Pharmaceutical Research, Vol. 26, No. 2, February 2009 (© 2008)
DOI: 10.1007/s11095-008-9737-6

Research Paper

High-Throughput Self-Interaction Chromatography: Applications in Protein Formulation Prediction

David H. Johnson,[1,4] Arun Parupudi,[2] W. William Wilson,[2] and Lawrence J. DeLucas[3]

Received July 8, 2008; accepted September 24, 2008; published online October 16, 2008

Purpose. Demonstrate the ability of an artificial neural network (ANN), trained on a formulation screen of measured second virial coefficients to predict protein self-interactions for untested formulation conditions.

Materials and Methods. Protein self-interactions, quantified by the second virial coefficient, B_{22}, were measured by self-interaction chromatography (SIC). The B_{22} values of lysozyme were measured for an incomplete factorial distribution of 81 formulation conditions of the screen components. The influence of screen parameters (pH, salt and additives) on B_{22} value was modeled by training an ANN using B_{22} value measurements. After training, the ANN was asked to predict the B_{22} value for the complete factorial of parameters screened (12,636 conditions). Twenty of these predicted values (distributed throughout the range of predictions) were experimentally measured for comparison.

Results. The ANN was able to predict lysozyme B_{22} values with a significance of $p<0.0001$ and RMSE of 2.6×10^{-4} mol ml/g^2.

Conclusions. The results indicate that an ANN trained on measured B_{22} values for a small set of formulation conditions can accurately predict B_{22} values for untested formulation conditions. As a measure of protein–protein interactions correlated with solubility, B_{22} value predictions based on a small screen may enable rapid determination of high solubility formulations.

KEY WORDS: artificial neural network; formulation development; physical protein stability; self-interaction chromatography; systematic screening.

INTRODUCTION

A protein's interaction with itself and with other proteins affects important characteristics such as its solubility (1), aggregation (2) and ability to crystallize (3). Measurement of second virial coefficients, B_{22} (4), provides one method to quantify protein interactions at the molecular level. B_{22} is a measure of the entirety of two body protein self-interactions that includes contributions from excluded volume, electrostatic factors (attractive and repulsive) and hydrophobic interactions. In terms of McMillan–Meyer solution theory (5), B_{22} is related to a potential of mean force which describes all of the interaction forces between protein molecules in a dilute solution. Positive B_{22} values correspond to net repulsive forces of the protein and are correlated with increased protein solubility in solution (1,6) whereas values in the negative range correspond to the net attractive forces required for protein insolubility (i.e. precipitation or crystallization conditions (3)). Identified as one indicator of the physical stability of proteins in solution (7), the second virial coefficient depends on a variety of solution formulation parameters including temperature, pH and the type and concentration of salts and excipients (additives). As these additives interact with a protein's surface, they naturally change that surface with respect to shape change and other interaction parameters.

The second virial coefficient can provide functional insight at various stages of the drug discovery process. The initial evaluation of a protein's function in human pathology is often facilitated by study of the protein's structure by means of x-ray diffraction. George and Wilson have shown (3) that proteins generally crystallize when their B_{22} values are in a "crystallization slot" ranging from approximately −0.2 to −8 ($\times10^{-4}$ mol ml/g^2). This B_{22} range, confirmed by several research groups, represents slightly to moderately attractive forces between proteins, a condition that appears to be important for nucleation and subsequent crystal formation (2,8–10).

The determination of solution conditions yielding diffraction quality crystals, as well as high protein solubility and/or low nonspecific aggregation of proteins expressed in prokaryotic and eukaryotic systems, represent major bottlenecks in high-throughput protein structure (11,12). Although there have been advances in the ability to recover bioactive protein from the inclusion bodies of various expression

[1] Department of Biomedical Engineering, University of Alabama at Birmingham, Birmingham, Alabama, USA.
[2] Department of Chemistry, Mississippi State University, Starkville, Mississippi, USA.
[3] Center for Biophysical Sciences and Engineering, 1530 Third Avenue South, Birmingham, Alabama 35294, USA.
[4] To whom correspondence should be addressed. (e-mail: dhj@uab.edu)

systems (13), these techniques require customization to the protein of interest, a requirement that is not conducive to high throughput methods. The mathematical relationship between the B_{22} value and solubility, derived by Haas *et al.* (1), indicates a marked increase in solubility with increasing B_{22} value. This relationship has been validated experimentally for a variety of proteins (1,6,8). Thus, a second application of the second virial coefficient involves its use as a diagnostic for protein solubility.

Protein solubility and stability are as important in the evaluation of therapeutic proteins as they are in the study of proteins involved in disease pathology. The Food and Drug Administration's (FDA) evaluation of a drug candidate includes two primary criteria: solubility and membrane permeability (17). In a recent overview of pharmaceutical drug screening techniques (18), three methods of solubility screening were identified: UV absorption, nephelometry and flow cytometry. These methods, developed for analysis of small molecules, are used to calculate current or potential solubility of a specific drug formulation and can be performed in a high throughput manner. However, they do not directly quantify the protein self-interactions that influence solubility and aggregation of protein therapeutic molecules.

Measurement of the second virial coefficient, performed using static light scattering (SLS) (3), consumes a significant amount of protein and time (multiple light scattering readings are necessary to calculate one B_{22} value) and it requires careful attention to solution clarity. In contrast, a second method for determining the second virial coefficient known as self-interaction chromatography (SIC) provides advantages to each of the constraints of the SLS method as referenced by Tessier *et al.* (20). SIC initially requires chemical coupling of protein to a solid support followed by careful packing of the support in a small chromatography column. Once prepared, however, the column is stable and can be repeatedly used to measure B_{22} values, making it more applicable to high-throughput techniques. Each measurement consists of flowing a mobile microgram injection of the protein across the immobilized protein particles using an HPLC. The retention time of the mobile protein is directly related to its interaction with immobilized column protein (19), thereby providing a direct measurement of how two proteins (bound and injected) interact with one another. Formulas relating the chromatographic retention time to B_{22} values can be found in the paper of Tessier *et al.* (20). This technique has been successfully used with low throughput screens (16 conditions) to measure the interactive effects of two formulation parameters on B_{22} (21).

