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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
24 October 2013 (24.10.2013)

WIPO | PCT

(10) International Publication Number
WO 2013/156552 A1

(51) **International Patent Classification**:
C12N 15/00 (2006.01) ***G06F 19/16*** (2011.01)
G01N 33/68 (2006.01) ***G06F 19/18*** (2011.01)

(21) **International Application Number**:
PCT/EP2013/058052

(22) **International Filing Date**:
18 April 2013 (18.04.2013)

(25) **Filing Language**: English

(26) **Publication Language**: English

(30) **Priority Data**:
61/635,208 18 April 2012 (18.04.2012) US

(71) **Applicants**: **VIB VZW** [BE/BE]; Rijvisschestraat 120, B-9052 Gent (BE). **KATHOLIEKE UNIVERSITEIT LEUVEN, K.U.LEUVEN R&D** [BE/BE]; Waaistraat 6 - bus 5105, B-3000 Leuven (BE).

(72) **Inventors**: **ROUSSEAU, Frederic**; Stichelberg 18, B-1702 Groot-Bijgaarden (BE). **SCHYMKOWITZ, Joost**; Attenrodestraat 15, B-3391 Meensel-Kiezegem (BE). **GANESAN, Ashok**; 35/8A LGGS Nagar, Tiruvannamalai 606601 (IN). **SIEKIERSKA, Aleksandra**; Ladderstraat 8, B-1000 Brussel (BE). **DE SMET, Frederik**; Lennekemare 10, B-3020 Winksele (BE). **VAN DURME, Joost**; Oeselgemstraat 33, B-8720 Wakken (BE).

(74) **Common Representative**: **VIB VZW**; Rijvisschestraat 120, B-9052 Gent (BE).

(81) **Designated States** *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, 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, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, 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, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Published:

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

— *with sequence listing part of description (Rule 5.2(a))*

WO 2013/156552 A1

(54) **Title**: MEANS AND METHODS FOR GENERATING IMPROVED PROTEINS

(57) **Abstract**: The present invention provides a general method for the production of protein variants with a reduced aggregation propensity without affecting the thermodynamic stability of the variant with respect to the wild type protein.

MEANS AND METHODS FOR GENERATING IMPROVED PROTEINS

Field of the invention

The present invention relates to the field of protein design and more particularly to methods to improve the design of recombinant proteins. Even more particularly the invention provides a method for generating variants of existing proteins which have a reduced protein aggregation.

Introduction to the invention

Protein aggregation is mediated by short aggregation prone sequences that assemble into intermolecular beta-structures, which form the core of the aggregate. In native conditions, these stretches are buried inside the globular structure of the protein and are hence protected from aggregation by the thermodynamic stability of the fold. Although the vast majority of proteins contain at least one such aggregation prone region, protein aggregation in healthy cells is effectively suppressed by a number of mechanisms, which cooperate to maintain proteostasis (Balch, Morimoto et al. 2008). One of them are gatekeeper residues, strongly enriched at the flanks of the aggregating regions, that slow down the aggregation reaction (Otzen, Kristensen et al. 2000; Richardson and Richardson 2002; Rousseau, Serrano et al. 2006; Monsellier and Chiti 2007). Moreover, molecular chaperones, such as Hsp70, bind to exposed aggregating regions, preventing intermolecular assembly to nucleate (Van Durme, Maurer-Stroh et al. 2009). Finally, protein turnover rates (De Baets, Reumers et al. 2011) and protein expression levels (Tartaglia, Pechmann et al. 2009) are tuned to minimise problems with protein aggregation. During normal ageing, these cellular defence mechanisms have been shown to erode (Kikis, Gidalevitz et al. 2010) and many proteins have been observed to break through the proteostasis boundary in ageing tissues (Lee, Weindruch et al. 2000; Zou, Meadows et al. 2000; Lund, Tedesco et al. 2002; Pletcher, Macdonald et al. 2002; Lu, Pan et al. 2004; Ben-Zvi, Miller et al. 2009; Bishop, Lu et al. 2010), often without apparent adverse effects. On the other hand, aggregation of specific proteins has been convincingly linked to a number of age-related human diseases, including neurodegenerative disorders such as Alzheimer Disease and Parkinson Disease as well as cancer (Xu, Reumers et al. 2011) and metabolic diseases (Ishii, Kase et al. 1996; Soong, Brender et al. 2009). In these cases the aggregation problem is often exacerbated through mutations, which increase the solvent exposure of the aggregation prone regions by thermodynamically destabilising the native structure (Dobson 2004).

However, when proteins are employed for research, therapy or industrial applications, they need to withstand artificial conditions for which evolution has poorly equipped them. Given the ubiquitous nature of aggregation prone sequences in the proteome, it is not surprising that protein aggregation is often observed when proteins are expressed far beyond their normal

concentration in conditions with no or insufficient molecular chaperones. Moreover, once purified the proteins are expected to last far beyond their natural lifetime, allowing the critical nucleating events to start the protein aggregation reaction. Several methods have been developed to reduce the aggregation problem, for example by using cell lines with increased chaperone content (Schlieker, Bukau et al. 2002), by generating fusion proteins with solubilising tags (Zhang, Howitt et al. 2004; Park, Han et al. 2008; Song, Lee et al. 2011) or by careful formulation of buffers (Wang 1999). Another approach would be to adapt the primary sequence to the new requirements through carefully selected mutations. Although this has the disadvantage of altering the protein sequence, this is often not a prohibitive consideration. In the present invention we have developed a rational design strategy, which we designated the Solubis method, which produces reduced aggregating variants of proteins by simultaneously reducing the aggregation tendency of the variant and at the same time preserving the thermodynamic stability and structural integrity. In exemplary embodiments the method employs the FoldX (Schymkowitz, Borg et al. 2005) and TANGO (Fernandez-Escamilla, Rousseau et al. 2004) algorithms to identify selected mutations that render a protein less aggregation prone, while maintaining or even improving its intrinsic stability and function. Specific examples are presented for the generation of variant proteins of industrial utility.

Figure legends

Figure 1: Structure and aggregation propensity of human wild type α-Galactosidase.
(A) Human α-Gal is a homodimer (PDB 3gxp) (Lieberman, D'Aquino J et al. 2009) in which each monomer contains a (β/α)8 domain (central part, yellow and green), where the active site is located and an antiparallel β domain (central part, orange and blue). The structure was visualized with YASARA (Krieger, Koraimann et al. 2002). (B) The intrinsic aggregation propensity of the α-Gal sequence as predicted by the TANGO algorithm reveals three strongly aggregation-prone regions: (1) M_{284}ALWAIMA$_{291}$, (2) L_{347}AWAVAMI$_{355}$ and (3) Y_{365}TIAVAS$_{371}$. The regions predicted by TANGO were indicated in the structure with numbers 1, 2 and 3 and were coloured in red (Figure 1A). (C) Scatter plots representing the results of computational gatekeeper scans for each of the aggregation prone regions of α-Gal (ΔΔG FoldX versus ΔTANGO). For the TANGO region 2 mutation A348R could be identified (green amino acid residue in Figure 1A), whereas for the TANGO region 3 mutations A368R and A368P (green amino acid residue in Figure 1A). (D) A table representing the mutations with a predicted beneficial effect on thermodynamic stability of greater than 2 kcal/mol obtained by FoldX analysis (yellow amino acid residue in Figure 1A).

Figure 2: Effect of the single improving mutation on α-Gal aggregation and activity.
(A) Western blot of size exclusion chromatography fractions (SEC) of wild type α-Gal, selected aggregating mutants and improving mutants in transiently transfected Hela cells. WT α-Gal and single improving mutants eluted in later fractions (12.5-14.5 ml) than the aggregating mutants, corresponding to the active soluble form of the protein. (B) Quantification of the solubility of α-Gal mutants in transiently transfected Hela cells. The bands densities from Western blot of SEC fractions from several experiments were quantified. Fractions from 6.5-10.5 ml elution were considered as insoluble whereas from 12.5-14.5ml as soluble. Wild type α-Gal showed approximately 70% of solubility, whereas aggregating mutants were highly insoluble. Single improving mutants reached around 80-90% of total solubility (with the exception of A368R mutant). Statistical significance was calculated in comparison to wild type α-Gal and means "*" $p< 0.05$ and "****" $p<0.0001$. (C) Quantification of the enzymatic activity of wild type α-Gal and single improving mutations in transiently transfected Hela cells. The activity of the wild type and the mutants was comparably similar with a slight reduction in the case of A368R mutant. (D) A scatter plot of the predicted change of the thermodynamic stability (ΔΔG FoldX) associated with the improving mutations versus the experimental values of enzyme activity. For all of the mutations there is a good correlation between predicted and experimental values.

Figure 3: Effect of the double and triple improving mutation on α-Gal aggregation and activity.
(A) Western blot of size exclusion chromatography fractions (SEC) of wild type α-Gal and double and triple improving mutants in transiently transfected Hela cells. WT α-Gal and improving mutants eluted in later fractions (12.5-14.5 ml) corresponding to the active soluble form of the protein. (B) Quantification of the solubility of α-Gal mutants in transiently transfected Hela cells. The bands densities from Western blot of SEC fractions from several experiments were quantified. Fractions from 6.5-10.5 ml elution were considered as insoluble whereas from 12.5-14.5ml as soluble. Wild type α-Gal and improving mutants showed similar activity reaching from 70 till 90% of total solubility. (C) Quantification of the enzymatic activity of wild type α-Gal and double and triple improving mutations in transiently transfected Hela cells. The activity of the improved mutants was approximately 2 times fold higher than the wild type protein. Statistically significant values in comparison to wild type were indicated with asterisks: "***" $p< 0.001$ and "****" $p<0.0001$. (D) Western blot of the expression level of wild type α-Gal and improved mutants in transiently transfected HeLa cells. All the improving mutants expressed significantly higher than the wild type α-Gal. This unequal expression of the constructs explains the differences in the enzymatic activity between the constructs.

Figure 4: Structure and aggregation propensity of yellow florescent protein.

(A) A beta-barrel fold, in which center a chromophore is localized, represents the structure of YFP. (B) The intrinsic aggregation propensity of YFP predicted by the TANGO algorithm showed three strongly aggregation-prone regions. The regions predicted by TANGO were indicated in the structure with numbers 1, 2 and 3 and coloured in red (Figure 4A). (C) Scatter plots representing the results of computational gatekeeper scans for each of the aggregation prone regions of YFP (ΔΔG FoldX versus ΔTANGO). For the TANGO region 2 mutations Y151E, M153K and A154P were identified (green amino acids residues in Figure 4A), whereas for the TANGO region 3 – T225E and A227D (green amino acids residues in Figure 4A).

Figure 5: A model for troublesome fusion proteins based on YFP

(A) Our model for aggregating fusion proteins that was established by fusing to N terminus of YFP an aggregating peptide LLRLTGW (SS7-YFP). (B) Images of wild type YFP, SS7-YFP and SS7-YFP Y151E transiently expressed in Hela cells. YFP was equally distributed in the whole cell, whereas SS7-YFP formed bright aggregates (highlighted with red arrows in the image) in the cytoplasm. Solubis mutant SS7-YFP Y151E, similarly to WT YFP, was evenly distributed in the whole cell and the aggregates formation was significantly reduced as compared to SS7-YFP. The nuclei were visualized with DAPI.

Figure 6: Aggregation levels and single cell data of SS7-YFP model fusion protein and its improving mutations in HeLa and U2OS cells. (A) A bar chart showing the effect of the improving mutations on the aggregation rate in transiently transfected Hela cells. SS7-YFP expression resulted with aggregation in 50% of cells with the baseline aggregation of YFP on 10% level. Mutation Y151E completely abrogated the aggregation of SS7-YFP model construct. Mutation M153K did not have any effect on the aggregation level, whereas A154P mutant affected the chromophore center resulting in poor construct expression. Mutations T225E and A227D significantly reduced the aggregation till level 35% and 38% respectively and in combination till 30%. Combining single mutations from 2nd regions with single mutations from 3rd region gave a significant reducing effect from 35% (mutants M153K/A227D and A154P/A227D) up to 44% (mutants M153K/T225E and A154P/T225E). Asterisks over the error bars indicating the p-value mean: "*" $p< 0.05$, "**" $p< 0.01$, "***" $p< 0.001$ and "****" $p<0.0001$.

(B) A bar chart showing the effect of the improving mutations on the aggregation rate in transiently transfected U2OS cells. SS7-YFP model construct aggregated in 65% and YFP alone in 16% examined cells. The highest significant reduction of the aggregation level was observed in the case of Y151E mutant. Mutation M153K had a very minor effect on the aggregation rate. Mutation A227D resulted in a significant decrease of the aggregation rate (till

50%), which was even more pronounced when combined with T225E mutant (till 40%). Combining mutations from different aggregation-prone regions resulted in the strongest decrease of aggregation for mutants M153KT225E and M153KA227D (till 45%). Asterisks over the error bars indicating the p-value mean: "*" p< 0.05, "**" p< 0.01, "***" p< 0.001 and "****" p<0.0001.

(C) A plot showing number of aggregates per cell for YFP, SS7-YFP and selected improving mutations in transiently transfected HeLa cels. A decrease in number of aggregates for per single cell in comparison to SS7-YFP could be observed for mutants Y151E, T225EA227D, M153KA227D and A154PA227D.

(D) A plot illustrating the total area of the aggregates per cell for YFP, SS7-YFP and selected improving mutations in transiently transfected HeLa cells. The smaller aggregates size was observed in the case of mutants Y151E, T225EA227D and M153KA227D. (E) A plot showing number of aggregates per cell for YFP, SS7-YFP and selected improving mutations in transiently transfected U2OS cells. Mutations Y151E and T225EA227D resulted in reducing the number of aggregates per cell. (F) A plot illustrating the total area of the aggregates per cell for YFP, SS7-YFP and selected improving mutations in transiently transfected U2OS cells. The smaller aggregates size was observed in the case of mutants T225EA227D, Y151E and M153KA227D.

Figure 7: Results from Solubis analysis run on selected protein structures.

(A) A heat map version of the MASS plot (for Mutant Aggregation & Stability Spectrum) for more than 70.000 mutations generated during the Solubis analysis for the 585 high quality PDB structures. The X-axis shows the change in TANGO score, the Y-axis shows the the change in FoldX structural stability associated with the mutation and the colour code indicates the frequency of occurring mutations with that ΔTANGO and $\Delta\Delta$G profile. The mutations that maximally reduce aggregation while preserving thermodynamic stability are indicated on the map with a box.

(B) A bar chart demonstrating the percentage of Solubis mutations per structural class: (α) all α helical, (β) all β sheet, (α/β) mixed α helix and β sheet and (α+β) combined α helix and β sheet. The percentage of mutations identified using Solubis is very similar for all SCOP classes.

(C) A bar chart showing the percentage of different domains within SCOP classes: with at least five Solubis mutations. All SCOP classes have a similar percentage of Solubis mutants per class.

Figure 8: Differential scanning calorimetric (DSC) data which show that the mutant PA (S588L/T605E) aggregates at higher temperature compared to the wild type PA.

Figure 9: Mouse macrophage cells are treated with wild type PA and lethal factor and compared to the treatment of mutant PA (S588L/T605E) and lethal factor. It is shown that not only the biological activity is preserved for the mutant PA but that the activity is also higher.

