Published in final edited form as:
*Comput Chem Eng*. 2013 November 11; 58(2013): 369–377.

# Protein aggregation and lyophilization: Protein structural descriptors as predictors of aggregation propensity

**Brock C. Roughton**[a], **Lavanya K. Iyer**[b], **Esben Bertelsen**[b,c], **Elizabeth M. Topp**[b], and **Kyle V. Camarda**[d,*]

[a]Bioengineering Graduate Program, University of Kansas, Lawrence, KS, United States

[b]Department of Industrial and Physical Pharmacy, Purdue University, West Lafayette, IN, United States

[c]Department of Pharmaceutics and Analytical Chemistry, University of Copenhagen, Copenhagen, Denmark

[d]Department of Chemical and Petroleum Engineering, University of Kansas, Lawrence, KS, United States

## Abstract

Lyophilization can induce aggregation in therapeutic proteins, but the relative importance of protein structure, formulation and processing conditions are poorly understood. To evaluate the contribution of protein structure to lyophilization-induced aggregation, fifteen proteins were co-lyophilized with each of five excipients. Extent of aggregation following lyophilization, measured using size-exclusion chromatography, was correlated with computational and biophysical protein structural descriptors via multiple linear regression. Descriptor selection was performed using exhaustive search and forward selection. The results demonstrate that, for a given excipient, extent of aggregation is highly correlated by eight to twelve structural descriptors. Leave-one-out cross validation showed that the correlations were able to successfully predict the aggregation for a protein “left out” of the data set. Selected descriptors varied with excipient, indicating both protein structure and excipient type contribute to lyophilization-induced aggregation. The results show some descriptors used to predict protein aggregation in solution are useful in predicting lyophilized protein aggregation.

### Keywords

Protein formulation; Biologics; Aggregation prediction; Lyophilization; Multiple linear regression; Structural descriptors

## 1. Introduction

Aggregates are formed during the manufacture and storage of protein drugs, and are associated with an increased risk of immunogenicity and therapeutic failure in patients (Rosenberg, 2006). Understanding the structural properties of proteins that lead to aggregation is critical to the design of safe and effective protein drug products, and an ability to predict aggregation propensity (i.e., the likelihood and extent to which a protein will aggregate) with reasonable accuracy would accelerate development. Several approaches



*Corresponding author at: Department of Chemical and Petroleum Engineering, University of Kansas, Room 4132 Learned Hall, 1530W. 15th, Lawrence, KS 66045, United States. Tel.: +1 785 864 4965; fax: +1 785 864 4967. camarda@ku.edu (K.V. Camarda).

have been developed to estimate aggregation propensity for a given protein, which can classified into two main methods: heuristic-based methods and simulation-based methods.

Heuristic-based approaches attempt to use prior history on aggregation or causes of aggregation in proteins to develop predictors for aggregation propensity. The aim of a heuristic-based approach is to relate protein properties to experimental data on protein aggregation, with the end result being a predictive model or algorithm that returns aggregation propensity given a measure of protein structure. Several algorithms have been developed to predict protein aggregation in solution as a function of structural parameters. For example, AGGRESCAN utilizes the intrinsic aggregation propensity of amino acids obtained from an experimental aggregation database of mutated β-amyloid peptides (Conchillo-Sole et al., 2007). PASTA predicts the likelihood of amino acid sequences being involved in intermolecular β-sheet formation, based on minimization of β-pairing energies (Trovato, Seno, & Tosatto, 2007). Zyggregator uses factors such as protein hydrophobicity, electrostatic interactions and alternating stretches of polar and non-polar residues to predict aggregation propensity (Tartaglia & Vendruscolo, 2008). For all of these methods, protein primary structure (amino acid sequence) is used to return one or more scoring parameters which are indicative of the propensity of a protein to aggregate. For instance, AGGRESCAN returns the number of aggregation prone regions, or “hot spots” in a protein. The number of hot spots is then used to qualitatively indicate the likelihood of protein aggregation occurring, with a larger number of hot spots corresponding to a higher likelihood. Therefore, a hallmark of current methods is qualitative results in the form of aggregation predictors that must be interpreted.

Simulation-based methods use any of the many available molecular simulation software packages or newly developed tools to investigate interactions between protein molecules or dynamics within a single protein molecule. The aim of simulation-based methods is to determine if aggregation is likely to happen based on the energetics of protein–protein interactions (Ma & Nussinov, 2006). Alternatively, simulation-based methods can investigate the dynamics of a single protein molecule to determine if the properties of the protein could become amenable to aggregation (Irbäck & Mohanty, 2006). For example, the spatial aggregation propensity (SAP) algorithm uses molecular simulations to determine the average exposed hydrophobic surface area for a given protein, with larger exposed hydrophobic surface areas representing increased aggregation propensity (Chennamsetty, Voynov, Kayser, Helk, & Trout, 2009). In general, simulations are more computationally expensive than use of a model or algorithm to predict aggregation propensity. Simulations are usually required for every system of interest. Simulation-based methods necessitate three-dimensional structure of a protein for determination of aggregation propensity and thus require more structural information than the heuristic-based methods described previously. Simulation-based approaches offer advantages over current heuristic-based approaches due to the ability for qualitative assessments (e.g., free energy calculations of protein-protein interactions) and inclusion of formulation conditions via explicit solvent and solute modeling. Recently, hybrid approaches have been developed to combine simulation results with heuristic model-based predictions. The Developability Index has been constructed for monoclonal antibodies utilizing net charge and spatial aggregation propensity (SAP) (Lauer et al., 2012). Additionally, the osmotic second virial coefficient ($B_{22}$) has also been used to predict protein self-association in aggregation (Chi, Krishnan, Randolph, & Carpenter, 2003; Printz, Kalonia, & Friess, 2012), though it is based on experimental measurement and not on a priori descriptors of protein structure.

The described approaches to predicting aggregation all assume a solution environment. Approximately 40% of current protein drug products are marketed as solids, many as lyophilized powders for reconstitution. In addition, protein drugs that have been expressed

and purified are often stored in lyophilized form prior to final formulation and packaging. Whether aggregation predictors developed for solutions can be applied effectively to lyophilized solids is unknown. Lyophilization involves freezing a protein solution followed by removal of ice by sublimation. The process subjects the protein to stresses such as denaturation at the ice surface, pH shifts and freeze concentration (Anchordoquy & Carpenter, 1996; Bhatnagar, Bogner, & Pikal, 2007; Chang, Kendrick, & Carpenter, 1996b). Removal of water and loss of hydrogen bonds during the drying stage can produce intra- and intermolecular interactions that differ from those in solution (Chang, Beauvais, Dong, & Carpenter, 1996a; Prestrelski, Pikal, & Arakawa, 1995). Lyophilized proteins are reconstituted prior to administration or formulation, and may or may not regain their original conformation and activity upon rehydration. Since this array of stresses and environments differs considerably from aqueous solution, it is reasonable to question whether properties that predispose proteins to aggregate in solution are also important in lyophilized solids.

