may include natural nucleosides (*i.e.*, adenosine, thymidine, guanosine, cytidine, uridine, deoxyadenosine, deoxythymidine, deoxyguanosine, and deoxycytidine), nucleoside analogs (*e.g.*, 2-aminoadenosine, 2-thiothymidine, inosine, pyrrolo-pyrimidine, 3-methyl adenosine, C5-propynylcytidine, C5-propynyluridine, C5-bromouridine, C5-fluorouridine, C5-iodouridine, C5-methylcytidine, 7-deazaadenosine, 7-deazaguanosine, 8-oxoadenosine, 8-oxoguanosine, O(6)-methylguanine, and 2-thiocytidine), chemically modified bases, biologically modified bases (*e.g.*, methylated bases), intercalated bases, modified sugars (*e.g.*, 2′-fluororibose, ribose, 2′-deoxyribose, arabinose, and hexose), and/or modified phosphate groups (*e.g.*, phosphorothioates and 5′-N-phosphoramidite linkages).

[0019] "Small molecule": The term "small molecule," as used herein, refers to a non-peptidic, non-oligomeric organic compound either prepared in the laboratory or found in nature. Small molecules, as used herein, can refer to compounds that are "natural product-like," however, the term "small molecule" is not limited to "natural product-like" compounds. Rather, a small molecule is typically characterized in that it contains several carbon-carbon bonds, and has a molecular weight of less than 1500, although this characterization is not intended to be limiting for the purposes of the present invention. In certain other preferred embodiments, natural-product-like small molecules are utilized.

[0020] "Stable": The term "stable" as used herein to refer to a protein refers to any aspect of protein stability. The stable modified protein as compared to the original wild type protein possesses any one or more of the following characteristics: more soluble, more resistant to

aggregation, more resistant to denaturation, more resistant to unfolding, more resistant to improper or undesired folding, greater ability to renature, increased thermal stability, increased stability in a variety of environments (*e.g.*, pH, salt concentration, presence of detergents, presence of denaturing agents, *etc.*), and increased stability in non-aqueous environments. In certain embodiments, the stable modified protein exhibits at least two of the above characteristics. In certain embodiments, the stable modified protein exhibits at least three of the above characteristics. Such characteristics may allow the active protein to be produced at higher levels. For example, the modified protein can be overexpressed at a higher level without aggregation than the unmodified version of the protein. Such characteristics may also allow the protein to be used as a therapeutic agent or a research tool.

Brief Description of the Drawing

[0001] *Figure 1.* Supercharged green fluorescent proteins (GFPs). (a) Protein sequences of GFP variants, with fluorophore-forming residues highlighted green, negatively charged residues highlighted red, and positively charged residues highlighted blue. (b) Electrostatic surface potentials of sfGFP (left), GFP(+36) (middle), and GFP(–30) (right), colored from –25 kT/e (red) to +25 kT/e (blue).

[0002] *Figure 2.* Intramolecular properties of GFP variants. (a) Staining and UV fluorescence of purified GFP variants. Each lane and tube contains 0.2 μg of protein. (b) Circular dichroism spectra of GFP variants. (c) Thermodynamic stability of GFP variants, measured by guanidinium-induced unfolding.

[0003] *Figure 3.* Intermolecular properties of supercharged proteins. (a) UV-illuminated samples of purified GFP variants ("native"), those samples heated 1 min at 100 °C ("boiled"), and those samples subsequently cooled for 2 h at 25°C ("cooled"). (b) Aggregation of GFP variants was induced with 40% TFE at 25 °C and monitored by right-angle light scattering. (c) Supercharged GFPs adhere reversibly to oppositely charged macromolecules. Sample 1: 6 μg of GFP(+36) in 30 μl of 25 mM Tris pH 7.0 and 100 mM NaCl. Sample 2: 6 μg of GFP(–30) added to sample 1. Sample 3: 30 μg of salmon sperm DNA added to sample 1. Sample 4: 20 μg of *E. coli* tRNA added to sample 1. Sample 5: Addition of NaCl to 1 M to sample 4. Samples 6-8: identical to samples 1, 2, and 4, respectively, except using sfGFP instead of GFP(+36). All samples were spun briefly in a microcentrifuge and visualized under UV light. (d) Enzymatic assays of GST variants.

Reactions contained 0.5 mg/mL of GST variant, 20 mM chlorodinitrobenzene, 20 mM glutathione, and 100 mM potassium phosphate pH 6.5. Product formation was monitored at 340 nm, resulting in observed reaction rates (k_{obs}) of 6 min^{-1} for wild-type GST, 2.2 min^{-1} for GST(–40), and 0.9 min^{-1} for GST(–40) after being boiled and cooled.

[0004] *Figure 4.* (a) Excitation and (b) emission spectra of GFP variants. Each sample contained an equal amount of protein as quantitated by chromophore absorbance at 490 nm.

[0005] *Figure 5.* Biotin-binding activity of streptavidin variants, measured as described previously (Kada *et al.*, Rapid estimation of avidin and streptavidin by fluorescence quenching or fluorescence polarization. *Biochim. Biophys. Acta* **1427**, 44-48 (1999); incorporated herein by reference) by monitoring binding-dependent of biotin-4-fluorescein (Invitrogen). Protein samples were titrated into 0.3 μM biotin-4-fluorescein (B4F), 100 mM NaCl, 1 mM EDTA, 0.1 mg/mL bovine serum albumin (BSA), 50 mM potassium phosphate pH 7.5. Quenching of fluorescence at 526 nm was measured on a Perkin-Elmer LS50B luminescence spectrometer with excitation at 470 nm. Measurements were normalized to control titrations that contained a 600-fold excess of non-fluorescent biotin. The three proteins in the bottom of the legend are included as negative controls.

Detailed Description of Certain Preferred Embodiments of the Invention

[0021] The invention provides a system for modifying proteins to be more stable. The system is thought to work by changing non-conserved amino acids on the surface of a protein to more polar or charged amino acid residues. The amino acids residues to be modified may be hydrophobic, hydrophilic, charged, or a combination thereof. Any protein may be modified using the inventive system to produce a more stable variant. These modifications of surface residues have been found to improve the extrathermodynamic properties of proteins. As proteins are increasingly used as therapeutic agents and as they continue to be used as research tools, a system for altering a protein to make it more stable is important and useful. Proteins modified by the inventive method typically are resistant to aggregation, have an increased ability to refold, resist improper folding, have improved solubility, and are generally more stable under a wide range of conditions including denaturing conditions such as heat or the presence of a detergent.

[0022] Any protein may be modified to create a more stable variant using the inventive system. Natural as well as unnatural proteins (*e.g.*, engineered proteins) may be modified. Example of proteins that may be modified include receptors, membrane bound proteins, transmembrane proteins, enzymes, transcription factors, extracellular proteins, therapeutic proteins, cytokines, messenger proteins, DNA-binding proteins, RNA-binding proteins, proteins involved in signal transduction, structural proteins, cytoplasmic proteins, nuclear proteins, hydrophobic proteins, hydrophilic proteins, *etc.* The protein to be modified may be derived from any species of plant, animal, or microorganism. In certain embodiments, the protein is a mammalian protein. In certain embodiments, the protein is a human protein. In certain embodiments, the proteins is derived from an organism typically used in research. For example, the protein to be modified may be from a primate (*e.g.*, ape, monkey), rodent (*e.g.*, rabbit, hamster, gerbil), pig, dog, cat, fish (*e.g., zebrafish)*, nematode (*e.g.*, *C. elegans*), yeast (*e.g.*, *Saccharomyces cervisiae*), or bacteria (*e.g.*, *E. coli*).

[0023] The inventive system is particularly useful in modifying proteins that are susceptible to aggregation or have stability issues. The system may also be used to modify proteins that are being overexpressed. For example, therapeutic proteins that are being produced recombinantly may benefit from being modified by the inventive system. Such modified therapeutic proteins are not only easier to produce and purify but also may be more stable with respect to storage and use of the protein.

[0024] The inventive system involves identifying non-conserved surface residues of a protein of interest and replacing some of those residues with a residue that is hydrophilic, polar, or charged at physiological pH. The inventive system includes not only methods for modifying a protein but also reagents and kits that are useful in modifying a protein to make it more stable.

[0025] The surface residues of the protein to be modified are identified using any method(s) known in the art. In certain embodiments, the surface residues are identified by computer modeling of the protein. In certain embodiments, the three-dimensional structure of the protein is known and/or determined, and the surface residues are identified by visualizing the structure of the protein. In other embodiments, the surface residues are predicted using computer software. In certain particular embodiments, Average Neighbor Atoms per Sidechain Atom (AvNAPSA) is used to predict surface exposure. AvNAPSA is an automated measure of surface exposure which has been implemented as a computer

program. See Appendix A. A low AvNAPSA value indicates a surface exposed residue, whereas a high value indicates a residue in the interior of the protein. In certain embodiments, the software is used to predict the secondary structure and/or tertiary structure of a protein and the surface residues are identified based on this prediction. In other embodiments, the prediction of surface residues is based on hydrophobicity and hydrophilicity of the residues and their clustering in the primary sequence of the protein. Besides *in silico* methods, the surface residues of the protein may also be identified using various biochemical techniques, for example, protease cleavage, surface modification, *etc.*

[0026] Of the surface residues, it is then determined which are conserved or important to the functioning of the protein. The identification of conserved residues can be determined using any method known in the art. In certain embodiments, the conserved residues are identified by aligning the primary sequence of the protein of interest with related proteins. These related proteins may be from the same family of proteins. For example, if the protein is an immunoglobulin, other immunoglobulin sequences may be used. The related proteins may also be the same protein from a different species. For example, the conserved residues may be identified by aligning the sequences of the same protein from different species. To give but another example, proteins of similar function or biological activity may be aligned. Preferably, 2, 3, 4, 5, 6, 7, 8, 9, or 10 different sequences are used to determine the conserved amino acids in the protein. In certain embodiments, the residue is considered conserved if over 50%, 60%, 70%, 75%, 80%, or 90% of the sequences have the same amino acid in a particular position. In other embodiments, the residue is considered conserved if over 50%, 60%, 70%, 75%, 80%, or 90% of the sequences have the same or a similar (*e.g.*, valine, leucine, and isoleucine; glycine and alanine; glutamine and asparagine; or aspartate and glutamate) amino acid in a particular position. Many software packages are available for aligning and comparing protein sequences as described herein. As would be appreciated by one of skill in the art, either the conserved residues may be determined first or the surface residues may be determined first. The order does not matter. In certain embodiments, a computer software package may determine surface residues and conserved residues simultaneously. Important residues in the protein may also be identified by mutagenesis of the protein. For example, alanine scanning of the protein can be used to determine the important amino acid residues in the protein. In other embodiments, site-directed mutagenesis may be used.

[0027] Once non-conserved surface residues of the protein have been identified, each of the residues is identified as hydrophobic or hydrophilic. In certain embodiments, the residues is assigned a hydrophobicity score. For example, each non-conserved surface residue may be assigned an octanol/water logP value. Other hydrophobicity parameters may also be used. Such scales for amino acids have been discussed in: Janin, "Surface and Inside Volumes in Globular Proteins," *Nature* 277:491-92, 1979; Wolfenden *et al.*, "Affinities of Amino Acid Side Chains for Solvent Water," *Biochemistry* 20:849-855, 1981; Kyte *et al.*, "A Simple Method for Displaying the Hydropathic Character of a Protein," *J. Mol. Biol.* 157:105-132, 1982; Rose *et al.*, "Hydrophobicity of Amino Acid Residues in Globular Proteins," *Science* 229:834-838, 1985; Cornette *et al.*, "Hydrophobicity Scales and Computational Techniques for Detecting Amphipathic Structures in Proteins," *J. Mol. Biol.* 195:659-685, 1987; Charton and Charton, "The Structure Dependence of Amino Acid Hydrophobicity Parameters," *J. Theor. Biol.* 99:629-644, 1982; each of which is incorporated by reference. Any of these hydrophobicity parameters may be used in the inventive method to determine which non-conserved residues to modify. In certain embodiments, hydrophilic or charged residues are identified for modification.

[0028] At least one identified non-conserved or non-vital surface residue is then chosen for modification. In certain embodiments, hydrophobic residue(s) are chosen for modification. In other embodiments, hydrophilic and/or charged residue(s) are chosen for modification. In certain embodiments, more than one residue is chosen for modification. In certain embodiments, 1, 2, 3, 4, 5, 6, 7, 8, 9, or 10 of the identified residues are chosen for modification. In certain embodiments, over 10, over 15, or over 20 residues are chosen for modification. As would be appreciated by one of skill in the art, the larger the protein the more residues that will need to be modified. Also, the more hydrophobic or susceptible to aggregation or precipitation the protein is, the more residues will need to be modified. In certain embodiments, multiple variants of the protein, each with different modifications, are produced and tested to determine the best variant in terms of biological activity and stability.