Figure 10: Effect of the toxicity of wild type PA and lethal factor compared to mutant PA (S588L/T605E) and lethal factor after heat stress treatment on 45°C. It is shown that the mutant PA has a significant preservation of the activity when subjected to heat stress.

Detailed description of the invention

The present invention will be described with respect to particular embodiments and with reference to certain drawings but the invention is not limited thereto but only by the claims. Any reference signs in the claims shall not be construed as limiting the scope. The drawings described are only schematic and are non-limiting. In the drawings, the size of some of the elements may be exaggerated and not drawn on scale for illustrative purposes. Where the term "comprising" is used in the present description and claims, it does not exclude other elements or steps. Where an indefinite or definite article is used when referring to a singular noun e.g. "a" or "an", "the", this includes a plural of that noun unless something else is specifically stated.

Furthermore, the terms first, second, third and the like in the description and in the claims, are used for distinguishing between similar elements and not necessarily for describing a sequential or chronological order. It is to be understood that the terms so used are interchangeable under appropriate circumstances and that the embodiments of the invention described herein are capable of operation in other sequences than described or illustrated herein.

The following terms or definitions are provided solely to aid in the understanding of the invention. Unless specifically defined herein, all terms used herein have the same meaning as they would to one skilled in the art of the present invention. Practitioners are particularly directed to Sambrook et al., Molecular Cloning: A Laboratory Manual, 2nd ed., Cold Spring Harbor Press, Plainsview, New York (1989); and Ausubel et al., Current Protocols in Molecular Biology (Supplement 47), John Wiley & Sons, New York (1999), for definitions and terms of the art. The definitions provided herein should not be construed to have a scope less than understood by a person of ordinary skill in the art.

In the present application, the standard one letter notation of amino acids will be used. Typically, the term “amino acid” will refer to “proteinogenic amino acid”, i.e. those amino acids that are naturally present in proteins.

The term "sequence identity" as used herein refers to the extent that sequences are identical on a nucleotide-by-nucleotide basis or an amino acid-by-amino acid basis over a window of comparison. Thus, a "percentage of sequence identity" is calculated by comparing two optimally aligned sequences over the window of comparison, determining the number of positions at which the identical nucleic acid base (e.g., A, T, C, G, I) or the identical amino acid residue (e.g., Ala, Pro, Ser, Thr, Gly, Val, Leu, Ile, Phe, Tyr, Trp, Lys, Arg, His, Asp, Glu, Asn, Gln, Cys and Met) occurs in both sequences to yield the number of matched positions, dividing the number of matched positions by the total number of positions in the window of comparison (i.e., the window size), and multiplying the result by 100 to yield the percentage of sequence identity. For the purposes of the present invention, "sequence identity" will be understood to mean the "match percentage" calculated by the DNASIS computer program (Version 2.5 for windows; available from Hitachi Software engineering Co., Ltd., South San Francisco, Calif., USA) using standard defaults as used in the reference manual accompanying the software. "Similarity" refers to the percentage number of amino acids that are identical or constitute conservative substitutions. Similarity may be determined using sequence comparison programs such as GAP (Deveraux et al. 1984, Nucleic Acids Research 12, 387-395). In this way, sequences of a similar or substantially different length to those cited herein might be compared by insertion of gaps into the alignment, such gaps being determined, for example, by the comparison algorithm used by GAP.

The terms ‘aggregation nucleating regions’ or ‘aggregation prone regions’ or ‘beta-aggregation regions’ or ‘self-association regions’ are equivalent and refer to regions identified in proteins which are responsible for inducing the aggregation of the proteins. In what follows the sequence constraints of these regions are further clarified. Mutational studies of the kinetics of aggregation of full-length proteins revealed simple correlations between aggregation and physico-chemical properties such as β-sheet propensity, hydrophobicity and charge. This stimulated the development of computer algorithms that can identify said aggregation nucleating regions in the amino acid sequence of a protein. One of these is the Zyggregator algorithm of Dobson et al. (Pawar et al., *J Mol Biol* **350:** 379–392 (2005)), which identifies aggregation-prone sequences by comparing the aggregation propensity score of a given amino acid sequence with an average propensity calculated for a set of sequences of similar length. The statistical mechanics algorithm TANGO (Fernandez-Escamilla et al., *Nat Biotechnol* **22**:1302–1306 (2004)), on the other hand, balances the physico-chemical

parameters mentioned above, supplemented by the assumption that an amino acid is fully buried in the aggregated state: this means it becomes fully desolvated and entropically restricted. From an input sequence, TANGO generates an extensive sample of fragments for which competing structural propensities, such as helix or hairpin formation, are considered. All the fragments are then balanced in a global partition sum, which allows the identification of sequence regions that predominantly form aggregates. The TANGO algorithm has an accuracy of more than 90% for a set of 176 experimentally validated peptides (Fernandez-Escamilla et al., *Nat Biotechnol* **22**:1302–1306 (2004)). Importantly, both the Zyggregator algorithm and TANGO perform well for peptides and denatured proteins. For globular proteins, a partly folded molecule can either refold to the native state or misfold into an aggregated state. As a result, both reactions are in competition and a precise understanding of the kinetics is essential to predict the final outcome in terms of folding or misfolding/aggregation. Hence, in the context of the present invention, it is important to identify sequences in globular proteins that kinetically favour the reduction of aggregation. The Tango algorithm has been described in more detail elsewhere (particularly Fernandez-Escamilla et al., Nat. Biotechnol. 22:1302-1306, 2004, especially the Methods section on pages 1305 and 1306 are herein specifically incorporated by reference. See also the Supplementary Notes 1 and 2 of the same article for further details on the methods and the data sets used for the calibration and the testing of the TANGO algorithm. Briefly, to predict aggregation nucleating regions of a protein (or polypeptide), TANGO simply calculates the partition function of the phase-space. To estimate the aggregation tendency of a particular amino acid sequence, the following assumptions are made: (i) in an ordered beta-sheet aggregate, the main secondary structure is the beta-strand. (ii) the regions involved in the aggregation process are fully buried, thus paying full solvation costs and gains, full entropy and optimizing their H-bond potential (that is, the number of H-bonds made in the aggregate is related to the number of donor groups that are compensated by acceptors. An excess of donors or acceptors remains unsatisfied). (iii) complementary charges in the selected window establish favorable electrostatic interactions, and overall net charge of the peptide inside but also outside the window disfavors aggregation. TANGO can be accessed on the World Wide Web.

A high Tango score of a sequence stretch typically corresponds to a sequence with high (and kinetically favourable) beta-aggregation propensity. In the present invention the sequence space of “the lowest tango-scoring sequences” of a particular variant of a protein – generated in the context of the present invention - are preferred.

It can be calculated that more than 80 % of all proteins have at least one aggregation nucleating segment within their primary sequence. As a result protein aggregation is often encountered when proteins are overexpressed or recombinantly produced. Moreover

aggregation represents a major liability with respect to the immunogenicity of biotherapeutics. However, redesigning globular proteins to eliminate aggregation is not a straightforward task as most aggregation nucleating sequences are part of the hydrophobic core and therefore difficult to mutate without disrupting protein structure and function. In the present invention we have developed a minimal redesign method, termed as SolubiS, to abrogate aggregation by silencing aggregation nucleating sequences through the introduction of specific mutations, which are selected to maximally reduce the intrinsic aggregation propensity of the sequence whilst preserving thermodynamic stability of the functional protein. The present method allows sifting hundreds to thousands of mutations, simultaneously evaluating protein aggregation and stability, typically producing 1 to 5 appropriate mutations per target protein. In the appended examples the method is exemplified for three relevant proteins: i) human α-Galactosidase, which is currently used in enzyme replacement therapy for Fabry disease, ii) Yellow Fluorescent Protein (YFP), an important research biologic and iii) Anthrax Protective Antigen (PA) which is a key toxin secreted by *Bacillus anthracis*. In each case we identified mutants that displayed a marked reduction in protein aggregation upon overexpression whilst preserving both stability and functionality. Furthermore an *in silico* analysis of a non-redundant set of 443 high-resolution crystallographic structures shows that 75 % of globular proteins with a high aggregation propensity are amenable to our redesign strategy, showing that the invented method is broadly applicable for the improvement of globular proteins.

Accordingly the invention provides in a first embodiment a method for the production of a reduced aggregating variant of a protein that has at least two aggregation nucleating regions, said method comprising the following steps: a) determining the aggregation nucleating region in said protein, b) generating a list of variant proteins wherein each variant protein has on at least one amino acid position in said determined aggregation nucleating regions a changed amino acid to either R, K, E, D or P, c) calculating for each of said variants the predicted aggregation score and the predicted change in thermodynamic stability with respect to the wild type protein, and d) producing a reduced aggregating variant, which is derived from said list, wherein said variant has at the same time a maximally reduced aggregation score, a maximal preservation of thermodynamic stability and no structural changes with respect to the wild type protein.

The term 'reduced aggregating variant of a protein' refers to a variant protein (or a mutant protein) which has with respect to the wild type protein (i.e. the naturally occurring protein) a 10%, a 20%, a 30%, a 40%, a 50%, a 60%, a 70%, an 80%, a 90% or even higher % of reduced aggregation. Non-limiting methods for measuring a reduced aggregation are herein further provided in the appended examples. The term 'aggregating nucleating regions' is

herein described before and non-limiting examples of methods are described how to identify (or to determine which is an equivalent word) 'aggregating nucleating regions' in a protein. In a particular embodiment the protein from which it is started to develop a reduced aggregating variant has at least two, at least three, at least four or more aggregating nucleating regions. The aggregating nucleating regions which are identified are 'in silico' modified wherein at least one of the amino acid positions present in said aggregating nucleating region are changed towards either an R, a K, an E, a D or a P. Thus a list of variant proteins is generated wherein each amino acid position of the aggregating nucleating regions is changed in five different amino acids (i.e. an R, a K, an E, a D or a P). Thus for each specific amino acid position present in said aggregating nucleating region five different variants are generated. In another particular embodiment at least two of the amino acid positions present in said aggregating nucleating region are changed towards either an R, a K, an E, a D or a P. In another particular embodiment at least two of the amino acid positions can be changed towards an R, a K, an E, a D or a P and said at least two amino acid positions are modified in two different aggregating nucleating regions of the protein.

In a particular embodiment a reduced aggregating variant of a protein comprises at least 1 mutation in one of its aggregation nucleating regions. In another particular embodiment a reduced aggregating variant of a protein comprises at least 2 mutations in in one of its aggregation nucleating regions. In yet another particular embodiment a reduced aggregating variant of a protein comprises at least 2 mutations, each mutation in a different aggregation nucleating region.

For each variant protein generated the predicted aggregation score is calculated by use of algorithms described herein before. In addition, for each variant protein generated the predicted thermodynamic stability is calculated using methods described herein (e.g. the FoldX algorithm). Other algorithms to calculate the thermodynamic stability are known to the person skilled in the art. A non-limiting example to determine the thermodynamic stability is the molecular modeling software Rosetta (Das R and Baker D (2008) *Annual Rev. Biochemistry* 28, 363-382).

The present method hinges on the availability of the three dimensional structure of the protein one aims to modify into a reduced aggregating variant protein. Therefore in the method of the invention the most optimal reduced aggregating variant protein which is produced, is derived from the generated list of all the variants, and needs to have at the same time a maximally reduced aggregation score, a maximal preservation of the thermodynamic stability and in addition has no structural predicted changes with respect to the wild type protein. The appended examples show that only a very limited amount of variant proteins are produced which fulfill the above conditions. In a particular embodiment 1, 2, 3, 4, 5, 6, 7, 8, 9 or 10

variant proteins are produced with the methods of the invention. In another particular embodiment less aggregating variants of a protein which are produced by introducing a mutation in one aggregation nucleating region can be combined with variants which are produced by introducing a mutation in another aggregation nucleating region, i.e. by combining two identified separate mutant variants into one new variant. Specific examples are provided in the appended examples.

In yet another embodiment the invention provides a method for the production of a reduced aggregating variant of a protein that has at least two aggregation nucleating regions, said method comprising the following steps: a) determining the aggregation nucleating region in said protein, b) generating a list of variant proteins wherein each variant protein has on at least one amino acid position in said determined aggregation nucleating regions a changed amino acid to either R, K, E, D or P, c) calculating for each of said variants the predicted aggregation score and the predicted change in thermodynamic stability with respect to the wild type protein, d) introducing at least one compensatory mutation outside at least one of the aggregation nucleating regions with the purpose of preserving the thermodynamic stability of said protein and e) producing a reduced aggregating variant, wherein at least one mutation is derived from said list and wherein additionally also at least one compensatory mutation is present in said variant, wherein said variant has at the same time a maximally reduced aggregation score, a maximal preservation of thermodynamic stability and no structural changes with respect to the wild type protein.

The term “compensatory mutation” refers to an amino acid change in the protein which is introduced in addition to at least one change of one of the amino acids of the aggregation nucleating region into either R, K, E, D or P. Often this change in the aggregation nucleating region of said protein reduces the thermodynamic stability of the resulting variant protein and a compensatory mutation needs to be introduced to compensate for said reduction of thermodynamic stability. Typically, a compensatory mutation is situated outside the aggregation nucleating region which comprises a mutation.

In yet another embodiment the present invention provides a reduced aggregation variant of a protein which has at least two aggregation nucleating regions which is obtainable by a) determining at least two aggregation nucleating region in said protein, b) generating a list of variant proteins wherein each variant protein has on at least one amino acid position in said determined aggregation nucleating regions a changed amino acid to either R, K, E, D or P, c) calculating for each of said variants the predicted aggregation score and the predicted change in thermodynamic stability with respect to the wild type protein, and d) producing a reduced aggregating variant, which is derived from said list, wherein said variant has at the same time

a maximally reduced aggregation score, a maximal preservation of thermodynamic stability and no structural changes with respect to the wild type protein.

In yet another embodiment the present invention provides a reduced aggregation variant of a protein which has at least two aggregation nucleating regions which is obtainable by a) determining at least two aggregation nucleating region in said protein, b) generating a list of variant proteins wherein each variant protein has on at least one amino acid position in said determined aggregation nucleating regions a changed amino acid to either R, K, E, D or P, c) calculating for each of said variants the predicted aggregation score and the predicted change in thermodynamic stability with respect to the wild type protein, and d) introducing a compensatory mutation outside at least one of the aggregation nucleating regions with the purpose of preserving the thermodynamic stability of said protein and e) producing a reduced aggregating variant, which is derived from said list, wherein said variant has at the same time a maximally reduced aggregation score, a maximal preservation of thermodynamic stability and no structural changes with respect to the wild type protein.