In this study we used self-interaction chromatography to rapidly measure the B_{22} value of hen egg-white lysozyme in 81 solution formulations. The screen measures the pair-wise effects of nine different additives on the self-interaction of the lysozyme protein. The well known incomplete factorial experimental design technique, applied to crystallization screening by Carter and Carter (22), is used to ensure wide coverage of the search space with a reduced number of test conditions. The incomplete factorial design is accomplished by mapping the parameters of interest (pH, salts, additives, concentrations) onto an orthogonal array (23,24). Mapping parameters to an orthogonal array allows equal representation of parameter levels throughout the search space while reducing the 12,636 possible parameter combinations down to a reasonable screen size of 81 conditions. The B_{22} values are measured to quantify the degree of lysozyme self-interaction in each of the formulations.

The results of the screen are first analyzed by manually examining the linear and quadratic trends of each formulation parameter on B_{22} value. Parameters with the most statistically significant effect on protein-protein interaction (B_{22} value) of lysozyme are identified within the screen. These parameters with strong influence on protein interactions (such as NaCl) are shown to have an effect on B_{22} value regardless of the presence of other additives in varying formulations. This allows for the rapid identification of additives that could be used to modify protein–protein interactions.

While a manual examination of parameter effects can identify the strong correlations of single parameters, this initial analysis does not examine the effect of parameter interactions. To analyze the effect of additive combinations on protein–protein interaction we modeled the results of the B_{22} value screen using an artificial neural network (ANN). Artificial neural networks have utility when the effect of specific combinations of a large number of variables/parameters, as well as each variable's level (i.e. concentration of various chemicals), must be analyzed to determine the optimal combination to yield a desired outcome. The large number of potential additive combinations and their possible levels defines a search space that precludes manual inspection of the data as a reasonable method for finding the optimum parameters and parameter concentrations. Artificial neural networks are able to utilize an incomplete factorial subset of parameter combinations to determine correlations between discrete variable combinations and their respective levels. Neural network modeling has been used to predict novel crystallization conditions (25) and to confirm theoretical calculations of B_{22} for very small molecules (26).

An artificial neural network is essentially a set of non-linear weighted functions which map input variables (screen parameters) into output variables (B_{22} value) (27). The weights are initialized to random values resulting in a random mapping of the screen parameters onto B_{22} values. The subsequent training process to determine optimal weights is performed by iteratively updating the weights to reduce error between the ANN output and observed values (B_{22} screen). For each iteration, the ANN attempts to produce B_{22} values closer to the observed B_{22} values for the given input parameters. After the training process is complete, the neural network model is used to produce B_{22} value predictions of lysozyme for all possible formulations of one or two additives.

ANN B_{22} value predictions of lysozyme for 20 different formulation conditions were experimentally validated via SIC B_{22} values of lysozyme dissolved in each condition. The chosen conditions included 10 from the most positive and negative B_{22} value predictions combined with 10 spread throughout the range of predicted B_{22} values. The results demonstrate that an artificial neural network, trained using an incomplete factorial additive screen, can accurately predict the second virial coefficient of the protein in previously untested formulations.

Finally, the ANN model is compared with a more traditional generalized linear model (GLM). Identical parameters used as inputs for the artificial neural network are

included for consideration by the GLM. In the stepwise procedure the GLM uses an iterative process to determine which parameters significantly influence the second virial coefficient. For GLM analysis, the gradual process of ANN weight determination during each iteration is replaced by linear regression to calculate optimal linear model coefficients. The significance of each parameter is considered during each iteration, with new parameters added or removed based on a predetermined alpha value threshold. After significant parameters are identified by this stepwise process, linear model coefficients are calculated by linear regression. The GLM, like the trained ANN, can be used to predict the B_{22} values of the protein for untested formulation conditions. Comparison of the GLM predictions to the ANN predictions indicates that, for this application, the ANN produces a more accurate and robust model than the GLM.

MATERIALS AND METHODS

Screen Conditions

Hen egg-white lysozyme was purchased from Calbiochem. The chromatography particles, Toyopearl AF-Formyl-650M (65 μm diameter particle, 0.1 μm diameter pore), were purchased from Tosoh Bioscience. Buffer formulation chemicals include glycerol, glycine, glutamic acid, mannitol, sodium citrate, sodium acetate and acetic acid; all purchased from Fisher Scientific. Additional formulation chemicals PEG4000, MPD and trehalose were purchased from Sigma-Aldrich under the Fluka brand name. Sigma-Aldrich was also the source for chromatography bead capping agent, ethanolamine, as well as the formulation chemicals Na_2SO_4, Na HEPES, HEPES acid and citric acid. The final two formulation chemicals, succinic acid and arginine, were purchased from Acros Organics.

Each of the 81 solution formulations contain buffer, salt and one or two co-solvents listed in Table I. The salt and each co-solvent can appear in low, medium or high concentration which varies depending on the solubility of the individual salt or co-solvent added. If all combinations of two solvents and four salts at three individual levels of concentration are combined with four pH levels represented by the buffers then the full factorial of 12,636 conditions is determined. To reduce the number of conditions in which the B_{22} of lysosyme is measured, the identity of each screen formulation was determined by mapping the parameters onto an orthogonal array design as described by Sloane (28). This mapping produces formulation targets in which each pair of variables are equally represented throughout the screen (thereby producing a balanced screen with respect to the influence of individual parameters).