[0029] In certain embodiments, the residues chosen for modification are mutated into more hydrophilic residues (including charged residues). Typically, the residues are mutated into more hydrophilic natural amino acids. In certain embodiments, the residues are mutated into amino acids that are charged at physiological pH. For example, the residue may be changed to an arginine, aspartate, glutamate, histidine, or lysine. In certain embodiments, all

the residues to be modified are changed into the same different residue. For example, all the chosen residues are changed to a glutamate residue. In other embodiments, the chosen residues are changed into different residues; however, all the final residues may be either positively charged or negatively charged at physiological pH. In certain embodiments, to create a negatively charged protein, all the residues to be mutated are converted to glutamate and/or aspartate residues. In certain embodiments, to create a positively charged protein, all the residues to be mutated are converted to lysine residues. For example, all the chosen residues for modification are asparagine, glutamine, lysine, and/or arginine, and these residues are mutated into aspartate or glutamate residues. To give but another example, all the chosen residues for modification are aspartate, glutamate, asparagine, and/or glutamine, and these residues are mutated into lysine. This approach allows for modifying the net charge on the protein to the greatest extent.

[0030] In other embodiments, the protein may be modified to keep the net charge on the modified protein the same as on the unmodified protein. In still other embodiments, the protein may be modified to decrease the overall net charge on the protein while increasing the total number of charged residues on the surface. In certain embodiments, the theoretical net charge is increased by at least +1, +2, +3, +4, +5, +10, +15, +20, +25, +30, or +35. In certain embodiments, the theoretical net charge is decreased by at least -1, -2, -3, -4, -5, -10, -15, -20, -25, -30, or -35. In certain embodiments, the chosen amino acids are changed into non-ionic, polar residues (*e.g.*, cysteine, serine, threonine, tyrosine, glutamine, asparagine).

[0031] These modification or mutations in the protein may be accomplished using any technique known in the art. Recombinant DNA techniques for introducing such changes in a protein sequence are well known in the art. In certain embodiments, the modifications are made by site-directed mutagenesis of the polynucleotide encoding the protein. Other techniques for introducing mutations are discussed in *Molecular Cloning: A Laboratory Manual*, 2nd Ed., ed. by Sambrook, Fritsch, and Maniatis (Cold Spring Harbor Laboratory Press: 1989); the treatise, *Methods in Enzymology* (Academic Press, Inc., N.Y.); Ausubel *et al. Current Protocols in Molecular Biology* (John Wiley & Sons, Inc., New York, 1999); each of which is incorporated herein by reference. The modified protein is expressed and tested. In certain embodiments, a series of variants is prepared and each variant is tested to determine its biological activity and its stability. The variant chosen for subsequent use may be the most stable one, the most active one, or the one with the greatest overall combination

of activity and stability. After a first set of variants is prepared an additional set of variants may be prepared based on what is learned from the first set. The variants are typically created and overexpressed using recombinant techniques known in the art.

[0032] The inventive system has been used to created variants of GFP. These variants have been shown to be more stable and to retain their fluorescence. A GFP from *Aequorea victoria* is described in GenBank Accession Number P42212, incorporated herein by reference. The amino acid sequence of this wild type GFP is as follows:

MSKGEELFTGVVPILVELDGDVNGHKFSVSGEGEGDATYGKLTLKFICTTGKLPVPW
PTLVTTFSYGVQCFSRYPDHMKQHDFFKSAMPEGYVQERTIFFKDDGNYKTRAEVK
FEGDTLVNRIELKGIDFKEDGNILGHKLEYNYNSHNVYIMADKQKNGIKVNFKIRHNI
EDGSVQLADHYQQNTPIGDGPVLLPDNHYLSTQSALSKDPNEKRDHMVLLEFVTAA
GITHGMDELYK (SEQ ID NO: 1)

Wild type GFP has a theoretical net charge of -7. Using the inventive system, variants with a theoretical net charge of -29, -30, -25, +36, +48, and +49 have been created. Even after heating the +36 GFP to 95 °C, 100% of the variant protein is soluble and the protein retains ≥70% of its fluorescence.

[0033] The amino acid sequences of the variants of GFP that have been created include:

GFP-NEG25

MGHHHHHHGGASKGEELFTGVVPILVELDGDVNGHEFSVRGEGEGDATEGELTLKF
ICTTGELPVPWPTLVTTLTYGVQCFSRYPDHMKQHDFFKSAMPEGYVQERTISFKDD
GTYKTRAEVKFEGDTLVNRIELKGIDFKEDGNILGHKLEYNFNSHDVYITADKQENGI
KAEFEIRHNVEDGSVQLADHYQQNTPIGDGPVLLPDDHYLSTESALSKDPNEDRDHM
VLLEFVTAAGIDHGMDELYK (SEQ ID NO: 2)

GFP-NEG29

MGHHHHHHGGASKGEELFDGEVPILVELDGDVNGHEFSVRGEGEGDATEGELTLKF
ICTTGELPVPWPTLVTTLTYGVQCFSRYPDHMDQHDFFKSAMPEGYVQERTISFKDD
GTYKTRAEVKFEGDTLVNRIELKGIDFKEDGNILGHKLEYNFNSHDVYITADKQENGI
KAEFEIRHNVEDGSVQLADHYQQNTPIGDGPVLLPDDHYLSTESALSKDPNEDRDHM
VLLEFVTAAGIDHGMDELYK (SEQ ID NO: 3)

GFP-NEG30

MGHHHHHHGGASKGEELFDGVVPILVELDGDVNGHEFSVRGEGEGDATEGELTLKF
ICTTGELPVPWPTLVTTLTYGVQCFSDYPDHMDQHDFFKSAMPEGYVQERTISFKDD
GTYKTRAEVKFEGDTLVNRIELKGIDFKEDGNILGHKLEYNFNSHDVYITADKQENGI

KAEFEIRHNVEDGSVQLADHYQQNTPIGDGPVLLPDDHYLSTESALSKDPNEDRDHM
VLLEFVTAAGIDHGMDELYK (SEQ ID NO: 4)

GFP-POS36)

MGHHHHHHGGASKGERLFRGKVPILVELKGDVNGHKFSVRGKGKGDATRGKLTLK
FICTTGKLPVPWPTLVTTLTYGVQCFSRYPKHMKRHDFFKSAMPKGYVQERTISFKK
DGKYKTRAEVKFEGRTLVNRIKLKGRDFKEKGNILGHKLRYNFNSHKVYITADKRK
NGIKAKFKIRHNVKDGSVQLADHYQQNTPIGRGPVLLPRNHYLSTRSKLSKDPKEKR
DHMVLLEFVTAAGIKHGRDERYK (SEQ ID NO: 5)

GFP-POS42

MGHHHHHHGGRSKGKRLFRGKVPILVELKGDVNGHKFSVRGKGKGDATRGKLTLK
FICTTGKLPVPWPTLVTTLTYGVQCFSRYPKHMKRHDFFKSAMPKGYVQERTISFKK
DGKYKTRAEVKFEGRTLVNRIKLKGRDFKEKGNILGHKLRYNFNSHKVYITADKRK
NGIKAKFKIRHNVKDGSVQLADHYQQNTPIGRGPVLLPRKHYLSTRSKLSKDPKEKR
DHMVLLEFVTAAGIKHGRKERYK (SEQ ID NO: 6)

GFP-POS49

MGHHHHHHGGRSKGKRLFRGKVPILVKLKGDVNGHKFSVRGKGKGDATRGKLTLK
FICTTGKLPVPWPTLVTTLTYGVQCFSRYPKHMKRHDFFKSAMPKGYVQERTISFKK
DGKYKTRAEVKFKGRTLVNRIKLKGRDFKEKGNILGHKLRYNFNSHKVYITADKRK
NGIKAKFKIRHNVKDGSVQLAKHYQQNTPIGRGPVLLPRKHYLSTRSKLSKDPKEKR
DHMVLKEFVTAAGIKHGRKERYK (SEQ ID NO: 7)

As would be appreciated by one of skill in the art, homologous proteins are also considered to be within the scope of this invention. For example, any protein that includes a stretch of 20, 30, 40, 50, or 100 amino acids which are 60%, 70%, 80%, 90%, 95%, or 100% homologous to any of the above sequences is considered part of the invention. In addition, addition and deletion variants are also contemplated by the invention. In certain embodiments, any GFP with a mutated residue as shown in any of the above sequences is considered part of the invention. In certain embodiments, the sequence includes 2, 3, 4, 5, 6, 7, 8, 9, 10, or more mutations as shown in any of the sequences above.

[0034] Any DNA sequence that encodes the above GFP variants is also include within the scope of the invention. Exemplary DNA sequences which encode each of the variants above are as follows:

GFP-NEG25

ATGGGGCATCACCATCATCATCATGGCGGTGCGTCTAAGGGGGAGGAGTTATTTA

CGGGTGTGGTGCCGATCCTGGTGGAGCTTGATGGCGATGTTAACGGCCATGAATT
TTCTGTCCGCGGTGAAGGGGAGGGTGATGCCACGGAAGGGGAGCTGACACTTAA
ATTTATTTGCACCACCGGTGAACTCCCGGTCCCGTGGCCGACCCTGGTGACCACC
CTGACCTACGGCGTTCAATGCTTTTCACGTTATCCGGATCACATGAAGCAACACG
ACTTCTTTAAAAGCGCGATGCCTGAAGGCTATGTTCAAGAACGTACAATTAGTTT
TAAAGATGACGGCACCTACAAGACCCGTGCGGAAGTAAAATTTGAAGGGGACAC
TTTAGTGAACCGCATCGAGCTGAAAGGGATCGATTTTAAAGAAGATGGGAATAT
CCTGGGACACAAACTTGAATACAACTTTAATAGTCATGACGTCTATATCACGGCG
GACAAACAGGAAAACGGAATTAAGGCAGAATTTGAGATTCGGCATAATGTCGAA
GATGGCTCGGTACAGTTGGCTGATCACTATCAGCAGAATACGCCGATTGGAGAT
GGTCCGGTTTTATTACCAGACGATCACTATCTGTCCACCGAATCCGCCCTGAGCA
AAGATCCGAATGAAGACCGGGACCATATGGTTCTGCTGGAATTTGTTACGGCGG
CTGGTATTGACCATGGCATGGATGAGCTGTATAAGTAG (SEQ ID NO: 8)

GFP-NEG29

ATGGGGCATCACCATCATCATCATGGCGGTGCGTCTAAGGGGGAGGAGTTATTTG
ATGGTGAAGTGCCGATCCTGGTGGAGCTTGATGGCGATGTTAACGGCCATGAATT
TTCTGTCCGCGGTGAAGGGGAGGGTGATGCCACGGAAGGGGAGCTGACACTTAA
ATTTATTTGCACCACCGGTGAACTCCCGGTCCCGTGGCCGACCCTGGTGACCACC
CTGACCTACGGCGTTCAATGCTTTTCACGTTATCCGGATCACATGGACCAACACG
ACTTCTTTAAAAGCGCGATGCCTGAAGGCTATGTTCAAGAACGTACAATTAGTTT
TAAAGATGACGGCACCTACAAGACCCGTGCGGAAGTAAAATTTGAAGGGGACAC
TTTAGTGAACCGCATCGAGCTGAAAGGGATCGATTTTAAAGAAGATGGGAATAT
CCTGGGACACAAACTTGAATACAACTTTAATAGTCATGACGTCTATATCACGGCG
GACAAACAGGAAAACGGAATTAAGGCAGAATTTGAGATTCGGCATAATGTCGAA
GATGGCTCGGTACAGTTGGCTGATCACTATCAGCAGAATACGCCGATTGGAGAT
GGTCCGGTTTTATTACCAGACGATCACTATCTGTCCACCGAATCCGCCCTGAGCA
AAGATCCGAATGAAGACCGGGACCATATGGTTCTGCTGGAATTTGTTACGGCGG
CTGGTATTGACCATGGCATGGATGAGCTGTATAAGTAG (SEQ ID NO: 9)

GFP-NEG30

ATGGGGCATCACCATCATCATCATGGCGGTGCGTCTAAGGGGGAGGAGTTATTTG
ATGGTGTGGTGCCGATCCTGGTGGAGCTTGATGGCGATGTTAACGGCCATGAATT
TTCTGTCCGCGGTGAAGGGGAGGGTGATGCCACGGAAGGGGAGCTGACACTTAA
ATTTATTTGCACCACCGGTGAACTCCCGGTCCCGTGGCCGACCCTGGTGACCACC
CTGACCTACGGCGTTCAATGCTTTTCAGATTATCCGGATCACATGGACCAACACG
ACTTCTTTAAAAGCGCGATGCCTGAAGGCTATGTTCAAGAACGTACAATTAGTTT
TAAAGATGACGGCACCTACAAGACCCGTGCGGAAGTAAAATTTGAAGGGGACAC
TTTAGTGAACCGCATCGAGCTGAAAGGGATCGATTTTAAAGAAGATGGGAATAT
CCTGGGACACAAACTTGAATACAACTTTAATAGTCATGACGTCTATATCACGGCG
GACAAACAGGAAAACGGAATTAAGGCAGAATTTGAGATTCGGCATAATGTCGAA
GATGGCTCGGTACAGTTGGCTGATCACTATCAGCAGAATACGCCGATTGGAGAT
GGTCCGGTTTTATTACCAGACGATCACTATCTGTCCACCGAATCCGCCCTGAGCA
AAGATCCGAATGAAGACCGGGACCATATGGTTCTGCTGGAATTTGTTACGGCGG
CTGGTATTGACCATGGCATGGATGAGCTGTATAAGTAG (SEQ ID NO: 10)