Current heuristic-based aggregation prediction models are computationally efficient and easy to use, with the main drawback being a disconnect between the calculated propensity score and experimental aggregation measurement. Additionally, the effects of the formulation are not incorporated in aggregation propensity calculation and the suitability for use with lyophilized proteins is not established. The proposed work proposes three objectives for the improvement of model-based prediction of aggregation propensity: (1) build predictive quantitative models by correlating different aggregation predictors with experimental measures of aggregation, as opposed to a vague or qualitative aggregation propensity score. (2) Determine the suitability of aggregation predictors derived from solution conditions for predicting protein aggregation in lyophilized solids. (3) Investigate the effects of formulation on the aggregation predictors selected in a predictive model.

## 2. Materials and methods

### 2.1. Overview

In the studies reported here, four formulations of fifteen different proteins were prepared, subjected to lyophilization, then reconstituted and assayed for aggregate content. Various multiple linear regression analyses were then performed to test the hypothesis that the experimental aggregate content in lyophilized proteins is correlated with descriptors of protein structure, and that these descriptors can be used to predict aggregation propensity of lyophilized proteins. Fig. 1 provides an overview of the experimental procedure and subsequent model development stage. Percent monomer remaining after lyophilization was used as the experimental quantitative measure of aggregation propensity, with values less than 100% indicating loss of monomeric protein due to aggregation. Descriptors obtained from computational predictions were limited to approaches based solely on primary sequence (AGGRESCAN and PASTA). Both a forward selection method and an exhaustive search method were used to identify and select descriptors that strongly correlated with experimental data. Descriptor selection was performed to identify structural properties of proteins that were good predictors of aggregation propensity in lyophilized systems, similar to other data-mining approaches used to pinpoint useful descriptors (Zhang & Huan, 2010). The analysis demonstrates that, for a given formulation condition, the extent of aggregation as measured by percent monomer remaining shows strong correlation and good prediction power when eight to twelve structural descriptors are selected. Leave-one-out cross-validation (LOOCV) shows that the correlations generally are able to successfully predict the aggregation for a protein “left out” of the data set, using regression for the remaining fourteen. Overall, the proposed correlations are able to predict protein aggregation after lyophilization for a variety of proteins on a formulation-by-formulation basis, showing that some heuristic-based descriptors developed for use with proteins in solution have applications for proteins in a lyophilized state.

### 2.2. Materials

Fifteen commercially available proteins were selected to represent a wide range of possible molecular weights and structural properties (see Table 1). All proteins were purchased from Sigma–Aldrich (St. Louis, MO) except for trypsin inhibitor (Worthington Biochemical Corporation, Lakewood, NJ). Proteases were avoided in the study as they are subject to autocatalytic degradation.

Sucrose, glycine, urea and guanidine hydrochloride (Gdn HCl) were used as excipients (i.e., additives) to develop four different formulations for lyophilization. All excipients and buffer materials (monobasic and dibasic potassium hydrogen phosphate) were obtained from Sigma–Aldrich. Proteins were dialyzed prior to formulation to remove unknown excipients that may have been added by the manufacturer, as described below. All other chemicals and reagents were used as received.

### 2.3. Preparation of solutions containing protein and excipient

Lyophilized samples were prepared from solutions containing a protein and one of several excipients. To prepare the protein solutions, each protein was dissolved in potassium phosphate buffer (20 mM, pH 7.4) to give a stock solution with 2.0 mg/mL protein. The solutions were dialyzed using Biotech Cellulose Ester dialysis tubing (MWCO 8000–10,000 Da, Spectrum Laboratories, Rancho Dominguez, CA) against phosphate buffer (20 mM, pH 7.4) at 4 °C for 24 h with two changes of dialysis buffer. After dialysis, the solutions were filtered through a 0.2 μm syringe filter (Gelman Nylon Acrodisc 13, Sigma–Aldrich) and stored at 4 °C for use within 24 h.

Excipient solutions were prepared by dissolving the excipient in phosphate buffer. This solution was then filtered using a 0.2 μm syringe filter and stored at 4 °C. Sucrose (2.0 mg/mL), glycine (2.0 mg/mL), urea (2.0 M) and Gdn HCl (2.0 M) solutions were prepared accordingly.

### 2.4. Lyophilization

Five lyophilized formulations were prepared for each protein (Table 1) to generate 75 different types of lyophilized samples (15 unlyophilized samples + 75 lyophilized samples = 90 total samples). Prior to lyophilization, the protein stock solutions were mixed with excipient or potassium phosphate buffer (20 mM, pH 7.4) solutions in lyophilization vials to give a final protein concentration of 1.0 mg/mL, with each lyophilization vial containing 400 μL of the formulation. The protein: excipient ratio was 1:1 by weight for sucrose and glycine, while the final concentration of urea and Gdn HCl was 1.0 M. The formulations were lyophilized in quadruplicate using a VirTis adVantage Plus freeze dryer (SP Industries, Inc., Gardiner, NY). The same conservative lyophilization cycle was used for all sample types. Shelves were first pre-cooled to −2 °C (15 min), followed by sample freezing at −40 °C (50 min) and drying under vacuum (70 mTorr) at −35 °C for 10 h, −20 °C for 8 h, −5 °C for 6 h, with continued drying (100 mTorr) at 10 °C for 6 h, 25 °C for 6 h and 4 °C for 0.5 h (Sophocleous et al., 2012). No attempt was made to optimize the cycle for the individual combinations of protein and excipient (e.g., using $T_g'$). To produce a non-lyophilized control solution, the remaining protein stock solution was diluted 1:1 with potassium phosphate buffer to give a final protein concentration of 1.0 mg/mL.

The lyophilized powders were reconstituted in 400 μL DDW and allowed to dissolve at room temperature. The solutions were then transferred to 1.5 mL Eppendorf tubes and spun at 12,000 rpm for 10 min, and any visible pellet or particles were noted. An adequate volume of the supernatant was removed for analysis using high performance size exclusion

chromatography (HP-SEC). Next, most of the supernatant was removed and any pellet re-suspended in potassium phosphate buffer for analysis using SDS-PAGE.

### 2.5. High performance size exclusion chromatography

The non-lyophilized control solutions and the supernatant from the lyophilized, reconstituted samples were analyzed using HP-SEC. Analysis was performed on an Agilent LC system (1200 Series, Agilent Technologies, Santa Clara, CA) with a TSKgel G3000SWxI column (Tosoh Bioscience LLC, King of Prussia, PA). The mobile phase was 50 mM potassium phosphate buffer, pH 7.0 containing 200 mM NaCl. A flow rate of 0.8–1.0 mL/min was used and UV signals were collected at 215 nm and at 280 nm. The peak area was calculated for each protein from the chromatogram. The areas were correlated to % monomeric protein content, normalizing against the areas of the corresponding non-lyophilized control solutions which were assumed to contain 100% monomer.

### 2.6. SDS-PAGE

The supernatant (and re-suspended pellet, when produced) from lyophilized and non-lyophilized solutions were analyzed for the presence of large aggregates using SDS-PAGE. Protein samples were mixed with non-reducing or reducing (containing β-mercaptoethanol as a reducing agent) loading buffer, stained with bromophenol blue, vortexed and heated at 95 °C for 5 min. The samples were cooled and then loaded onto 10% or 12% polyacrylamide gels. Low molecular weight markers (GE Healthcare, Waukesha, WI) were used as a reference ladder. SDS-PAGE analysis was performed on a Mini-PROTEAN Tetra cell electrophoresis instrument attached to a PowerPac Basic power supply (Bio-Rad Laboratories, Hercules, CA) using 10% gels for all proteins except α-amylase, BSA, ovalbumin and catalase. These proteins were analyzed on 12% gels and referenced against broad range molecular weight markers (Bio-Rad Laboratories), owing to their larger molecular weight. Gels were stained with Coomassie Brilliant Blue R-250 staining solution for 30–60 min on a rocking platform (VWR International, Radnor, PA), followed by destaining for approximately 2 days.

### 2.7. UV–visible spectroscopy

UV–visible spectra were obtained for all protein samples, lyophilized as well as non-lyophilized, using an Agilent 8453 UV–vis spectrophotometer. 400 μL of protein solution was added to a very low volume cuvette and spectra were collected in the wavelength range 200–600 nm with an integration time of 10 s and an interval of 1 nm. The aggregation index (A.I.) was used as a measure of aggregates in solution, since increased absorbance at 350 nm has been associated with presence of larger particles. A.I. was calculated using optical densities at 280 nm and 350 nm according to the method described by Katayama et al. (2005) as follows:

$$\text{A.I.} = \frac{\text{OD}_{350}}{\text{OD280} - \text{OD}_{350}} \times 100$$

### 2.8. Protein descriptors

Linear regressions were developed to relate percent monomeric protein recovered after lyophilization and reconstitution, as measured by HP-SEC, to protein structural descriptors. Three sets of descriptors were used in the correlations: (i) physical descriptors based on structure and reported melting temperatures ($T_m$), (ii) AGGRESCAN descriptors and (iii) PASTA descriptors (Table 1 and S1). Both AGGRESCAN and PASTA use protein primary structure to predict aggregation propensity, particularly the potential of forming β-amyloid

structures. AGGRESCAN descriptors were obtained using the AGGRESCAN online server (Conchillo-Sole et al., 2007). AGGRESCAN predicts an aggregation profile and locates regions of the protein above a threshold value, defined as hot spots (Conchillo-Sole et al., 2007). PASTA generates pairings between residues of varying length that have the lowest predicted energies for self-interaction and generates an aggregation profile with peaks representing regions of high aggregation propensity (Trovato et al., 2007). The AGGRESCAN and PASTA descriptors used for correlations are defined in Table S1. Correlations were generated using all descriptor sets at once on a per formulation basis.

### 2.9. Descriptor selection

Two methods were used to select optimum descriptor sets for the correlations: exhaustive search and forward selection. In exhaustive search, for a defined model size (i.e., number of descriptors), all descriptors were evaluated and the descriptors that minimized the Akaike Information Criterion (AIC) score were chosen. The AIC score is equal to the error of the fit plus a penalty for the number of descriptors used (Wasserman, 2004). As descriptors are added, the error of the fit decreases but the penalty is increased. As the model size was increased, the optimal set of descriptors was selected de novo, and descriptors were not guaranteed to be retained as model size was increased. Exhaustive search was performed using the LEAPS package (Lumley, 2013) in the statistical software R (Dalgaard, 2008). The final correlation was selected as the model size that returned the minimal AIC score as compared to all other model sizes. LEAPS uses Mallow's $C_p$ score in exhaustive search, which yields the same results as use of AIC for linear correlations (Wasserman, 2004).

In the forward selection method, once a descriptor is selected it is retained as model size increases, and each additional descriptor is added only if it reduces the AIC score. If the AIC score cannot be reduced by adding another descriptor, the selection process ends and the resulting correlation becomes final. In the forward selection method, it is possible that no descriptors will be selected; this occurs if no single descriptor is strongly correlated to the data. Forward selection was performed using the DAAG package in R (Dalgaard, 2008; Maindonald & Braun, 2010).

### 2.10. Evaluation of predictive power

Once a final correlation was selected by either forward search or exhaustive search, LOOCV was performed to evaluate the predictive power of the correlation. Prediction errors can become large if the dataset is overfit by using too many descriptors. In LOOCV, one data point (i.e., one protein) was removed and a correlation generated using the same descriptors selected for the final correlation. The new correlation was then used to predict the extent of aggregation for the protein that was left out. The LOOCV procedure was repeated for every protein, resulting in 15 total predictions. LOOCV is used in the validation of quantitative structure–activity relationships (QSARs) and quantitative structure-property relationships (QSPRs) (Bolboaca & Jäntschi, 2008; Wu et al., 2010), and has been used in pharmaceutical applications. LOOCV was performed using the DAAG package in R (Dalgaard, 2008; Maindonald & Braun, 2010). The error between the predicted values and actual values was represented as the prediction sum of the squares (PRESS) (Quan, 1988):

$$\mathrm{PRESS}=\sum_{i=1}^{n}\left(Y_i-\widehat{Y}_{\langle i\rangle}\right)^2$$

where $n$ is the number of observations, $Y_i$ is the observed value for observation $i$, and $\hat{Y}_{\langle i\rangle}$ is the predicted value for observation $i$ when the observation is left-out. The PRESS value was used to calculate a $Q^2$ value, which is indicative of the predictive power of the equation. The value for $Q^2$ is less than or equal to the value of $R^2$, with better predictive power being

provided the closer $Q^2$ is to $R^2$ (Quan, 1988). In general, $R^2 - Q^2 < 0.3$ indicates that the model is not overfit and does not contain outliers (Bolboacă & Jäntschi, 2008). Previous QSARs and QSPRs have used $Q^2$ to evaluate predictive power (Eslick et al., 2009). $Q^2$ is related to the PRESS value:

$$Q^2 = 1 - \frac{\mathrm{PRESS}}{\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2}$$

where $\overline{Y}_i$ is the average value for all observations.

## 3. Results

### 3.1. Experimental measures of protein aggregation

For the fifteen proteins and five lyophilized formulations studied here, aggregation varied with protein, with formulation and with the analytical method used to assess aggregation (Table S2). The proteins can be grouped according to aggregation tendency. Five proteins (lysozyme, ovalbumin, cytochrome C, α-amylase, BSA) showed *high aggregation tendency* across the formulation types as indicated by low (<80%) recovery of monomeric protein by HP-SEC, high aggregation index (>100) and/or the presence of high molecular weight bands on SDS-PAGE. Six proteins (RNAse A, α-chymotrypsinogen, ConA, α-lactoglobulin, SOD, trypsin inhibitor) showed *low aggregation tendency* using these metrics, while the remaining four proteins (myoglobin, DNAse I, catalase, β-lactoglobulin) showed *intermediate aggregation tendency*. Greater than 100% recovery of monomeric protein by HP-SEC was observed for some samples and could reflect incomplete separation of aggregate from monomeric protein or protein unfolding. While the assignment of proteins to these groups is somewhat arbitrary, it is clear that the proteins selected show a range of aggregation propensities on lyophilization. The data set is therefore suitable for assessing the effects of protein structure on lyophilization-induced aggregation within the parameter space defined by their structural descriptors. Note that, since the largest protein in the data set (BSA, 66 kD) is considerably smaller than monoclonal antibodies, these and other large proteins are not expected to be well-described by the correlations developed here.

With regard to formulation, those containing buffer, sucrose or glycine all produced aggregates following lyophilization for some of the proteins studied (Table S2). Compared to these excipients, urea formulations produced a greater extent of aggregation for a greater number of proteins, as expected for this denaturant (Table S2). Formulations containing Gdn HCl showed no retention of monomeric protein by HP-SEC for 11 of the 15 proteins, and pellets and/or high molecular weight bands on SDS-PAGE for 8 of 15. Because the observed extent of aggregation was very high and relatively insensitive to protein structure in Gdn HCl formulations, this formulation was omitted in developing correlations. The correlations thus were developed using the four remaining excipients (i.e., buffer, sucrose, glycine or urea).

Of the three methods used to assess aggregation (SDS-PAGE, AI, HP-SEC), only AI and HP-SEC were used quantitatively; therefore, only results from these two methods can be used to develop quantitative correlations with protein structural descriptors. AI values were not considered quantitatively reliable. For example, some formulations for proteins such as concanavalin A, cytochrome-c, β-lactoglobulin and trypsin inhibitor showed large AI values but had large errors. In other cases, proteins with low AI values showed loss of monomeric protein by HP-SEC and formation of a pellet on SDS-PAGE (e.g., catalase in urea, Table S2). This may be due to the formation of insoluble precipitates that settle out of solution and are not detected on UV. Furthermore, RNase, lysozyme α-chymotrypsinogen and many

other proteins did not show significant differences in AI values across formulations. As a result, correlations were developed based on the % retention of monomer as measured by HP-SEC and AI values were not used further.

### 3.2. Development of correlations

Two methods were used to develop correlations relating protein descriptors to percent monomer retained following lyophilization: exhaustive search and forward selection. For both methods, the descriptor set used to generate correlations for each formulation was comprised of physical descriptors, AGGRESCAN descriptors and PASTA descriptors. The following subsections detail and compare the results for each method.

**3.2.1. Exhaustive search method—**The exhaustive search method was performed using all available descriptors for each formulation. Good fits, as determined by minimum AIC scores, were obtained with model sizes between eight and twelve descriptors (Table 2). The descriptors selected for each formulation are listed in Table 2, together with statistical measures of goodness-of-fit ($R^2$) and predictive power ($Q^2$ and $R^2 - Q^2$).

In general, the descriptors selected differed from formulation to formulation. Across all formulations, each descriptor type was selected with similar frequencies: physical descriptors were selected 16 times, AGGRESCAN descriptors were selected 12 times and PASTA descriptors were selected 11 times. No single descriptor was selected for all formulations. The most commonly selected descriptors were % *β-sheet*, $T_m$, $E_{avg}$, and *Peaks*, which were all selected for three of the four formulations. At least one descriptor of each type was selected for each formulation.

The correlations for all four formulations had small ($R^2 - Q^2$) values and $R^2$ values close to 1, indicating that they provide a reasonable tool for predicting the percent retained monomeric protein after lyophilization within each formulation type. The correlation for the buffer formulation had the best fit and best predictive power, having the highest $R^2$ and $Q^2$ values and the lowest ($R^2 - Q^2$) values. The correlation for the glycine formulation provided the poorest fit and lowest $Q^2$ value, and also provided the poorest predictive power as indicated by the largest ($R^2 - Q^2$) value. A summary of the regression for the four formulations, together with values of the regression coefficients, is presented in Table 3.

**3.2.2. Forward selection method—**Forward selection was also used to build correlations using all available descriptors. Due to the nature of the selection method, the final correlations differ in the number of descriptors selected (Table 4). Physical descriptors were selected most frequently with this method, accounting for 9 out of the 11 descriptors selected (Table 4). Only physical descriptors were selected for urea and sucrose formulations and four out of the five descriptors selected for the buffer formulation were physical descriptors. The most commonly selected descriptor was *pI*, which was selected first for the buffer and sucrose formulations and second for the urea formulation. The early selection of *pI* indicates that this descriptor provides a superior fit to the experimental data for the buffer and sucrose formulations and a very good fit for the urea formulation when compared to the other descriptors.

With forward selection, all of the ($R^2 - Q^2$) values were large and no correlation provided a good fit to the data, as indicated by the low $R^2$ values. The correlation for the buffer formulation had the highest number of descriptors and yielded the highest $R^2$ value. However, the predictive power of the correlation was unsatisfactory and provided the largest ($R^2 - Q^2$) value among the four formulations. The sucrose formulation provided a slightly higher $R^2$ value than the urea formulation, despite using one less descriptor. The correlation

for the sucrose formulation had the lowest ($R^2 - Q^2$) value among the correlations generated by forward selection.

**3.2.3. Comparison of methods—**Models generated by exhaustive search were superior to those generated by forward selection, having better fits and greater predictive power as indicated by the higher $R^2$, higher $Q^2$ and lower ($R^2 - Q^2$) values (Tables 2 and 4). Forward selection is less computationally expensive when compared to exhaustive search. For development with models that involve large sets of possible descriptors, use of exhaustive search may be infeasible due to computation requirements. However, the time needed for descriptor selection was comparable for both methods using the descriptor set in this model. Additionally, the results indicate that use of a forward search is insufficient in developing a predictive model with sufficient accuracy. As a result, the forward selection method was not pursued and models generated by exhaustive search are emphasized in the results and discussion below.

### 3.3. Predictive power of correlations

Within a formulation, correlations showed good fits ($R^2 > 0.98$) and satisfactory predictive power ($R^2 - Q^2 < 0.2$) using the exhaustive search method (Table 2). Parity plots comparing the predicted percentage of monomeric protein to the experimental value are shown in Fig. 2. Good agreement between predicted and actual values is observed for all four formulations, with the greatest deviation observed for the glycine formulation (Fig. 2). The data for the urea formulation is spread fairly evenly and validation resulted in a high $Q^2$ value. For the other formulations, one protein had a substantially lower observed and predicted percent monomer values than the other proteins. However, this outlying observation resulted in lower $Q^2$ values only for the glycine formulations (Fig. 2), as high prediction error was found for the outlier when the point was left out during cross-validation. High $Q^2$ values were obtained for both the buffer and sucrose formulations, despite the outlier. The results suggest that the descriptors selected for the buffer and sucrose formulation are able to account for the structural differences in the outlying protein sufficiently, yielding a low prediction error when the protein was left-out during cross-validation.

### 3.4. Performance of individual descriptor sets

The models presented in Tables 2–4 were generated by pooling all of the available descriptors from three descriptor sets: (i) physical descriptors, (ii) AGGRESCAN descriptors, (iii) PASTA descriptors. Correlations were also developed for each individual descriptor set in isolation, using the exhaustive search method (data not shown). At small model sizes, physical descriptors provided the best fit for the buffer, urea and sucrose formulations. The glycine formulation showed similar fits for model sizes of one descriptor, regardless of the descriptor set used. At larger model sizes, no single descriptor set could provide a fit comparable to that given by pooling all available descriptors.

Overall, physical descriptors performed better across all model sizes than the other individual descriptor sets. Thus, while reasonable fits could be obtained using only one descriptor set in isolation ($R^2 \approx 0.7–0.8$; data not shown), pooling the descriptors provided better fits ($R^2 \geq 0.98$; Table 2).

### 3.5. Protein descriptor covariance

The descriptors used in developing the correlations were taken from several different sources without regard to possible covariance, either within a given descriptor set or among the pooled descriptors. Analysis of covariance was performed to determine which descriptors were correlated strongly with one another. Moderate to high covariance (≥|0.7|)

was observed for some descriptor pairs taken from different descriptor sets, as expected (Table S3). Within a given descriptor set, AGGRESCAN descriptors showed moderate to high covariance (≥|0.7|), as did PASTA descriptors. Some pairs of physical descriptors also showed high covariance (e.g., % *α-helix* vs % &*beta;-sheet*). For any given correlation developed through multiple linear regression (Tables 2 and 4), few or no descriptors were selected that show moderate to high covariance (≥|0.7|).

## 4. Discussion

The results presented here demonstrate that, for a given type of formulation, the extent of protein aggregation on lyophilization is strongly correlated with both physical and heuristic-based computational descriptors of protein structure. The best correlations (Tables 2 and 3) were achieved using an exhaustive search method and descriptors pooled from the AGGRESCAN and PASTA algorithms along with selected physical descriptors (Table 1 and S1). LOOCV demonstrated that the resulting correlations were able to provide good predictions of aggregation propensity. The results suggest that protein structure determines aggregation propensity during lyophilization and can be used for prediction purposes when the formulation components are held constant.

Independently, each of the heuristic-based algorithms provided considerably poorer correlations with lower predictive power than those built from pooled set of descriptors. The descriptors from both the AGGRESCAN and PASTA sets showed high covariance (see Table S3). As a result, the amount of structural information captured by either method is limited despite the large number of descriptors obtained from both methods. The addition of physical descriptors in the pooled set allows more structural features of the protein to be represented and thus provides better fits.

Descriptors selected varied between formulations and no single protein descriptor could account for the extent of aggregation across all formulations. This indicates that, for lyophilized formulations, the excipient and its interactions with the protein are important contributors to aggregation. The heuristic-based algorithms used here do not explicitly include excipient or medium effects. However, the heuristic-based algorithms were developed using data from proteins in solution. As both AGGRESCAN and PASTA descriptors were frequently selected, the algorithms are shown to be useful in prediction of aggregation under lyophilized conditions.

The most commonly selected descriptors provide insight into the factors contributing to lyophilization-induced aggregation. In the eight correlations presented in Tables 2 and 4, *pI*, % *β-sheet*, and $T_m$ were selected five times and were the most commonly selected descriptors. All three have been implicated in aggregation induced by colloidal interactions or protein unfolding. The PASTA descriptors *Peaks* and $E_{avg}$ were selected for three of the four correlations generated by exhaustive search (Table 2). Interestingly, the percent monomer *increased* with increasing *Peaks* values for the buffer and sucrose formulations. While the reason for this is not clear, it may reflect a decrease in the size of each aggregation prone region as the number of regions increases. The PASTA descriptor $E_{avg}$ describes the average interaction energies between residue pairings for a given protein, with lower energies indicating stronger interactions. As the average energies across all pairings for a protein ($E_{avg}$) were more highly selected than the pairing resulting in the minimum energy ($E_{min}$), the presence of several moderately aggregation-prone regions may increase the propensity toward aggregation more than the presence of one highly aggregation-prone region. Also, the two descriptors showed a high covariance (0.99), which may explain why only one was selected. The descriptors *# of free SH* and # S–S combined to be selected in four of the eight correlations. The frequent selection of thiol/disulfide related descriptors is

not surprising, since free thiol groups are reactive and can lead to the formation of disulfide-linked covalent aggregates. SDS-PAGE results confirmed that reducible aggregates were observed for proteins containing four or more free thiol groups (Table S2).

Examination of the descriptors that were *not* selected is also instructive. Apolar surface area (apolar) and fractional apolar surface area ($f_{apolar}$) were not highly selected. The lack of selection of apolar suggests that aggregation during lyophilization is not strongly correlated to total apolar surface area. Furthermore, larger percentages of apolar surface area do not appear to affect aggregation as $f_{apolar}$ was not chosen for any of the correlations.

The correlations developed here can be used as formulation design tools, albeit with limited scope. For example, the correlation for the buffer formulation could be used to assess whether a new protein is likely to be destabilized during aggregation in the absence of excipients, and the correlations for the glycine and sucrose formulations could be used to select the better of these two excipients. The predictive ability of the correlations is expected to be greatest for proteins whose properties fall within the structural space defined by the 15 proteins studied here. Perhaps more importantly, the correlations are limited in that the effects of excipients on aggregation are not included quantitatively, since the number of excipients tested was small. Previous studies by the authors have related protein aggregation on lyophilization to molecular descriptors of structure for carbohydrate excipients (Roughton, Topp, & Camarda, 2012). Broader correlations that address both protein structure and excipient effects would be more useful in formulation design, and could direct the development of new excipients that better prevent aggregation for a given protein.

## 5. Conclusions

Within a given type of formulation, the percent monomer remaining following lyophilization is highly correlated to descriptors of protein structure pooled from the AGGRESCAN and PASTA algorithms along with a limited list of physical descriptors. In general, no one descriptor proved useful for all correlations; rather the work has established that the descriptor classes investigated (AGGRESCAN, PASTA and physical descriptors) are all useful for correlation of aggregation propensity to protein structure. The correlations developed are able to predict an experimental measure of aggregation (%*Monomer* from HP-SEC), offering an easily understandable result for identifying the aggregation propensity of a given protein. Prediction accuracy for a protein “left out” of the data set is reasonable, suggesting the models can be used as a predictive tool. The correlations show that descriptors obtained from solution-based heuristic methods can be used to quantitatively predict aggregation of proteins in the lyophilized state. The descriptors that best correlate with percent monomer vary from formulation to formulation, indicating that model-based approaches should account for formulation to allow for wider applicability. With key protein descriptors identified here, the development of correlations incorporating both protein and excipient properties is warranted and will be the direction of future work. Such correlations would provide a framework for rational formulation design as a function of both protein and excipient.

## Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

## Acknowledgments

The authors gratefully acknowledge financial support through NIH RO1 GM085293 (PI: E. Topp) and from the College of Pharmacy at Purdue University. The authors also thank Haider Tarar for his assistance in preliminary linear regression using R.

## References


Anchordoquy TJ, Carpenter JF. Polymers protect lactate dehydrogenase during freeze-drying by inhibiting dissociation in the frozen state. Archives of Biochemistry and Biophysics. 1996; 332:231–238. [PubMed: 8806730]

Arakawa T, Kita Y. Stabilizing effects of caprylate and acetyltryptophanate on heat-induced aggregation of bovine serum albumin. Biochimica et Biophysica Acta. 2000; 1479:32–36. [PubMed: 10862953]

Bhatnagar BS, Bogner RH, Pikal MJ. Protein stability during freezing: Separation of stresses and mechanisms of protein stabilization. Pharmaceutical Development and Technology. 2007; 12:505–523. [PubMed: 17963151]

Biliaderis CG, Weselake RJ, Petkau A, Friesen AD. A calorimetric study of human CuZn superoxide dismutase. Biochemical Journal. 1987; 248:981–984. [PubMed: 3435496]

Blanco E, Ruso JM, Prieto G, Sarmiento F. Different thermal unfolding pathways of catalase in the presence of cationic surfactants. The Journal of Physical Chemistry B. 2007; 111:2113–2118. [PubMed: 17284066]

Bolboacă SD, Jäntschi L. Modelling the property of compounds from structure: Statistical methods for models validation. Environmental Chemistry Letters. 2008; 6:175–181.

Borrebaeck C, Mattiasson B. A study of structurally related binding properties of concanavalin A using differential scanning calorimetry. European Journal of Biochemistry. 1980; 107:67–71. [PubMed: 7398639]

Branchu S, Forbes RT, York P, Nyqvist H. A central composite design to investigate the thermal stabilization of lysozyme. Pharmaceutical Research. 1999; 16:702–708. [PubMed: 10350014]

Chalikian TV, Volker J, Anafi D, Breslauer KJ. The native and the heat-induced denatured states of alpha-chymotrypsinogen A: thermodynamic and spectroscopic studies. Journal of Molecular Biology. 1997; 274:237–252. [PubMed: 9398530]

Chang BS, Beauvais RM, Dong A, Carpenter JF. Physical factors affecting the storage stability of freeze-dried interleukin-1 receptor antagonist: Glass transition and protein conformation. Archives of Biochemistry and Biophysics. 1996; 331:249–258. [PubMed: 8660705]

Chang BS, Kendrick BS, Carpenter JF. Surface-induced denaturation of proteins during freezing and its inhibition by surfactants. Journal of Pharmaceutical Sciences. 1996; 85:1325–1330. [PubMed: 8961147]

Chen WJ, Lee IS, Chen CY, Liao TH. Biological functions of the disulfides in bovine pancreatic deoxyribonuclease. Protein Science. 2004; 13:875–883. [PubMed: 15044724]

Chennamsetty N, Voynov V, Kayser V, Helk B, Trout BL. Design of therapeutic proteins with enhanced stability. Proceedings of the National Academy of Sciences of the United States of America. 2009; 106:11937–11942. [PubMed: 19571001]

Chi EY, Krishnan S, Randolph TW, Carpenter JF. Physical stability of proteins in aqueous solution: Mechanism and driving forces in nonnative protein aggregation. Pharmaceutical Research. 2003; 20:1325–1336. [PubMed: 14567625]

Conchillo-Sole O, de Groot NS, Aviles FX, Vendrell J, Daura X, Ventura S. AGGRESCAN: A server for the prediction and evaluation of hot spots of aggregation in polypeptides. BMC Bioinformatics. 2007; 8:65. [PubMed: 17324296]

Dalgaard, P. Introductory Statistics with R. 2. New York, NY: Springer Science + Business Media, LLC; 2008.

Duy C, Fitter J. Thermostability of irreversible unfolding alpha-amylases analyzed by unfolding kinetics. Journal of Biological Chemistry. 2005; 280:37360–37365. [PubMed: 16150692]

Eslick JC, Ye Q, Park J, Topp EM, Spencer P, Camarda KV. A computational molecular design framework for crosslinked polymer networks. Computers and Chemical Engineering. 2009; 33:954–963. [PubMed: 23904665]

Hendrix TM, Griko Y, Privalov P. Energetics of structural domains in alpha-lactalbumin. Protein Science. 1996; 5:923–931. [PubMed: 8732764]

Irbäck A, Mohanty S. PROFASI: A Monte Carlo simulation package for protein folding and aggregation. Journal of Computational Chemistry. 2006; 27:1548–1555. [PubMed: 16847934]

Jain RK, Hamilton AD. Designing Protein Denaturants: Synthetic Agents Induce Cytochrome c Unfolding at Low Concentrations and Stoichiometries. Angewandte Chemie International Edition. 2002; 41:641–643.

Katayama DS, Nayar R, Chou DK, Campos J, Cooper J, Vander Velde DG, Villarete L, Liu CP, Cornell Manning M. Solution behavior of a novel type 1 interferon, interferon-tau. Journal of Pharmaceutical Sciences. 2005; 94:2703–2715. [PubMed: 16258985]

Kella NK, Kinsella JE. Enhanced thermodynamic stability of beta-lactoglobulin at low pH. A possible mechanism. Biochemical Journal. 1988; 255:113–118. [PubMed: 3196307]

Kelly L, Holladay LA. A comparative study of the unfolding thermodynamics of vertebrate metmyoglobins. Biochemistry. 1990; 29:5062–5069. [PubMed: 2378866]

Lauer TM, Agrawal NJ, Chennamsetty N, Egodage K, Helk B, Trout BL. Developability index: A rapid in silico tool for the screening of antibody aggregation propensity. Journal of Pharmaceutical Sciences. 2012; 101:102–115. [PubMed: 21935950]

Lumley, T. Package ‘leaps’. 2013. http://cran.r-project.org/web/packages/leaps/leaps.pdf

Ma B, Nussinov R. Simulations as analytical tools to understand protein aggregation and predict amyloid conformation. Current Opinion in Chemical Biology. 2006; 10:445–452. [PubMed: 16935548]

Maindonald, J.; Braun, WJ. Data analysis and graphics using R: An example-based approach. 3. New York. NY: Cambridge University Press; 2010.

Prestrelski SJ, Pikal KA, Arakawa T. Optimization of lyophilization conditions for recombinant human interleukin-2 by dried-state conformational analysis using Fourier-transform infrared spectroscopy. Pharmaceutical Research. 1995; 12:1250–1259. [PubMed: 8570516]

Printz M, Kalonia DS, Friess W. Individual second virial coefficient determination of monomer and oligomers in heat-stressed protein samples using size-exclusion chromatography-light scattering. Journal of Pharmaceutical Sciences. 2012; 101:363–372. [PubMed: 21938728]

Quan NT. The prediction sum of squares as a general measure for regression diagnostics. Journal of Business and Economic Statistics. 1988; 6:501–504.

Rosenberg A. Effects of protein aggregates: An immunologic perspective. The AAPS Journal. 2006; 8:E501–E507. [PubMed: 17025268]

Roughton BC, Topp EM, Camarda KV. Use of glass transitions in carbohydrate excipient design for lyophilized protein formulations. Computers and Chemical Engineering. 2012; 36:208–216.

Roychaudhuri R, Sarath G, Zeece M, Markwell J. Reversible denaturation of the soybean Kunitz trypsin inhibitor. Archives of Biochemistry and Biophysics. 2003; 412:20–26. [PubMed: 12646263]

Sophocleous AM, Zhang J, Topp EM. Localized effects of hydration on lyophilized myoglobin by hydrogen/deuterium exchange mass spectrometry 1. Exchange mapping. Molecular Pharmaceutics. 2012 (submitted for publication).

Takahashi T, Irie M, Ukita T. A comparative study on enzymatic activity of bovine pancreatic ribonuclease A, ribonuclease S and ribonuclease S′. Journal of Biochemistry. 1969; 65:55–62. [PubMed: 5771710]

Tani F, Shirai N, Onishi T, Venelle F, Yasumoto K, Doi E. Temperature control for kinetic refolding of heat-denatured ovalbumin. Protein Science. 1997; 6:1491–1502. [PubMed: 9232650]

Tartaglia GG, Vendruscolo M. The Zyggregator method for predicting protein aggregation propensities. Chemical Society Reviews. 2008; 37:1395–1401. [PubMed: 18568165]

Trovato A, Seno F, Tosatto SC. The PASTA server for protein aggregation prediction. Protein Engineering Design & Selection. 2007; 20:521–523.

Wasserman, L. All of statistics: A concise course in statistical inference. New York: Springer; 2004.