The present method offers a variety of possible applications. One application is for example in the field of enzyme replacement therapy. Several proteins can be optimized into reduced aggregating variants. Non-limiting examples of such proteins are for example glucocerebrosidase, alfa-galactosidase A, alpha-galactosidase, alpha-L-iduronidase and GlcNAc phosphotransferase. Yet another application is the generation of so called 'biobetters' which are improved (i.e. reduced aggregating variants of existing biological). Non-limiting examples are known in the art as 'biobetters' which can be produced from interferon-beta, insulin, granulocyte macrophage stimulating factors and members of the interleukin family. Yet another application is in the field of affinity chromatography. For example reduced aggregation binding proteins can be designed which have a reduced aggregation in apolar solvents. Yet another application is in the field of agrobiotechnology. It can be envisaged that certain crucial proteins suffer from aggregation when crops encounter conditions of abiotic stress such as heat, cold or salt stress. By generating variants of such crucial proteins which are less prone to aggregation, novel transgenic crops can be generated which are resistant to abiotic stress. Yet another application is in the field of enzymology. Novel enzyme variants can be produced with the current method which are less prone to aggregation and hence remain active for a longer period than the corresponding wild type enzyme. Yet another application is in the field of protein production. Less aggregation prone variants will show an increased expression level and makes the downstream purification processing easier.

In addition, the present invention also provides specific examples. In one specific example reduced aggregation variants of alpha-galactosidase A are generated. The amino acid sequence of alpha-galactosidase A is depicted in SEQ ID NO: 1 (Alpha-Gal A (genbank identifier NP_000169)).

Thus in a specific embodiment a reduced aggregation variant of the alpha-galactosidase A protein (wild type is depicted in SEQ ID NO: 1) is provided which is selected from the list consisting of i) A348R/A368R, ii) A348R/A368P, iii) A348R/A368R/S405L and iv) A348R/A368P/S405L.

Accordingly, in a specific embodiment the invention provides a reduced aggregation variant of the alpha-galactosidase A protein (wild type is depicted in SEQ ID NO: 1) which is selected from the list consisting of i) A348R/A368R, ii) A348R/A368P, iii) A348R/A368R/S405L and iv) A348R/A368P/S405L for the treatment of Fabry disease.

In yet another specific example reduced aggregation variants of yellow fluorescent protein, citrine variant, are generated. The amino acid sequence of yellow fluorescent protein, citrine variant is depicted in SEQ ID NO: 2 (Yellow Fluorescent Protein (YFP) citrine variant)

In another specific embodiment a reduced aggregation variant of the yellow fluorescent protein (wild type sequence is depicted in SEQ ID NO: 2) selected from the list consisting of i) M153K/T225E, ii) M153K/A227D, iii) Y151E, iv) M153K/A227D and v) T225E/A227D.

In yet another specific example a reduced aggregation variant of *Bacillus anthracis* Protective Antigen is provided. The amino acid sequence of the *Bacillus anthracis* Protective Antigen is depicted in SEQ ID NO: 3.

In another specific embodiment a reduced aggregation variant of the *Bacillus anthracis* Protective Antigen (wild type sequence is depicted in SEQ ID NO: 3) is S588L/T605E.

It is to be understood that although particular embodiments, specific configurations as well as materials and/or molecules, have been discussed herein for cells and methods according to the present invention, various changes or modifications in form and detail may be made without departing from the scope and spirit of this invention. The following examples are provided to better illustrate particular embodiments, and they should not be considered limiting the application. The application is limited only by the claims.

Examples

1.The Solubis method

Protein aggregation nucleating regions can be identified using specialised software, which have been reviewed elsewhere (Belli, Ramazzotti et al. 2011). In the present invention we employed the statistical thermodynamics algorithm TANGO (Fernandez-Escamilla, Rousseau et al. 2004) to detect aggregation nucleating regions in the target sequence. We selected proteins for which high-resolution crystallographic structures are available so that the topological position of the aggregating regions can be visualised using atomic structure viewers. The structural information also enables the use of an atomic force field to eliminate mutations that thermodynamically destabilise the native structure. Again, methods for predicting the mutational effects on protein stability have been reviewed elsewhere (Chen and Shen 2009) and the results shown here were obtained with the FoldX forcefield (Schymkowitz, Borg et al. 2005). Two classes of mutations can be designed to reduce protein aggregation: (i) Mutations that eliminate or strongly reduce the intrinsic aggregation propensity of the sequence, thereby slowing down the aggregation reaction and (ii) Mutations that stabilise the interaction of the aggregating region with the rest of the structural domain in which it resides, thus providing additional protection from solvent exposure. In the ideal case, mutations can be identified that unify both goals, but often a combination of mutations is required to maximally suppress aggregation. Reduction of intrinsic aggregation is usually achieved by the introduction of aggregation breaking residues, called gatekeepers (Rousseau, Serrano et al. 2006; Monsellier and Chiti 2007), in the aggregation nucleating sequences. Since the gatekeepers consist of the charged amino acids (Arg, Lys, Glu, Asp) and proline, most often they need to be placed in exposed regions in order not to disturb the hydrophobic core of the protein. The Solubis method thus consists in systematically mutating the residues residing within a mostly structurally buried aggregation prone region (or TANGO zone) to each of the gatekeeper residues and calculating the consequent change in TANGO score as well as the change in the thermodynamic stability of the protein using FoldX (this process will be called gatekeeper scan in what follows). In the case where the gatekeeper residues can only be placed by compromising the thermodynamic stability of the protein, we scan for compensatory mutations using the FoldX algorithm. During the whole process, we avoided mutation of residues that are known to be involved in catalysis or binding.

2. Generation of less aggregating variants of alpha-galactosidase A (alpha-Gal)

Human α-Gal is a lysosomal hydrolase that cleaves neutral glycosphingolipids with terminal α-linked galactosyl moieties, mainly globotriaosylceramides (Gb3). Deficiency in the activity of this glycoprotein results in accumulation of the enzyme's substrates, leading to Fabry disease

(FD) (OMIM 301500), a metabolic X-linked inherited lysosomal storage disorder (LSD) (Brady, Gal et al. 1967; Eng and Desnick 1994). The structure is a homodimer in which each monomer contains a (β/α) domain (Figure 1A, central part, yellow and green parts) where the active site is found, and an antiparallel β domain (Figure 1A, orange and blue). The regions predicted by TANGO (Figure 1B) to be highly aggregation prone (indicated with 1, 2 and 3 and coloured in red in Figure 1A, upper left), cluster in the β-domain and the interface between the domains. In particular, region 3 is likely to be at risk of nucleating aggregation from visual inspection of the structure given its edge position in the beta-sheet. The results of computational gatekeeper scan of each of the aggregation prone regions of α-Gal is shown as a MASS-plot (Mutant Aggregation & Stability Spectrum), i.e. a scatter plot (Figure 1C) of the change in thermodynamic stability (ΔΔG values calculated by FoldX in kcal/mol) versus change in the intrinsic aggregation propensity (values calculated by TANGO, range between 0 and 100 per amino acid residue) associated to each aggregation nucleating region. These plots allow to easily identify ideal mutations, with large negative values on both axes, i.e. mutations that reduce the intrinsic aggregation propensity while increasing the thermodynamic stability of an aggregation nucleating region. For TANGO region 1, no such mutations could be identified, owing mainly to the fact that this region is completely buried inside the tightly packed domain interface. TANGO regions 2 and 3 display one such mutation each (A348R and A368P), as well as the possibility for stabilising region 3 with little predicted effect of intrinsic aggregation (A368R) (Figure 1C).

In addition, we performed an exhaustive mutation scan throughout the β domain and we listed the mutations with a predicted beneficial effect on thermodynamic stability of greater than 2 kcal/mol in the table set in Figure 1D. The single most stabilising mutation (S405L, $\Delta\Delta G_{FoldX}$=-3.34 kcal/mol) tightens the interaction of the edge beta strand (the site of TANGO region 3) with the rest of the domain. Based on this rationale, we decided to investigate the effect of these mutations in an experimental setup. For this purpose the full length cDNA sequence encoding human α-Gal was cloned into the pcDNA4/TO/myc-His vector and the individual mutations were generated using site directed mutagenesis. For comparison, we also generated D165V and A288D mutants, which are associated with Fabry disease and were previously found to be particularly aggregation prone. We transiently overexpressed the entire set of mutant and the wild type proteins in HeLa cells and observed the solubility of α-Gal in the lysates using size exclusion chromatography on a S200 column, followed by western blot (see material and methods for experimental details). As is shown in Figure 2A, the wild type protein elutes mainly as a dimer but also shows faint bands in the SEC fractions that elute to volumes corresponding to large molecular sizes (exclusion limit of this column is 600 kDa), consistent with partial misfolding and aggregation of wild type a-Gal. As expected, the disease

associated mutations D165V and A288D show a significant increase of these high molecular weight assemblies. The individual mutations selected by SolubiS to reduce aggregation show a modest decrease in the high-molecular fractions when compared to wild type a-Gal, that is most pronounced for A348R mutant (Figure 2A). Quantification of the solubility revealed that aggregating mutants (D165V and A288D) were highly insoluble (<10%), whereas single improving mutants reached around 80-90% of total solubility in comparison to 65% for wild type (Figure 2B). We also determined the enzymatic activity α-Gal by following the conversion of the fluorogenic substrate 4-methylumbelliferyl-α-D-galactopyranoside (4-MU-α-Gal) over time. Consistently, single improving mutants showed similar activity in comparison to wild type (only slightly reduced in the case of A368R mutant) (Figure 2C) suggesting that our selected mutations do not affect the active site of the enzyme and therefore do not interfere with its enzymatic function. Interestingly, plotting experimental values of enzyme activity versus predicted change in thermodynamic stability showed very good correlation (Figure 2D) showing that the predicted increased stability of the protein actually correlates to improved activity.

Overall, the effects of single SolubiS mutants show a decrease of misfolding and aggregation, an improved solubility while leaving enzymatic activity unharmed. The fact that the observed improvements are overall modest is explained by the fact that a-Gal possesses 3 aggregation nucleating regions: improving one region by a single mutation leaves it susceptible to aggregation by the other regions. It is therefore expected that targeting several zones in parallel by multiple mutants should have a synergistic effect on the solubility and enzymatic activity of a-Gal.

In order to determine the best combinations of mutations, we generated several double (A348R/A368P and A348R/A368R) and triple mutants (A348R/A368P/S405L and A348R/A368R/S405L) consisting of the single mutations in TANGO region 2 and TANGO region 3 as well as the stabilizing mutant S405L. Interestingly we observed a significant increase in the enzymatic activity for all mutants compared to wild type or the single mutants (Figure 3C). Such an increase in activity could be caused by an increase in intrinsic activity or by more efficient protein folding resulting in higher expression of native a-Gal. When investigated by western blot, it appeared that, when equal amounts of plasmid DNA were transfected, this resulted in increased expression levels between the constructs (Figure 3D) consistent with an increase in the protein folding efficiency and a more stable (less degradable) protein configuration, leading into more enzymatic activity. This is further confirmed by SEC fractionation: the double and triple mutants result in a higher yield of low molecular weight a-Gal. This is particularly striking for mutant A348R/A368R/S405L for which higher molecular forms are almost not detectable. Together these data show that the rational

redesign of a-Gal by SolubiS is able to generate double and triple mutants optimizing enzymatic activity by displaying improved foldability and expression.

3. Generation of less aggregating variants of Yellow Fluorescent Protein (YFP)

Fusion proteins often display loss of proper folding after fusion to certain targets resulting in mislocalisation or functional inactivation. In order to investigate if our protein improvement method could reduce this problem, we selected the Aequorea Yellow Fluorescent Protein (YFP) citrine, a bright intrinsically fluorescent protein with a known high-resolution atomic structure (Griesbeck, Baird et al. 2001). The protein folds into the typical beta-barrel structure with a chromophore running through the centre which is formed by the cyclisation of the backbone of residues 65-67 to form an imidazolidone structure (Figure 4A). Analysis with TANGO reveals three aggregation prone regions, indicated in red in the structure (Figure 4B) and gatekeeper scans with FoldX and TANGO of the three regions (Figure 4C) revealed possible improving mutations in each TANGO region. As one of the TANGO regions corresponds to the chromophore center and it is known that mutation near this region alters the spectral properties of the fluorescent protein, we avoided mutations in this site and focussed on optimising the two remaining regions. Given the fact that the chromophore region is fully buried inside the beta barrel lends further support to this strategy. Hence, we generated Y151E, M153K and A154D in TANGO region 2 and T225E and A227D in TANGO region 3. Though the latter region has a relatively low TANGO score, it has a C-terminal rendering it susceptible to breathing motions, thereby probably facilitating aggregation. In order to mimic YFP misfolding by fusion we constructed an N-terminal fusion of an aggregating peptide (LLRLTGW) to citrine (Figure 5A): in effect this models the effect of a strongly aggregating protein fully exposing an aggregation nucleating region. Quenching of fluorescence by aggregation has previously been used to screen for soluble Alzheimer's b-peptide variants. Here we keep the aggregating peptide as a constant handicap and aim at increasing the capacity of YFP to cope with this additional burden. Normally, YFP expressed in mammalian cells distributes equally throughout the whole cell, however combined with the aggregating peptide it forms bright aggregates located in the cytoplasm (Figure 5B) validating our fusion model. Using our model for aggregating fusion proteins (named SS7-YFP), we decided to investigate the effect of selected improving mutations on the aggregation rate of YFP. For this purpose, we generated single and double mutants in YFP using site directed mutagenesis and transiently overexpressed them in HeLa and U2OS cell lines. In order to obtain the most complete and detailed information about the aggregation pattern of SS7-YFP and its mutants (not only the number of cells with aggregates but also the number of aggregates per cell and the area of the aggregates) we employed the high-content analysis microscope system (IN Cell Analyzer 2000). Figures 6 A and B summarize the results from automated cell counting in Hela and U2OS cells. In the case of Hela cells, approximately 50% of all the cells formed

aggregates after expression of SS7-YFP (Figure 6A). The baseline for YFP aggregation itself was around 10%. In the second aggregation prone region the mutation Y151E was found to completely abrogate the aggregation. Mutation M153K had no effect whatsoever on the aggregation rate, whereas mutant A154P gave very poor expression levels (and could not be quantified) suggesting that it interferes with the properties of the chromophore center. Mutations in the third aggregation prone region, T225E and A227D, significantly reduced the aggregation rate to 35% and 38% respectively and when combined to 30%. Combining single mutations from the 2nd region with single mutations from the 3rd region gave a significant reducing effect from 35% (mutants M153K/A227D and A154P/A227D) up to 44% (mutants M153K/T225E and A154P/T225E). For U2OS cells we had similar observations as for HeLa cells. The highest reduction (>90%) in aggregation was observed for the Y151E mutant (Figure 6B). The other mutations from the 2nd aggregating region (M153K and A154P) had a very minor effect on the aggregation rate. The A227D mutation from the 3rd region resulted in a significant decrease of the aggregation rate, which was even more pronounced when combined with the T225E mutant. Combining mutations from different regions for M153K/T225E and M153K/A227D gave the highest decrease to 45%. The strongest aggregation reducing effect of the Y151E mutant in two different/independent cell lines suggests the universal nature of the SolubiS method.

Looking at a number of the aggregates per cell, both in Hela (Figure 6C) and U2OS (Figure 6E), we observed a decrease in the proportion of cells with high numbers of aggregates for mutants Y151E, T225E/A227D and M153K/A227D in comparison to the model aggregating protein SS7-YFP. Furthermore, there was a reduction of the total aggregates area per cell for mutants Y151E and T225E/A227D for both Hela (Figure 6D) and U2OS cells (Figure 6F).