The water source for formulations was pre-filtered at 18 MΩ by a Millipore MilliQ system with trace sodium azide added to retard bacteria growth. Sodium and acid forms of 0.1 M buffers are mixed at their pKa in the presence of co-solvents (except in the case of the succinic buffer which was adjusted to pH with NaOH). The pH of each solution was confirmed via a Corning 430 pH meter with the final pH adjusted solutions filtered (0.22 μm (Fisher Scientific) syringe filter) and stored at room temperature.

Protein Immobilization

Lysozyme (LYZ) was immobilized to AF-Formyl-650M beads as described by Valente *et al.* (29) with only slight modification. One ml of 1 M K_2HPO_4 at pH 7.0 was added to 350 μl of AF-Formyl-650M beads followed by centrifugation (bench-top, 30 s 7k rpm). The wash was performed an additional two times to remove excess packing buffer. LYZ (5 mg) was dissolved in the phosphate buffer and incubated with the beads. Fifteen mg of sodium cyanoborohydride was added to the bead mixture to activate the binding chemistry and mixed via rotary mixer at room temperature for 90 min. A 5 μl sample of the supernatant containing unbound LYZ was diluted with 45 μl of 0.1 M sodium acetate buffer pH 4.7 and assayed via a bicinchoninic acid (BCA) assay (Thermo Scientific). The beads were centrifuged and washed twice with phosphate buffer plus 5% (*w*/*v*) NaCl and twice with phosphate buffer sans NaCl to remove any remaining LYZ. After binding and washing, unreacted formyl groups were capped by adding 1 ml of 1 M ethanolamine at pH 8.0 and 10 mg sodium cyanoborohydride, followed by additional rotary mixing for 90 min. After this final step of immobilization the beads were washed twice with 1 mL of the sodium acetate buffer.

Self-Interaction Chromatography

Immobilized beads were packed into a micro-column consisting of teflon FEP tubing (i.d. 0.03″, o.d. 1/16″) and were blocked at one end by a 2 μm stainless steel screen (Valco). Two 1.1 cm lengths of packed tubing (~5 μl each) were cut from the packing end, diluted with 45 μl of sodium

Table I. Formulation Parameters

Buffers	Salts	Additives		
Acetate (p*K*a 4.7)	NaCl[a]	Arginine[b]	Sucrose[c]	MPD[d]
Succinate (p*K*a 5.6)	NaCitrate[a]	Glutamic Acid[b]	Mannitol[c]	PEG4000[d]
MES (p*K*a 6.1)	Na_2SO_4[a]	Glycine[b]	Trehalose[c]	Glycerol[f]
HEPES (p*K*a 7.5)				

A list of additives, salts and buffers utilized in the formulation screen
[a] Low: 0.1 M; Medium: 0.3 M; High: 0.5 M
[b] Low: 0.02 M; Medium: 0.04 M; High: 0.06 M
[c] Low: 0.1 M; Medium: 0.2 M; High: 0.3 M
[d] Low: 5%; Medium: 10%; High: 15% (*w*/*v*)
[f] Low: 3%; Medium: 6%; High: 9% (*w*/*v*)

acetate buffer and assayed using the BCA assay (Pierce Biotechnology) to determine protein binding density on the column. The packing end of the column was then cut to 18 cm length, sealed with an additional screen. When not in use the column was stored with 0.1 M sodium acetate buffer pH 4.7 at 4°C. A second column, referred to as the dead column, was packed with beads that have been subjected to only the capping portion of the immobilization procedure. Acetone was used as a non-interacting void-volume marker and was dissolved in water at 3% (*v*/*v*) for injections. The protein injection solution consists of 5 mg of lysozyme dissolved in 1 ml of each of the four separate 0.1 M buffers (Table I).

All chromatograms were generated using a high performance liquid chromatography (Shimadzu) system consisting of two pumps, an auto-sampler for sample injection, column oven, 280 nm UV detector and software for automatic retention-time calculation. Each screen formulation was run through the column at 60 μl/min and the auto-sampler was used to inject 1 μl of the 5 mg/ml LYZ solution in buffer identical to the formulation buffer applied to the column. Column temperature was maintained at 23°C. Injections were performed in triplicate over the same column and B_{22} values measured for the entire 81-condition screen on two columns with the final B_{22} values averaged over both column. Solutions with outlying (1.5*IQR) variance between two columns (N=9) were measured on a third column. If the B_{22} of two columns were within the average standard deviation between two columns (1.7 B_{22} units) the disagreeing measurement was excluded. Sample chromatograms shown in Fig. 1 demonstrate the influence of NaCl on retention time measured at peak elution. In the primary equation used to calculate B_{22} values, Eq. 1, NA is Avagadro's number and MW is the molecular weight of the protein.

$$B = \frac{N_A}{MW^2}\left(B_{HS} - \frac{k'}{\phi\rho}\right) \tag{1}$$

The phase ratio, ϕ, is the ratio of the available surface area per unit of null volume and has been calculated for a variety of different chromatography particles (30). The density of protein immobilized on the column is ρ. In this study binding density varied from 17.5 to 22.4 mg/ml (measured by Pierce BCA assay). The variation in protein binding determines the magnitude with which variations in protein retention time affect B_{22} value. The variable k' is the chromatographic retention factor calculated from the protein retention time (t_r) and acetone retention time (t_0) given by the equation:

$$k' = \frac{t_r - t_0}{t_0} \tag{2}$$

In this equation, Eq. 2, the acetone retention time (t_0) acts as a non-interacting marker to establish the relationship between non-interacting molecules with bound protein compared to interacting molecules with bound protein.