GFP-POS36

ATGGGGCATCATCATCATCACCACGGCGGGGCGTCTAAGGGAGAGCGCTTGTTTC
GCGGCAAAGTCCCGATTCTTGTGGAGCTCAAAGGTGATGTAAATGGTCATAAATT
TAGTGTGCGCGGGAAAGGGAAAGGAGATGCTACGCGGGGCAAGCTCACCCTGAA
ATTTATTTGCACAACCGGCAAACTGCCAGTGCCGTGGCCTACATTAGTCACTACT
CTGACGTACGGTGTTCAGTGCTTTTCTCGCTATCCCAAACACATGAAACGCCATG
ATTTCTTCAAGAGCGCGATGCCAAAAGGTTATGTGCAGGAACGCACCATCAGCTT
TAAAAAAGACGGCAAATATAAAACCCGTGCAGAAGTTAAATTCGAAGGCCGCAC
CCTGGTCAACCGCATTAAACTGAAAGGTCGTGACTTCAAAGAGAAAGGTAATAT
TCTTGGTCACAAACTGCGCTATAATTTCAACTCTCACAAAGTTTATATTACGGCG
GATAAACGTAAAAACGGGATTAAAGCGAAATTTAAGATTCGTCATAATGTTAAA
GACGGCAGTGTGCAGTTAGCGGATCATTATCAGCAGAATACCCCAATTGGTCGC
GGTCCAGTGCTGCTGCCGCGTAACCATTATCTGTCGACCCGCAGCAAACTCAGCA
AAGACCCGAAAGAAAAACGTGACCACATGGTATTACTGGAATTTGTGACCGCAG
CAGGCATTAAACATGGCCGCGATGAACGTTACAAATAG (SEQ ID NO: 11)

GFP-POS44

ATGGGCCATCATCATCACCACCACGGCGGCCGCTCAAAAGGTAAACGCTTGTTCC
GTGGTAAAGTACCGATCTTAGTGGAGCTCAAAGGGGATGTGAATGGCCATAAGT
TCTCGGTTCGTGGCAAAGGTAAGGGAGATGCGACGCGCGGCAAATTAACGCTGA
AATTCATTTGTACTACAGGTAAACTGCCGGTGCCATGGCCTACTCTCGTCACCAC
GTTGACCTATGGGGTTCAATGCTTCAGCCGGTACCCTAAACACATGAAGCGCCAC
GATTTCTTCAAATCGGCGATGCCAAAGGGGTATGTCCAGGAACGCACTATCAGCT
TCAAAAAAGACGGTAAGTATAAAACTCGTGCTGAAGTTAAATTCGAAGGACGCA
CACTGGTAAATCGCATTAAATTGAAGGGGCGCGACTTTAAGGAAAAAGGTAATA
TCTTAGGTCACAAATTGCGCTACAACTTCAACTCTCATAAAGTTTACATTACAGC
AGATAAGCGTAAAAATGGCATCAAAGCGAAATTCAAAATTCGTCACAATGTGAA
AGATGGTAGCGTGCAATTAGCCGATCATTACCAGCAGAATACGCCGATCGGTCG
CGGCCCAGTACTGTTGCCGCGCAAACATTACTTATCTACCCGGAGTAAACTGTCT
AAAGACCCAAAAGAGAAGCGCGACCATATGGTTCTCCTGGAGTTTGTCACCGCC
GCCGGAATTAAACACGGCCGCAAAGAGCGCTATAAATAG (SEQ ID NO: 12)

GFP-POS49

ATGGGCCACCATCATCATCACCACGGGGGACGCTCTAAAGGTAAACGTCTGTTTC
GTGGAAAGGTGCCCATTCTGGTTAAACTCAAAGGTGATGTCAACGGCCATAAGTT
TTCGGTTCGTGGCAAAGGTAAAGGTGATGCGACGCGCGGGAAATTAACACTGAA
ATTTATTTGCACAACCGGAAAACTCCCTGTGCCGTGGCCGACTTTGGTGACCACA
TTAACCTATGGTGTTCAATGCTTCTCACGTTATCCGAAGCATATGAAACGTCATG
ATTTTTTCAAATCGGCTATGCCGAAAGGTTACGTCCAGGAGCGCACCATCTCATT
TAAGAAAGACGGTAAGTATAAAACCCGTGCTGAAGTAAAATTCAAAGGACGCAC
CCTGGTGAATCGCATTAAACTGAAAGGTCGTGATTTCAAAGAAAAGGGAAATAT
TTTAGGGCATAAGCTCCGTTATAATTTTAACAGTCATAAGGTGTATATTACCGCT
GATAAACGCAAAAACGGAATCAAAGCGAAATTTAAGATCCGTCATAATGTAAAA
GATGGCTCAGTCCAACTGGCAAAACATTACCAGCAGAATACCCCGATCGGCCGC
GGTCCTGTGCTTCTGCCGCGTAAACACTACTTGTCGACCCGGTCAAAATTGAGTA

AAGATCCGAAGGAAAAGCGTGATCACATGGTCTTGAAGGAATTTGTAACTGCAG
CAGGTATTAAACACGGGCGCAAAGAACGTTACAAATAG (SEQ ID NO: 13)

[0035] Polynucleotide sequence homologous to the above sequences are also within the scope of the present invention. In certain embodiments, the polynucleotide sequence include a stretch of 50, 100, or 150 nucleotides that are 60%, 70%, 80%, 90%, 95%, 98%, 99%, or 100% homologous to any one of the above sequence. The present invention also includes sequence where one or more nucleotides is inserted or deleted from one of the above sequences. Any polynucleotide sequence with a mutation as shown in any of the sequences above is considered part of the invention. In certain embodiments, the sequence includes 2, 3, 4, 5, 6, 7, 8, 9, 10, or more mutations as shown in any of the sequences above.

[0036] The present invention also provides vector (*e.g.*, plasmids, cosmids, viruses, *etc.*) that comprise any of the inventive sequences herein or any other sequence (DNA or protein) modified using the inventive system. In certain embodiments, the vector includes elements such as promoter, enhancer, ribosomal binding sites, *etc.* sequences useful in overexpressing the inventive GFP variant in a cell. The invention also includes cells comprising the inventive sequences or vectors. In certain embodiments, the cells overexpress the variant GFP. The cells may be bacterial cells (*e.g.*, *E. coli*), fungal cells (*e.g.*, *P. pastoris*), yeast cells (*e.g.*, *S. cerevisiae*), mammalian cells (*e.g.*, CHO cells), or human cells.

[0037] The inventive system has been used to created variants of streptavidin. These variants have been shown to form soluble tetramers that bind biotin. The amino acid sequence of this wild type streptavidin is as follows:

AAEAGITGTWYNQLGSTFIVTAGADGALTGTYESAVGNAESRYVLTGRYDSAPATD
GSGTALGWTVAWKNNYRNAHSATTWSGQYVGGAEARINTQWLLTSGTTEANAWK
STLVGHDTFTKVKPSAAS (SEQ ID NO: XX)

Wild type streptavidin has a theoretical net charge of -4. Using the inventive system, variants with a theoretical net charge of -40 and +52 have been created. Even after heating the variants to 100 °C, the proteins remained soluble.

[0038] The amino acid sequences of the variants of streptavidin that have been created include:

SAV-NEG40

MGHHHHHHGGAEAGITGTWYNQLGSTFIVTAGADGALTGTYESAVGDAESEYVLT
GRYDSAPATDGSGTALGWTVAWKNDYENAHSATTWSGQYVGGAEARINTQWLLT
SGTTEADAWKSTLVGHDTFTKVEPSAAS (SEQ ID NO: XX)

SAV-POS52

MGHHHHHHGGAKAGITGTWYNQLGSTFIVTAGAKGALTGTYESAVGNAKSRYVLT
GRYDSAPATKGSGTALGWTVAWKNKYRNAHSATTWSGQYVGGAKARINTQWLLT
SGTTKAKAWKSTLVGHDTFTKVKPSAAS (SEQ ID NO: XX)

As would be appreciated by one of skill in the art, homologous proteins are also considered to be within the scope of this invention. For example, any protein that includes a stretch of 20, 30, 40, 50, or 100 amino acids which are 60%, 70%, 80%, 90%, 95%, or 100% homologous to any of the above sequences is considered part of the invention. In addition, addition and deletion variants are also contemplated by the invention. In certain embodiments, any streptavidin with a mutated residue as shown in any of the above sequences is considered part of the invention. In certain embodiments, the sequence includes 2, 3, 4, 5, 6, 7, 8, 9, 10, or more mutations as shown in any of the sequences above.

[0039] Any DNA sequence that encodes the above streptavidin variants is also included within the scope of the invention. Exemplary DNA sequences which encode each of the variants above are as follows:

SAV-NEG40

GGTTCAGCCATGGGTCATCACCACCACCATCACGGTGGCGCCGAAGCAGGTATT
ACCGGTACCTGGTATAACCAGTTAGGCTCAACCTTTATTGTGACCGCGGGAGCGG
ACGGCGCCTTAACCGGTACCTACGAATCAGCTGTAGGTGACGCGGAATCAGAGT
ACGTATTAACCGGTCGTTATGATAGCGCGCCGGCGACTGACGGTAGCGGTACTGC
TTTAGGTTGGACCGTAGCGTGGAAGAATGATTATGAAAACGCACATAGCGCAAC
AACGTGGTCAGGGCAGTACGTTGGCGGAGCTGAGGCGCGCATTAACACGCAGTG
GTTATTAACTAGCGGCACCACTGAAGCTGATGCCTGGAAGAGCACGTTAGTGGG
TCATGATACCTTCACTAAAGTGGAACCTTCAGCTGCGTCATAATAATGACTCGAG
ACCTGCA (SEQ ID NO: XX)

SAV-POS52

GGTTCAGCCATGGGTCATCACCACCACCATCACGGTGGCGCCAAAGCAGGTATT
ACCGGTACCTGGTATAACCAGTTAGGCTCAACCTTTATTGTGACCGCGGGAGCGA
AAGGCGCCTTAACCGGTACCTACGAATCAGCTGTAGGAAACGCAAAATCACGCT
ACGTATTAACCGGTCGTTATGATAGCGCGCCGGCGACTAAAGGTAGCGGTACTG
CTTTAGGTTGGACCGTAGCGTGGAAGAATAAGTATCGTAATGCGCACAGTGCTAC
CACTTGGTCAGGGCAGTACGTAGGGGGAGCCAAAGCACGTATCAACACGCAGTG

GTTATTAACATCAGGTACCACCAAAGCGAAAGCCTGGAAGAGCACGTTAGTGGG
TCATGATACCTTCACTAAAGTGAAACCTTCAGCTGCGTCATAATAATGACTCGAG
ACCTGCA (SEQ ID NO: XX)

[0040] Polynucleotide sequence homologous to the above sequences are also within the scope of the present invention. In certain embodiments, the polynucleotide sequence include a stretch of 50, 100, or 150 nucleotides that are 60%, 70%, 80%, 90%, 95%, 98%, 99%, or 100% homologous to any one of the above sequence. The present invention also includes sequence where one or more nucleotides is inserted or deleted from one of the above sequences. Any polynucleotide sequence with a mutation as shown in any of the sequences above is considered part of the invention. In certain embodiments, the sequence includes 2, 3, 4, 5, 6, 7, 8, 9, 10, or more mutations as shown in any of the sequences above.

[0041] The present invention also provides vector (*e.g.*, plasmids, cosmids, viruses, *etc.*) that comprise any of the inventive sequences herein or any other sequence (DNA or protein) modified using the inventive system. In certain embodiments, the vector includes elements such as promoter, enhancer, ribosomal binding sites, *etc.* sequences useful in overexpressing the inventive streptavidin variant in a cell. The invention also includes cells comprising the inventive sequences or vectors. In certain embodiments, the cells overexpress the variant streptavidin. The cells may be bacterial cells (*e.g.*, *E. coli*), fungal cells (*e.g.*, *P. pastoris*), yeast cells (*e.g.*, *S. cerevisiae*), mammalian cells (*e.g.*, CHO cells), or human cells.

[0042] The inventive system has been used to created variants of glutathione-*S*-transferase (GST). These variants have been shown to retain the catalytic activity of wild type GST. The amino acid sequence of this wild type GST is as follows:

MGHHHHHHGGPPYTITYFPVRGRCEAMRMLLADQDQSWKEEVVTMETWPPLKPSC
LFRQLPKFQDGDLTLYQSNAILRHLGRSFGLYGKDQKEAALVDMVNDGVEDLRCKY
ATLIYTNYEAGKEKYVKELPEHLKPFETLLSQNQGGQAFVVGSQISFADYNLLDLLRI
HQVLNPSCLDAFPLLSAYVARLSARPKIKAFLASPEHVNRPINGNGKQ (SEQ ID NO:
XX)

Wild type GST has a theoretical net charge of +2. Using the inventive system, a variant with a theoretical net charge of -40 has been created. This variant catalyzes the addition of glutathione to chloronitrobenzene with a specific activity only 2.7-fold lower than that of wild type GST. Even after heating the variant to 100 °C, the protein remained soluble, and the protein recovered 40% of its catalytic activity upon cooling.