Overall, these data demonstrate that extensive gatekeeper scans using TANGO together with an assessment of the effect on thermodynamic stability by FoldX allows to identify the structural features of a globular fold that are amenable to improvement.

4. Generation of a less aggregation variant of the *Bacillus anthracis* Protective Antigen

Anthrax infection caused by *Bacillus anthracis* may be classified based on the portal of entry into the host (cutaneous, gastrointestinal, or pulmonary), and symptoms may include fever with mild to severe systemic symptoms of malaise and headache. In severe forms of anthrax, general toxemia with shock, sepsis, and death may occur. The major virulence factor of *B. anthracis* consists of three proteins, edema factor, protective antigen (PA), and lethal factor (LF). The combination of PA and LF produces lethal toxin (LeTx) that is lethal in several animal models including mice. Recombinant PA (rPA) is currently being explored as a vaccine candidate but the protein suffers from poor stability and efficacy. Two aggregation prone regions were identified in the PA protein (one in domain d3 and another one in domain d4).

The sequence comprising the TANGO zone (underlined in NATNIYTVLDKIK (SEQ ID NO: 4)) was based for the generation of a list of mutants. Mutant T605E was selected for introducing the compensatory mutation S588L for amino acid sequence numbering see SEQ ID NO: 3). Differential scanning calorimetry (DSC) shows in Figure 8 that the mutant PA (S588L/T605E) only starts to aggregate at higher temperature as compared to the wild type PA.

In a next step we studied the in vitro biological activity of the mutant PA (S588L/T605E). Thereto murine macrophage cells were treated with different concentrations of wild type PA and mutant PA (S588L/T605E) in combination with lethal factor (for the macrophage toxicity assay see Price B. *et al* (2001) *Infect. Immun.* 69:4509-4515). We conclude that the biological activity of the mutant PA (S588L/T605E) is not only conserved but is also slightly improved as compared to the wild type PA (see Figure 9). Importantly when the macrophage toxicity assay is carried out with wild type and mutant PA (S588L/T605E), after a heat stress challenge at 45°C, the percentage of toxicity of mutant PA is preserved for a much longer time than the wild type PA (see Figure 10). In a next step we investigated the effect of the protective effect of antisera derived from mice immunized with wild type PA or mice immunized with mutant PA (S588L/T605E). Our data indicate that antisera derived from mice immunized with mutant PA (S588L/T605E) protect macrophages from toxicity when they are challenged with wild type PA.

5. General applicability of the Solubis method

An obvious limitation of our method is the requirement of high resolution structural information, which for human proteins is available for 20-30% of the cases and is significantly lower for other species (Edwards 2009). If homology modeling is taken into account, the coverage could go up to 60-70% (Edwards 2009), albeit with a significant drop in accuracy on the $\Delta\Delta G$ calculation with FoldX. In order to investigate the applicability and scope of the SolubiS method, we here decided to run the analysis on a non-redundant set (sequence identity below 30%) of 585 protein domains for which high quality structures are available (R-factor better than 0.19, resolution better than 1.5 A), which were selected by the WHATIF consortium (Hooft, Sander et al. 1996). For the current analysis, the structures were mapped to the SCOP structural classification of protein domains. In this set, our algorithm identified the aggregation nucleating regions with TANGO and performed a systematic mutation screen to aggregation gatekeeper residues Arg, Lys, Pro, Asp and Glu of all amino acids belonging to an aggregating sequence (Rousseau, Serrano et al. 2006) and the resulting mutations were evaluated using both FoldX and TANGO. Figure 7A shows a heat map version of the MASS plot for the >70.000 mutations generated for the 585 high quality PDB structures in the current analysis in which the X-axis shows the change in TANGO score associated with the mutation, the Y-axis shows the $\Delta\Delta G$ score associated with the mutation and the colour code indicates the frequency of occurring of mutations with that ΔTANGO and $\Delta\Delta G$ profile. The region of interest,

i.e. mutations that maximally reduce aggregation while preserving thermodynamic stability, is indicated by a box (Fig 7A), showing only a minority of mutants are fulfilling the SolubiS criteria. In Figure 7B the frequency of mutations that satisfy stringent Solubis criteria ($\Delta\Delta G < 1$ kcal/mol and ΔTANGO < -300) is shown per structural class within SCOP ((α) all α-helical, (β) all β-sheet, (α/β) mixed α-helix and β-sheet and (α+β) combined α-helix and β-sheet. The percentage of mutations identified using Solubis is very similar for all SCOP classes and average 4.5% of all gatekeeper mutations tested (71025). The scope of the method would be determined by the number of domains in which improving mutations could be found. Accepting some inaccuracy in the algorithms used, we here employed the following conservative statistic: we counted the percentage of domains in which at least five candidate mutations could be identified that satisfy the aforementioned criteria, which makes it safe to assume that at least one of these will stand up to experimental validation (Figure 7C). From the total of 585 analysed proteins, 142 proteins do not carry a strong aggregating region and hence do not need to be improved. The analysis shows that of the remaining 443 proteins 49% of protein domains in our dataset are amenable to mutational improvement by SolubiS, with minor variations between the SCOP classes. When considering less rigorous criteria, so counting at least one SolubiS mutation per protein domain, 75% of them could be subjected to our redesign method.

Materials and methods

1.In silico analysis of aggregation, stability and structure of α-Galactosidase and YFP

The aggregation propensities of α-Gal, YFP and their mutants were analysed with TANGO (Fernandez-Escamilla, Rousseau et al. 2004), an algorithm to predict aggregation-nucleating sequences in proteins. The effect of the mutations on α-Gal and YFP stability was analysed by calculating the change in free energy ($\Delta\Delta G$) upon mutation with the FoldX forcefield (Schymkowitz, Borg et al. 2005). Structural changes of α-Gal and YFP due to mutations were analyzed with YASARA (Krieger, Koraimann et al. 2002).

2.Plasmid construction and mutagenesis

The full-length cDNA sequence encoding human α-Gal A (NM_000169) was cloned into the pcDNA4/TO/myc-His vector (Invitrogen). The insert was amplified using primers specific for the human α-Gal gene on Gene Pool cDNA template from human normal skeletal muscle (Invitrogen) with Phusion polymerase (Finnzymes). Then, the PCR product was digested with restriction enzymes Hind III and Xho I and cloned in pcDNA4/TO/myc-His vector to generate an open reading frame encoding α-Gal with a C-terminal Myc-tag. Expression vectors containing single, double and triple mutated α-Gal (D165V, A288D, A346P, A368P, A368R

and S405L) were generated by site-directed mutagenesis using sequence-specific primers and PWO DNA polymerase (Roche).

YFP vector was kindly provided by Sam Lievens from VIB Department for Molecular Biomedical Research, UGent, Belgium. YFP model for aggregating proteins was established by adding to its N-terminal part the Hsp70 binding sequence (LLRLTGW (SEQ ID NO: 5)) obtained from LIMBO algorithm. This sequence was cloned into pcDNA5/FRT/TO-Gateway-EYFP-FLAG vector using HindIII and KpnI restriction sites. Single and double mutations in YFP (Y151E, M153K, A154P, T225E and A227D) were introduced by site-directed mutagenesis using sequence-specific primers and PWO DNA polymerase (Roche).

3.Cell culture and transient transfection

Human cervical cancer cell line HeLa and human osteosarcoma cell line U2OS (used for maximum 20 passages) were cultured in DMEM/F12 medium (Gibco) supplemented with 10% FCS and 1% antibiotics (penicillin/streptomycin) at 37°C in 5% CO2. For transient transfection in 6-well culture plates, 350.000 of HeLa cells were plated per well in the medium without antibiotics. 1 µg of plasmid DNA was transfected into HeLa cells using FuGENE HD transfection reagent (Roche) according to manufacturer's protocol. For transient transfection in 96-well culture plates, 6.000 of HeLa and U2OS cells were plated per well in the medium without antibiotics. 0,1 µg of plasmid DNA was transfected into the cells using FuGENE HD transfection reagent (Roche) according to manufacturer's protocol. 48 h after transfection, cells were removed from the incubator and examined.

4.SDS-PAGE and Western Blot

48 h after transfection HeLa cells were lysed in RIPA buffer (1% octylphenoxypolyethoxyethanol (IGEPAL), 0.5% sodium deoxycholate and 0.1% sodium dodecyl sulfate (SDS)) (Pierce) supplemented with protease inhibitors (Roche) and fractionated by SDS-PAGE (NuPAGE system, Invitrogen). For Western blot the scraped cells were heated with 2% SDS buffer at 99°C for 10 min, separated using a 10% Bis-Tris gel in MES running buffer and subsequently transferred by electroblotting (fixed current 0.4 A) on a nitrocellulose membrane (Millipore). The membrane was incubated in 5% dried non-fat milk powder dissolved in 0.2% Tris Buffer Saline-Tween (TBST) for 1 h at room temperature (RT) and subsequently incubated with primary mouse anti-myc antibody (Invitrogen) followed by incubation by secondary goat HRP-conjugated anti-mouse IgG (Promega). Proteins were visualized using chemiluminescence immunoblotting detection reagent (ECL, Millipore).

5.Size exclusion chromatography (SEC)

For the analysis of the α-Gal aggregation state by SEC, transfected HeLa cells were lysed in RIPA buffer supplemented with protease inhibitors, centrifuged 5 min at 3000 rpm and 400 µl of the supernatant was subsequently loaded onto a Superdex S200 HR10/30 column (GE Healthcare) equilibrated in hypotonic buffer (20 mM HEPES, 10 mM KCl, 1 mM MgCl2, 1 mM EDTA, 1 mM EGTA, 1 mM DTT, pH 7.5). Eluted fractions were concentrated by 20% trichloroacetic acid precipitation, washed with acetone and analyzed by SDS-PAGE. The bands densities were quantified using the Quantity One program from the ChemiDoc System (Bio-Rad). A mixture of molecular weight markers (Bio-Rad) was injected onto the column as a gel filtration standard.

6.Enzymatic assay

The activity of α-Gal was determined by fluorogenic substrate 4-methylumbelliferyl-α-D-galactopyranoside (5 mM 4-MU-α-Gal) as described previously (Mayes, Scheerer et al. 1981). N-acetylgalactosamine (D-GalNAc) was used as an inhibitor of α-Gal B activity. α-Gal B is a second α-Gal in the cells that hydrolyzes the artificial substrate but its activity in FD patients is normal or increased. In brief, HeLa cells transfected with wild type or mutant α-Gal were harvested and lysed in PBS by 3 cycles of freezing/thawing in acetone-dry ice water bath. The supernatant obtained by centrifugation at 10.000 x g was incubated with substrate solution (5 mM 4-MU-α-Gal and 100 mM D-GalNAc in 0.1 M citrate buffer pH 4.5) at 37°C and the fluorescence was measured in a plate reader (POLARstar OPTIMA, BMG Labtech) within an hour. The slope of the linear part of the substrate conversion curve was a measure of the concentration of active enzyme in the lysates. Additionally, α-Gal concentration in the whole cell lysates was determined by Western blot. To determine the enzymatic activity, the assays were performed in three independent experiments.

7.Analysis of the aggregation of YFP mutants

Hela and U2OS cells were transfected in a 96-well plate, as described above. 48 h after transfection, the cells were washed in Phosphate Buffer Saline pH 7.4 (PBS) and fixed with 4% formaldehyde (20 min, RT). Nuclei were stained with DAPI diluted 1:10.000 in PBS. In order to count the cells with aggregates we used the IN Cell analyser 2000 (GE Healthcare), a high-content analysis system. Image acquisition was done using a 20X objective. For image analysis, we employed the IN Cell Developer Toolbox (GE Healthcare).

8.Statistical analysis

To confirm the consistency of the results, all described experiments were performed in minimum three separate replicates. For statistical evaluation of the determined averages and

standard deviations of the mean, data were analysed for significant differences using unpaired Student's t-test with a p-value less than 0.05 ($P<0.05$). Asterisks indicating the level of the p-value centered over the error bar mean: "*" $p< 0.05$, "**" $p< 0.01$, "***" $p< 0.001$ and "****" $p<0.0001$.

5

References

Balch, W. E., R. I. Morimoto, et al. (2008). "Adapting proteostasis for disease intervention." Science 319(5865): 916-919.

Belli, M., M. Ramazzotti, et al. (2011). "Prediction of amyloid aggregation in vivo." EMBO Rep 12(7): 657-663.

Ben-Zvi, A., E. A. Miller, et al. (2009). "Collapse of proteostasis represents an early molecular event in Caenorhabditis elegans aging." Proc Natl Acad Sci U S A 106(35): 14914-14919.

Benichou, B., S. Goyal, et al. (2009). "A retrospective analysis of the potential impact of IgG antibodies to agalsidase beta on efficacy during enzyme replacement therapy for Fabry disease." Mol Genet Metab 96(1): 4-12.

Benjamin, E. R., R. Khanna, et al. (2012). "Co-administration With the Pharmacological Chaperone AT1001 Increases Recombinant Human alpha-Galactosidase A Tissue Uptake and Improves Substrate Reduction in Fabry Mice." Mol Ther.

Bishop, N. A., T. Lu, et al. (2010). "Neural mechanisms of ageing and cognitive decline." Nature 464(7288): 529-535.

Brady, R. O., A. E. Gal, et al. (1967). "Enzymatic defect in Fabry's disease. Ceramidetrihexosidase deficiency." N Engl J Med 276(21): 1163-1167.

Chen, J. and B. Shen (2009). "Computational Analysis of Amino Acid Mutation: A Proteome Wide Perspective." Current Proteomics 6: 228-234.

De Baets, G., J. Reumers, et al. (2011). "An evolutionary trade-off between protein turnover rate and protein aggregation favors a higher aggregation propensity in fast degrading proteins." PLoS Comput Biol 7(6): e1002090.

Dobson, C. M. (2004). "Principles of protein folding, misfolding and aggregation." Seminars in Cell & Developmental Biology 15(1): 3-16.

Edwards, A. (2009). "Large-scale structural biology of the human proteome." Annu Rev Biochem 78: 541-568.

Eng, C. M., M. Banikazemi, et al. (2001). "A Phase 1/2 Clinical Trial of Enzyme Replacement in Fabry Disease: Pharmacokinetic, Substrate Clearance, and Safety Studies." The American Journal of Human Genetics 68(3): 711-722.

Eng, C. M. and R. J. Desnick (1994). "Molecular basis of Fabry disease: mutations and polymorphisms in the human alpha-galactosidase A gene." Hum Mutat 3(2): 103-111.

Fernandez-Escamilla, A. M., F. Rousseau, et al. (2004). "Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins." Nat Biotechnol 22(10): 1302-1306.

Futerman, A. H. and G. van Meer (2004). "The cell biology of lysosomal storage disorders." Nat Rev Mol Cell Biol 5(7): 554-565.

Griesbeck, O., G. S. Baird, et al. (2001). "Reducing the environmental sensitivity of yellow fluorescent protein. Mechanism and applications." J Biol Chem 276(31): 29188-29194.

Hooft, R. W. W., C. Sander, et al. (1996). "Verification of protein structures: Side-chain planarity. ." J Appl Cryst 29: 714-716.

