Published in final edited form as:

Comput Chem Eng. 2012 January 10; 36(10): . doi:10.1016/j.compchemeng.2011.07.018.

Use of glass transitions in carbohydrate excipient design for lyophilized protein formulations

Brock C. Roughton^a, E.M. Topp^b, and Kyle V. Camarda^{c,*}

^a University of Kansas, Bioengineering Graduate Program, 1520 W. 15th Street, Lawrence, KS 66045, United States

^b Purdue University, Department of Industrial and Physical Pharmacy, 575 Stadium Mall Drive, West Lafayette, IN 47907, United States

^c University of Kansas, Department of Chemical and Petroleum Engineering, 1530 W. 15th Street, Lawrence, KS 66045, United States

Abstract

This work describes an effort to apply methods from process systems engineering to a pharmaceutical product design problem, with a novel application of statistical approaches to comparing solutions. A computational molecular design framework was employed to design carbohydrate molecules with high glass transition temperatures and low water content in the maximally freeze-concentrated matrix, with the objective of stabilizing lyophilized protein formulations. Quantitative structure–property relationships were developed for glass transition temperature of the anhydrous solute, glass transition temperature of the maximally concentrated solute, melting point of ice and Gordon–Taylor constant for carbohydrates. An optimization problem was formulated to design an excipient with optimal property values. Use of a stochastic optimization algorithm, Tabu search, provided several carbohydrate excipient candidates with statistically similar property values, as indicated by prediction intervals calculated for each property.

Keywords

Molecular design; Excipient; Lyophilization; Protein aggregation; Stochastic optimization

1. Introduction

When a protein is identified as a therapeutic candidate, a formulation must be developed. Additives, or excipients, are included in the formulation to stabilize and improve the final drug product. Stability is a major concern, as many degradation routes exist for proteins. For example, degradation can occur via aggregation, deamidation, isomerization, oxidation, glycation, and thioldisulphide exchange (Cleland et al., 1994; Manning, Chou, Murphy, Payne, & Katayama, 2010). Degradation not only results in product loss, but also can lead to issues in regulatory approval. In most cases, the FDA requires pharmaceutical product degradation to be below 10% of the product's final weight (Cleland et al., 1994).

For proteins that prove unstable under aqueous conditions, lyophilization is often employed. Lyophilization removes water from the formulation by sublimation, first through a freezing step and then through primary and secondary drying steps (Cleland et al., 1994; Costantino & Pikal, 2004). By removing water, the mobility of the protein is reduced and stability is improved. The resulting product is an amorphous solid with minimal water content. Lyophilization is the most common formulation choice for protein drugs, representing 46% of the biopharmaceuticals approved by the FDA through December 2003 (Costantino & Pikal, 2004). However, even lyophilized protein formulations can be subject to degradation, notably via aggregation.

In the lyophilized form, the protein can undergo aggregation, which is an often irreversible self-association resulting in an inactive protein complex (Cleland et al., 1994; Wang, 2005; Wang, Nema, & Teagarden, 2010). The aggregation process can occur due to physical interactions between protein surfaces or can result from chemical interactions of the amino acids, forming covalent bonds between proteins (Wang, 2005). The detection of aggregates is difficult, as the protein complexes may be insoluble (Wang, 2005). Aggregation not only lowers the efficacy of the protein drug but can also lead to severe and life threatening immunogenic responses (Rosenberg, 2006). Excipients in a lyophilized formulation should be selected to minimize aggregation, ensuring the safety and efficacy of the final product.

Carbohydrate excipients, such as sucrose and trehalose, have been shown to stabilize protein structure (Fung, Darabie, & McLaurin, 2005; Li, Williams, & Topp, 2008; Sinha, Li, Williams, & Topp, 2008). Many lyophilized protein formulations utilize simple sugars, disaccharides, oligosaccharides, or sugar alcohols as stabilizers (Cleland et al., 1994). Two main theories have been proposed for describing an excipient's ability to stabilize biomolecules during lyophilization: water replacement and vitrification. In the water replacement theory, stabilizing excipients are those that can substitute for water in the dried state through hydrogen bonding with the protein (Cleland et al., 1994). The vitrification hypothesis proposes that stabilizing excipients are those that form glasses during lyophilization (Crowe, Carpenter, & Crowe, 1998).

The vitrification approach was considered when designing an optimal carbohydrate excipient in this work. During lyophilization, a concentrated amorphous glass is produced during the freezing step and a mostly water free glass is produced during the drying steps (Costantino & Pikal, 2004). Initially, as a solution with excipients is cooled, a concentrated supercooled liquid or rubbery state is formed. The melting point of ice occurs first and represents the point where ice begins to form in the concentrated rubbery phase (Roos, 1997). Upon further cooling, the concentrated rubbery phase transitions into a concentrated glass phase, marked by the glass transition temperature of the maximally concentrated solute (Roos, 1997). The ice is removed via sublimation as a vacuum is applied to the system (Costantino & Pikal, 2004). The amount of water remaining in the maximally freeze-concentrated glass matrix can be estimated by the Gordon–Taylor equation (Costantino & Pikal, 2004; Roos, 1993). The lyophilized product temperature is then raised and residual water in the maximally freeze-concentrated glass matrix is removed during the drying steps. The phase transitions that occur in a lyophilized formulation are summarized in Fig. 1.