[0043] The amino acid sequences of variants of GST include:

GST-NEG40

MGHHHHHHGGPPYTITYFPVRGRCEAMRMLLADQDQSWEEEVVTMETWPPLKPSC
LFRQLPKFQDGDLTLYQSNAILRHLGRSFGLYGEDEEEAALVDMVNDGVEDLRCKY
ATLIYTDYEAGKEEYVEELPEHLKPFETLLSENEGGEAFVVGSEISFADYNLLDLLRIH
QVLNPSCLDAFPLLSAYVARLSARPEIEAFLASPEHVDRPINGNGKQ (SEQ ID NO:
XX)

GST-POS50

MGHHHHHHGGPPYTITYFPVRGRCEAMRMLLADQKQSWKEEVVTMKTWPPLKPSC
LFRQLPKFQDGKLTLYQSNAILRHLGRSFGLYGKKQKEAALVDMVNDGVEDLRCKY
ATLIYTKYKAGKKKYVKKLPKHLKPFETLLSKNKGGKAFVVGSKISFADYNLLDLLR
IHQVLNPSCLKAFPLLSAYVARLSARPKIKAFLASPEHVKRPINGNGKQ (SEQ ID NO:
XX)

As would be appreciated by one of skill in the art, homologous proteins are also considered to be within the scope of this invention. For example, any protein that includes a stretch of 20, 30, 40, 50, or 100 amino acids which are 60%, 70%, 80%, 90%, 95%, or 100% homologous to any of the above sequences is considered part of the invention. In addition, addition and deletion variants are also contemplated by the invention. In certain embodiments, any streptavidin with a mutated residue as shown in any of the above sequences is considered part of the invention. In certain embodiments, the sequence includes 2, 3, 4, 5, 6, 7, 8, 9, 10, or more mutations as shown in any of the sequences above.

[0044] Any DNA sequence that encodes the above GST variants is also included within the scope of the invention. Exemplary DNA sequences which encode each of the variants above are as follows:

GST-NEG40

GGTTCAGCCATGGGTCATCACCACCACCATCACGGTGGCCCGCCGTACACCATTA
CATACTTTCCGGTACGTGGTCGTTGTGAAGCGATGCGTATGTTATTAGCGGACCA
GGACCAATCATGGGAAGAAGAAGTAGTGACAATGGAAACCTGGCCGCCGTTAAA
GCCTAGCTGTTTATTCCGTCAATTACCGAAGTTTCAGGATGGTGATTTAACCTTAT
ACCAGTCTAACGCGATCTTACGTCATTTAGGTCGCTCATTTGGTTTATACGGTGA
AGATGAAGAAGAAGCAGCCTTAGTGGATATGGTGAATGATGGCGTGGAAGACTT
ACGTTGTAAATACGCGACGTTAATTTACACTGATTATGAAGCCGGTAAAGAGGA
GTACGTGGAAGAATTACCTGAACACCTGAAGCCGTTTGAAACATTACTGAGCGA
AAATGAAGGAGGTGAGGCGTTCGTAGTTGGTAGCGAAATTAGCTTCGCTGATTAT

AACTTATTAGACTTATTACGCATTCACCAGGTTTTAAATCCTAGCTGTTTAGACGC
TTTCCCGTTACTGAGCGCATATGTAGCGCGCCTGAGCGCCCGTCCGGAAATTGAA
GCTTTCTTAGCGTCACCTGAACACGTAGACCGCCCGATTAACGGAAACGGCAAG
CAGTAATAATGAGGTACCACCTGCA (SEQ ID NO: XX)

GST-POS50

GGTTCAGCCATGGGTCATCACCACCACCATCACGGTGGCCCGCCGTACACCATTA
CATACTTTCCGGTACGTGGTCGTTGTGAAGCGATGCGTATGTTATTAGCGGACCA
GAAACAATCATGGAAAGAAGAAGTAGTGACAATGAAGACCTGGCCGCCGTTAAA
GCCTAGCTGTTTATTCCGTCAATTACCGAAGTTTCAGGATGGTAAATTAACCTTAT
ACCAGTCTAACGCGATCTTACGTCATTTAGGTCGCTCATTTGGTTTATACGGTAA
GAAGCAGAAAGAAGCAGCCTTAGTGGATATGGTGAATGATGGCGTGGAAGACTT
ACGTTGTAAATACGCGACGTTAATTTACACTAAATATAAAGCCGGTAAAAAGAA
GTACGTGAAAAAATTACCTAAACACCTGAAGCCGTTTGAAACATTACTGAGCAA
AAATAAAGGAGGTAAGGCGTTCGTAGTTGGTAGCAAGATTAGCTTCGCTGATTAT
AACTTATTAGACTTATTACGCATTCACCAGGTTTTAAATCCTAGCTGTTTAAAGGC
TTTCCCGTTACTGAGCGCATATGTAGCGCGCCTGAGCGCCCGTCCGAAGATCAAA
GCTTTCTTAGCGTCACCTGAACACGTGAAGCGCCCGATTAACGGAAACGGCAAG
CAGTAATAATGAGGTACCACCTGCA (SEQ ID NO: XX)

[0045] The present invention also provides vector (*e.g.*, plasmids, cosmids, viruses, *etc.*) that comprise any of the inventive sequences herein or any other sequence (DNA or protein) modified using the inventive system. In certain embodiments, the vector includes elements such as promoter, enhancer, ribosomal binding sites, *etc.* sequences useful in overexpressing the inventive GST variant in a cell. The invention also includes cells comprising the inventive sequences or vectors. In certain embodiments, the cells overexpress the variant GST. The cells may be bacterial cells (*e.g.*, *E. coli*), fungal cells (*e.g.*, *P. pastoris*), yeast cells (*e.g.*, *S. cerevisiae*), mammalian cells (*e.g.*, CHO cells), or human cells.

[0046] The present invention also includes kits for modifying proteins of interest to produce more stable variants of the protein. These kits typically include all or most of the reagents needed create a more stable variant of a protein. In certain embodiments, the kit includes computer software to aid a researcher in designing the more stable variant protein based on the inventive method. The kit may also include all of some of the following: reagents, primers, oligonucleotides, nucleotides, enzymes, buffers, cells, media, plates, tubes, instructions, vectors, *etc.* The research using the kit typically provides the DNA sequence for mutating to create the more stable variant. The contents are typically packaged for convenience use in a laboratory.

[0047] These and other aspects of the present invention will be further appreciated upon consideration of the following Examples, which are intended to illustrate certain particular embodiments of the invention but are not intended to limit its scope, as defined by the claims.

Examples

Example 1 – Supercharging Proteins Can Impart Extraordinary Resilience

[0048] Protein aggregation, a well known culprit in human disease (Cohen, F. E.; Kelly, J. W., *Nature* 2003, 426, (6968), 905-9; Chiti, F.; Dobson, C. M., *Annu Rev Biochem* 2006, 75, 333-66; each of which is incorporated herein by reference), is also a major problem facing the use of proteins as therapeutic or diagnostic agents (Frokjaer, S.; Otzen, D. E., *Nat Rev Drug Discov* 2005, 4, (4), 298-306; Fowler, S. B.; Poon, S.; Muff, R.; Chiti, F.; Dobson, C. M.; Zurdo, J., *Proc Natl Acad Sci U S A* 2005, 102, (29), 10105-10; each of which is incorporated herein by reference). Insights into the protein aggregation problem have been garnered from the study of natural proteins. It has been known for some time that proteins are least soluble at their isoelectric point, where they bear a net charge of zero (Loeb, J., *J Gen Physiol* 1921, 4, 547-555; incorporated herein by reference). More recently, small differences in net charge (±3 charge units) have been shown to predict aggregation tendencies among variants of a globular protein (Chiti, F.; Stefani, M.; Taddei, N.; Ramponi, G.; Dobson, C. M., *Nature* 2003, 424, (6950), 805-8; incorporated herein by reference), and also among intrinsically disordered peptides (Pawar, A. P.; Dubay, K. F.; Zurdo, J.; Chiti, F.; Vendruscolo, M.; Dobson, C. M., *J Mol Biol* 2005, 350, (2), 379-92; incorporated herein by reference). Together with recent evidence that some proteins can tolerate significant changes in net charge (for example, the finding that carbonic anhydrase retains catalytic activity after exhaustive chemical acetylation of its surface lysines (Gudiksen *et al.*, *J Am Chem Soc* 2005, 127, (13), 4707-14; incorporated herein by reference)), these observations led us to conclude that the solubility and aggregation resistance of some proteins might be significantly enhanced, without abolishing their folding or function, by extensively mutating their surfaces to dramatically increase their net charge, a process we refer to herein as "supercharging".

[0049] We began with a recently reported state-of-the-art variant of green fluorescent protein (GFP) called "superfolder GFP" (sfGFP), which has been highly optimized for folding efficiency and resistance to denaturants (Pedelacq *et al.*, *Nat Biotechnol* 2006, 24, (1),

79-88; incorporated herein by reference). Superfolder GFP has a net charge of –7, similar to that of wild-type GFP. Guided by a simple algorithm to calculate solvent exposure of amino acids (see *Materials and Methods*), we designed a supercharged variant of GFP having a theoretical net charge of +36 by mutating 29 of its most solvent-exposed residues to positively charged amino acids (Figure 1). The expression of genes encoding either sfGFP or GFP(+36) yielded intensely green-fluorescent bacteria. Following protein purification, the fluorescence properties of GFP(+36) were measured and found to be very similar to those of sfGFP. Encouraged by this finding, we designed and purified additional supercharged GFPs having net charges of +48, –25, and –30, all of which were also found to exhibit sfGFP-like fluorescence (Figure 2a). All supercharged GFP variants showed circular dichroism spectra similar to that of sfGFP, indicating that the proteins have similar secondary structure content (Figure 2b). The thermodynamic stabilities of the supercharged GFP variants were only modestly lower than that of sfGFP (1.0–4.1 kcal/mol, Figure 2c and Table 1) despite the presence of as many as 36 mutations.

[0050] Although sfGFP is the product of a long history of GFP optimization (Giepmans *et al.*, *Science* 2006, 312, (5771), 217-24; incorporated herein by reference), it remains susceptible to aggregation induced by thermal or chemical unfolding. Heating sfGFP to 100°C induced its quantitative precipitation and the irreversible loss of fluorescence (Figure 3a). In contrast, supercharged GFP(+36) and GFP(–30) remained soluble when heated to 100°C, and recovered significant fluorescence upon cooling (Figure 3a). Importantly, while 40% 2,2,2-trifluoroethanol (TFE) induced the complete aggregation of sfGFP at 25 °C within minutes, the +36 and –30 supercharged GFP variants suffered no significant aggregation or loss of fluorescence under the same conditions for hours (Figure 3b).

[0051] In addition to this remarkable aggregation resistance, supercharged GFP variants show a strong, reversible avidity for highly charged macromolecules of the opposite charge (Figure 3c). When mixed together in 1:1 stoichiometry, GFP(+36) and GFP(–30) immediately formed a green fluorescent co-precipitate, indicating the association of folded proteins. GFP(+36) similarly co-precipitated with high concentrations of RNA or DNA. The addition of NaCl was sufficient to dissolve these complexes, consistent with the electrostatic basis of their formation. In contrast, sfGFP was unaffected by the addition of GFP(–30), RNA, or DNA (Figure 3c).

[0052] We next sought to determine whether the supercharging principle could apply to proteins other than GFP, which is monomeric and has a well-shielded fluorophore. To this end, we applied the supercharging process to two proteins unrelated to GFP. Streptavidin is a tetramer with a total net charge of –4. Using the solvent-exposure algorithm, we designed two supercharged streptavidin variants with net charges of –40 or +52. Both supercharged streptavidin variants were capable of forming soluble tetramers that bind biotin, albeit with reduced affinity.

[0053] Glutathione-*S*-transferase (GST), a dimer with a total net charge of +2, was supercharged to yield a dimer with net charge of –40 that catalyzed the addition of glutathione to chlorodinitrobenzene with a specific activity only 2.7-fold lower than that of wild-type GST (Figure 3d). Moreover, the supercharged streptavidins and supercharged GST remained soluble when heated to 100°C, in contrast to their wild-type counterparts, which, like sfGFP, precipitated quantitatively and irreversibly (Table 1). In addition, GST(–40) recovered 40% of its catalytic activity upon cooling (Figure 3d).

[0054] In summary, we have demonstrated that monomeric and multimeric proteins of varying structures and functions can be "supercharged" by simply replacing their most solvent-exposed residues with like-charged amino acids. Supercharging profoundly alters the intermolecular properties of proteins, imparting remarkable aggregation resistance and the ability to associate in folded form with oppositely charged macromolecules like "molecular Velcro." We note that these unusual intermolecular properties arise from high net charge, rather than from the total number of charged amino acids, which was not significantly changed by the supercharging process (Table 1).

[0055] In contrast to these dramatic intermolecular effects, the intramolecular properties of the seven supercharged proteins studied here, including folding, fluorescence, ligand binding, and enzymatic catalysis, remained largely intact. Supercharging therefore may represent a useful approach for reducing the aggregation tendency and improving the solubility of proteins without abolishing their function. These principles may be particularly useful in *de novo* protein design efforts, where unpredictable protein handling properties including aggregation remain a significant challenge. In light of the above results of supercharging natural proteins, it is tempting to speculate that the aggregation resistance of designed proteins could also be improved by biasing the design process to increase the

frequency of like-charged amino acids at positions predicted to lie on the outside of the folded protein.