Ishii, S., R. Kase, et al. (1996). "Aggregation of the inactive form of human alpha-galactosidase in the endoplasmic reticulum." Biochemical and biophysical research communications 220(3): 812-815.

Kikis, E. A., T. Gidalevitz, et al. (2010). "Protein homeostasis in models of aging and age-related conformational disease." Advances in experimental medicine and biology 694: 138-159.

Krieger, E., G. Koraimann, et al. (2002). "Increasing the precision of comparative models with YASARA NOVA--a self-parameterizing force field." Proteins 47(3): 393-402.

Lee, C. K., R. Weindruch, et al. (2000). "Gene-expression profile of the ageing brain in mice." Nat Genet 25(3): 294-297.

Lieberman, R. L., A. D'Aquino J, et al. (2009). "Effects of pH and iminosugar pharmacological chaperones on lysosomal glycosidase structure and stability." Biochemistry 48(22): 4816-4827.

Lu, T., Y. Pan, et al. (2004). "Gene regulation and DNA damage in the ageing human brain." Nature 429(6994): 883-891.

Lund, J., P. Tedesco, et al. (2002). "Transcriptional profile of aging in C. elegans." Curr Biol 12(18): 1566-1573.

Mayes, J. S., J. B. Scheerer, et al. (1981). "Differential assay for lysosomal alpha-galactosidases in human tissues and its application to Fabry's disease." Clinica chimica acta; international journal of clinical chemistry 112(2): 247-251.

Monsellier, E. and F. Chiti (2007). "Prevention of amyloid-like aggregation as a driving force of protein evolution." EMBO reports 8(8): 737-742.

Otzen, D. E., O. Kristensen, et al. (2000). "Designed protein tetramer zipped together with a hydrophobic Alzheimer homology: a structural clue to amyloid assembly." Proc Natl Acad Sci U S A 97(18): 9907-9912.

Park, J. S., K. Y. Han, et al. (2008). "Solubility enhancement of aggregation-prone heterologous proteins by fusion expression using stress-responsive Escherichia coli protein, RpoS." BMC Biotechnol 8: 15.

Pletcher, S. D., S. J. Macdonald, et al. (2002). "Genome-wide transcript profiles in aging and calorically restricted Drosophila melanogaster." Curr Biol 12(9): 712-723.

Porto, C., M. Cardone, et al. (2009). "The pharmacological chaperone N-butyldeoxynojirimycin enhances enzyme replacement therapy in Pompe disease fibroblasts." Mol Ther 17(6): 964-971.

Richardson, J. S. and D. C. Richardson (2002). "Natural beta-sheet proteins use negative design to avoid edge-to-edge aggregation." Proc Natl Acad Sci U S A 99(5): 2754-2759.

Rousseau, F., L. Serrano, et al. (2006). "How evolutionary pressure against protein aggregation shaped chaperone specificity." J Mol Biol 355(5): 1037-1047.

Schlieker, C., B. Bukau, et al. (2002). "Prevention and reversion of protein aggregation by molecular chaperones in the E. coli cytosol: implications for their applicability in biotechnology." J Biotechnol 96(1): 13-21.

Schymkowitz, J., J. Borg, et al. (2005). "The FoldX web server: an online force field." Nucleic Acids Res 33(Web Server issue): W382-388.

Shen, J. S., N. J. Edwards, et al. (2008). "Isofagomine increases lysosomal delivery of exogenous glucocerebrosidase." Biochem Biophys Res Commun 369(4): 1071-1075.

Song, J. A., D. S. Lee, et al. (2011). "A novel Escherichia coli solubility enhancer protein for fusion expression of aggregation-prone heterologous proteins." Enzyme Microb Technol 49(2): 124-130.

Soong, R., J. R. Brender, et al. (2009). "Association of highly compact type II diabetes related islet amyloid polypeptide intermediate species at physiological temperature revealed by diffusion NMR spectroscopy." J Am Chem Soc 131(20): 7079-7085.

Tartaglia, G. G., S. Pechmann, et al. (2009). "A relationship between mRNA expression levels and protein solubility in E. coli." J Mol Biol 388(2): 381-389.

Tesmoingt, C., O. Lidove, et al. (2009). "Enzyme therapy in Fabry disease: severe adverse events associated with anti-agalsidase cross-reactive IgG antibodies." Br J Clin Pharmacol 68(5): 765-769.

Thurberg, B. L., H. Rennke, et al. (2002). "Globotriaosylceramide accumulation in the Fabry kidney is cleared from multiple cell types after enzyme replacement therapy." Kidney Int 62(6): 1933-1946.

Van Durme, J., S. Maurer-Stroh, et al. (2009). "Accurate prediction of DnaK-peptide binding via homology modelling and experimental data." PLoS Comput Biol 5(8): e1000475.

Wang, W. (1999). "Instability, stabilization, and formulation of liquid protein pharmaceuticals." Int J Pharm 185(2): 129-188.

Xu, J., J. Reumers, et al. (2011). "Gain of function of mutant p53 by coaggregation with multiple tumor suppressors." Nature chemical biology 7(5): 285-295.

Zhang, Y. B., J. Howitt, et al. (2004). "Protein aggregation during overexpression limited by peptide extensions with large net negative charge." Protein Expr Purif 36(2): 207-216.

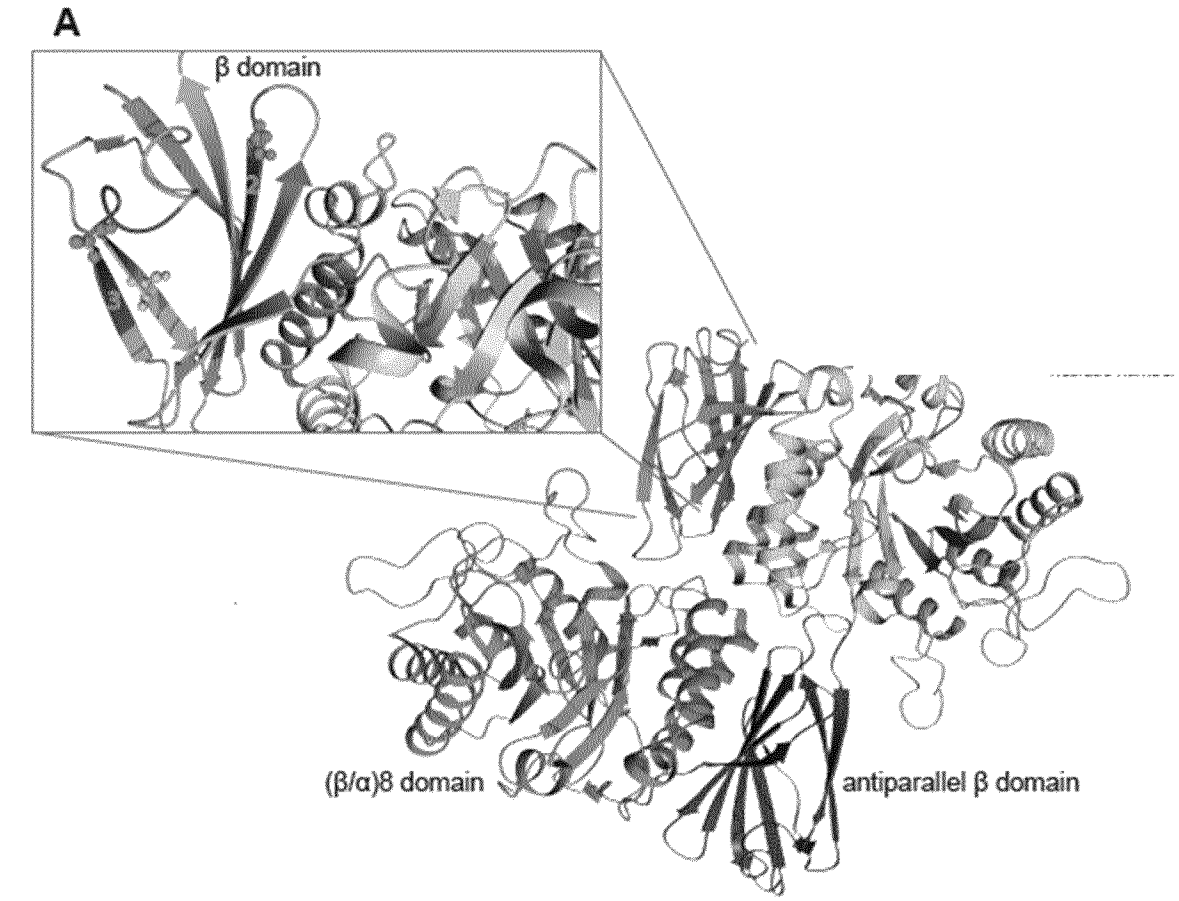
Zou, S., S. Meadows, et al. (2000). "Genome-wide study of aging and oxidative stress response in Drosophila melanogaster." Proc Natl Acad Sci U S A 97(25): 13726-13731.

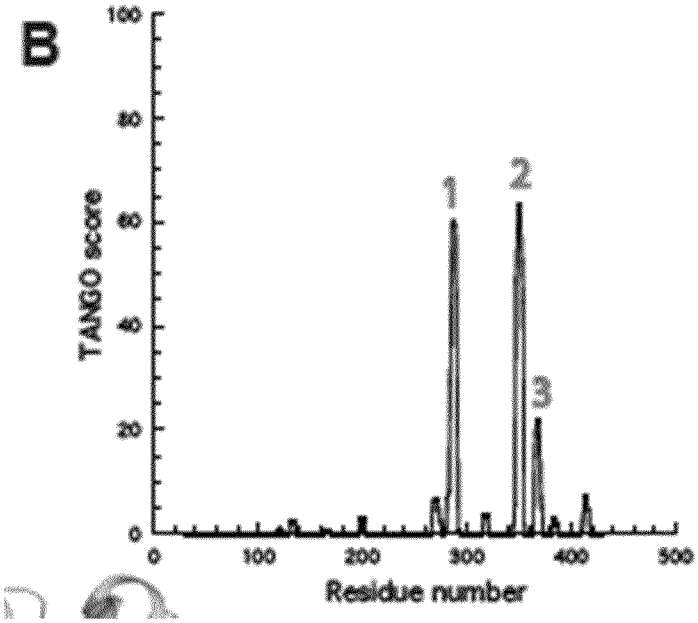
Claims

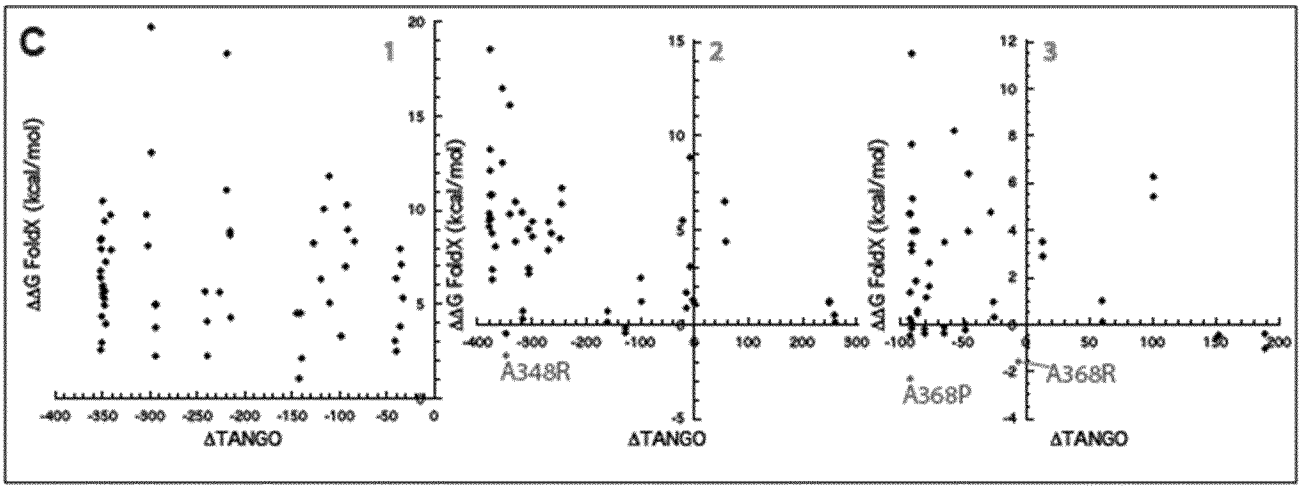
1. A method to produce a reduced aggregating variant of a protein that comprises at least two aggregation nucleating regions, said method comprising the following steps
   a) determining the aggregation nucleating region in said protein,
   b) generating a list of variant proteins wherein each variant protein has on at least one amino acid position in said determined aggregation nucleating regions a changed amino acid to either R, K, E, D or P,
   c) calculating for each of said variants the predicted aggregation score and the predicted change in thermodynamic stability with respect to the wild type protein,
   d) producing a reduced aggregating variant, which is derived from said list, wherein said variant has at the same time a maximally reduced aggregation score, a maximal preservation of thermodynamic stability and no structural changes with respect to the wild type protein.
2. A method according to claim 1 which further comprises after step c) the introduction of a compensatory mutation outside at least one of the aggregation nucleating regions with the purpose of preserving the thermodynamic stability of said protein.
3. A reduced aggregation variant of a protein which comprises at least two aggregation nucleating regions which is obtainable by
   a) determining at least two aggregation nucleating region in said protein,
   b) generating a list of variant proteins wherein each variant protein has on at least one amino acid position in said determined aggregation nucleating regions a changed amino acid to either R, K, E, D or P,
   c) calculating for each of said variants the predicted aggregation score and the predicted change in thermodynamic stability with respect to the wild type protein, and
   d) producing a reduced aggregating variant, which is derived from said list, wherein said variant has at the same time a maximally reduced aggregation score, a maximal preservation of thermodynamic stability and no structural changes with respect to the wild type protein.
4. A reduced aggregation variant of a protein which is obtainable by the steps defined in claim 3 but wherein after step c) introducing a compensatory mutation outside at least one of the aggregation nucleating regions with the purpose of preserving the thermodynamic stability of said protein.
5. A reduced aggregation variant of the alpha-galactosidase A protein (SEQ ID NO: 1) selected from the list consisting of i) A348R/A368R, ii) A348R/A368P, iii) A348R/A368R/S405L and iv) A348R/A368P/S405L.

6. A reduced aggregation variant of the alpha-galactosidase A protein according to claim 5 for the treatment of Fabry disease.
7. A reduced aggregation variant of the yellow fluorescent protein (SEQ ID NO: 2) selected from the list consisting of i) M153K/T225E, ii) M153K/A227D, iii) Y151E, iv) M153K/A227D and v) T225E/A227D.
8. A reduced aggregation variant of the *Bacillus anthracis* Protective Antigen (SEQ ID NO: 3) which is mutant PA (S588L/T605E).