Wu J, Mei J, Wen S, Liao S, Chen J, Shen Y. A self-adaptive genetic algorithm-artificial neural network algorithm with leave-one-out cross validation for descriptor selection in QSAR study. Journal of Computational Chemistry. 2010; 31:1956–1968. [PubMed: 20512843]

Zhang J, Huan J. Comparison of chemical descriptors for protein-chemical interaction prediction. International Journal of Computational Bioscience. 2010; 1:13–21.

## Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.compchemeng.2013.07.008.

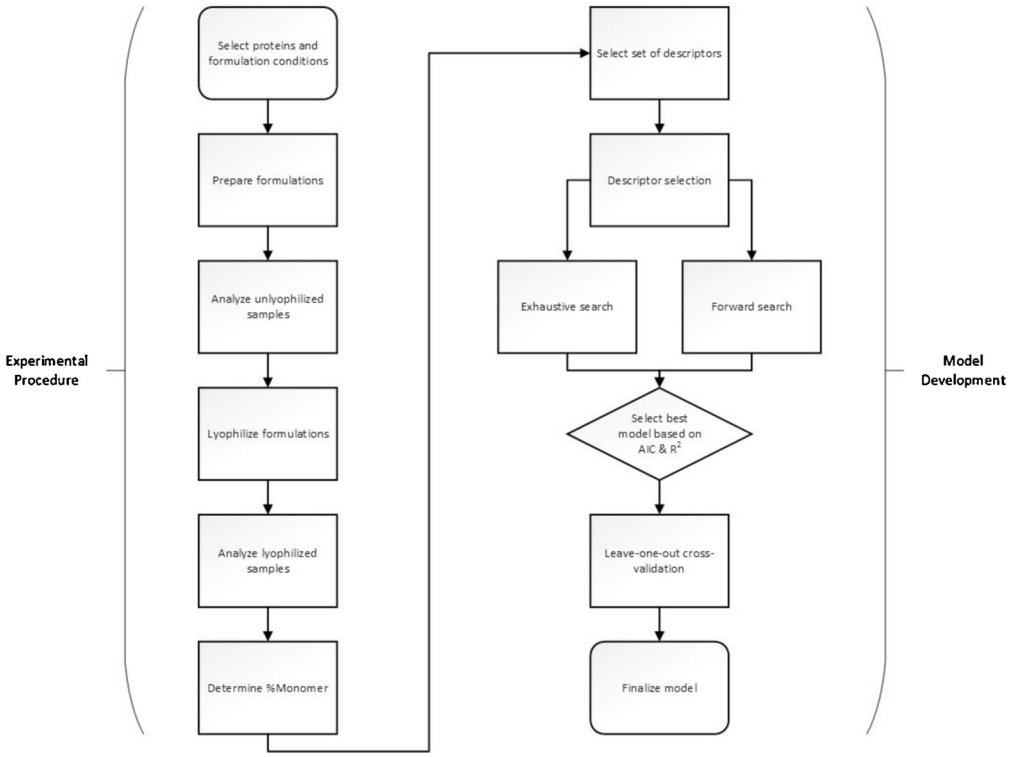


**Fig. 1.**
Overview of experimental procedure and model development. Experimental results were taken in triplicate.

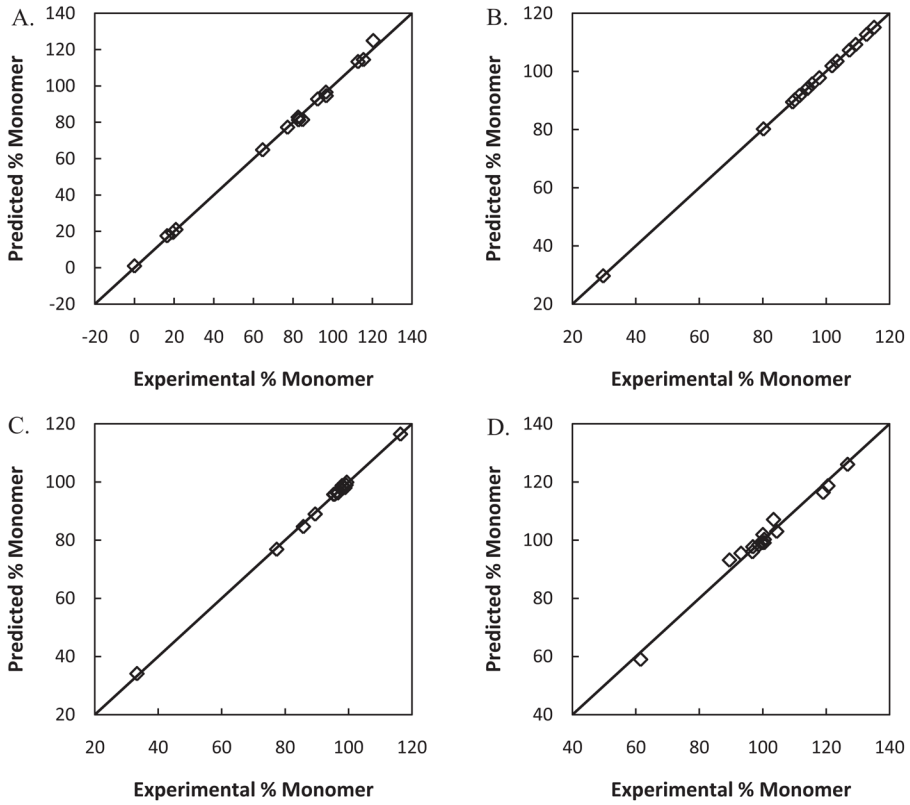


**Fig. 2.**

Parity plots of experimental percent monomeric protein values (%Monomer) from HP-SEC versus predicted percent monomeric protein values from developed correlations for (A) urea, (B) buffer (potassium phosphate 20 mM, pH 7.4), (C) sucrose, and (D) glycine formulations.

*Comput Chem Eng*. Author manuscript; available in PMC 2014 November 11.

**Table 1**

Proteins used to relate lyophilization-induced aggregation to protein structure, with physical descriptors of their structure.

| **Protein**[a] | **PDB code** | **MW (kDa)** | ***pI*** | **% α helix** | **% β sheet** | **# SS bonds** | **# free thiols** | $T_m$ **(° C)**[b] |
|---|---|---|---|---|---|---|---|---|
| Myoglobin | 1wla | 16.9 | 7.2 | 75.8 | 0 | 0 | 0 | 81.6 (Kelly & Holladay, 1990) |
| Lysozyme | 1lyz | 14.3 | 9.36 | 41.9 | 6.2 | 4 | 0 | 75.7 (Branchu et al., 1999) |
| Ovalbumin | 1ova | 44.3 | 5.19 | 30.8 | 31.3 | 1 | 4 | 76 (Tani et al., 1997) |
| RNase A | 5rsa | 16.5 | 8.93 | 21 | 33 | 4 | 0 | 62.5 (Takahashi et al., 1969) |
| DNase I | 3dni | 31.3 | 5.34 | 28 | 28 | 2 | 0 | 65 (Chen et al., 2004) |
| α-CT A | 2cga | 25.7 | 8.52 | 13.5 | 32 | 5 | 0 | 50 (Chalikian et al., 1997) |
| Cytochrome C | 2b4z | 12.3 | 10.02 | 41 | 0 | 0 | 0 | 85 (Jain & Hamilton, 2002) |
| Con A | 3cna | 25.5 | 5.47 | 3.8 | 46.4 | 0 | 0 | 101 (Borrebaeck & Mattiasson, 1980) |
| Catalase | 4blc | 59.9 | 6.79 | 29.6 | 19.7 | 0 | 4 | 56.2 (Blanco et al., 2007) |
| α-amylase | 1bli | 58 | 6.33 | 26.2 | 25.6 | 0 | 0 | 102 (Duy & Fitter, 2005) |
| α-lactalbumin | 1f6s | 14.2 | 4.92 | 45 | 7.1 | 4 | 0 | 68.9 (Hendrix et al., 1996) |
| β-lactoglobulin | 1cj5 | 19.9 | 4.93 | 11 | 31 | 2 | 1 | 64.8 (Kella & Kinsella, 1988) |
| SOD | 1cbj | 16.3 | 5.85 | 4 | 31 | 1 | 1 | 96 (Biliaderis et al., 1987) |
| BSA | 3v03 | 66 | 5.82 | 67 | 0 | 17 | 1 | 59 (Arakawa & Kita, 2000) |
| Trypsin inhibitor | 1avu | 20.1 | 4.95 | 1.4 | 97.2 | 2 | 0 | 65 (Roychaudhuri et al., 2003) |