In the present work, glass transitions that occur during lyophilization are used as targets for the optimal design of a carbohydrate excipient to be used as a stabilizer in protein formulations. Roos suggested that the increase of transition temperatures could be “used in product development to improve freeze-drying behavior and stability of dehydrated materials” (1997). Experimental trial-and-error approaches could be used to identify such compounds, but the approaches are usually expensive and time-consuming. The literature has shown glass transitions to be dependent on chemical structure (Slade & Levine, 1995).

A lyophilized product must reach a temperature below both the glass transition temperature of the maximally concentrated solute and the melting point of ice to ensure minimal water content and glass formation, restricting protein mobility. By restricting mobility, the protein's potential to aggregate is reduced. An appreciable temperature difference between the glass transition temperature of the maximally concentrated solute and the melting point of ice is also desired, as the freeze-concentrated solution is annealed between these temperatures to ensure maximal solute concentration and minimal water content (Roos, 1997). The glass transition temperature of the anhydrous solute is important for long-term storage stability as well, as lyophilized formulations are usually stored at temperatures at least 50 °C below their glass transition temperature (Costantino & Pikal, 2004). The glass transition temperature of the maximally concentrated solute is used with the Gordon–Taylor constant and the glass transition temperature of the anhydrous solute to estimate the water content in the final freeze-concentrated matrix (Costantino & Pikal, 2004; Roos, 1993). An ideal excipient will form a freeze-concentrate with minimal water content and will remain a glass during drying and storage, restricting protein mobility and reducing the potential for aggregation.

Computational methods are commonly employed in the design of pharmaceutically relevant molecules. The most frequently utilized methods involve fragment-based or structure-based drug design, where a molecule is designed to interact with a particular biological target based on binding affinity (Congreve, Chessari, Tisi, & Woodhead, 2008; Zoete, Grosdidier, & Michielin, 2009). An especially common and successful application has been the design of new antibiotic candidates (Cherkasov et al., 2008; Haddad et al., 2002). An alternative methodology known as computational molecular design has been used in the design of drug molecules (Siddhaye, Camarda, Topp, & Southard, 2000). Recently computational molecular design has been extended to design alternative excipients for drug formulations (Solvason, Chemmangattuvalappil, & Eden, 2010).

In this work, computational molecular design (CMD) was applied to the design of novel carbohydrate excipients with optimal properties for minimizing aggregation in lyophilized protein formulations. CMD offers a quicker and cheaper method to identify candidate excipients when compared to trial-and-error methods. CMD methodology consists of a forward and a backward problem (Venkatasubramanian, Chan, & Caruthers, 1994). For the forward problem, quantitative structure–property relationships (QSPRs) have been developed for the glass transition temperature of the anhydrous solute, glass transition temperature of the maximally concentrated solute, melting point of ice and Gordon–Taylor constant for carbohydrates. The experimental data were collected from published literature (Roos, 1993). Although more recent data exists (Costantino & Pikal, 2004; Crowe, Reid, & Crowe, 1996; Her & Nail, 1994), the data set selected is preferable as it includes property values for the largest number of carbohydrate molecules. Additionally, there is some debate about where the glass transitions occur (Roos, 1997), so using one data set ensures that all the properties were determined in the same manner. In the relationships, the properties were correlated to the topology of the molecule as represented by connectivity indices (Bicerano, 2002; Kier & Hall, 1986; Randic, 1975). Error estimation for the properties predicted by the QSPRs was performed using prediction intervals (Wasserman, 2004). For solution of the inverse (product design) problem, the QSPRs were used in a CMD framework utilizing a stochastic optimization method known as Tabu search to select candidate molecules with property values estimated to match prespecified targets (Eslick et al., 2009; Glover, 1990; Lin, Chavali, Camarda, & Miller, 2005). Prediction intervals were used to compare locally optimal solutions generated from Tabu search, to the authors' knowledge providing the first attempt at statistical comparison of locally optimal solutions in computational molecular design. While the properties of the excipients can be predicted with the QSPRs developed, the actual properties of the lyophilized product may be slightly different due to the presence

of the protein. By designing carbohydrates with high glass transition temperatures and low water content in the maximally freeze-concentrated matrix, several optimal excipients for lyophilized protein formulations have been proposed.

2. Methods

Computational molecular design (CMD) methodology consists of a forwards and a backwards problem that must be solved to provide an optimal molecule of interest. CMD requires a set of QSPRs to describe the properties of interest for the class of compounds to be designed for the forward problem. The chemical structure or topology must be classified by molecular descriptors, which are used with experimental data to develop the set of QSPRs. The best molecular descriptors should be selected to ensure a good predictive model is developed. Error in the predicted properties can be estimated using prediction intervals. The QSPRs are used with any other available property equations in the CMD framework to design a molecule with optimal properties.

2.1. Calculation of molecular descriptors

Several types of molecular descriptors have been considered for use in property prediction. Group contribution methods identify common molecular groups, which are then used to build the molecule of interest. Group contribution (GC) methods have been utilized successfully to predict many physical properties of organic compounds, including boiling point and freezing point (Joback & Reid, 1987). GC approaches have also been used to predict vapor liquid equilibria (Gani, Tzouvaras, Rasmussen, & Fredenslund, 1989). Quantum approaches by use of COSMO-RS have been proposed for predicting extraction of drug molecules (Lei, Chen, & Li, 2007).

Topological descriptors are a class of molecular descriptors which identify individual atoms and their bonding configuration in a molecule. Connectivity indices are a type of topological descriptors first proposed by Randic (1975). Later work extended the use of connectivity indices to pharmaceutical product property prediction (Kier & Hall, 1986) and polymer property prediction (Bicerano, 2002). The use of connectivity indices have been proposed to describe missing groups for GC models (Gani, Harper, & Hostrup, 2005; Satyanarayana, Abildskov, & Gani, 2009).