Figure 1


A
β domain
(β/α)8 domain
antiparallel β domain

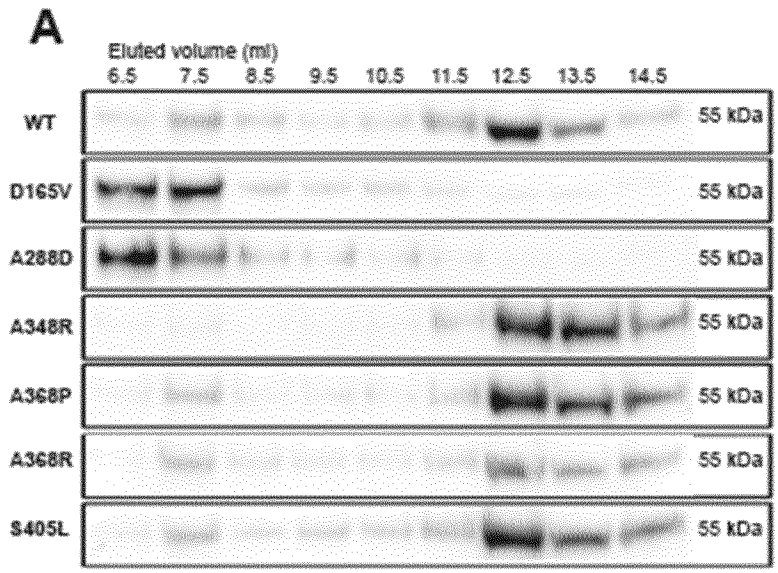

B
TANGO score
Residue number
1
2
3
0
20
40
60
80
100
0
100
200
300
400
500

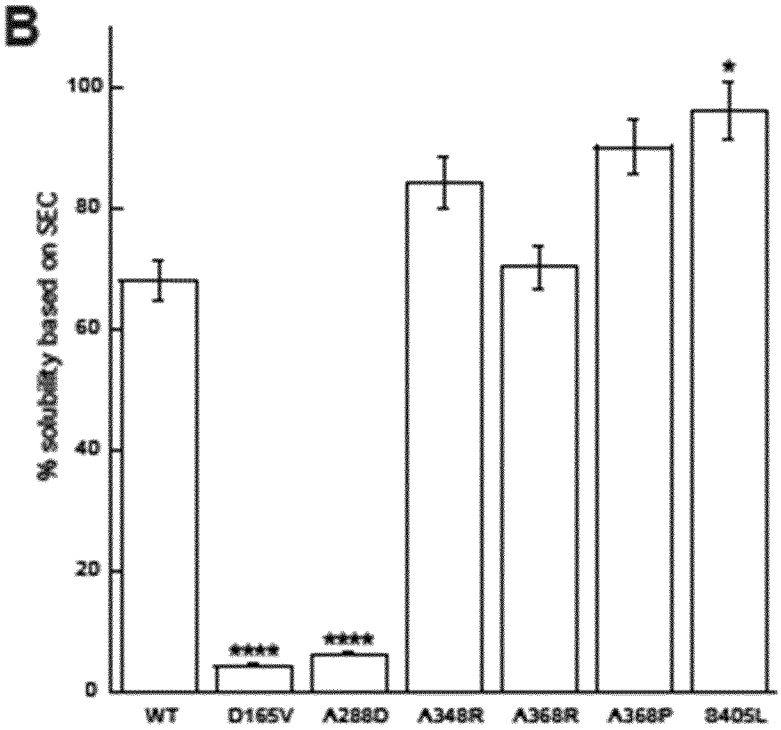


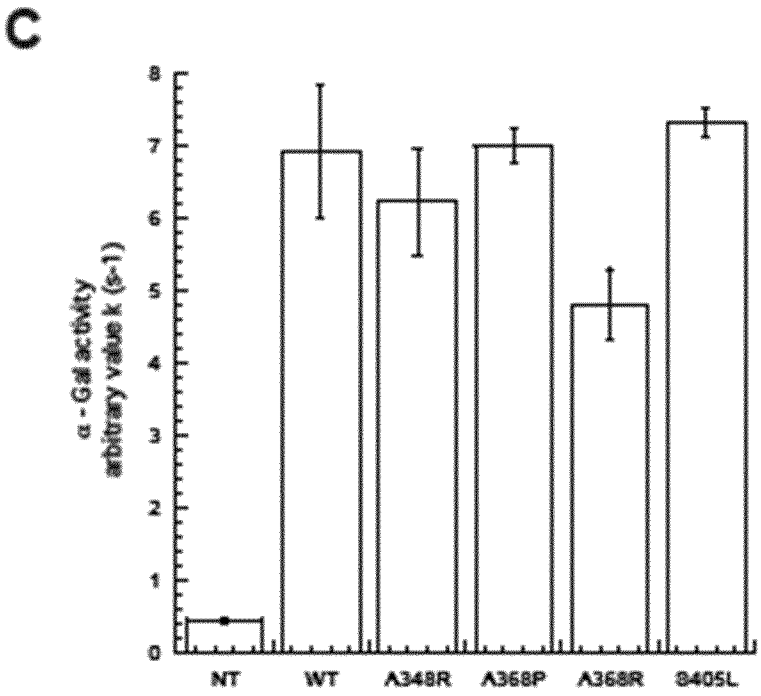
D

pos	wt	mut	difTANGO	difstab
405	S	L	-0,03	-3,34
405	S	M	-0,02	-3,09
333	Q	R	1,8	-2,62
368	A	P	-91,8	-2,31
408	N	P	0,1	-2,27
327	Q	M	0,0	-2,20

Figure 2




B
% solubility based on SEC
100
80
60
40
20
0
****
****
*
WT
D165V
A288D
A348R
A368R
A368P
S405L


C
α - Gal activity
arbitrary value k (s-1)
8
7
6
5
4
3
2
1
0
NT
WT
A348R
A368P
A368R
S405L

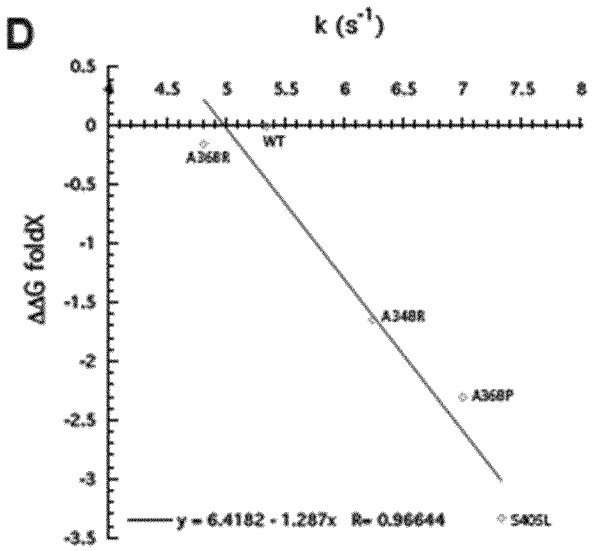
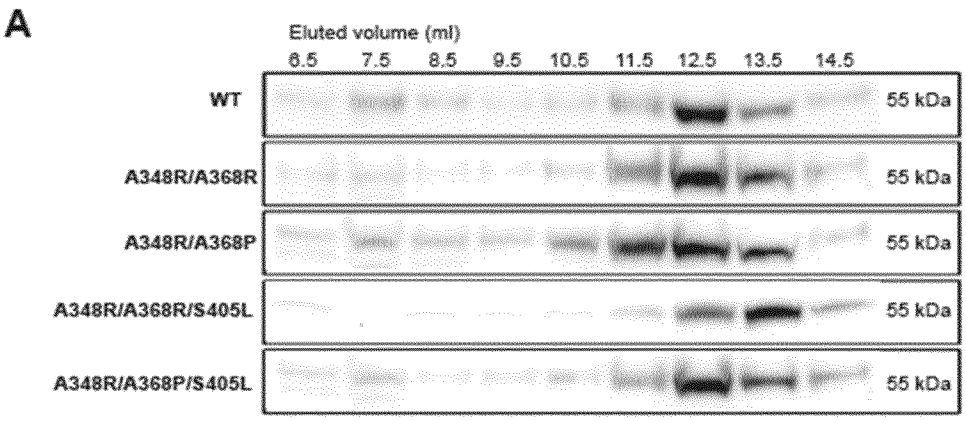
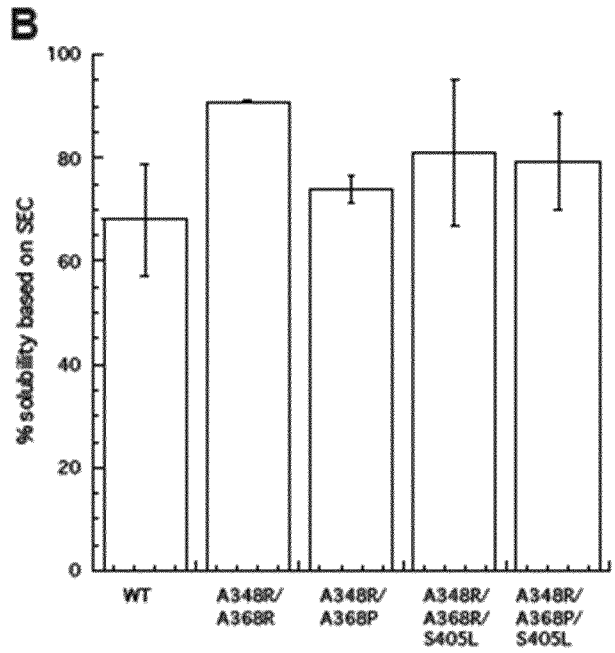
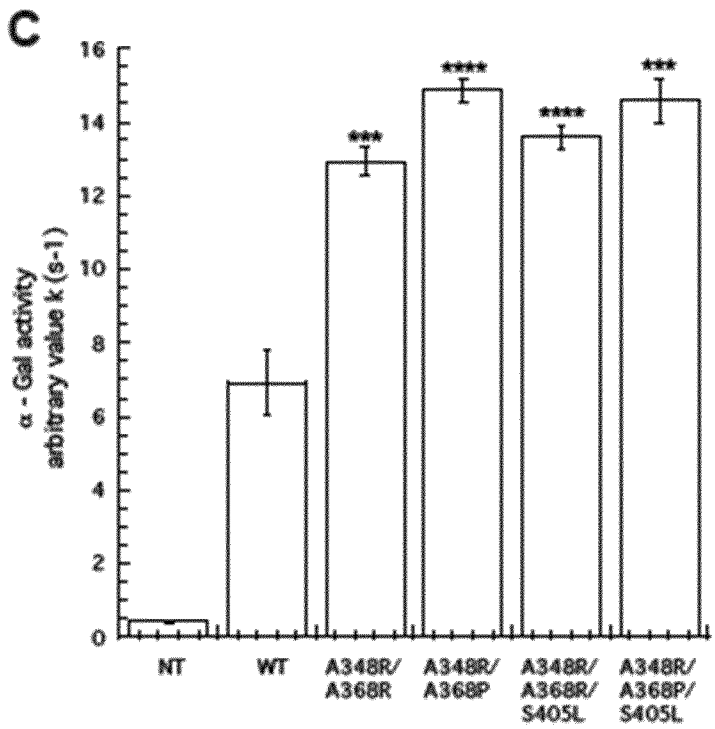

D
k (s-1)
ΔΔG foldX
WT
A368R
A348R
A368P
S405L
y = 6.4182 - 1.287x R= 0.96644

Figure 3


A
Eluted volume (ml)
6.5 7.5 8.5 9.5 10.5 11.5 12.5 13.5 14.5
WT
A348R/A368R
A348R/A368P
A348R/A368R/S405L
A348R/A368P/S405L
55 kDa


B
% solubility based on SEC
WT
A348R/ A368R
A348R/ A368P
A348R/ A368R/ S405L
A348R/ A368P/ S405L


C
α - Gal activity
arbitrary value k (s-1)
0
2
4
6
8
10
12
14
16
***
****
****
***
NT
WT
A348R/ A368R
A348R/ A368P
A348R/ A368R/ S405L
A348R/ A368P/ S405L

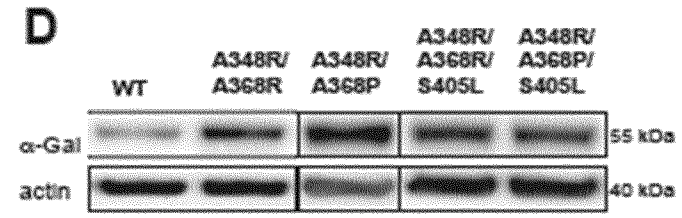
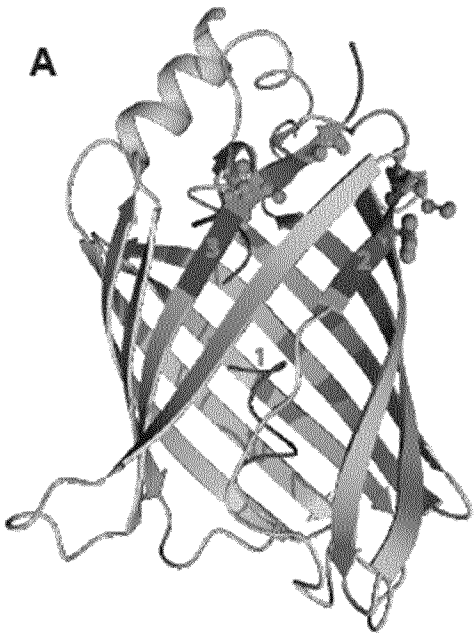
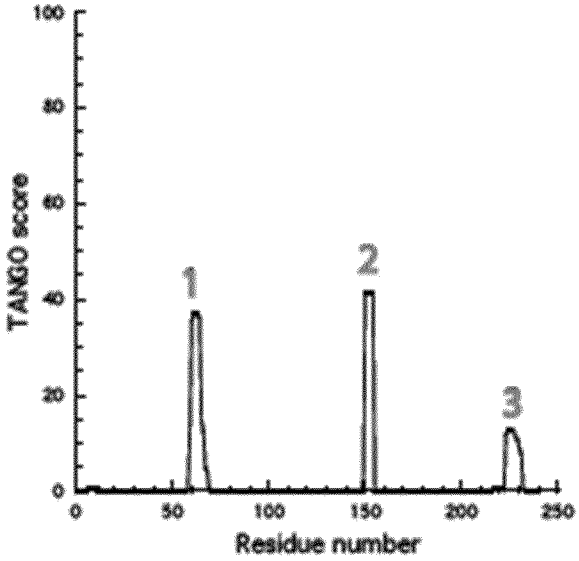

D
WT
A348R/ A368R
A348R/ A368P
A348R/ A368R/ S405L
A348R/ A368P/ S405L
α-Gal
actin
55 kDa
40 kDa