The chromatograms shown in Fig. 1 hold additional importance as the method by which column integrity is verified throughout the screening process. The B_{22} value of lysozyme in NaCl concentrations of 5% and 0% (*w*/*v*) was measured after every eight formulation conditions, consisting of three injections each, or every 24 protein injections. The column was expected to be fairly stable because protein was covalently bound to chromatography media and unbound active groups were rendered relatively inert by the capping process. Regular validation of the column ensured that the addition or loss of protein from the column did not significantly alter B_{22} measurements throughout the screening process. The standard deviation of lysozyme B_{22} value for NaCl conditions was only 1.1 B_{22} units throughout the lifetime of the column (81 screen formulation conditions). This gave assurance that the chromatography column does not experience a significant change in activity due to the addition or subtraction of protein to the column. To ensure the dead volume of the column was not significantly altered from packing of column material, acetone retention time was also measured after every eight formulation conditions. At a fixed protein retention time the standard deviation of B_{22} measurements due to variation in acetone retention time (including effects from column packing) was 0.8 B_{22} units.

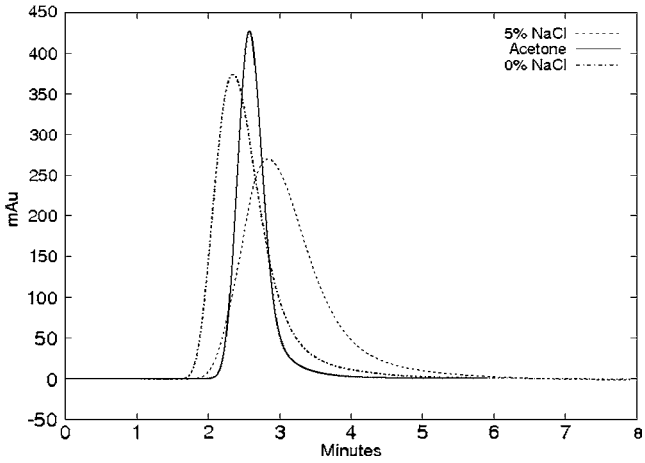

Fig. 1. Retention times for lysozyme in 5% NaCl and 0% NaCl in 0.1 M sodium acetate buffer demonstrates the affect of NaCl on lysozyme self-interaction. The retention time for 3% acetone in the same buffer with 5% NaCl provides a reference point for conversion of retention times to B_{22} values.

Static Light Scattering

The traditional static light scattering (SLS) experiment requires measurement of the scattered light intensity from a protein solution in excess of background as a function of protein concentration. The traditional SLS experiment was modified in two important ways in order to minimize both time and protein required for a single B_{22} measurement (31). The first modification is the incorporation of a low volume (~1 μL) scattering cell. The second modification is a configuration allowing the simultaneous measurement of scattering intensity and protein concentration. This is accomplished by using a bifurcated fiber to deliver both the incident laser beam for scattering and the incident UV beam for absorption (protein concentration) measurements. The advantage of this configuration is that the simultaneous measurement of light scattering intensity and protein concentration allows the determination of the second virial coefficient from a single injection of protein sample into a flow system. Typically, 5–10 μl of protein solution at 1–2 mg/ml protein concentration were required for a single B_{22} measurement.

The intensity and concentration data were treated according to the SLS working equation (32):

$$\frac{Kc}{R_{90}} = \frac{1}{M} + 2B_{22}c \tag{3}$$

where K is an optical constant (cm^2 mol g^{-2}) given by $K=4\pi^2 (dn/dc)^2 n_o^2/(N_A \lambda^4)$, c is the protein concentration (g cm^{-3}), R_{90} is the Rayleigh factor (cm^{-1}) at angle 90°, M is the molecular weight of the protein (g mol^{-1}), B_{22} is the second virial coefficient (mol ml g^{-2}), dn/dc is the refractive index increment (cm^3 g^{-1}), n_0 is the solvent refractive index, N_A is Avogadro's number (mol^{-1}), and λ is the wavelength (cm) of the incident light in a vacuum. According to Eq. 3, a plot of Kc/R_{90} vs c (often called a single angle Zimm plot) linearizes the SLS data and B_{22} is determined from the limiting slope.

Artificial Neural Network (ANN)

Artificial neural network modeling was performed using the Java Object Oriented Neural Engine (JOONE) (33). Fig. 2a shows the overall network topology of the neural network used in this study including inputs, node configuration and B_{22} value output. Each node represents a nonlinear transformation of inputs and is grouped into one of two layers according to distance from the input parameters. Regardless of position in the topology, the output of each node is calculated by two steps shown in Fig. 2b. First, a weighted sum of inputs to the node is calculated, z. The hyperbolic tangent is taken of this weighted sum to calculate node output. Each node in layer 1 takes as input all formulation parameters while each node in layer 2 takes all outputs from layer 1 as input. The final B_{22} value output is calculated as a simple weighted sum of layer 2 outputs without a nonlinear transformation. This permits the range of output values to match the range of screened B_{22} values rather than the (−1,1) range of the hyperbolic tangent function. Through calculation of each layer's outputs in sequence this architecture is able to estimate the protein B_{22} value for a given set of condition formulation parameters. The weights associated with each input node are the variables subject to training, thereby creating a network function that most accurately represents protein B_{22} values over all given formulation parameters.

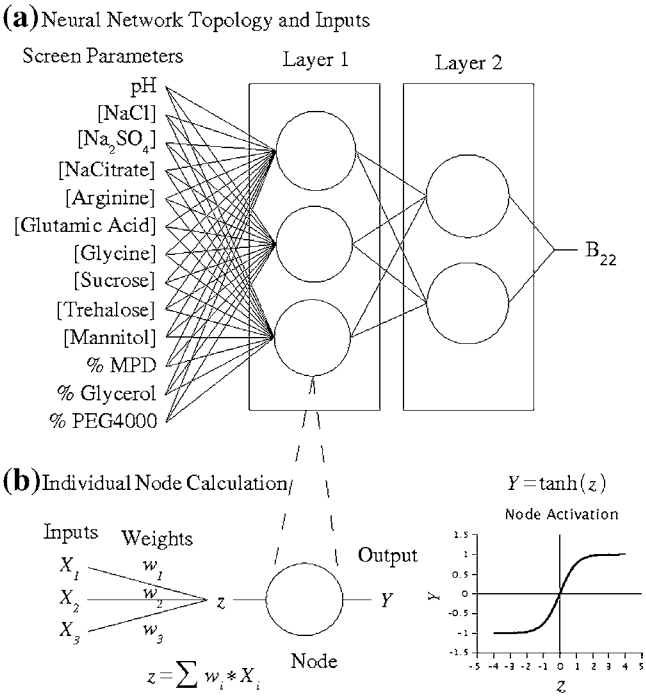

Fig. 2. The artificial neural network topology (**a**) uses parameters of a single formulation as input to each node in Layer 1. Each node's output (**b**) is calculated by an activation function (tanh) whose input is a weighted sum of the node input. The output of nodes in layer one are forwarded as the input to Layer 2. The output of nodes in Layer 2 are weighted and summed to produce a B_{22} value prediction based on the input formulations.

This architecture (input vector, layers and output) is generally referred to as a feed-forward multilayer perceptron and is capable of modeling a continuous function to arbitrary accuracy given a sufficient number of nodes (27). Arbitrary accuracy is apparent if one considers a network topology containing one node for each formulation condition (N=81). After training the weight parameters, the response of each node could represent the measured B_{22} value for each specific formulation condition. Such an exact fit to the screen would result in over-fitting to the error inherent to the screen and would not provide a good, generalized response to formulation conditions outside those on which it was trained.

When the training algorithm is responsible for adjusting neural network weights to over-fit output to a specific training set it is referred to as over-training. To address the problem of over-training of the neural network, we split the set of screen conditions into a training set (90%) and a validation set (10%) and used a technique called early-termination to determine when to stop the training procedure. During training, the weights are iteratively adjusted using the gradient decent algorithm of back-propagation. This algorithm assigns an error contribution and updates each weight based on the root mean square error (RMSE) between the neural network output and the measured protein second virial coefficient for each formulation condition in the training set. RMSE is also calculated between the neural network output and measured B_{22} values in the validation set for each iteration. The validation set RMSE is not used to improve weight values, but instead acts as the basis for deciding when to terminate the training procedure. The network weights are fixed at the minimum validation RMSE over a set number of iterations (1,000). Validation set RMSE is also used as a measure of how well a network topology is able to generalize to untested formulation conditions. All network topologies from 1×1 to 6×6 nodes were evaluated by a validation set RMSE to determine the 3×2 network topology used for this study. Further details about neural network algorithms and methods can be found in Bishop's review (27) of the subject as well as in the JOONE software documentation (33).

Stepwise Generalized Linear Model (GLM)

The stepwise generalized linear model was performed using the JMP (34) statistical software package. The neural network inputs shown in Fig. 2 were also the parameters used for the GLM. The GLM algorithm requires explicit identification of interaction and high order terms for consideration. In addition to the neural network inputs, all pairwise

interactions and square terms of the formulation screen were included for consideration. The stepwise algorithm was configured to include terms with a significance of alpha <0.20 with higher order and interaction terms restricted to only those whose lower order terms were also significant.

Prediction Verification

The second virial coefficient for all combinations of buffer, salt and a maximum of two excipients (12,636 conditions) were predicted by the trained ANN. Five conditions from the most positive B_{22} values and five of the most negative B_{22} values as well as ten equally spaced throughout the range of predicted B_{22} values were selected for experimental confirmation. These 20 verification formulations (not included in the training process) were prepared and B_{22} values of lysozyme in each were experimentally measured using the identical method as the original 81 screen conditions.

The question of whether 81 screen conditions are necessary or if a smaller subset would suffice was addressed by evaluating the ability of the neural network to predict the verification B_{22} values while training on a reduced set of the initial screen. First a condition was randomly removed from the original training set of the neural network. The training process described above was repeated on the reduced training set with the same validation set size remaining constant (deemed a valid indication of the overall population). Then the neural network, trained on a reduced set of the original 81 condition screen, was used to calculate predictions for the verification B_{22} values. Progressively reducing the sample size, followed by training and prediction, allows error as a function of sample size to be evaluated.

To determine how sample size affects neural network B_{22} value predictions, the validation set was kept constant while iteratively removing a random condition from the training set. As there is no consensus in the literature as to how this type of analysis should be performed, a constant validation set was chosen as a good measure of the ability for the network to generalize. We were able to see how available training data influences accuracy by keeping the same validation set through repeated reductions in the training set size. After iterative removal of a random condition from the training set, the ANN is re-trained and then asked to predict the same 20 verification formulations chosen for the initial evaluation of ANN performance. This process was performed three times (with a different sequence of random removals each time) and the error for a given training set size was taken as the average of all three series of removals.

RESULTS AND DISCUSSION

Confirmation by Static Light Scattering

A strong correlation (r=0.97) between static light scattering and self-interaction chromatography was observed (Table II) for ten test conditions as has been previously reported by other laboratories (20,29). The primary differences between the two measurements were found at two of the most positive B_{22} values. B_{22} values in this range were expected to exhibit greater error since large positive B_{22} values have been shown to correspond to very high levels of solubility (1,6). Thus a small difference in B_{22} value represents a larger difference in solubility. Therefore, from a practical perspective, all high positive B_{22} values represent regions of high protein solubility even though individual B_{22} value errors are larger in this region.

Table II. Comparison Between B_{22} Values Measured by Static Light Scattering (SLS) and Self-Interaction Chromatography (SIC)

Condition ID	SLS B_{22} ($\times 10^{-4}$ mol ml/g^2)	SIC B_{22} ($\times 10^{-4}$ mol ml/g^2)
9	9	9.3
24	−1.4	−1.3
27	14	11
35	9	5.3
36	1.4	2.1
39	−1.4	−1.3
46	−0.5	0.0
60	7	7.7
72	3.7	3.8
79	−5	−3.6

SCREEN RESULTS

The B_{22} value results of lysozyme for the 81 formulation conditions demonstrates some characteristics expected of the protein. For example, the mean B_{22} of the screen is positive 1.1×10^{-4} mol ml/g^2 which is reflective of the general soluble nature of lysozyme. Additionally, a majority of the formulation conditions (55%) reside in the crystallization slot identified by George and Wilson (3) which is approximately $[-8,-0.2]\times10^{-4}$ mol ml/g^2. This is indicative of the ease with which lysozyme crystals are formed. It is also of interest to note that the average standard deviation between measurements was 1.7×10^{-4} mol ml/g^2. This suggests that B_{22} measurements produced using self-interaction chromatography are reproducible throughout a large range of different solution conditions.

Interesting trends are also observed when viewing the influence of a single parameter throughout the screen. Fig. 3 shows a graph of B_{22} value *versus* three individual parameter concentrations (NaCl, MPD, Glycine). The variation between plotted B_{22} values at a fixed concentration is due to the fact that other additives change with each condition. Error bars around each point indicate the error from measurement to measurement for each specific formulation. The increasing lysozyme self-interaction (decreasing B_{22}) with increased concentration of sodium chloride (Fig. 3a) is expected and has been demonstrated in other studies by both SIC and SLS (29). At the mid and high concentrations of NaCl, four of the five conditions with positive B_{22} values contain MPD. This combined with the fact that MPD shows a trend (Fig. 3b) of decreasing lysozyme self-interaction (increasing B_{22}) with increasing concentration identifies MPD as a potential solubilizing agent for lysozyme. Quadratic relationships between additive concentration and B_{22} value, such as that apparent in glycine (Fig. 3c) could also indicate an additive which might help stabilize protein self-interaction at a specific level. These single factor cross sections are useful for identifying individual additives which have a strong influence

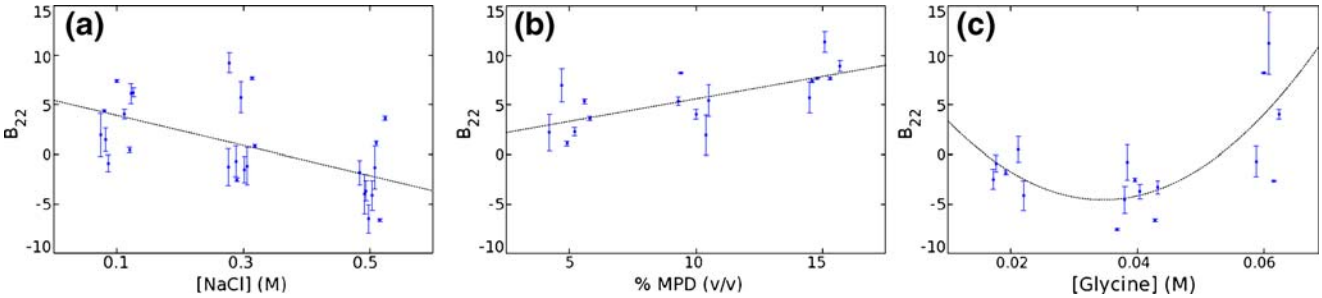

Fig. 3. Response of B_{22} value for lysozyme by **a** NaCl (F test; df=1; p=0.0006), **b** MPD (F test; df=1; p=0.001) and **c** glycine (F test; df=2; p=0.006) throughout all screen conditions containing the additive of interest. *Error bars* represent standard error between SIC measurements between whereas variability between points at a fixed additive concentration is attributed to changes in formulation parameters outside the additive of interest. Scatter along the abscissa is added to prevent overlapping of error bars.

on B_{22} value. However, the prediction capability of single variable linear and quadratic regression models is obviously not sufficient to capture the variability in protein-protein interactions caused by formulations with multiple additives.

Modeling and Prediction Results

The neural network trained on all conditions, except for nine (10%) reserved for validation, produces a model which predicts the original screen with a RMSE of 1.7×10^{-4} mol ml/g^2. This is equal to the observed average standard deviation between measured B_{22} values and reinforces the notion that early termination of training based on the validation set error prevents over-training of the neural network to the screen results. Upon completion of training, the neural network is used to predict B_{22} values for all possible variable combinations with one or two additives (12,636 formulation conditions). From this entire number of predictions, 20 predictions were chosen for verification. These 20 conditions were chosen to represent the entire solubility range, with some from the most positive and negative predicted B_{22} values. These formulation conditions and their predicted second virial coefficients are shown in Table III. The experimental formulations in Table III were prepared and their effect on lysozyme's second virial coefficient measured via SIC. A plot of measured B_{22} values *versus* ANN predicted values are shown in Fig. 4. This figure demonstrates that the neural network is able to predict second virial coefficients with an accuracy of 2.6×10^{-4} mol ml/g^2.

Screen sample size plays a role in how accurately the ANN model is able to predict untested formulation conditions. Fig. 5 shows the relationship between screen sample size and the prediction error of the ANN. As the size of the training set decreases the prediction error of the ANN increases. However, a B_{22} value prediction error of 3 B_{22} units is still attainable with a training set size of 45 screen conditions. The addition of formulation conditions to the training set provides a diminishing improvement to the prediction error. It is interesting to note in Fig. 5 that the error curve does not completely flatten at a screen sample

Table III. ANN Predictions of 20 Formulations Selected for Verification

Buffer	Salt	Excipient 1	Excipient 2	Predicted B_{22}[a]	Measured B_{22}[a]
0.1 M HEPES	0.5 M NaCl	0.04 M Glycine	0.04 M Arginine	−6.0	−8.45
0.1 M Succinate	0.5 M Na_2SO_4	0.06 M Glycine	0.2 M Mannitol	−5.8	−6.21
0.1 M HEPES	0.5 M NaCl	0.04 M Glutamic Acid	0.3 M Mannitol	−5.5	−7.68
0.1 M Acetate	0.5 M Na_2SO_4	0.06 M Glycine	0.1 M Sucrose	−5.1	−1.42
0.1 M MES	0.5 M NaCl	0.3 M Mannitol	0.2 M Trehalose	−5.1	−2.38
0.1 M Succinate	0.5 M NaCl	0.06 M Glycine	0.3 M Sucrose	−3.6	−3.2
0.1 M HEPES	0.5 M Na_2SO_4	0.06 M Arginine	0.1 M Mannitol	−2.6	−2.03
0.1 M HEPES	0.1 M Na_2SO_4	0.02 M Arginine	0.1 M Mannitol	−1.8	−0.8
0.1 M Succinate	0.5 M NaCl	0.3 M Trehalose	5% MPD	−1.1	3.51
0.1 M Succinate	0.5 M NaCitrate	0.02 M Glycine	0.3 M Sucrose	−0.4	−1.34
0.1 M Acetate	0.3 M Na_2SO_4	0.06 M Arginine	0.2 M Trehalose	0.4	1.34
0.1 M Acetate	0.1 M NaCl	0.06 M Arginine	–	1.5	3.08
0.1 M Succinate	0.1 M Na_2SO_4	0.2 M Mannitol	10% PEG4000	2.8	−0.6
0.1 M MES	0.3 M NaCl	15% MES	–	4.4	−0.6
0.1 M HEPES	0.3 M NaCl	10% MPD	9% Glycerol	6.8	6.18
0.1 M MES	0.3 M NaCitrate	9% Glycerol	10% PEG4000	9.1	6.61
0.1 M MES	0.3 NaCl	0.04 M Glycine	15% PEG4000	9.3	9.18
0.1 M MES	0.1 M Na_2SO_4	0.06 M Glycine	15% MPD	9.8	7.42
0.1 M HEPES	0.3 M NaCitrate	0.1 M Trehalose	15% MPD	13	9.54
0.1 M HEPES	0.3 M NaCitrate	10% MPD	10% PEG4000	14	7.85

[a] ($\times10^{-4}$ mol ml/g^2)

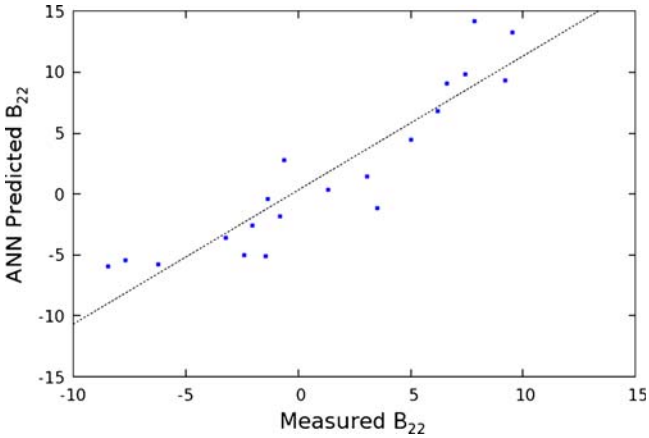

Fig. 4. ANN predicted B_{22} value *vs* measured B_{22} values of the 20 verification formulations (F test; df=1; p<0.0001; RMSE=2.6×10^{-4} mol ml/g^2).

size of 81 formulations. An extrapolation of this suggests a screen of over 100 formulation conditions could permit ANN B_{22} predictions with an error close to 1.7×10^{-4} mol ml/g^2; the variability between B_{22} values measured on separate columns.

The standard generalized linear model provides a comparison of the ANN with a standard linear regression technique used for data analysis/predictions. The terms of the GLM were determined by considering all single terms, interaction terms and square terms and incrementally adding the most significant remaining parameter until there are no more parameters with a significance of alpha <0.20. The GLM parameters and their significance level generated by this method are listed in Table IV. This table demonstrates one benefit of the GLM over ANN. Incremental analysis of each parameter produces a list of factors and their p-value significance. This helps identify specific formulation parameters which could increase solubility. However, when predicting the second virial coefficient of protein in previously unformulated conditions the GLM does not perform as well as the ANN. The

Table IV. Additives with Statistically Significant Influence as Determined by Stepwise GLM

Factor	p value	Magnitude×10^{-4} mol ml/g^2
Glycine	<0.0001	−1.5
MPD	<0.0001	2.5
NaCl	<0.0001	1.4
PEG4000	<0.0001	2.1
Arginine	0.0001	−1.3
Citrate	0.0004	0.5
Mannitol	0.0127	−0.6
Glycerol	0.0562	−1.4
Glycine × PEG4000	<0.0001	1.7
Arginine × Glycine	0.0002	−2.2
NaCl × MPD	0.0099	−0.5
NaCl × Mannitol	0.0154	−0.4
Citrate × Glycine	0.0272	0.3
Citrate × Mannitol	0.1573	−0.2
Glycine × Glycerol	0.1685	−0.5
NaCl^2	<0.0001	−0.7
Glycerol^2	0.0108	1.0

plot in Fig. 6 shows the same 20 measured B_{22} values for ANN validation *versus* the GLM predictions. Although both predictions are statistically significant (F test; df=1; p<0.0001), the GLM is accurate with a RMSE of 3.3×10^{-4} mol ml/g^2 which implies the ANN is approximately 25% more accurate than the GLM. However both techniques are useful for formulation prediction based on a small subset of conditions.