Connectivity indices are employed here due to the ability to calculate indices and to store bonding information for any excipient proposed, regardless of the molecular groups present. A hydrogen-suppressed graph is used to represent the topology of the molecule in connectivity index calculations. Each non-hydrogen atom is represented as a vertex and each bond between two non-hydrogen atoms is represented as an edge. The omission of hydrogen atoms is accounted for in the calculation of the connectivity indices (Bicerano, 2002). The carbohydrate trehalose is used as an example of graph representation in Fig. 2.

The simple connectivity index (δ) for any non-hydrogen atom is equal to the number of non-hydrogen atoms bonded to the given non-hydrogen atom. In terms of the molecular graph, the δ of each vertex is equal to the number of edges adjacent to the vertex. The simple valence connectivity index (δ^v) incorporates the electronic configuration of an atom. The equation of δ^v is given below, where Z^v is the number of valence electrons of an atom, N_H is the number of hydrogen atoms bonded to an atom, and Z is the atomic number of an atom.

$$\delta^v = \frac{Z^v - N_H}{Z - Z^v - 1} \quad (1)$$

The simple and simple valence connectivity indices are used in the calculation of n th-order connectivity (${}^n\chi$) and valence connectivity (${}^n\chi^v$) indices. The zeroth-order ($n = 0$) connectivity indices are known as atomic connectivity indices, representing all atoms in a molecule. The first-order ($n = 1$) connectivity indices are known as bond connectivity indices, representing all the bonds in a molecule (Bicerano, 2002). The second-order ($n = 2$) connectivity indices represent all paths in a molecule that are two bonds in length. It is possible to calculate connectivity indices at higher orders, but this work uses only zeroth-, first-, and second-order connectivity indices to simplify calculations. Eq. (2) gives the generalized equation for n th-order connectivity indices, while the form for n th-order valence connectivity indices is given by Eq. (3) (Consonni & Todeschini, 2000). The order is given by n , N_s represents the number of subgraphs of n th order, and k is an index indicating the atom of interest.

$${}^n\chi = \sum_{i=1}^{N_s} \left(\prod_{k=1}^{n+1} \frac{1}{\delta_k} \right)_i^{1/2} \quad (2)$$

$${}^n\chi^v = \sum_{i=1}^{N_s} \left(\prod_{k=1}^{n+1} \frac{1}{\delta_k^v} \right)_i^{1/2} \quad (3)$$

This work also employs the use of n th-order average connectivity (${}^n\xi$) and average valence connectivity (${}^n\xi^v$) indices, given by Eqs. (4) and (5). The n th-order average connectivity index is calculated by dividing the n th-order connectivity index by the total number of n th-order subgraphs present in the molecular graph. In molecular terms, the zeroth-order is normalized by the number of atoms, the first order is normalized by the number of bonds, and the second-order is normalized by the number of paths that are two bonds in length. For all connectivity indices utilized, the second-order was assumed to be the highest order necessary to represent the properties under consideration. The fit provided by the QSPRs developed show that zeroth-, first-, and second-order indices are sufficient to correlate the properties of interest. Table 1 gives the values of all zeroth-, first-, and second-order connectivity indices described in this work for trehalose.

$${}^n\xi = \frac{{}^n\chi}{N_s} \quad (4)$$

$${}^n\xi^v = \frac{{}^n\chi^v}{N_s} \quad (5)$$

2.2. Selection of molecular descriptors for QSPRs

When developing property correlations such as QSPRs, a trade-off between accuracy and simplicity must be made. Any property data set can be correlated with a perfect fit by selecting as many descriptors as the number of data points minus one. While such a correlation offers high accuracy for the compounds used to generate the correlation, the correlation may perform poorly when predicting the property for a compound not used to generate the correlation due to the complexity. When selecting the number of descriptors to use for a property correlation, both the accuracy and the simplicity must be taken into account. Several methods may be employed to determine the best number of descriptors.

Some examples include the training error, Mallows's C_p statistic, cross-validation, and the Akaike Information Criterion (Wasserman, 2004).

Mallows's C_p statistic offers an estimate of the accuracy and simplicity of a correlation and has been employed in this work. The C_p statistic assigns a score to a given QSPR based upon lack of fit plus a penalty for complexity (Wasserman, 2004). The C_p value for some correlation S is calculated as follows:

$$C_p(S) = \left(\sum_{j=1}^m (P_j(S) - P_j)^2 \right) + 2l\hat{\sigma}^2 \quad (6)$$

where $P_j(S)$ is the predicted property at observation j given by correlation S , P_j is the actual measured property at observation j , m is the number of observations, and l is the number of terms in the correlation S (equal to the number of descriptors used plus one, for the intercept). The value given by $\hat{\sigma}^2$ is an unbiased estimate of the variance (Wasserman, 2004), shown below:

$$\hat{\sigma}^2 = \left(\frac{1}{m-l} \right) \sum_{j=1}^m (P_j(S) - P_j)^2 \quad (7)$$

In this work, the statistical program R (R-Development-Core-Team, 2010) was used to linearly regress the properties of interest to the connectivity indices selected for use, yielding the desired QSPR models. Experimental data for the properties was collected from literature (Roos, 1993). The method of least-squares fitting was used (Wasserman, 2004). The leaps package (Lumley, 2004) was utilized for descriptor selection in the QSPR models. For each model size, indicated by the number of connectivity indices used, leaps was used to conduct an exhaustive search looking at all possible descriptor combinations for a given model size to determine which set of connectivity indices yielded the lowest value for Mallows's C_p statistic. The procedure was repeated for each model size from one descriptor to twelve descriptors. The C_p values for all the models were compared and the model size with the lowest value was selected for use as a QSPR for the property of interest. The C_p statistic was only used as a criterion in the generation of the QSPRs.

2.3. Quantification of error in predicted properties

A QSPR's use lies in predicting a property for a new molecule of interest, represented by the molecular descriptors. The resulting predicted value has some error term involved. A novel approach to solution comparison using prediction intervals was utilized. Prediction intervals allow for the error from the fitted model as well as any error in a future observation to be quantified (Wasserman, 2004). A confidence interval only accounts for error in the correlation, so a prediction interval is always larger than a confidence interval.

A $1 - \alpha$ prediction interval for a predicted property of interest P_* is given by Eq. (8).

$$P_* \pm t_{\alpha/2} \sqrt{\hat{\sigma}^2 \left(x_*^T (X^T X)^{-1} X_* + 1 \right)} \quad (8)$$

where $t_{\alpha/2}$ is student's t -test value for $m - l$ degrees of freedom, x_* is the vector of descriptors used in the prediction, x_*^T is the transposed vector of descriptors used in the prediction, X is the matrix of descriptors used to build the correlation, and X^T is the transposed matrix of descriptors used to build the correlation.

Given the connectivity indices for a designed carbohydrate excipient, new property values were predicted. Using the connectivity index values, R was used to calculate prediction intervals at a 95% level. The prediction intervals were included to provide a reasonable range for the expected properties of the designed excipient molecule. The prediction intervals were also used to determine if two solutions from the optimal design were statistically different once the solutions were found using Tabu search. When comparing the predicted properties of two designed molecules, overlapping prediction intervals indicate that the solutions are not statistically different. The use of prediction intervals to compare solutions represents a novel approach to evaluating solutions generated in computational molecular design.

2.4. Calculation of solute concentration of maximally freeze-concentrated matrix

The amount of water remaining in the maximally freeze-concentrated glass matrix is an important parameter for a lyophilized protein formulation. Water can allow for increased molecular mobility within the glass. Protein–protein interactions occur more readily with higher mobility, increasing the likelihood of aggregation. The water content in the maximally freeze-concentrated matrix can be estimated by the Gordon–Taylor equation, given below (Costantino & Pikal, 2004; Roos, 1993).

$$T_g = \frac{w_1 T_{g,1} + k w_2 T_{g,2}}{w_1 + k w_2} \quad (9)$$

where T_g is the glass transition temperature of the mixture of compounds 1 and 2, w_1 and w_2 are the weight fractions of compounds 1 and 2, $T_{g,1}$ and $T_{g,2}$ are the glass transition temperatures of the individual compounds, and k is the Gordon–Taylor constant. Compound 1 is the solute and compound 2 is the solvent. There is a unique Gordon–Taylor constant value for any given two compounds comprising a mixture, which is determined experimentally (Roos, 1993).

For use with the maximally freeze-concentrated glass, the glass transition temperature of the maximally freeze-concentrated solute is used for the glass transition temperature of the mixture. The solute compound is the carbohydrate excipient and the solvent is water. The glass transition temperature of the anhydrous solute is used for $T_{g,1}$ and the glass transition temperature of water (-135°C) is used for $T_{g,2}$ (Roos, 1993). By rearranging the Gordon–Taylor equation, the solute concentration (weight fraction) of the maximally freeze concentrated matrix (C'_g) can be solved for, as shown in Eq. (10).

$$C'_g = \frac{k (T_{g,\text{water}} - T'_g)}{T'_g - T_{g,\text{carb}} + k (T_{g,\text{water}} - T'_g)} \quad (10)$$

2.5. CMD problem formulation and solution

As noted above, the CMD methodology consists of a forward and an inverse problem (Venkatasubramanian et al., 1994). The forward problem requires property prediction, fulfilled by development of the QSPR models. The inverse or product design problem designs a novel molecule based on target property values.

Optimization methods are needed to find feasible, near optimal solutions to the inverse problem. The QSPR models, along with structural requirements, provide constraints for the design. The objective function used minimizes the difference between the target property

values and the predicted property values of the molecule designed. Eq. (11) provides the optimization problem formulation (Eslick et al., 2009; Siddhaye et al., 2000).

$$\begin{aligned} \min \quad & s = \sum_M \frac{1}{P_m^{scale}} |P_m - P_m^{target}| \\ & P_m = f_m(y) \\ & y = g(a_{ijk}, w_i) \\ & h_c(a_{ijk}, w_i) \geq 0 \end{aligned} \quad (11)$$

P_m is a predicted value of property m , which is part of the set of properties M . The target value for the property m is P_m^{target} and P_m^{scale} is the scaling factor. The QSPR for property m is f_m and the descriptor values are contained in vector y . The structural descriptors (connectivity indices) are defined by a_{ijk} and w_i in vector g . The identity of group i is given by w_i and a_{ijk} is a binary variable indicating whether a bond of type k connects groups i and j . The constraints ensuring a feasible structure are given by h_c .

When all the equations given in Eq. (11) are linear, the optimization problem is formulated as a mixed integer linear program (MILP). A mixed integer non-linear program (MINLP) results when any of the equations used in the problem formulation are nonlinear. In this work, a MINLP form results from the problem formulation due to the use of second order connectivity indices and average connectivity indices in the QSPRs.

The optimization problem can be solved using either deterministic or stochastic methods. Deterministic approaches which solve for a global optimum have been used in CMD successfully (Maranas, 1996; Sahinidis, Tawarmalani, & Yu, 2003). For an MINLP, global optimality is not guaranteed. Due to inherent error in the QSPRs, based on experimental error and statistical error in correlation and quantified by the prediction intervals, determination of the global optimum is not necessary. Many local optima could yield molecules exhibiting property values that are not statistically different. In this work, a stochastic approach is utilized to solve the CMD framework.

Tabu Search was used as the stochastic approach, which generates many locally optimal solutions (Glover, 1990). During the implementation of the algorithm, previous solutions are stored in a Tabu list. If a new solution is too similar to a previous solution, it is considered Tabu and is discarded. By comparing solutions to the Tabu list, solutions will not be centered on a local optimum and will instead be chosen from new areas in the solution space (Eslick et al., 2009). A list of the solutions that provide the best match to the target property values is stored. Solutions are generated by changing the molecular groups based upon rules for addition or removal. By properly defining rules for the addition or removal of groups, structural constraints and requirements for valency do not need to be explicitly stated. A previously developed CMD framework was used to implement Tabu search and design the optimal carbohydrate excipient (Eslick et al., 2009; McLeese, Eslick, Hoffmann, Scurto, & Camarda, 2010).

3. Results and discussion

3.1. Summary of QSPR model development

The R statistical program (R-Development-Core-Team, 2010) was used for linear regression of the desired properties to the connectivity indices of the excipients in the training set. The training set data was obtained from literature (Roos, 1993). The carbohydrate excipients used in the training set were monosaccharides (ribose, xylose, fructose, fucose, glucose, and sorbose), disaccharides (lactose, lactulose, melibiose, sucrose, and trehalose), oligosaccharides (raffinose), and sugar alcohols (maltitol, sorbitol, and xylitol). The leaps

package in R (Lumley, 2004) was used to conduct an exhaustive search to determine which combination of descriptors provided the lowest C_p value for each number of possible descriptors that could be used for the QSPR (ranging from one to twelve). As an example, Fig. 3 shows the results from development of the model for glass transition temperature of maximally freeze-concentrated solute (T'_g). The lowest C_p value was observed for the model using six connectivity indices and the model was selected to be used as the QSPR.

A summary of the QSPRs developed in this work is provided in Table 2. The individual QSPRs are further detailed in the following sections.

3.2. QSPR for glass transition temperature of anhydrous solute

The glass transition temperature of the anhydrous solute provides an indication of the stability of the carbohydrate excipient during storage conditions. In addition to being used as a criterion for storage stability, T_g is used in the calculation of the excipient concentration in a maximally freeze-concentrated matrix.

The correlation developed for the T_g of carbohydrate excipients is:

$$T_g (^{\circ}\text{C}) = -243.13^0\chi + 1314.26^1\chi + 184.37^2\chi - 1444.52^1\chi^v - 270.82^2\chi^v + 6028.51^0\xi - 12674.4^1\xi - 6833.65^2\xi + 29314.02^2\xi^v - 1539.27 \quad (12)$$

The correlation provides a good fit for the experimental measurements, providing a coefficient of determination (R^2) value of 0.998. The predicted T_g values are compared to the experimental values using a parity plot in Fig. 4.

3.3. QSPR for glass transition temperature of maximally concentrated solute

In addition to being used as a criterion for ensuring formation of a maximally freeze-concentrated glass matrix, the glass transition temperature of the maximally concentrated solute (T'_g) is used in the calculation of the excipient concentration in the glass matrix.

The correlation developed for the T'_g of carbohydrate excipients is:

$$T'_g (^{\circ}\text{C}) = 40.79^0\chi + 16.63^2\chi - 70.27^0\chi^v - 373.79^1\xi - 815.24^2\xi + 2028.90^2\xi^v + 7.12 \quad (13)$$

The correlation provides a good fit for the experimental measurements, providing an R^2 value of 0.994. The predicted T'_g values are compared to the experimental values using a parity plot in Fig. 5.

3.4. QSPR for the melting point of ice

The melting point of ice represents the onset of ice formation during the freezing part of the lyophilization process. The concentrated solution must be annealed between T'_g and T'_m to ensure that the resulting freeze-concentrated glass matrix is maximally concentrated.

The correlation developed for the T'_m of carbohydrate excipients is given below:

$$T'_m (^{\circ}\text{C}) = 65.63^0\chi + 79.76^1\chi + 14.73^2\chi - 175.32^0\chi^v - 60.7.09^1\xi + 1591.25^0\xi^v + 442.48^1\xi^v - 759.39 \quad (14)$$

The correlation provides a good fit for the experimental measurements, providing a R^2 value of 0.997. The predicted T'_m values are compared to the experimental values using a parity plot in Fig. 6.

3.5. QSPR for Gordon–Taylor constant

The Gordon–Taylor constant is used in the calculation of the solute concentration from the Gordon–Taylor equation. The Gordon–Taylor constant for a compound is usually derived from glass transition measurements (Roos, 1993). It should be noted that through the descriptor selection method, the same connectivity indices were found to provide the best model as those used in the model for glass transition temperature of the anhydrous solute.

The developed correlation for the Gordon–Taylor constant of carbohydrate excipients is given below:

$$k = -7.13^0\chi + 38.50^1\chi + 5.45^2\chi - 42.26^1\chi^v - 8.03^2\chi^v + 177.64^0\xi - 371.72^1\xi - 200.43^2\xi + 858.52^2\xi^v - 42.00 \quad (15)$$

The correlation provides a good fit for the experimental measurements, providing a R^2 value of 0.998. The predicted k values are compared to the experimental values using a parity plot in Fig. 7.

3.6. Design of a novel carbohydrate with target properties

According to vitrification theory, a protein stabilizing excipient should be able to form a glass during lyophilization and subsequent storage of the therapeutic product. The glass transition temperature of the maximally freeze-concentrated solute and the melting point of ice for a carbohydrate excipient must be high enough to be feasibly reached during the lyophilization process. The protein drug product must first be annealed at a temperature between T'_g and T_m and then reduced below T'_g to yield a maximally freeze-concentrated matrix (Roos, 1997). The melting temperature of ice is higher than the glass transition temperature of the maximally concentrated solute. The glass transition temperature of the anhydrous solute must be sufficiently high such that the carbohydrate remains a glass during storage of the protein drug product. A common heuristic is that the storage temperature of an amorphous drug formulation should be 50 °C below the anhydrous glass transition temperature (Costantino & Pikal, 2004). For a formulation to be stored at room temperature, a T_g of at least 80 °C is desired. This criterion is very important, as more than 70% of commercial lyophilized products cannot be stored at room temperature and must be refrigerated to maintain stability, complicating the use of the drugs (Costantino & Pikal, 2004). All three phase transition temperatures are used along with the Gordon–Taylor constant to determine the excipient concentration (weight fraction) of a maximally freeze-concentrated matrix (Eq. (10)), without protein present. A higher excipient concentration corresponds to lower residual water in the matrix, which lowers the mobility of the protein and thus reduces the potential for aggregation. Table 3 lists the design targets for the five excipient properties of interest.

No target is placed on the actual value of the Gordon–Taylor constant. The Gordon–Taylor constant is calculated only for use in subsequent calculation of the excipient concentration in a maximally freeze-concentrated matrix. The described molecular design framework was employed using a program previously used to design various molecules, including dental polymers (Eslick et al., 2009) and ionic liquids (McLeese et al., 2010). The results for the three best excipient candidates are given in Table 4, determined by the solutions with the lowest objective function scores. Seven total candidates were generated. Error was calculated using prediction integrals at a 95% level for the property values predicted using

the QSPR models. The structures of the proposed carbohydrate excipients are given in Figs. 8–10.

The proposed excipients are similar to disaccharide and oligosaccharide molecular topologies. Candidates 1, 2, and 3 are isomers. The property values of the proposed excipients show that the computationally designed excipient molecules should stabilize protein formulations. Protein mobility should be limited due to the high solute concentration of the maximally freeze-concentrated matrix. The values for the glass transition temperature of the maximally freeze-concentrated solute and the melting point of ice are high enough that they can be reached during lyophilization. Additionally, the gap in the two temperatures is large enough to allow for annealing between the two temperatures, ensuring that the maximum solute concentration is reached in the freeze-concentrated matrix. The high glass transition temperatures of the anhydrous solute are high enough that drying and long term storage conditions will not change the desired glass structure of the protein formulation.

The prediction intervals show that for all three candidates, all four properties predicted by the QSPR models have overlapping prediction intervals. Due to the overlapping prediction intervals, the predicted property values of all the candidates are not statistically different; all three candidates are valid solutions for the optimization problem. Use of Tabu search was able to provide several optimal excipient candidates with statistically similar property values, where a deterministic method would have only provided one candidate. The prediction intervals for the glass transition temperature of the anhydrous solute were large for all three candidates. The large magnitude of the prediction interval is likely due to the target value being close to the upper limit of the property data used to build the correlation.

One should note that since not all possible properties of importance for an excipient have been included, these structures should be considered candidates for protein drug excipients, and not finalized designs to be immediately utilized. Since all of the structures designed in this work are similar to disaccharides, it is likely that they can be synthesized. However, further studies would be required to ensure that the excipients themselves exhibit sufficient stability to be used in drug formulations.

4. Conclusions

Computational molecular design has been applied to the design of carbohydrate excipients for lyophilized protein formulations. Several novel excipients were identified, without the need for costly and time consuming trial-and-error experiments. The vitrification approach to excipients aiding in protein stability was considered. Property models were generated to relate the glass transition temperature of the anhydrous solute, glass transition temperature of the maximally concentrated solute, melting point of ice and Gordon–Taylor constant to the molecular topology for carbohydrates. Tabu search was able to provide several optimal excipient candidates with statistically similar property values, as determined by the novel application of prediction intervals to computational molecular design. From the design targets, excipients have been optimally designed to have high glass transition temperatures of the anhydrous solute, high glass transition temperatures of the maximally concentrated solute, high melting points of ice, and low water concentrations in the freeze-concentrated glass matrix. From the properties, the proposed excipients should limit protein mobility during lyophilization and thus minimize aggregation. While the properties of the excipients can be predicted with the QSPRs developed, the actual properties of the lyophilized product may be slightly different due to the presence of the protein, and will depend on protein structure and concentration. Future work will investigate other desirable properties to design for, as vitrification is a necessary, yet not sufficient, condition for stable lyophilized protein formulations (Crowe et al., 1998). The following classes of molecules will also be

investigated in addition to carbohydrates: polymers, amino acids, and amino acid derivatives. Such compounds are currently used as stabilizers in lyophilized formulations (Costantino & Pikal, 2004). The work here has provided a starting point for computational molecular design of aggregation-reducing excipients for lyophilized protein formulations.

Acknowledgments

The authors would like to thank the NIH for support from RO1 GM085293-01. The authors would also like to thank T. Steele Reynolds and Anthony I. Pokphanh for their work on the project. Brock C. Roughton would like to thank J.R. Hacker for assistance with the implementation of the computational molecular design framework employed in this work, and J.C. Eslick for initial development of the program used for the computational molecular design.

References

- Bicerano, J. Prediction of polymer properties. 3rd ed.. Marcel Dekker; New York: 2002.
- Cherkasov A, Hilpert K, Jenssen H, Fjell CD, Waldbrook M, Mullaly SC, et al. Use of artificial intelligence in the design of small peptide antibiotics effective against a broad spectrum of highly antibiotic-resistant superbugs. *ACS Chemical Biology*. 2008; 4:65–74. [PubMed: 19055425]
- Cleland, J.L.; Langer, R.S.; American Chemical Society. Division of Biochemical Technology. Formulation and delivery of proteins and peptides. American Chemical Society; Washington, DC: 1994.
- Congreve M, Chessari G, Tisi D, Woodhead AJ. Recent developments in fragment-based drug discovery. *Journal of Medicinal Chemistry*. 2008; 51:3661–3680. [PubMed: 18457385]
- Consonni, V.; Todeschini, R. Handbook of molecular descriptors. Wiley-VCH; Weinheim/New York: 2000.
- Costantino, H.R.; Pikal, M.J. Lyophilization of biopharmaceuticals. AAPS Press; Arlington, VA: 2004.
- Crowe JH, Carpenter JF, Crowe LM. The role of vitrification in anhydrobiosis. *Annual Review of Physiology*. 1998; 60:73–103.
- Crowe LM, Reid DS, Crowe JH. Is trehalose special for preserving dry biomaterials? *Biophysical journal*. 1996; 71:2087–2093. [PubMed: 8889183]
- Eslick JC, Ye Q, Park J, Topp EM, Spencer P, Camarda KV. A computational molecular design framework for crosslinked polymer networks. *Computers & Chemical Engineering*. 2009; 33:954–963. [PubMed: 23904665]
- Fung J, Darabie AA, McLaurin J. Contribution of simple saccharides to the stabilization of amyloid structure. *Biochemical and Biophysical Research Communications*. 2005; 328:1067–1072. [PubMed: 15707986]
- Gani R, Harper PM, Hostrup M. Automatic creation of missing groups through connectivity index for pure-component property prediction. *Industrial & Engineering Chemistry Research*. 2005; 44:7262–7269.
- Gani R, Tzouvaras N, Rasmussen P, Fredenslund A. Prediction of gas solubility and vapor–liquid equilibria by group contribution. *Fluid Phase Equilibria*. 1989; 47:133–152.
- Glover F. Artificial intelligence, heuristic frameworks and tabu search. *Managerial and Decision Economics*. 1990; 11:365–375.
- Haddad J, Kotra LP, Llano-Sotelo B, Kim C, Azucena EF, Liu M, et al. Design of novel antibiotics that bind to the ribosomal acyltransfer site. *Journal of the American Chemical Society*. 2002; 124:3229–3237. [PubMed: 11916405]
- Her L-M, Nail SL. Measurement of glass transition temperatures of freeze-concentrated solutes by differential scanning calorimetry. *Pharmaceutical Research*. 1994; 11:54–59. [PubMed: 8140056]
- Joback KG, Reid RC. Estimation of pure-component properties from group-contribution. *Chemical Engineering Communications*. 1987; 57:233–243.
- Kier, L.B.; Hall, L.H. Molecular connectivity in structure–activity analysis. Research Studies Press; Wiley; New York: 1986.