Figure 4


A

B


TANGO score
Residue number
1
2
3

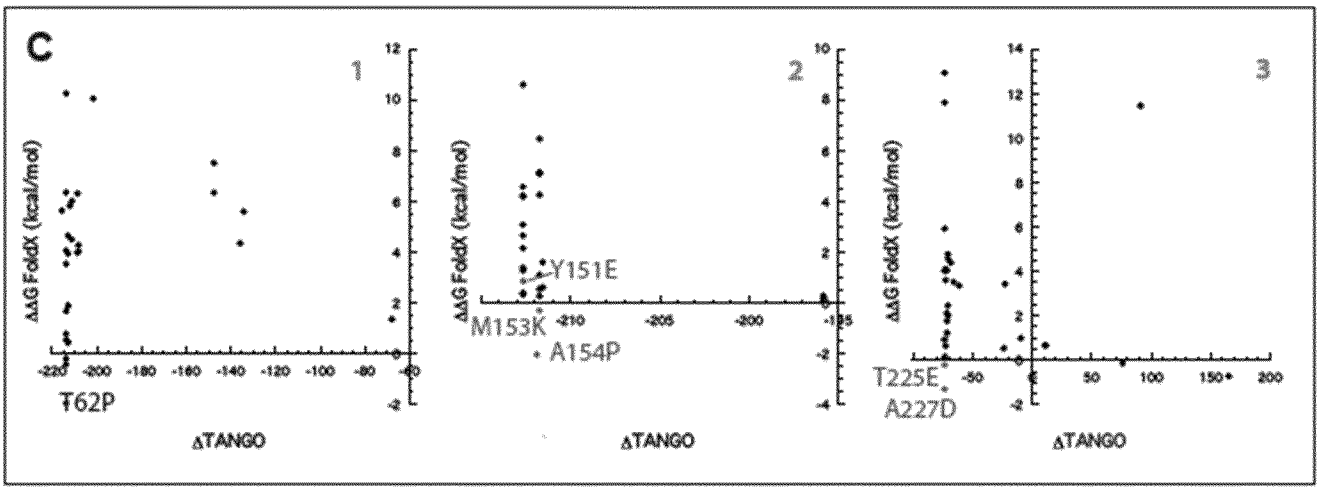

C
1
2
3
ΔΔG FoldX (kcal/mol)
ΔTANGO
T62P
Y151E
M153K
A154P
T225E
A227D

Figure 5

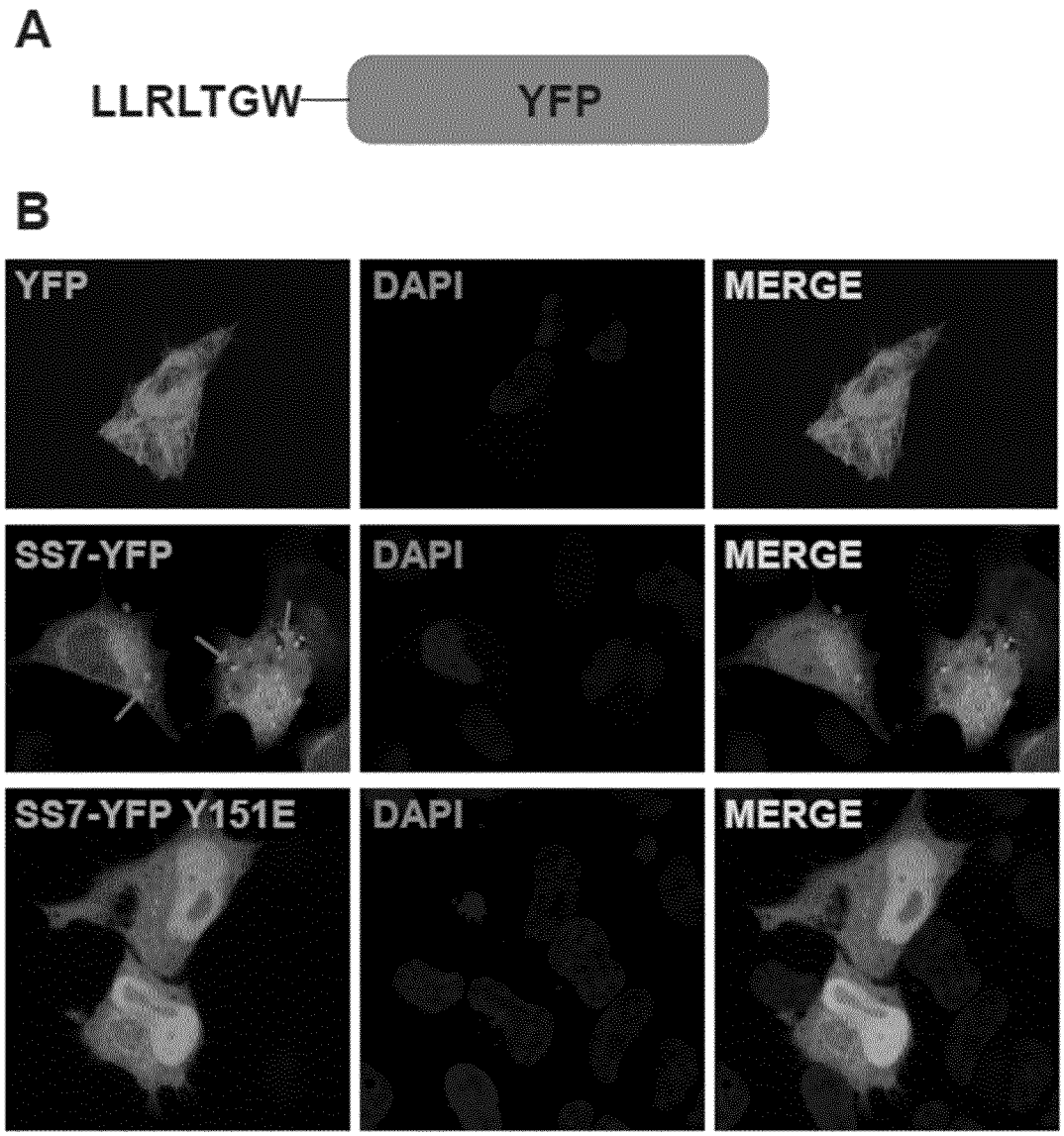

A
LLRLTGW
YFP
B
YFP
DAPI
MERGE
SS7-YFP
DAPI
MERGE
SS7-YFP Y151E
DAPI
MERGE

Figure 6

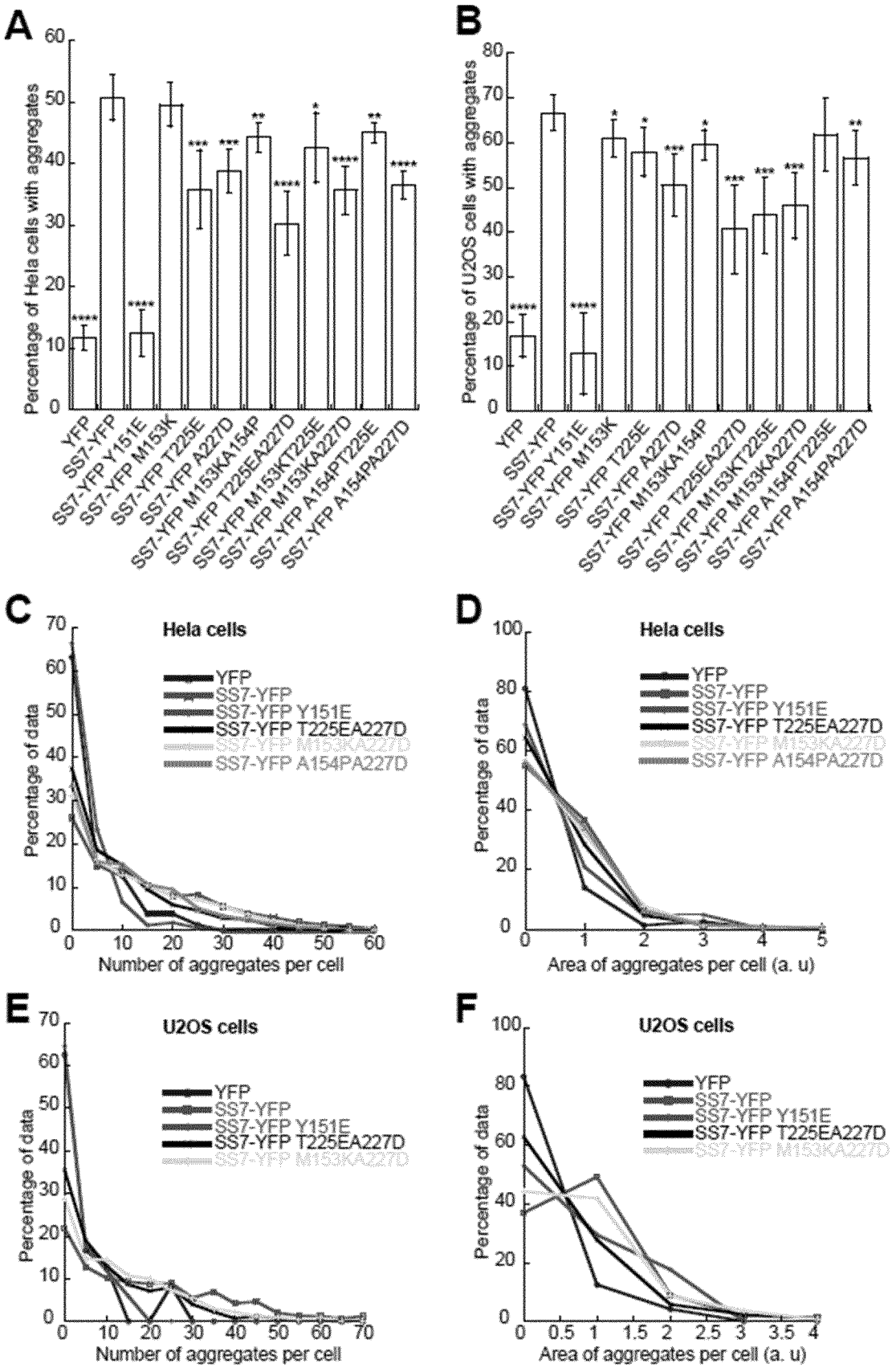

A
Percentage of Hela cells with aggregates
YFP
SS7-YFP
SS7-YFP Y151E
SS7-YFP M153K
SS7-YFP T225E
SS7-YFP A227D
SS7-YFP M153KA154P
SS7-YFP T225EA227D
SS7-YFP M153KT225E
SS7-YFP M153KA227D
SS7-YFP A154PT225E
SS7-YFP A154PA227D
B
Percentage of U2OS cells with aggregates
C
Hela cells
Percentage of data
Number of aggregates per cell
D
Hela cells
Area of aggregates per cell (a. u)
E
U2OS cells
F
U2OS cells