[0056] Protein supercharging illustrates the remarkable plasticity of protein surfaces and highlights the opportunities that arise from the mutational tolerance of solvent-exposed residues. For example, it was recently shown that the thermodynamic stability of some proteins can be enhanced by rationally engineering charge-charge interactions (Strickler *et al.*, *Biochemistry* 2006, 45, (9), 2761-6; incorporated herein by reference). Protein supercharging demonstrates how this plasticity can be exploited in a different way to impart extraordinary resistance to protein aggregation. Our findings are consistent with the results of a complementary study in which removal of all charges from ubiquitin left its folding intact but significantly impaired its solubility (Loladze *et al*, *Protein Sci* 2002, 11, (1), 174-7; incorporated herein by reference).

[0057] These observations may also illuminate the modest net-charge distribution of natural proteins (Knight *et al.*, *Proc Natl Acad Sci U S A* 2004, 101, (22), 8390-5; Gitlin *et al.*, *Angew Chem Int Ed Engl* 2006, 45, (19), 3022-60; each of which is incorporated herein by reference): the net charge of 84% of Protein Data Bank (PDB) polypeptides, for example, falls within ±10. Our results argue against the hypothesis that high net charge creates sufficient electrostatic repulsion to force unfolding. Indeed, GFP(+48) has a higher positive net charge than any polypeptide currently in the PDB, yet retains the ability to fold and fluoresce. Instead, our findings suggest that nonspecific intermolecular adhesions may have disfavored the evolution of too many highly charged natural proteins. Almost all natural proteins with very high net charge, such as ribosomal proteins L3 (+36) and L15 (+44), which bind RNA, or calsequestrin (–80), which binds calcium cations, associate with oppositely charged species as part of their essential cellular functions.

Materials and Methods

Design procedure and supercharged protein sequences. Solvent-exposed residues (shown in grey below) were identified from published structural data (Weber, P.C., Ohlendorf, D.H., Wendoloski, J.J. & Salemme, F.R. Structural origins of high-affinity biotin binding to streptavidin. *Science* **243**, 85-88 (1989); Dirr, H., Reinemer, P. & Huber, R. Refined crystal structure of porcine class Pi glutathione S-transferase (pGST P1-1) at 2.1 A resolution. *J Mol Biol* **243**, 72-92 (1994); Pedelacq, J.D., Cabantous, S., Tran, T., Terwilliger, T.C. & Waldo,

G.S. Engineering and characterization of a superfolder green fluorescent protein. *Nat Biotechnol* **24**, 79-88 (2006); each of which is incorporated herein by reference) as those having AvNAPSA < 150, where AvNAPSA is average neighbor atoms (within 10 Å) per sidechain atom. Charged or highly polar solvent-exposed residues (DERKNQ) were mutated either to Asp or Glu, for negative-supercharging (red); or to Lys or Arg, for positive-supercharging (blue). Additional surface-exposed positions to mutate in green fluorescent protein (GFP) variants were chosen on the basis of sequence variability at these positions among GFP homologues. The supercharging design process for streptavidin (SAV) and glutathione-*S*-transferase (GST) was fully automated: residues were first sorted by solvent exposure, and then the most solvent-exposed charged or highly polar residues were mutated either to Lys for positive supercharging, or to Glu (unless the starting residue was Asn, in which case to Asp) for negative supercharging.

```
                1
SAV(-40)        MGHHHHHHGGAEAGITGTWYNQLGSTFIVTAGADGALTGTYESAVGDAESRYVLTGRYDSAPATDGSGTA
wtSAV           ----------AAEAGITGTWYNQLGSTFIVTAGADGALTGTYESAVGNAESRYVLTGRYDSAPATDGSGTA
SAV(+52)        MGHHHHHHGGAKAGITGTWYNQLGSTFIVTAGAKGALTGTYESAVGNAKSRYVLTGRYDSAPATKGSGTA

                71
SAV(-40)        LGWTVAWKNDYENAHSATTWSGQYVGGAEARINTQWLLTSGTTEADAWKSTLVGHDTFTKVEPSAAS
wtSAV           LGWTVAWKNNYRNAHSATTWSGQYVGGAEARINTQWLLTSGTTEANAWKSTLVGHDTFTKVKPSAAS
SAV(+52)        LGWTVAWKNKYRNAHSATTWSGQYVGGAKARINTQWLLTSGTTKAKAWKSTLVGHDTFTKVKPSAAS


                1
GST(-40)        MGHHHHHHGGPPYTITYFPVRGRCEAMRMLLADQDQSWEEEVVTMETWPPLKPSCLFRQLPKFQDGDLTLYQSNA
wtGST           MGHHHHHHGGPPYTITYFPVRGRCEAMRMLLADQDQSWKEEVVTMETWPPLKPSCLFRQLPKFQDGDLTLYQSNA
GST(+50)        MGHHHHHHGGPPYTITYFPVRGRCEAMRMLLADQKQSWKEEVVTMKTWPPLKPSCLFRQLPKFQDGKLTLYQSNA

                75
GST(-40)        ILRHLGRSFGLYGEDEEEAALVDMVNDGVEDLRCKYATLIYTDYEAGKEEYVEELPEHLKPFETLLSENEGGEAF
wtGST           ILRHLGRSFGLYGKDQKEAALVDMVNDGVEDLRCKYATLIYTNYEAGKEKYVKELPEHLKPFETLLSQNQGGQAF
GST(+50)        ILRHLGRSFGLYGKKQKEAALVDMVNDGVEDLRCKYATLIYTKYKAGKKKYVKKLPKHLKPFETLLSKNKGGKAF

                151
GST(-40)        VVGSEISFADYNLLDLLRIHQVLNPSCLDAFPLLSAYVARLSARPEIEAFLASPEHVDRPINGNGKQ
wtGST           VVGSQISFADYNLLDLLRIHQVLNPSCLDAFPLLSAYVARLSARPKIKAFLASPEHVNRPINGNGKQ
GST(+50)        VVGSKISFADYNLLDLLRIHQVLNPSCLKAFPLLSAYVARLSARPKIKAFLASPEHVKRPINGNGKQ
```

Protein expression and purification. Synthetic genes optimized for *E. coli* codon usage were purchased from DNA 2.0, cloned into a pET expression vector (Novagen), and overexpressed in *E. coli* BL21(DE3)pLysS for 5–10 hours at 15°C. Cells were harvested by centrifugation and lysed by sonication. Proteins were purified by Ni-NTA agarose

chromotography (Qiagen), buffer-exchanged into 100 mM NaCl, 50 mM potassium phosphate pH 7.5, and concentrated by ultrafiltration (Millipore). All GFP variants were purified under native conditions. Wild-type streptavidin was purchased from Promega. Supercharged streptavidin variants were purified under denaturing conditions and refolded as reported previously for wild-type streptavidin (Thompson *et al.* Construction and expression of a synthetic streptavidin-encoding gene in Escherichia coli. *Gene* **136**, 243-246 (1993); incorporated herein by reference), as was supercharged GST. Wild-type GST was purified under either native or denaturing conditions, yielding protein of comparable activity.

Electrostatic surface potential calculations (Figure 1b). Models of –30 and +48 supercharged GFP variants were based on the crystal structure of superfolder GFP (Pedelacq *et al.*, Engineering and characterization of a superfolder green fluorescent protein. *Nat Biotechnol* **24**, 79-88 (2006); incorporated herein by reference). Electrostatic potentials were calculated using APBS (Baker *et al.*, Electrostatics of nanosystems: application to microtubules and the ribosome. *Proc Natl Acad Sci U S A* **98**, 10037-10041 (2001); incorporated herein by reference) and rendered with PyMol (Delano, W.L., The PyMOL Molecular Graphics System, www.pymol.org (2002); incorporated herein by reference) using a scale of –25 kT/e (red) to +25kT/e (blue).

Protein staining and UV-induced fluorescence (Figure 2a). 0.2 μg of each GFP variant was analyzed by electrophoresis in a 10% denaturing polyacrylamide gel and stained with Coomassie brilliant blue dye. 0.2 μg of the same protein samples in 25 mM Tris pH 8.0 with 100 mM NaCl was placed in a 0.2 mL Eppendorf tube and photographed under UV light (360 nm).

Thermal denaturation and aggregation (Figure 3a). Purified GFP variants were diluted to 2 mg/mL in 25 mM Tris pH 8.0, 100 mM NaCl, and 10 mM beta-mercaptoethanol (BME), then photographed under UV illumination ("native"). The samples were heated to 100 °C for 1 minute, then photographed again under UV illumination ("boiled"). Finally, the samples were cooled 2 h at room temperature and photographed again under UV illumination ("cooled").

Chemically induced aggregation (Figure 3b). 2,2,2-trifluoroethanol (TFE) was added to produce solutions with 1.5 mg/mL protein, 25 mM Tris pH 7.0, 10 mM BME, and 40% TFE. Aggregation at 25 °C was monitored by right-angle light scattering.

Size-exclusion chromotography (Table 1). The multimeric state of SAV and GST variants was determined by analyzing 20–50 μg of protein on a Superdex 75 gel-filtration column. Buffer was 100 mM NaCl, 50 mM potassium phosphate pH 7.5. Molecular weights were determined by comparison with a set of monomeric protein standards of known molecular weights analyzed separately under identical conditions.

Table 1. Calculated and experimentally determined protein properties.

name	MW (kD)	length (aa)	n_{pos}	n_{neg}	$n_{charged}$	Q_{net}	pI	ΔG (kcal/mol)[a]	native MW (kD)[b]	% soluble after boiling[c]
GFP (–30)	27.8	248	19	49	68	–30	4.8	10.2	n.d.	98
GFP (–25)	27.8	248	21	46	67	–25	5.0	n.d.	n.d.	n.d.
sfGFP	27.8	248	27	34	61	–7	6.6	11.2	n.d.	4
GFP (+36)	28.5	248	56	20	76	+36	10.4	8.8	n.d.	97
GFP (+48)	28.6	248	63	15	78	+48	10.8	7.1	n.d.	n.d.
SAV (–40)	14.3	137	5	15	20	–10	5.1	n.d.	55 ± 5 (tetramer)	99
wtSAV	13.3	128	8	9	17	–1	6.5	n.d.	50 ± 5 (tetramer)	7
SAV (+52)	14.5	137	16	3	19	+13	10.3	n.d.	55 ± 5 (tetramer)	97
GST (–40)	24.7	217	17	37	54	–20	4.8	n.d.	50 ± 5 (dimer)	96
wtGST	24.6	217	24	23	47	+1	7.9	n.d.	50 ± 5 (dimer)	3
GST (+50)[d]	24.7	217	39	14	53	+25	10.0	n.d.	n.d.	n.d.

n_{pos}, number of positively charged amino acids (per monomer)
n_{neg}, number of negatively charged amino acids
$n_{charged}$, total number of charged amino acids
Q_{net}, theroretical net charge at neutral pH
pI, calculated isoelectric point
n.d., not determined

[a]measured by guanidinium denaturation (Figure 2c).
[b]measured by size-exclusion chromatography.
[c]percent protein remaining in supernatant after 5 min at 100 °C, cooling to 25 °C, and brief centrifugation.
[d]protein failed to express in *E. coli*.

2007256780 29 Jul 2013

Other Embodiments

[0058] Those of ordinary skill in the art will readily appreciate that the foregoing represents merely certain preferred embodiments of the invention. Various changes and modifications to the procedures and compositions described above can be made without departing from the spirit or scope of the present invention, as set forth in the following claims.

[0059] Throughout this specification and the claims which follow, unless the context requires otherwise, the word "comprise", and variations such as "comprises" and "comprising", will be understood to imply the inclusion of a stated integer or step or group of integers or steps but not the exclusion of any other integer or step or group of integers or steps.

[0060] The reference in this specification to any prior publication (or information derived from it), or to any matter which is known, is not, and should not be taken as an acknowledgment or admission or any form of suggestion that that prior publication (or information derived from it) or known matter forms part of the common general knowledge in the field of endeavour to which this specification relates.