- Lei Z, Chen B, Li C. COSMO-RS modeling on the extraction of stimulant drugs from urine sample by the double actions of supercritical carbon dioxide and ionic liquid. *Chemical Engineering Science*. 2007; 62:3940–3950.
- Li Y, Williams T, Topp E. Effects of excipients on protein conformation in lyophilized solids by hydrogen/deuterium exchange mass spectrometry. *Pharmaceutical Research*. 2008; 25:259–267. [PubMed: 17597380]
- Lin B, Chavali S, Camarda K, Miller DC. Computer-aided molecular design using Tabu search. *Computers & Chemical Engineering*. 2005; 29:337–347.
- Lumley, T. The leaps package. 2004. www.cran.r-project.org/doc/packages/leaps.pdf
- Manning M, Chou D, Murphy B, Payne R, Katayama D. Stability of protein pharmaceuticals: An update. *Pharmaceutical Research*. 2010; 27:544–575. [PubMed: 20143256]
- Maranas CD. Optimal computer-aided molecular design: A polymer design case study. *Industrial & Engineering Chemistry Research*. 1996; 35:3403–3414.
- McLeese SE, Eslick JC, Hoffmann NJ, Scurto AM, Camarda KV. Design of ionic liquids via computational molecular design. *Computers & Chemical Engineering*. 2010; 34:1476–1480.
- R-Development-Core-Team. R: A language and environment for statistical computing. 2010. <http://www.R-project.org>
- Randic M. Characterization of molecular branching. *Journal of the American Chemical Society*. 1975; 97:6609–6615.
- Roos Y. Melting and glass transitions of low molecular weight carbohydrates. *Carbohydrate Research*. 1993; 238:39–48.
- Roos YH. Frozen state transitions in relation to freeze drying. *Journal of Thermal Analysis and Calorimetry*. 1997; 48:535–544.
- Rosenberg AS. Effects of protein aggregates: An immunologic perspective. *AAPS Journal*. 2006; 8:E501–E507. [PubMed: 17025268]
- Sahinidis NV, Tawarmalani M, Yu M. Design of alternative refrigerants via global optimization. *AIChE Journal*. 2003; 49:1761–1775.
- Satyanarayana KC, Abildskov J, Gani R. Computer-aided polymer design using group contribution plus property models. *Computers & Chemical Engineering*. 2009; 33:1004–1013.
- Siddhaye S, Camarda KV, Topp E, Southard M. Design of novel pharmaceutical products via combinatorial optimization. *Computers & Chemical Engineering*. 2000; 24:701–704.
- Sinha S, Li Y, Williams TD, Topp EM. Protein conformation in amorphous solids by FTIR and by hydrogen/deuterium exchange with mass spectrometry. *Biophysical Journal*. 2008; 95:5951–5961. [PubMed: 18835903]
- Slade L, Levine H. Water and the glass transition—Dependence of the glass transition on composition and chemical structure: Special implications for flour functionality in cookie baking. *Journal of Food Engineering*. 1995; 24:431–509.
- Solvason CC, Chemmangattuvalappil NG, Eden MR. Multi-scale chemical product design using the reverse problem formulation. *Computer Aided Chemical Engineering*. 2010; 28:1285–1290.
- Venkatasubramanian V, Chan K, Caruthers JM. Computer-aided molecular design using genetic algorithms. *Computers & Chemical Engineering*. 1994; 18:833–844.
- Wang W. Protein aggregation and its inhibition in biopharmaceutics. *International Journal of Pharmaceutics*. 2005; 289:1–30. [PubMed: 15652195]
- Wang W, Nema S, Teagarden D. Protein aggregation—Pathways and influencing factors. *International Journal of Pharmaceutics*. 2010; 390:89–99. [PubMed: 20188160]
- Wasserman, L. All of statistics: A concise course in statistical inference. Springer; New York: 2004.
- Zoete V, Grosdidier A, Michielin O. Docking, virtual high throughput screening and in silico fragment-based drug design. *Journal of Cellular and Molecular Medicine*. 2009; 13:238–248. [PubMed: 19183238]

Fig. 1.
The phase transitions that occur during the lyophilization process.

Fig. 2.
(a) The carbohydrate trehalose. (b) Hydrogen-suppressed graph representation of trehalose.

Fig. 3.

Values for Mallow's C_p statistic found using leaps for the T'_g models. The lowest value is observed when six connectivity indices are used, indicating the size of the model that should be used. The line is included to guide the eye.

Fig. 4. Comparison of measured experimental values to QSPR predicted values for the glass transition temperature of the anhydrous solute.

Fig. 5.
Comparison of measured experimental values to QSPR predicted values for the glass transition temperature of the maximally concentrated solute.

Fig. 6.
Comparison of measured experimental values to QSPR predicted values for the melting point of ice.

Fig. 7.
Comparison of measured experimental values to QSPR predicted values for the Gordon–Taylor constant.

Fig. 8.
Optimal carbohydrate excipient candidate 1 proposed by computational molecular design.
The objective function score is 0.00800.

Fig. 9.
Optimal carbohydrate excipient candidate 2 proposed by computational molecular design.
The objective function score is 0.01367.

Fig. 10.
Optimal carbohydrate excipient candidate 3 proposed by computational molecular design.
The objective function score is 0.01373.

Table 1

Values for connectivity indices used in this work for the carbohydrate trehalose.

Connectivity index	Value
${}^0\chi$	17.309
${}^1\chi$	10.811
${}^2\chi$	9.719
${}^0\chi^v$	11.99
${}^1\chi^v$	7.079
${}^2\chi^v$	5.531
${}^0\xi$	0.753
${}^1\xi$	0.450
${}^2\xi$	0.278
${}^0\xi^v$	0.521
${}^1\xi^v$	0.295
${}^2\xi^v$	0.158

Table 2

Summary of QSPRs developed.

QSPR	Number of descriptors used	R^2
T_g	9	0.998
T'_g	6	0.994
T'_m	7	0.997
k	9	0.998

Table 3

Property targets for the design of optimal carbohydrate excipient.

Property	Targets
T_g	100 °C
T'_g	−30 °C
T''_g	−25 °C
k	Not specified (used in calculation of C'_g)
C'_g	0.85

Table 4

Property values of candidate carbohydrate excipients designed using tabu search.

Property	Candidate 1	Candidate 2	Candidate 3
T_g	100.9 ± 12.7 °C	99.8 ± 15.0 °C	90.3 ± 20.6 °C
T'_g	-32.6 ± 6.5 °C	-33.1 ± 6.7 °C	-31.7 ± 5.0 °C
T'_m	-24.8 ± 3.2 °C	-23.7 ± 3.5 °C	-24.1 ± 4.1 °C
k	6.76 ± 0.37	6.73 ± 0.44	6.46 ± 0.61
C'_g	0.838	0.838	0.845
MW	372.3 g/mol	373.3g/mol	373.3 g/mol
Obj function	0.00800	0.01367	0.01373