Figure 7

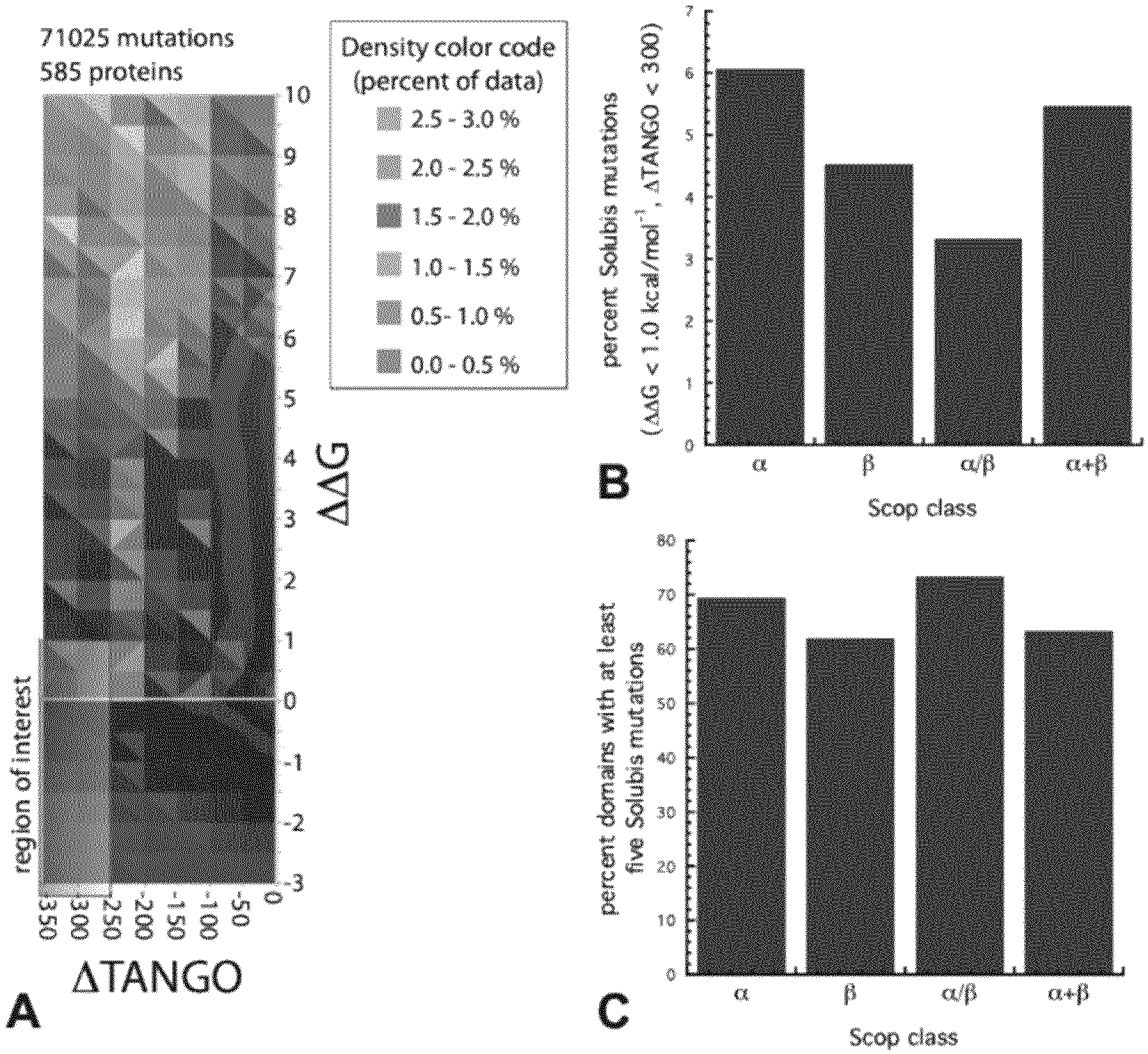


Figure 8

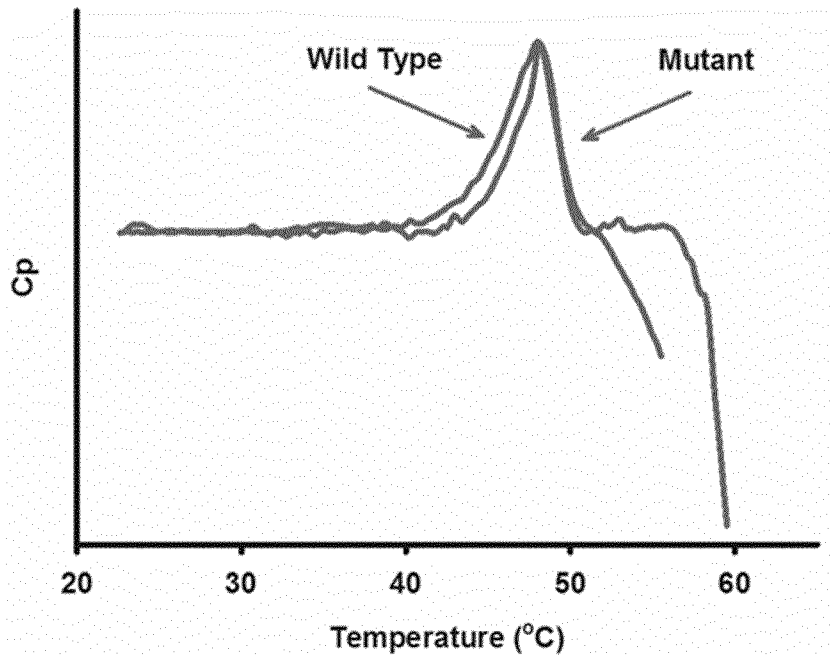
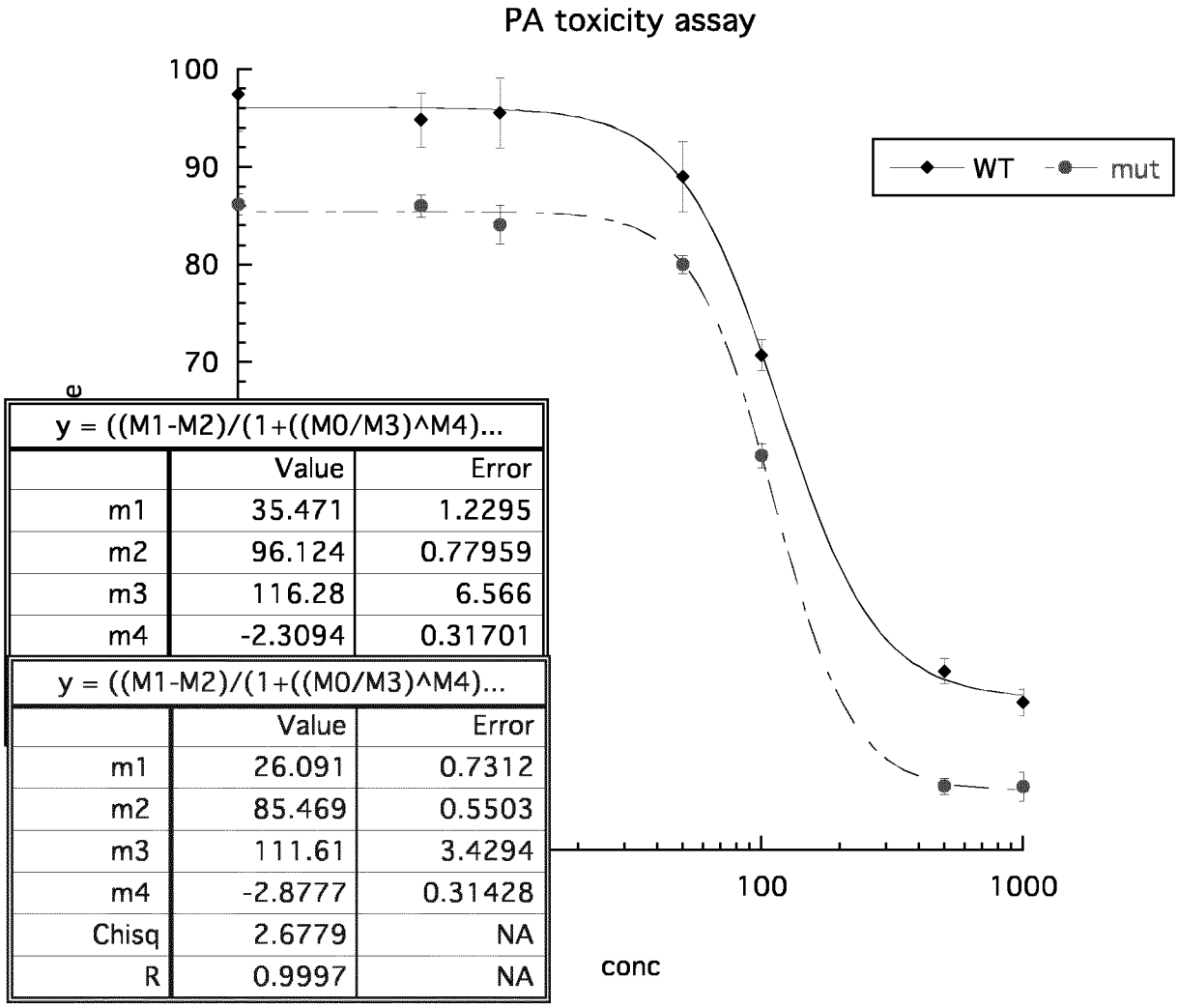


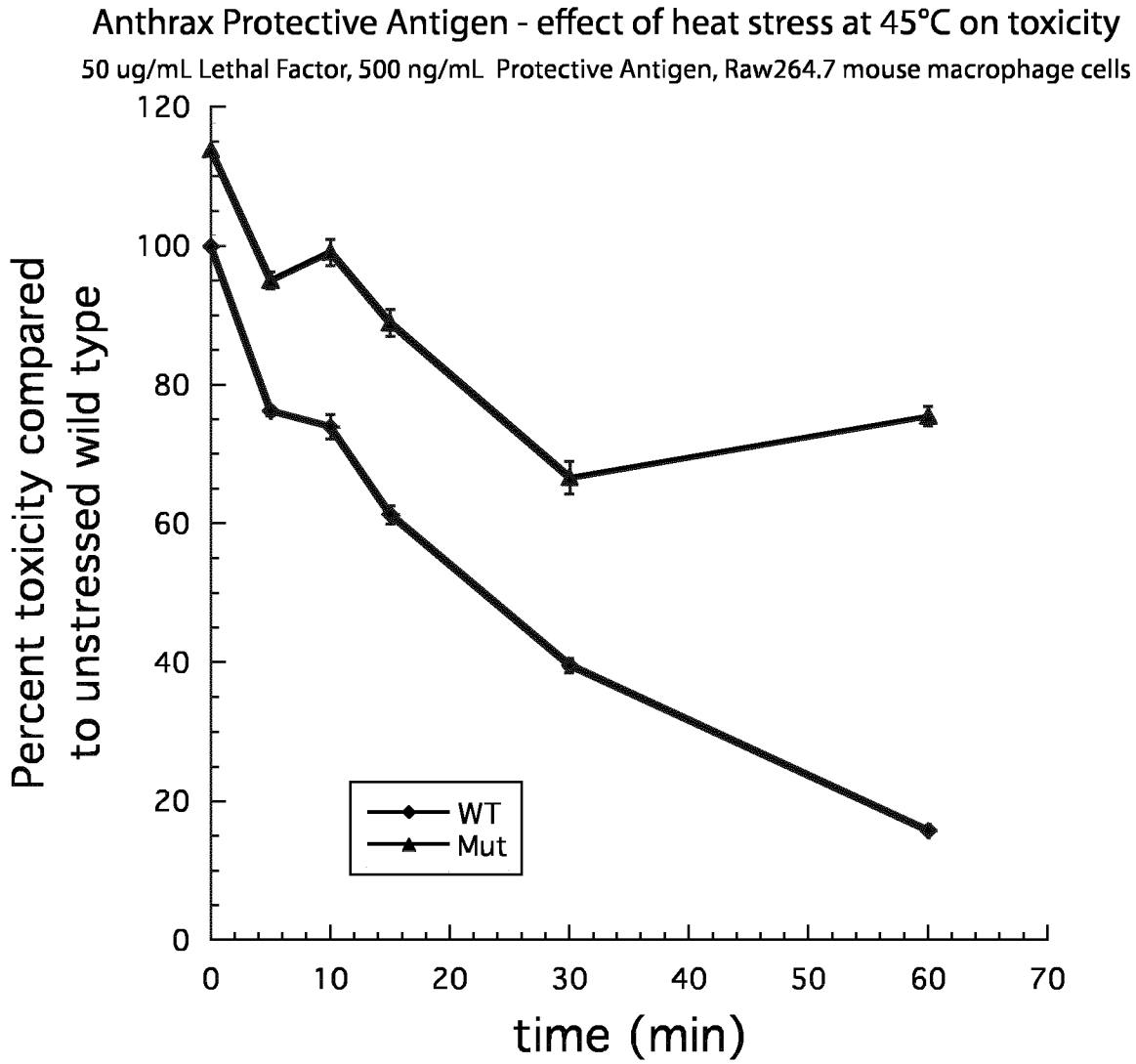
Figure 9



y = ((M1-M2)/(1+((M0/M3)^M4)...		
	Value	Error
m1	35.471	1.2295
m2	96.124	0.77959
m3	116.28	6.566
m4	-2.3094	0.31701

y = ((M1-M2)/(1+((M0/M3)^M4)...		
	Value	Error
m1	26.091	0.7312
m2	85.469	0.5503
m3	111.61	3.4294
m4	-2.8777	0.31428
Chisq	2.6779	NA
R	0.9997	NA

Figure 10


Anthrax Protective Antigen - effect of heat stress at 45°C on toxicity
50 ug/mL Lethal Factor, 500 ng/mL Protective Antigen, Raw264.7 mouse macrophage cells
Percent toxicity compared to unstressed wild type
120
100
80
60
40
20
0
0
10
20
30
40
50
60
70
time (min)
WT
Mut

INTERNATIONAL SEARCH REPORT

International application No
PCT/EP2013/058052

A. CLASSIFICATION OF SUBJECT MATTER
INV. C12N15/00 G01N33/68 G06F19/16 G06F19/18
ADD.

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

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
C12N G01N G06F

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

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)
EPO-Internal, BIOSIS, Sequence Search, EMBASE, WPI Data

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	NARESH CHENNAMSETTY ET AL: "Prediction of Aggregation Prone Regions of Therapeutic Proteins", THE JOURNAL OF PHYSICAL CHEMISTRY B, vol. 114, no. 19, 20 May 2010 (2010-05-20) , pages 6614-6624, XP055065027, ISSN: 1520-6106, DOI: 10.1021/jp911706q	3,4
Y	the whole document	1,2,5-8
	----- -/--	

[X] Further documents are listed in the continuation of Box C.

[X] See patent family annex.

* Special categories of cited documents :

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

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

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

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

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

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

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

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

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
11 June 2013	18/06/2013

Name and mailing address of the ISA/	Authorized officer
European Patent Office, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Fax: (+31-70) 340-3016	Offermann, Stefanie

2

Form PCT/ISA/210 (second sheet) (April 2005)

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	CHENNAMSETTY N ET AL: "Aggregation-Prone Motifs in Human Immunoglobulin G", JOURNAL OF MOLECULAR BIOLOGY, ACADEMIC PRESS, UNITED KINGDOM, vol. 391, no. 2, 14 August 2009 (2009-08-14), pages 404-413, XP026350710, ISSN: 0022-2836, DOI: 10.1016/J.JMB.2009.06.028 [retrieved on 2009-06-13]	3,4
Y	the whole document	1,2,5-8

X	CHENNAMSETTY NARESH ET AL: "Design of therapeutic proteins with enhanced stability", PROCEEDINGS OF THE NATIONAL ACADEMY OF SCIENCES, NATIONAL ACADEMY OF SCIENCES, US, vol. 106, no. 29, 21 July 2009 (2009-07-21), pages 11937-11942, XP002546638, ISSN: 0027-8424, DOI: 10.1073/PNAS.0904191106 [retrieved on 2009-07-01]	3,4
Y	the whole document	1,2,5-8

X	JOKE REUMERS ET AL: "Protein sequences encode safeguards against aggregation", HUMAN MUTATION, vol. 30, no. 3, 1 March 2009 (2009-03-01), pages 431-437, XP055064584, ISSN: 1059-7794, DOI: 10.1002/humu.20905	3,4
Y	the whole document	1,2,5-8

X	AGRAWAL NEERAJ J ET AL: "Aggregation in protein-based biotherapeutics: Computational studies and tools to identify aggregation-prone regions", AGGREGATION IN PROTEIN-BASED BIOTHERAPEUTICS: COMPUTATIONAL STUDIES AND TOOLS TO IDENTIFY AGGREGATION-PRONE REGIONS,, vol. 100, no. 12, 24 November 2011 (2011-11-24), pages 5081-5095, XP055064983, ISSN: 1520-6017, DOI: 10.1002/JPS.22705	3,4
Y	the whole document page 5085 page 5087 page 5090	1,2,5-8

	-/--	

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	JIE XU ET AL: "Gain of function of mutant p53 by coaggregation with multiple tumor suppressors", NATURE CHEMICAL BIOLOGY, NATURE PUBLISHING GROUP, NEW YORK, NY, US , vol. 7, no. 5 1 May 2011 (2011-05-01), pages 285-295, XP002676320, ISSN: 1552-4450, DOI: 10.1038/NCHEMBIO.546 Retrieved from the Internet: URL:http://www.nature.com/nchembio/journal/v7/n5/full/nchembio.546.html [retrieved on 2011-03-27]	3,4
Y	the whole document page 287, right-hand column	1,2,5-8

X	WO 2004/066168 A1 (UNIV CAMBRIDGE TECH [GB]; DOBSON CHRISTOPHER [GB]; CHITI FABRIZIO [IT]) 5 August 2004 (2004-08-05)	1,3,4
Y	claims 7,9; examples 1-4	2,5-8

T	JACINTE BEERTEN ET AL: "Aggregation prone regions and gatekeeping residues in protein sequences.", CURRENT TOPICS IN MEDICINAL CHEMISTRY, vol. 12, no. 22, 1 November 2012 (2012-11-01), pages 2470-2478, XP055064592, ISSN: 1568-0266	

T	A. SIEKIERSKA ET AL: "-Galactosidase Aggregation Is a Determinant of Pharmacological Chaperone Efficacy on Fabry Disease Mutants", JOURNAL OF BIOLOGICAL CHEMISTRY, vol. 287, no. 34, 6 July 2012 (2012-07-06) , pages 28386-28397, XP055064610, ISSN: 0021-9258, DOI: 10.1074/jbc.M112.351056	

T	J. BEERTEN ET AL: "Aggregation gatekeepers modulate protein homeostasis of aggregating sequences and affect bacterial fitness", PROTEIN ENGINEERING DESIGN AND SELECTION, vol. 25, no. 7, 15 June 2012 (2012-06-15), pages 357-366, XP055065014, ISSN: 1741-0126, DOI: 10.1093/protein/gzs031	

2

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No
PCT/EP2013/058052

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
WO 2004066168 A1	05-08-2004	AT 459056 T	15-03-2010
		DK 1586059 T3	14-06-2010
		EP 1586059 A1	19-10-2005
		ES 2341648 T3	24-06-2010
		PT 1586059 E	28-05-2010
		US 2006025977 A1	02-02-2006
		US 2008183455 A1	31-07-2008
		US 2008262742 A1	23-10-2008
		WO 2004066168 A1	05-08-2004

INTERNATIONAL SEARCH REPORT

International application No.
PCT/EP2013/058052

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

1. With regard to any nucleotide and/or amino acid sequence disclosed in the international application and necessary to the claimed invention, the international search was carried out on the basis of:

   a. (means)

   [ ] on paper

   [X] in electronic form

   b. (time)

   [X] in the international application as filed

   [ ] together with the international application in electronic form

   [ ] subsequently to this Authority for the purpose of search

2. [ ] In addition, in the case that more than one version or copy of a sequence listing and/or table relating thereto has been filed or furnished, the required statements that the information in the subsequent or additional copies is identical to that in the application as filed or does not go beyond the application as filed, as appropriate, were furnished.

3. Additional comments: