**Supplementary Information for**

"Protein Aggregation during Lyophilization: Protein Structural Descriptors as Predictors of Aggregation Propensity", Lavanya K. Iyer, Brock C. Roughton, Esben Bertelsen, Kyle V. Camarda and Elizabeth M. Topp

**Contents:**

**Table S1** – Protein structural descriptors and their definitions

**Table S2** – Summary of experimental measures of protein aggregation

**Table S3** – Summary of covariance analysis: Variable pairs with high degrees of covariance

**Table S1**. Protein structural descriptors and their definitions

| AGGRESCAN Descriptor Name | Definition |
|---|---|
| a3vSA | Sequence average amino acid aggregation propensity |
| nHS | Number of aggregation hot spots |
| NnHS | nHS normalized by number of residues in protein |
| AAT | Area of aggregation profile above hot spot threshold |
| THSA | Total area of aggregation profile comprising hot spots |
| TA | Total area of aggregation profile |
| AATr | AAT normalized by number of residues in protein |
| THSAr | THSA normalized by number of residues in protein |
| Na4vSS | Sliding window average of amino acid propensity values divided by number of amino acids in protein |
| **PASTA Descriptor Name** | |
| Emin | Minimum energy of PASTA pairings |
| Eavg | Average energy of PASTA pairings |
| Lmax | Average amino acid pair length of PASTA pairings |
| Lavg | Maximum amino acid pair length of PASTA pairings |
| (E/L)min | Minimum ratio of energy to length of PASTA pairings |
| (E/L)avg | Average ratio of energy to length of PASTA pairings |
| # of Peaks | Number of peaks in PASTA aggregation profile |
| **Physical Descriptor Name*** | |
| MW | Molecular weight (kDa) |
| pI | Isoelectric point |
| % a-helix | Percent a-helical secondary structure (PDB) |
| % b-sheet | Percent b-sheet secondary structure (PDB) |
| # S-S | Number of disulfide bonds (PDB) |
| # free thiols | Number of free thiol groups (PDB) |
| $T_m$ | Melting temperature |

*See also Table 1

**Table S2.** Summary of experimental measures of protein aggregation using SEC (% monomeric protein values), UV-Visible Spectroscopy (A.I. values) and SDS-PAGE

| Protein | Formulation | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Unlyophilized | | | Buffer | | | Sucrose | | | Glycine | | | Urea | | | Gdn HCl | | |
| | SEC[a] | A.I.[b] | SDS-PAGE[c] | SEC[a] | A.I.[b] | SDS-PAGE[c] | SEC[a] | A.I.[b] | SDS-PAGE[c] | SEC[a] | A.I.[b] | SDS-PAGE[c] | SECa | A.I.[b] | SDS-PAGE[c] | SEC[a] | A.I.[b] | SDS-PAGE[c] |
| Myoglobin | 100 (0) | 71 (31) | - | 96 (3) | 123 (59) | -, *dimer* | 99 (3) | 13 (1) | -, *dimer* | 99 (1) | 15 (2) | -, *dimer* | 85 (6) | 34 (15) | Pellet, *dimer* | 0 (0) | 24 (11) | Pellet, *dimer* |
| Lysozyme | 100 (0) | 5 (4) | -, *dimer* | 80 (12) | 11 (10) | -, *dimer* | 77 (18) | 7 (3) | -, *dimer* | 97 (2) | 1.9 (0.3) | -, *dimer* | 65 (23) | 18 (9) | -, *dimer* | 0 (0) | 2.1 (0.5) | -, *dimer* |
| Ovalbumin | 100 (0) | 43 (6) | +, *dimer, d* | 98 (2) | 245 (8) | +, *dimer, d* | 97 (1) | 5 (1) | +, *dimer, d* | 99.7 (1.7) | 113 (14) | +, *dimer, d* | 16.5 (0.6) | 413 (38) | Pellet, +, *dimer, d* | 0 (0) | 45 (14) | Pellet, +, *dimer* |
| RNase A | 100 (0) | 0.8 (0.3) | *dimer* | 113 (5) | 2.2 (0.2) | *dimer* | 89 (10) | 6 (1) | *dimer* | 97 (12) | 6.2 (0.1) | *dimer* | 97 (9) | 5.1 (0.3) | *dimer* | 0 (0) | 28 (19) | *dimer* |
| DNase I | 100 (0) | 6.6 (0.1) | - | 107 (15) | 109 (5) | - | 86 (12) | 9.9 (0.8) | - | 62 (35) | 59 (1) | - | 116 (12) | 239 (14) | - | 0 (0) | 44 (5) | - |
| α-chymotrypsinogen | 100 (0) | 7 (7) | - | 102 (6) | 13 (12) | - | 95.6 (0.5) | 1.05 (0.01) | - | 103 (2) | 24 (11) | - | 83 (4) | 12 (8) | - | 0 (0) | 9 (2) | Pellet, - |
| Cytochrome-C | 100 (0) | 159 (3) | -, *dimer* | 30 (23) | 3388 (1817) | -, *dimer* | 33 (24) | 120.4 (0.6) | -, *dimer* | 127 (90) | 61.5 (0.9) | -, *dimer* | 0 (0) | 193 (56) | -, *dimer* | 0 (0) | 65 (2) | +, *dimer* |
| Con A | 100 (0) | 23 (6) | - | 103 (20) | 9 (1) | - | 122 (19) | 20 (7) | - | 121 (15) | 26 (2) | - | 21 (9) | 80 (19) | - | 45 (12) | 388 (224) | - |
| Catalase | 100 (0) | 21 (13) | - | 90 (5) | 18 (12) | - | 98 (6) | 2.6 (0.2) | - | 100 (2) | 160 (16) | - | 19 (2) | 48 (22) | Pellet, +, d | 0 (0) | 10 (5) | Pellet, - |
| α-amylase | 100 (0) | 3.5 (0.5) | + | 94 (3) | 5 (1) | + | 97 (4) | 2.8 (0.2) | + | 104 (2) | 1.2 (0.4) | + | 92 (4) | 2.9 (0.6) | + | 2.77 (0.02) | 76 (6) | Pellet, + |
| α-lactalbumin | 100 (0) | 14 (10) | -, *dimer* | 109 (11) | 32.9 (0.5) | -, *dimer* | 99 (2) | 1.6 (0.5) | -, *dimer* | 100 (4) | 4.1 (0.8) | -, *dimer* | 97 (2) | 2.2 (0.2) | -, *dimer* | 0 (0) | 19 (5) | - |
| β-lactoglobulin | 100 (0) | 11 (9) | -, *dimer* | 92 (6) | 1.7 (0.4) | -, *Dimer* | 95 (3) | 2.5 (0.8) | -, *dimer* | 90 (6) | 1.7 (0.2) | -, *dimer* | 83 (3) | 4.8 (3.1) | -, *dimer* | 0 (0) | 924 (179) | Pellet, + |
| SOD | 100 (0) | 3.5 (0.3) | - | 89 (3) | 3.65 (0.04) | - | 99 (1) | 0.7 (0.1) | - | 100 (2) | 15.9 (0.2) | - | 113 (10) | 155 (24) | - | 48 (20) | 51 (10) | +, d |
| BSA | 100 (0) | 65.1 (0.5) | + | 94 (2) | 66.4 (0.5) | + | 96.5 (0.6) | 3.6 (0.5) | + | 93.2 (1.3) | 3.1 (0.8) | + | 77.2 (0.4) | 3.2 (0.6) | + | 117 (3) | 18 (6) | Pellet, + |
| Trypsin Inhibitor | 100 (0) | 2.5 (0.3) | - | 115 (16) | 7 (1) | - | 116 (15) | 192 (5) | - | 119 (13) | 29 (1) | - | 120 (15) | 45 (4) | - | 0 (0) | 304 (216) | - |

[a]SEC values represented as average % monomer ± SEM (n=3)

[b]A.I. values represented as average A.I. ± SEM (n=3)

[c]'Pellet' indicates formation of insoluble aggregates upon centrifugation of the reconstituted lyophilized formulation; presence of high molecular weight aggregate bands (greater than twice the molecular weight of the native protein) indicated by +; absence of high molecular weight aggregate bands indicated by -; presence of dimer band indicated by *dimer*, presence of disulfide-linked aggregates indicated by *d*

**Table S3**. Summary of covariance analysis: Variable pairs with high degrees of covariance

| $|\text{Covariance}(X,Y)| \geq 0.7$ | $|\text{Covariance}(X,Y)| \geq 0.8$ | $|\text{Covariance}(X,Y)| \geq 0.9$ |
|---|---|---|
| (*apolar*, *MW*) | | |
| (*$L_{max}$*, *$(E/L)_{avg}$*) | | |
| (*$(E/L)_{min}$*, *$(E/L)_{avg}$*) | | |
| (*a3vSA*, *$E_{min}$*) | | |
| (*a3vSA*, *$E_{avg}$*) | | |
| (*MW*, *AAT*) | | |
| (*AAT*, *$E_{min}$*) | | |
| (*AAT*, *$E_{avg}$*) | | |
| (*THSA*, *$E_{min}$*) | | |
| (*THSA*, *$E_{avg}$*) | (*% α-helix*, *% β-sheet*) | |
| (*a3vSA*, *THSA*) | (*apolar*, *# free SH*) | (*$E_{avg}$*, *$E_{min}$*) |
| (*a3vSA*, *TA*) | (*MW*, *nHS*) | (*$L_{avg}$*, *$L_{max}$*) |
| (*a3vSA*, *AATr*) | (*AAT*, *nHS*) | (*Na4vSS*, *a3vSA*) |
| (*TA*, *AATr*) | (*THSA*, *nHS*) | |
| (*a3vSA*, *THSAr*) | | |
| (*NnHS*, *THSAr*) | | |
| (*Na4vSS*, *$E_{min}$*) | | |
| (*Na4vSS*, *$E_{avg}$*) | | |
| (*Na4vSS*, *THSA*) | | |
| (*Na4vSS*, *TA*) | | |
| (*Na4vSS*, *AATr*) | | |
| (*Na4vSS*, *THSAr*) | | |