Appendix A

```
#!/usr/local/bin/perl

#########################################################################
#
# avnapsa
#
# prints list of AvNAPSA values for the specified PDB
#
# Mike Lawrence/Kevin Phillips  3/17/2006
#
#########################################################################

sub show_usage
{
print "\n",

"Usage: avnapsa <start_pdb> [params]\n",
"  -3  use 3-letter aa abbreviations (default)\n",
"  -1  use 1-letter aa abbreviations\n",
"  -onecol  print one column only (i.e. only the AvNAPSA results)\n\n";
}

################## global variables ###################################

@atoms;
  # fields loaded from PDB:
      # type
      # atomNum
      # atomName
      # resName
      # chain
      # resNum
      # x, y, z
  # computed fields
      # neighborCount

@distances;

@residues;
  # fields copied from PDB
      # resNum (PDB numbering)
      # resName
  # computed fields
      # avNapsa

#########################################################################

## parse command line

$use3or1 = 3;
$onecol_flag = 0;

$start_pdb = $ARGV[0];

for (my $a = 1; $a < @ARGV; ++$a)
{
       if ($ARGV[$a] eq "-1") { $use3or1 = 1; }
```

```
    elsif ($ARGV[$a] eq "-3") { $use3or1 = 3; }
    elsif ($ARGV[$a] eq "-onecol") { $onecol_flag = 1; }
    else { show_usage(); die "Invalid argument $ARGV[$a]\n"; }
}

unless (lc $start_pdb =~ /\.pdb/) { show_usage(); die "No starting pdb
specified.\n"; }

## read PDB and compute molecular parameters

read_PDB($start_pdb);
tabulate_residues();
$nres = @residues;

compute_distances();
compute_neighbor_counts();
compute_residue_avNapsa();

print_residues();
exit;


#
# print_residues
#
#

sub print_residues
{
    for (my $r = 0; $r < @residues; $r++)
    {
      my $name = $residues[$r]{resName};
        $name = toggle31($name) if ($use3or1 == 1);
        printf "%d %s AvNAPSA ", $residues[$r]{resNum}, $name unless
$onecol_flag;
        printf "%.0f\n",  $residues[$r]{avNapsa};
    }
    print "\nNum residues = ", $#residues+1, "\n\n" unless $onecol_flag;
}

#
# tabulate_residues
#
# goes through list of atoms and makes a list of amino acid residues
# and stores it in global variable @residues
#

sub tabulate_residues
{
    for ($a = 0; $a < @atoms; $a++)
    {
      $resNum = $atoms[$a]{resNum};
        if ( ! resNum_exists($resNum) )
        {
            push @residues,
          {
                resNum => $resNum,
```

```
                resName => $atoms[$a]{resName}
            };
        }
    }
}


#
# resNum_exists
#
# returns 1 if resNum is contained in @residues
#

sub resNum_exists($)
{
    my ($resNum) = @_;

    for ($r = 0; $r < @residues; $r++)
    {
      return 1 if ($residues[$r]{resNum} == $resNum);
    }
    return 0;
}


#
# resNum_to_resindex
#
# converts PDB numbering to index in @residues
#

sub resNum_to_resindex($)
{
    my ($resNum) = @_;

    for ($r = 0; $r < @residues; $r++)
    {
      return $r if ($residues[$r]{resNum} == $resNum);
    }
    return "none";
}


#
# readPDB(filename)
#
# reads the atoms from a PDB and returns them as an array of hashes
#

sub read_PDB($)
{
    my ($filename) = @_;

    open (PDB, $filename) or die("Could not open $filename\n");

    $#atoms = -1;          # clear atoms storage
```

```
# read the file

    foreach (<PDB>) {
      my $type = trim(substr($_, 0,6));              # RTyp field is columns 1-6
        next unless ($type eq "ATOM" || $type eq "HETATM");

      my $resName = trim(substr($_, 17, 3));         # Res field is columns 18-
20
        my $atomName = trim(substr($_, 12, 4));        # Atm field is columns
13-16

        next if uc $resName eq "HOH";                  # omit waters
        next if uc $atomName =~ /^[0-9]*H/;            # omit protons

        # add a hash to the array, containing data from this record of the PDB

        push @atoms, {
           type =>     $type,
           resName =>  $resName,
           atomName => $atomName,
           atomNum =>  trim(substr($_, 6,5)),    # Num field is columns 7-11
           chain =>    trim(substr($_, 21,1)),   # Chain field is column 22
           resNum =>   trim(substr($_, 22,4)),   # ResNo field is columns 23-26
           x =>        trim(substr($_, 30,8)),   # X field is columns 31-38
           y =>        trim(substr($_, 38,8)),   # Y field is columns 39-46
           z =>        trim(substr($_, 46,8))    # Z field is columns 37-54
         };
    }

    close(PDB);
}


#
# trim
#
# removes whitespace from start and end of string
#

sub trim($)
{
    my ($string) = @_;      # retrieve the passed argument
    $string =~ s/^\s+//;    # remove leading whitespace
    $string =~ s/\s+$//;    # remove trailing whitespace
    return $string;
}


#
# is_number
#
# returns 1 if passed argument is a number (allows whitespace, negative, and
decimal point)
# returns 0 if passed argument is blank or not a number
#
```

```
sub is_number($)
{
    $_ = shift;
    s/^\s+//;
    s/\s+$//;
    return 1 if /^-?[0-9]+$/ || /^-?[0-9]*\.[0-9]+$/ || /^-?[0-9]+\.[0-9]*$/;
    return 0;
}



#
# inter_residue_distance
#
# returns the minimum distance between any atoms of the specified residues
# (residues are specified according to index in @residues)
#

sub inter_residue_distance($, $)
{
    my ($r1, $r2) = @_;

    ## convert to PDB numbering

    my $resNum1 = $residues[$r1]{resNum};
    my $resNum2 = $residues[$r2]{resNum};

    my $min_dist = 1000000;

    for ($a1 = 0; $a1 < @atoms; ++$a1)
    {
        next unless ( $atoms[$a1]{resNum} == $resNum1 );
        for ($a2 = 0; $a2 < @atoms; ++$a2)
      {
            next unless ( $atoms[$a2]{resNum} == $resNum2 );
          my $dist = $distances[$a1][$a2];
          $min_dist = $dist if ($dist < $min_dist);
      }
    }
    return $min_dist;
}



#
# compute_distances
#
# computes the distances between all atoms
#

sub compute_distances
{
    for(my $atom1=0; $atom1 < @atoms; $atom1++)
    {
      for(my $atom2=$atom1; $atom2 < @atoms; $atom2++)
```

```
      {
          my ($x1,$y1,$z1) = ($atoms[$atom1]->{x}, $atoms[$atom1]->{y},
$atoms[$atom1]->{z});
          my ($x2,$y2,$z2) = ($atoms[$atom2]->{x}, $atoms[$atom2]->{y},
$atoms[$atom2]->{z});

          my $distance = sqrt(($x1-$x2)**2 + ($y1-$y2)**2 + ($z1-$z2)**2);

          $distances[$atom1][$atom2] = $distance;
          $distances[$atom2][$atom1] = $distance;
      }
    }
}


#
# compute_neighbor_counts
#
# computes the number of neighbors that each atom has.
# paramter is the cutoff, in Angstroms, for atomic neighborhood
#

sub compute_neighbor_counts
{
    $DISTANCE_CUTOFF = 10;      # criterion for neighborhood, in Angstroms

    for ($atom1=0; $atom1 < @atoms; $atom1++)
    {
      my $count = 0;
      for ($atom2=0; $atom2 < @atoms; $atom2++)
      {
            $count++ if ($distances[$atom1][$atom2] <= $DISTANCE_CUTOFF
                && $atom1 != $atom2);
        }
      $atoms[$atom1]{neighborCount} = $count;
    }
}



#
# compute_residue_avNapsa
#
# for each residue, compute
# Average Neighbor Atoms Per Sidechain Atom (AvNAPSA)
# (sidechain atoms are all those except N, C, O, CA)
# for glycines, just use CA
#

sub compute_residue_avNapsa
{
    for (my $r = 0; $r < @residues; $r++)
    {
      my $numSideChainAtoms = 0;
      my $totalNeighbors = 0;
      my $resName = $residues[$r]{resName};
      my $resNum = $residues[$r]{resNum};
```

```
      for (my $a = 0; $a < @atoms; $a++)
        {
        if ($atoms[$a]{resNum} == $resNum)
          {
          my $atomName = $atoms[$a]{atomName};
            if (
                 (    $atomName ne "C"
                   && $atomName ne "O"
                   && $atomName ne "N"
                   && $atomName ne "CA"
                  )
                    || ( $atomName eq "CA" && $resName eq "GLY")
                )
          {
                $numSideChainAtoms++;
                $totalNeighbors += $atoms[$a]{neighborCount};
          }
          }
        }
      my $avNapsa = $totalNeighbors / $numSideChainAtoms;
      $residues[$r]{avNapsa} = $avNapsa;
    }
}



#
# toggle31
#
# converts 3-letter abbrev to 1-letter
# or 1-letter abbrev to 3-letter
#


sub toggle31($)
{
    %conv3to1 = ( "ALA" => "A", "CYS" => "C", "SER" => "S", "LEU" => "L",
       "ILE" => "I", "PHE" => "F", "ARG" => "R", "ASN" => "N", "GLN" => "Q",
       "TYR" => "Y", "LYS" => "K", "ASP" => "D", "GLU" => "E", "VAL" => "V",
       "TRP" => "W", "MET" => "M", "HIS" => "H", "GLY" => "G", "PRO" => "P",
"THR" => "T" );

    %conv1to3 = reverse %conv3to1;

    my ($abbrev) = @_;

    $abbrev = uc $abbrev;
    return $conv1to3{$abbrev} if length ($abbrev) == 1;
    return $conv3to1{$abbrev} if length ($abbrev) == 3;
    die "in toggle31(): invalid amino acid abbreviation $abbrev\n";
}


#
# is_aa
#
```

```
# returns 1 if passed argument is a 1-letter amino acid
#

sub is_aa($)
{
    my ($string) = @_;

    return 1 if (length toggle31($string) == 3);
    return 0;
}
```

THE CLAIMS DEFINING THE INVENTION ARE AS FOLLOWS:

1. A method of improving the stability of a protein of interest, the method comprising steps of:

(i) identifying non-conserved surface residues of a protein of interest; and

(ii) replacing a plurality of non-conserved, surface residues with an amino acid residue that is positively charged at physiological pH.

2. A method of improving the stability of a protein of interest, the method comprising steps of:

(i) identifying non-conserved surface residues of a protein of interest; and

(ii) replacing a plurality of non-conserved, surface residues with an amino acid residue that is negatively charged at physiological pH.

3. A method of improving the stability of a protein of interest, the method comprising steps of:

(i) identifying surface residues of a protein of interest;

(ii) identifying non-conserved surface residues of the protein of interest;

(iii) assigning a hydrophobicity value to each of the identified non-conserved surface residues; and

(iv) replacing at least one surface residue with an amino acid residue that is charged at physiological pH, wherein the type of residue is either positively charged residues or negatively charged residues.

4. The method of claim 1 or 2, wherein the non-conserved, surface residues are hydrophobic.

5. The method of claim 1, wherein the non-conserved, surface residues are hydrophilic or negatively charged.

6. The method of claim 2, wherein the non-conserved, surface residues are hydrophilic or positively charged.

7. The method of claim 1, wherein the step of replacing comprises:
   (a) replacing at least one hydrophobic surface residue;
   (b) replacing at least one hydrophilic surface residue;
   (c) replacing at least one charged surface residue; or
   (d) replacing at least one surface residue with a lysine residue.

8. The method of claim 2, wherein the step of replacing comprises:
   (a) replacing at least one hydrophobic surface residue;
   (b) replacing at least one hydrophilic surface residue;
   (c) replacing at least one charged surface residue; or
   (d) replacing at least one surface residue with an aspartate or glutamate residue.

9. The method of claim 3, wherein the step of replacing comprises:
   (a) replacing at least one hydrophobic surface residue;
   (b) replacing at least one hydrophilic surface residue;
   (c) replacing at least one charged surface residue;
   (d) replacing at least one surface residue with a lysine residue; or
   (e) replacing at least one surface residue with an aspartate or glutamate residue.

10. The method of any one of claims 1 to 3, wherein all the residues being replaced are changed into the same residue.

11. The method of any one of claims 1 to 3, wherein the residues being replaced are changed into different residues.

12. The method of any one of claims 1 to 3, wherein the step of replacing comprises replacing at least two, at least five, at least ten, at least twenty or at least thirty surface residues.

13. The method of any one of claims 1 to 3, whereby the method creates a modified protein with a greater net charge at physiological pH than the original protein of interest.

14. The method of claim 2 or 3, whereby the method creates a modified protein of interest that is more negatively charged at physiological pH than the original protein of interest, preferably at least -5, at least -10, at least -15 or at least -20 more negatively charged at physiological pH than the original protein of interest.

15. The method of claim 1 or 3, whereby the method creates a modified protein of interest that is more positively charged at physiological pH than the original protein of interest, preferably at least +5, at least +10, at least +15 or at least +20 more positively charged at physiological pH than the original protein of interest.

16. The method of any one of claims 1 to 3, wherein the protein of interest is selected from the group consisting of a protein susceptible to aggregation, a hydrophobic protein, a membrane protein, a protein that is difficult to overexpress, a protein that is difficult to purify, a receptor, a transcription factor, an enzyme, a structural protein, a fluorescent protein, green fluorescent protein (GFP), an extracellular protein, streptavidin and glutathione-*S*-transferase.

17. The method of any one of claims 1 to 3, wherein the step of identifying the surface residues comprises:

(a) computer modeling the three-dimensional structure of the protein;

(b) predicting using algorithms whether a residue is found on the surface of a protein; or

(c) identifying residues with an avNAPSA value less than a threshold value.

18. The method of any one of claims 1 to 3, wherein the alignment is performed with at least 2, at least 3 or at least 5 other protein sequences.

19. The method of claim 3, wherein the step of assigning a hydrophobicity value comprises using octanol/water P values.

20. The method of any one of claims 1 to 3, wherein the step of replacing comprises

mutagenizing the sequence of the protein to replace the identified hydrophobic surface residue with a natural amino acid residue that is charged at physiological pH.

21. The method of claim 20, wherein the natural amino acid residue is selected from the group consisting of lysine, glutamate, aspartate, histidine, and arginine.

22. The method of any one of claims 1 to 3, wherein the step of replacing comprises:
(a) site-directed mutagenesis of the identified surface residues; or
(b) PCR mutagenesis of the identified surface residues.

23. A green fluorescent protein (+36 GFP) of amino acid sequence:
MGHHHHHHGGASKGERLFRGKVPILVELKGDVNGHKFSVRGKGKGDATRG
KLTLKFICTTGKLPVPWPTLVTTLTYGVQCFSRYPKHMKRHDFFKSAMPKGY
VQERTISFKKDGKYKTRAEVKFEGRTLVNRIKLKGRDFKEKGNILGHKLRYN
FNSHKVYITADKRKNGIKAKFKIRHNVKDGSVQLADHYQQNTPIGRGPVLLP
RNHYLSTRSKLSKDPKEKRDHMVLLEFVTAAGIKHGRDERYK (SEQ ID NO:
5).

24. A complex comprising the green fluorescent protein of claim 23 and RNA or DNA.

25. A polynucleotide encoding the green fluorescent protein of claim 23.

26. The polynucleotide of claim 25 of sequence:
ATGGGGCATCATCATCATCACCACGGCGGGGCGTCTAAGGGAGAGCGCTT
GTTTCGCGGCAAAGTCCCGATTCTTGTGGAGCTCAAAGGTGATGTAAATG
GTCATAAATTTAGTGTGCGCGGGAAAGGGAAAGGAGATGCTACGCGGGG
CAAGCTCACCCTGAAATTTATTTGCACAACCGGCAAACTGCCAGTGCCGT
GGCCTACATTAGTCACTACTCTGACGTACGGTGTTCAGTGCTTTTCTCGCT
ATCCCAAACACATGAAACGCCATGATTTCTTCAAGAGCGCGATGCCAAAA
GGTTATGTGCAGGAACGCACCATCAGCTTTAAAAAAGACGGCAAATATAA
AACCCGTGCAGAAGTTAAATTCGAAGGCCGCACCCTGGTCAACCGCATTA

AACTGAAAGGTCGTGACTTCAAAGAGAAAGGTAATATTCTTGGTCACAAA
CTGCGCTATAATTTCAACTCTCACAAAGTTTATATTACGGCGGATAAACGT
AAAAACGGGATTAAAGCGAAATTTAAGATTCGTCATAATGTTAAAGACGG
CAGTGTGCAGTTAGCGGATCATTATCAGCAGAATACCCCAATTGGTCGCG
GTCCAGTGCTGCTGCCGCGTAACCATTATCTGTCGACCCGCAGCAAACTC
AGCAAAGACCCGAAAGAAAAACGTGACCACATGGTATTACTGGAATTTGT
GACCGCAGCAGGCATTAAACATGGCCGCGATGAACGTTACAAATAG (SEQ
ID NO: 11).

27. A vector comprising the polynucleotide of claim 25 or 26.

28. A green fluorescent protein (+49 GFP) of amino acid sequence:
MGHHHHHHGGRSKGKRLFRGKVPILVKLKGDVNGHKFSVRGKGKGDATRG
KLTLKFICTTGKLPVPWPTLVTTLTYGVQCFSRYPKHMKRHDFFKSAMPKGY
VQERTISFKKDGKYKTRAEVKFKGRTLVNRIKLKGRDFKEKGNILGHKLRYN
FNSHKVYITADKRKNGIKAKFKIRHNVKDGSVQLAKHYQQNTPIGRGPVLLP
RKHYLSTRSKLSKDPKEKRDHMVLKEFVTAAGIKHGRKERYK (SEQ ID NO:
7).

29. A polynucleotide encoding the green fluorescent protein of claim 28.

30. The polynucleotide of claim 29 of sequence:
ATGGGCCACCATCATCATCACCACGGGGGACGCTCTAAAGGTAAACGTCT
GTTTCGTGGAAAGGTGCCCATTCTGGTTAAACTCAAAGGTGATGTCAACG
GCCATAAGTTTTCGGTTCGTGGCAAAGGTAAAGGTGATGCGACGCGCGGG
AAATTAACACTGAAATTTATTTGCACAACCGGAAAACTCCCTGTGCCGTG
GCCGACTTTGGTGACCACATTAACCTATGGTGTTCAATGCTTCTCACGTTA
TCCGAAGCATATGAAACGTCATGATTTTTTCAAATCGGCTATGCCGAAAG
GTTACGTCCAGGAGCGCACCATCTCATTTAAGAAAGACGGTAAGTATAAA
ACCCGTGCTGAAGTAAAATTCAAAGGACGCACCCTGGTGAATCGCATTAA
ACTGAAAGGTCGTGATTTCAAAGAAAAGGGAAATATTTTAGGGCATAAGC

2007256780 29 Jul 2013

TCCGTTATAATTTTAACAGTCATAAGGTGTATATTACCGCTGATAAACGCA
AAAACGGAATCAAAGCGAAATTTAAGATCCGTCATAATGTAAAAGATGGC
TCAGTCCAACTGGCAAAACATTACCAGCAGAATACCCCGATCGGCCGCGG
TCCTGTGCTTCTGCCGCGTAAACACTACTTGTCGACCCGGTCAAAATTGAG
TAAAGATCCGAAGGAAAAGCGTGATCACATGGTCTTGAAGGAATTTGTAA
CTGCAGCAGGTATTAAACACGGGCGCAAAGAACGTTACAAATAG (SEQ ID NO: 13).

31. A vector comprising the polynucleotide of claim 29 or 30.

FIG. 1a-1	FIG. 1a-2

FIG. 1a

```
GFP (-30)  MGHHHHHHGGASKGEELFDGVVPILVELDGDVNGHEFSVRGEGEGDATEG
GFP (-25)  MGHHHHHHGGASKGEELFTGVVPILVELDGDVNGHEFSVRGEGEGDATEG
sfGFP      MGHHHHHHGGASKGEELFTGVVPIILVELDGDVNGHKFSVRGEGEGDATNG
GFP (+36)  MGHHHHHHGGASKGERLFRGKVPILVELKGDVNGHKFSVRGKGKGDATRG
GFP (+48)  MGHHHHHHGGRSKGKRLERGKVPILVKLKGDVNGHKFSVRGKGKGDATRG

GFP (-30)  ELTLKFICTTGELPVPWPTLVTTLTYGVQCFSDYPDHMDQHDFFKSAMPE
GFP (-25)  ELTLKFICTTGELPVPWPTLVTTLTYGVQCFSRYPDHMKQHDFFKSAMPE
sfGFP      KLTLKFICTTGKLPVPWPTLVTTLTYGVQCFSRYPDHMKQHDFFKSAMPE
GFP (+36)  KLTLKFICTTGKLPVPWPTLVTTLTYGVQCFSRYPKHMKRHDFFKSAMPK
GFP (+48)  KLTLKFICTTGKLPVPWPTLVTTLTYGVQCFSRYPKHMKRHDFFKSAMPK

GFP (-30)  GYVQERTISFKDDGTYKTRAEVKFEGDTLVNRIELKGIDFKEDGNILGHK
GFP (-25)  GYVQERTISFKDDGTYKTRAEVKFEGDTLVNRIELKGIDFFKEDGNILGHK
sfGFP      GYVQERTISFKDDGTYKTRAEVKFEGDTLVNRIELKGIDFEKEDGNILGHK
GFP (+36)  GYVQERTISFKKDGKYKTRAEVKFEGRTLVNRIKLKGRDFKEKGNILGHK
GFP (+48)  GYVQERTISFKKDGKYKTRAEVKFKGRTLVNRIKLKGRDFKEKGNILGHK

GFP (-30)  LEYNFNSHDVYITADKQENGIKAEFEIRHNVEDGSVQLADHYQQNTPIGD
GFP (-25)  LEYNFNSHDVYITADKQENGIKAEFEIRHNVEDGSVQLADHYQQNTPIGD
sfGFP      LEYNFNSHNVYITADKQKNGIKANFKIRHNVEDGSVQLADHYQQNTPIGD
GFP (+36)  LRYNFNSHKVYITADKRKNGIKAKFKIRHNVKDGSVQLADHYQQNTPIGR
GFP (+48)  LRYNFNSHKVYITADKRKNGIKAKFKIRHNVKDGSVQLAKHYQQNTPIGR
```

FIG. 1a-1

```
GFP (-30)   GPVLLPDDHYLSTESALSKDPNEDRDHMVLLEFVTAAGITHGMDELYK
GFP (-25)   GPVLLPDDHYLSTESALSKDPNEDRDHMVLLEFVTAAGITHGMDELYK
sfGFP       GPVLLPDNHYLSTQSALSKDPNEKRDHMVLLEFVTAAGITHGMDELYK
GFP (+36)   GPVLLPRNHYLSTRSKLSKDPKEKRDHMVLLEFVTAAGIKHGRDERYK
GFP (+48)   GPVLLPRKHYLSTRSKLSKDPKEKRDHMVLLEFVTAAGIKHGRKERYK
```

FIG. 1a-2

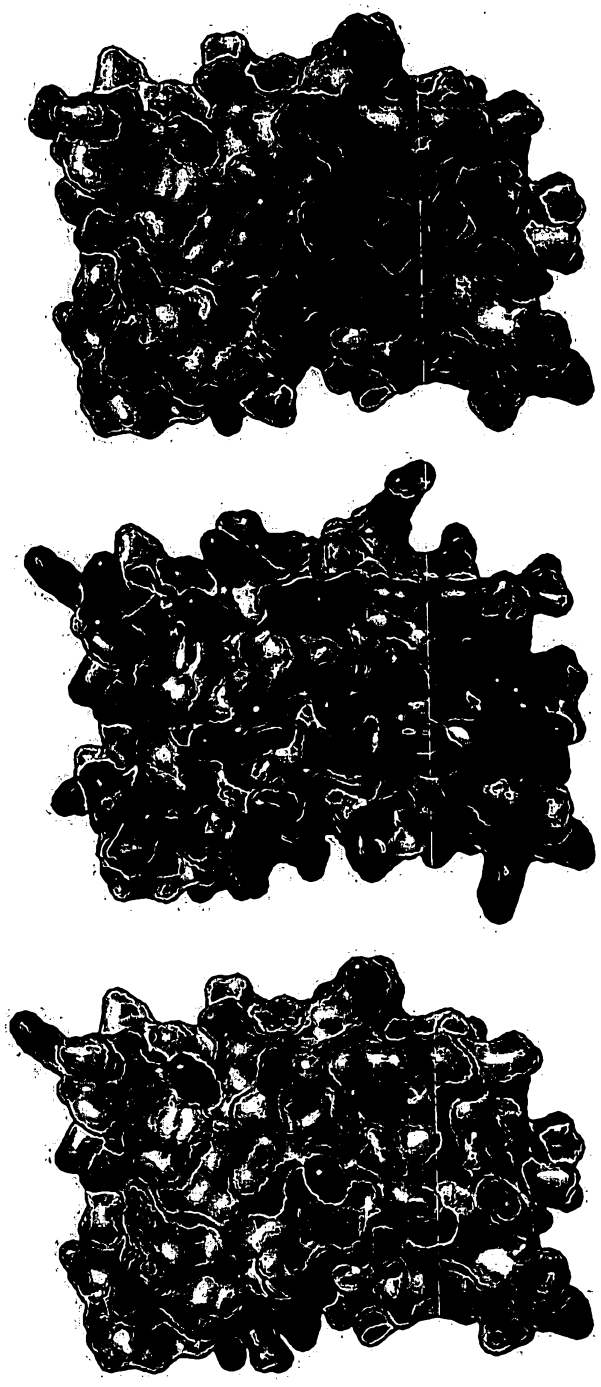


FIG. 1b

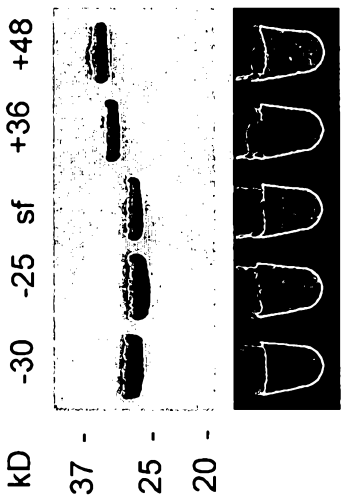


FIG. 2a

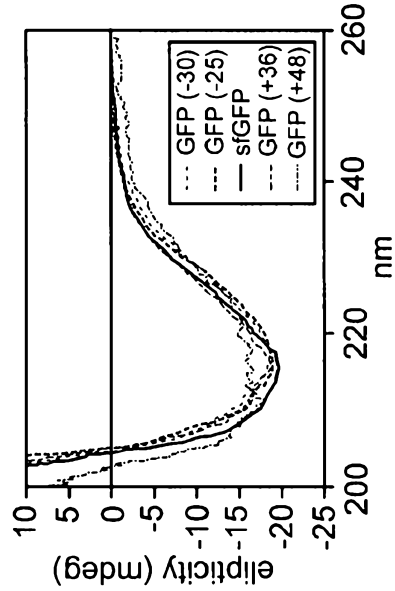


FIG. 2b

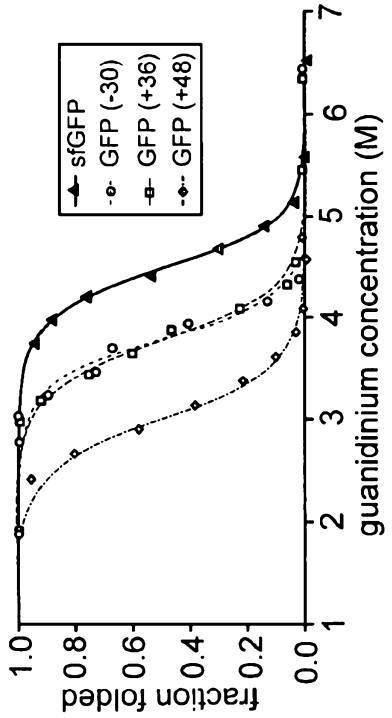


FIG. 2c

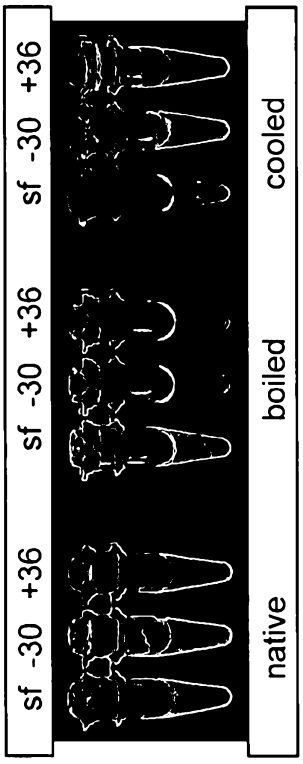


FIG. 3a

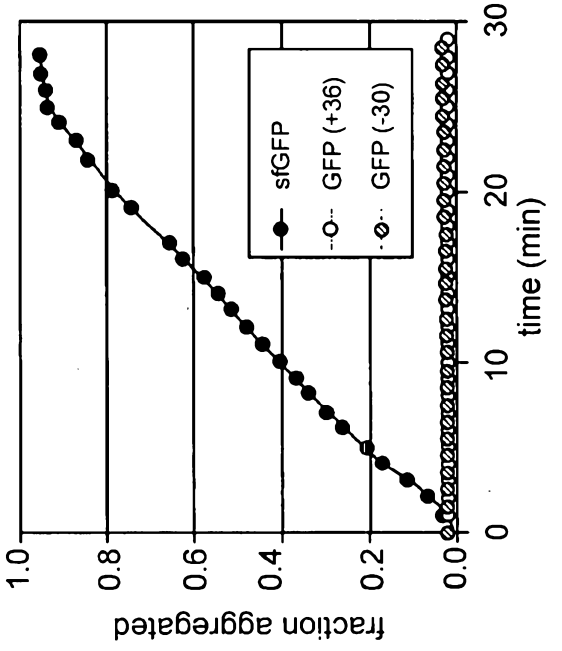


FIG. 3b

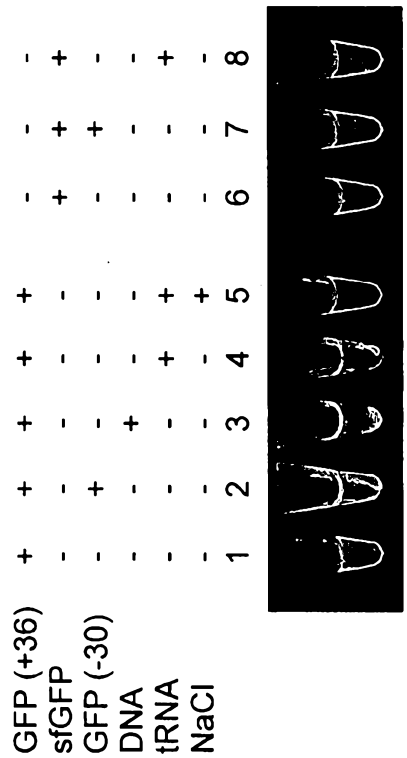


FIG. 3c

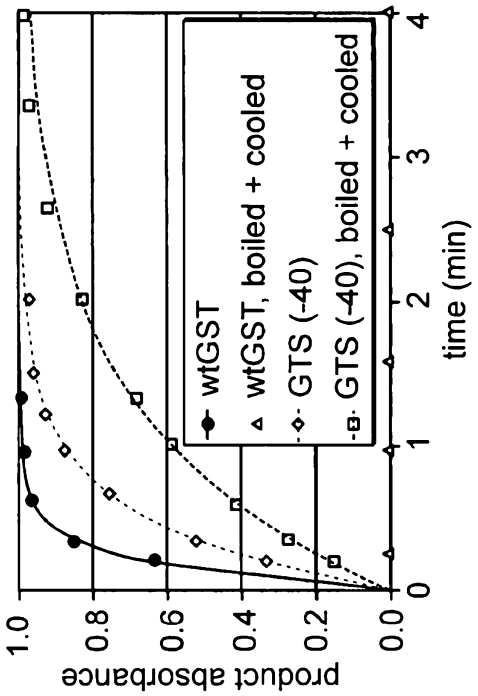


FIG. 3d

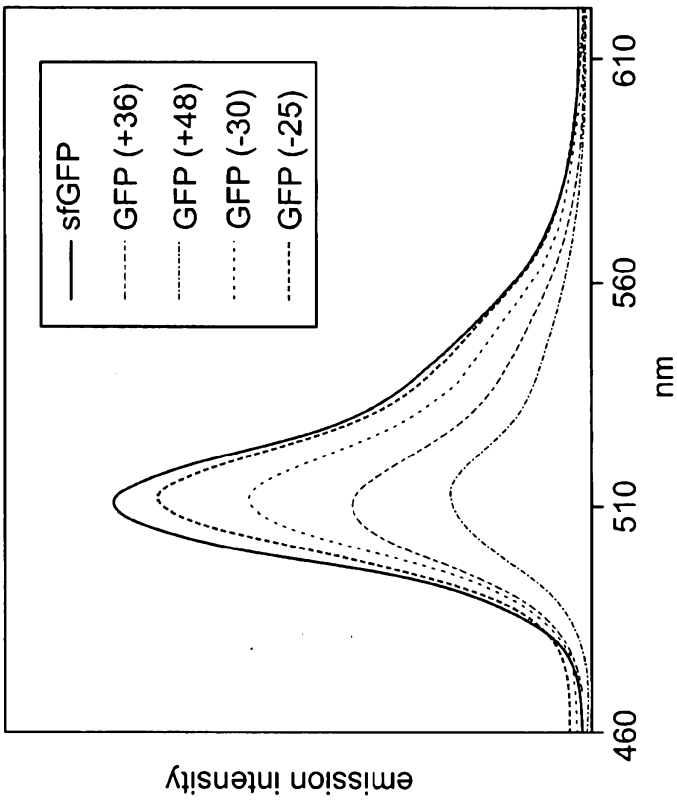

sfGFP
GFP (+36)
GFP (+48)
GFP (-30)
GFP (-25)
460
510
560
610
nm
emission intensity

FIG. 4b

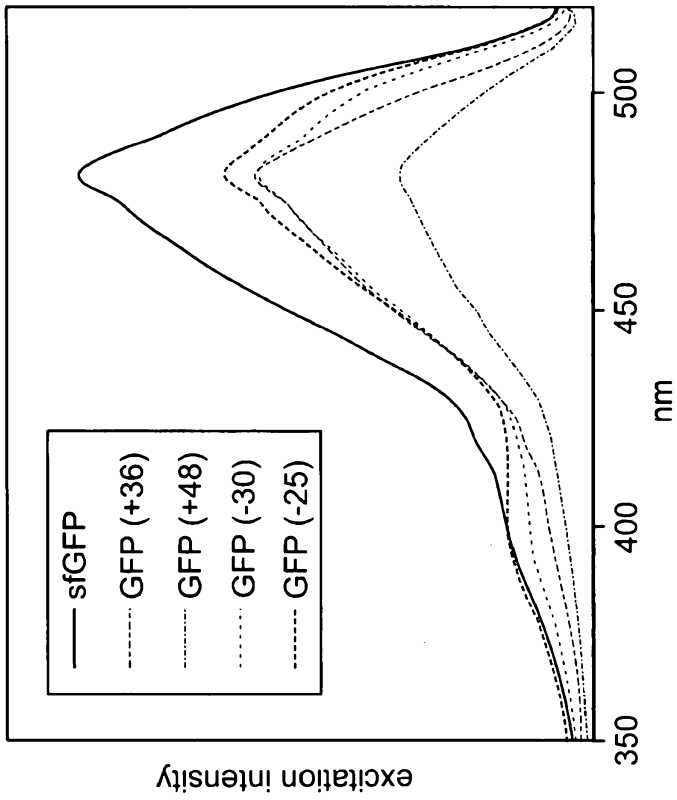
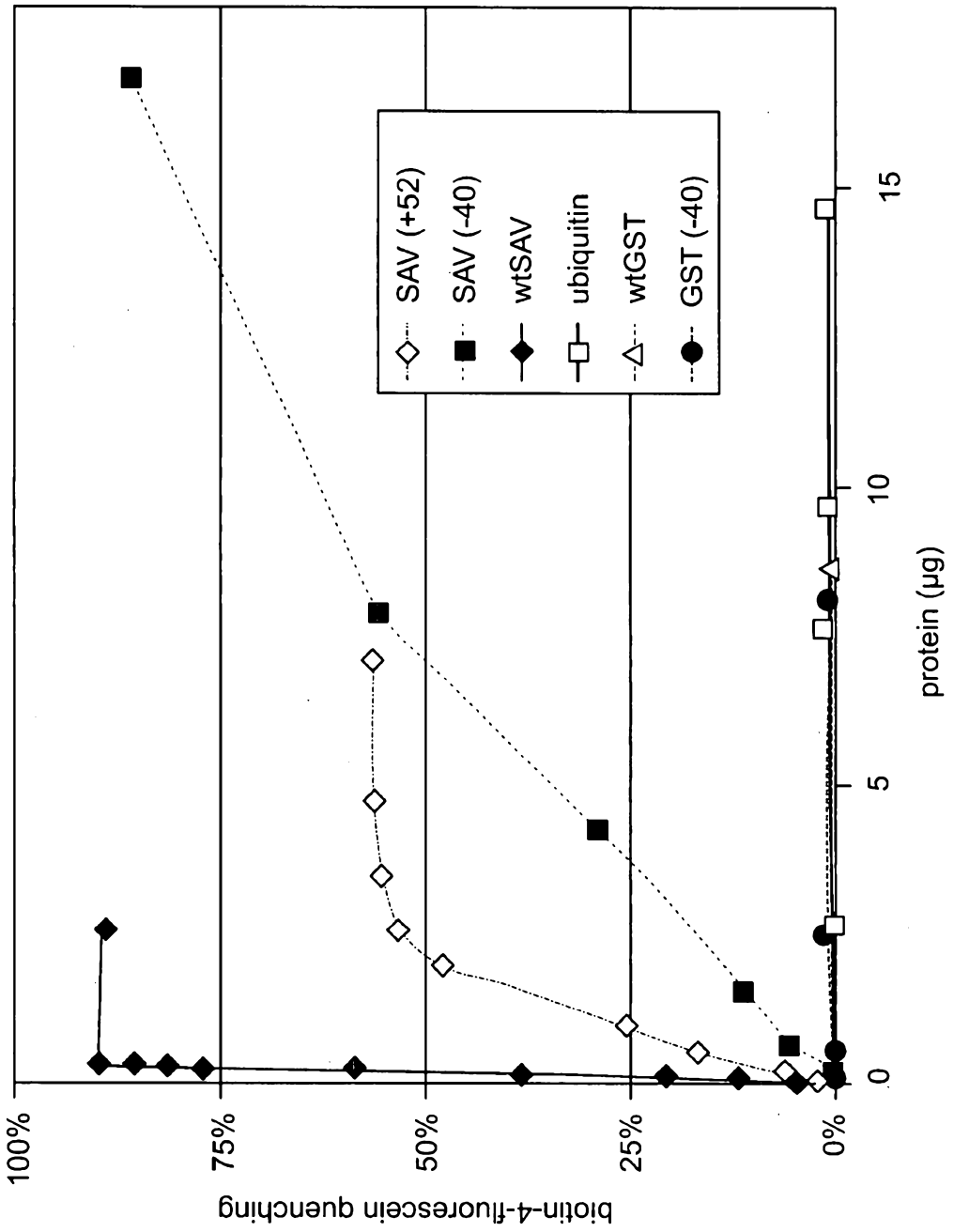

sfGFP
GFP (+36)
GFP (+48)
GFP (-30)
GFP (-25)
350
400
450
500
nm
excitation intensity

FIG. 4a

FIG. 5


biotin-4-fluorescein quenching
100%
75%
50%
25%
0%
0
5
10
15
protein (µg)
SAV (+52)
SAV (-40)
wtSAV
ubiquitin
wtGST
GST (-40)