[a] Abbreviations: RNase A = ribonuclease A; DNase I = deoxyribonuclease I; α-CT A = α-chymotrypsinogen A; Con A = concanavalin A; SOD = superoxide dismutase; BSA = bovine serum albumin.

[b] Protein melting temperatures ($T_m$) obtained from literature; citation shown in brackets, see References.

**Table 2**

Descriptors selected using an exhaustive search method with model size selected by minimizing AIC score.

| Formulation | Model Size | Descriptors selected | | | $R^2$ | $Q^2$ | $R^2 - Q^2$ |
|---|---|---|---|---|---|---|---|
| | | Physical | AGGRESCAN | PASTA | | | |
| Buffer[a] | 12 | %α-Helix, %β-Sheet, MW, # S–S, # free SH, $T_m$ | a3vSA, THSA | $E_{min}$, $E_{avg}$, $(E/L)_{min}$, Peaks | 1.000 | 0.999 | 0.001 |
| Urea | 10 | Apolar, *pI*, # S–S, $T_m$ | TA, Na4vSS | $E_{avg}$, $L_{avg}$, $(E/L)_{min}$, $(E/L)_{avg}$ | 0.998 | 0.976 | 0.022 |
| Sucrose | 9 | %β-Sheet, MW, *pI*, $T_m$ | a3vSA, NnHS, THSA, TA | Peaks | 0.999 | 0.987 | 0.012 |
| Glycine | 8 | %α-Helix, %β-Sheet | NnHS, AATr, THSAr, Na4vSS | $E_{avg}$, Peaks | 0.982 | 0.805 | 0.176 |

[a]Buffer used in the formulation was potassium phosphate buffer (20 mM, pH 7.4).

## Table 3

Correlation results for all four formulations. Descriptors were selected via exhaustive search with AIC evaluation.

| Formulation | Descriptor | Coefficient value | Standard error (*p*-value[a]) |
|---|---|---|---|
| *Buffer*[b] | (Intercept) | −0.25 | $0.48^{(0.65)}$ |
| | % α-Helix | −0.37 | $0.01^{***}$ |
| | % β-Sheet | −0.17 | $0.01^{**}$ |
| | MW | −0.53 | $0.01^{***}$ |
| | # S–S | −0.88 | $0.02^{***}$ |
| | # Free SH | −13.68 | $0.11^{***}$ |
| | $T_m$ | −0.50 | $0.01^{***}$ |
| | a3vSA | 155.36 | $2.70^{***}$ |
| | THSA | −0.07 | $0.03^{(0.13)}$ |
| | $E_{min}$ | −16.25 | $0.30^{***}$ |
| | $E_{avg}$ | 3.27 | $0.31^{**}$ |
| | $(E/L)_{min}$ | −53.63 | $0.34^{***}$ |
| | Peaks | 11.23 | $0.05^{***}$ |
| *Urea* | (Intercept) | 164.10 | $13.07^{***}$ |
| | Apolar | 4.66E – 03 | $1.57E-04^{***}$ |
| | *pI* | −7.16 | $0.80^{***}$ |
| | # S–S | 5.85 | $0.38^{***}$ |
| | $T_m$ | −2.17 | $0.09^{***}$ |
| | TA | −0.32 | $0.08^{*}$ |
| | Na4vSS | 5.13 | $0.47^{***}$ |
| | $E_{avg}$ | 12.53 | $1.64^{**}$ |
| | $L_{avg}$ | 6.37 | $0.59^{***}$ |
| | $(E/L)_{min}$ | −338.10 | $17.14^{***}$ |
| | $(E/L)_{avg}$ | 260.90 | $20.63^{***}$ |
| *Sucrose* | (Intercept) | 159.17 | $2.77^{***}$ |
| | % β-Sheet | 0.69 | $0.03^{***}$ |
| | MW | −2.84 | $0.08^{***}$ |
| | *pI* | −0.84 | $0.25^{*}$ |
| | $T_m$ | 0.62 | $0.03^{***}$ |
| | a3vSA | 686.56 | $17.31^{***}$ |
| | NnHS | −19.03 | $0.49^{***}$ |
| | THSA | 1.69 | $0.09^{***}$ |
| | TA | −2.44 | $0.06^{***}$ |

| Formulation | Descriptor | Coefficient value | Standard error (*p*-value[a]) |
|---|---|---|---|
| | Peaks | 3.89 | 0.20*** |
| *Glycine* | (Intercept) | 387.91 | 19.91*** |
| | % α-Helix | −0.65 | 0.11*** |
| | % β-Sheet | −0.53 | 0.17* |
| | NnHS | −23.98 | 1.88*** |
| | AATr | −2326.82 | 162.79*** |
| | THSAr | 2247.68 | 155.16*** |
| | Na4vSS | 3.04 | 0.31*** |
| | AvgE | 13.64 | 1.33*** |
| | Peaks | −3.73 | 0.60*** |

[a]Significance codes for the *p*-values are:

*** for <0.001,

** for <0.01,

* for <0.05.

[b]Buffer used in the formulation was potassium phosphate buffer (20 mM, pH 7.4).

**Table 4**

Descriptors selected using a forward search method with AIC evaluation. Numbering indicates order in which descriptors were selected. No emphasis indicates physical descriptors and bold text indicates AGGRESCAN descriptors. No PASTA descriptors were selected.

| Formulation | Descriptors selected | | | | | Regression performance | | |
|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | $R^2$ | $Q^2$ | $R^2 - Q^2$ |
| Buffer[a] | *pI* | $T_m$ | %β-Sheet | %α-Helix | THSA | 0.74 | 0.16 | 0.58 |
| Urea | # Free SH | *pI* | $T_m$ | – | – | 0.54 | 0.11 | 0.43 |
| Sucrose | *pI* | %β-Sheet | – | – | – | 0.57 | 0.19 | 0.38 |
| Glycine | a3vSA | – | – | – | – | 0.14 | −0.28 | 0.42 |

[a]Buffer used in the formulation was potassium phosphate buffer (20 mM, pH 7.4).