Limitation

A limitation of this screen and formulation prediction technique is in the ability to predict formulation conditions with parameter concentrations well outside the screened range. The inability for statistical models to extrapolate results outside their original input range is well known. This implies that the range of pH and salt/additives concentrations must be chosen based on an estimation of the effective range

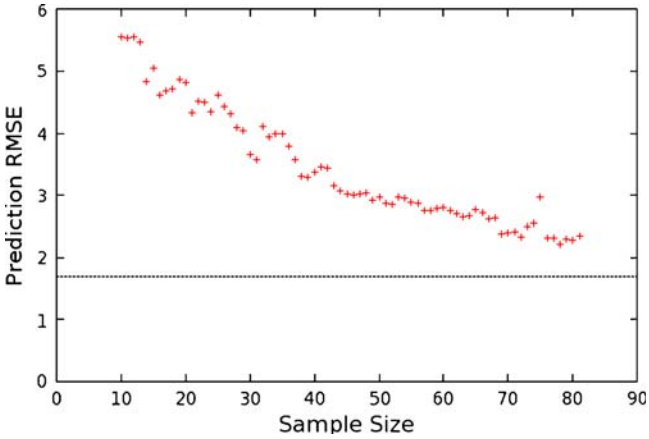

Fig. 5. ANN RMSE *vs* sample size. Incremental reduction in sample size shows an increase in error for artificial neural network predictions of the 20 verification formulations. *Dashed line* indicates the error between B_{22} value measurements by SIC between columns (1.7 mol ml/g^2).

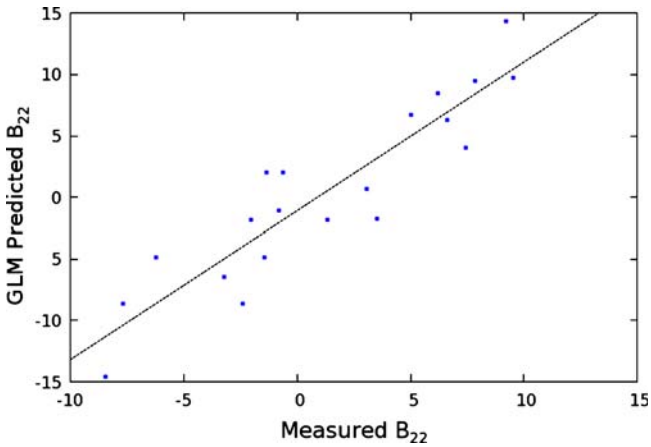

Fig. 6. GLM predicted B_{22} values vs measured B_{22} values of the 20 verification formulations (F test; df=1; p<0.0001; RMSE=3.3×10^{-4} mol ml/g^2).

for each parameter. For example the pH range of interest might be a region in relation to the expected p*I* of the protein. It is important to note that once parameter ranges are determined the screen and resulting statistical models will not be able to predict the B_{22} value of formulations with parameters significantly outside these ranges. However, this does not diminish the fact that the statistical models can accurately predict the B_{22} value of a large number of novel formulation conditions based on parameter combinations not measured in the original screen.

CONCLUSIONS

As hypothesized in previous publications (9,21,25), high throughput screening of second virial coefficients shows promise for evaluating the interactions of proteins in solution. We have demonstrated that an incomplete factorial screen combined with a neural network model can be used to accurately predict second virial coefficients for untested formulations. A B_{22} value screen of only 81 formulation conditions was used to predict the B_{22} values for 12,636 possible formulations with an accuracy of 2.6×10^{-4} mol ml/g^2. These preliminary studies suggest that a high-throughput chromatographic SIC system with increased automation may enhance and accelerate determinations of the optimum conditions that improve the physical solubility/stability of drug formulations. It also suggests this same strategy may be useful to predict formulation adjustments required for optimized protein expression and/or crystallization.

The strong correlation between SIC and SLS measurements of B_{22} value lends further evidence that SIC may be useful as a replacement for the SLS method. The use of SIC in lieu of SLS offers several significant advantages including: (1) SIC requires less protein per experiment, (2) SIC is easily performed with aqeous or membrane proteins whereas SLS is difficult or impossible to use with membrane proteins, (3) SIC is much faster than SLS, (4) SIC is useful with a wider variety of additives due to additive interference with the SLS signal, (5) SIC can be miniaturized and performed in a high-throughput manner thereby enabling studies on a large sample set (i.e. incomplete factorial).

The current time required to run self-interaction chromatography in triplicate is approximately 30 min. While 30 min per experiment by SIC is much faster than previous SLS methods (20), the use of B_{22} values for these applications would benefit significantly by increased throughput via parallelization, robotic automation and integration of analysis techniques into a single platform.

ACKNOWLEDGEMENTS

This work was supported by a grant from the NSF EPSCoR Graduate Research Scholars Program. Thank you to Dr. Lisa Nagy for help with the Shimadzu HPLC hardware.

REFERENCES

1. C. Haas, J. Drenth, and W. Wilson. Relation between the solubility of proteins in aqueous solutions and the second virial coefficient of the solution. *J. Phys. Chem. B.* **103**:2808–2811 (1999). doi:10.1021/jp9840351.
2. T. Ahamed *et al.* Design of self-interaction chromatography as an analytical tool for predicting protein phase behavior. *J. Chromatogr. A.* **1089**:111–124 (2005). doi:10.1016/j.chroma.2005.06.065.
3. A. George, and W. W. Wilson. Predicting protein crystallization from a dilute solution property. *Acta Crystallogr. D.* **50**:361–365 (1994). doi:10.1107/S0907444994001216.
4. B. L. Neal, D. Asthagiri, and A. M. Lenhoff. Molecular origins of osmotic second virial coefficients of proteins. *Biophys. J.* **75**:2469–2477 (1998).
5. J. McMillan, and J. E. Mayer. The statistical thermodynamics of multicomponent systems. *J. Chem. Phys.* **13**:276–305 (1945). doi:10.1063/1.1724036.
6. K. Demoruelle *et al.* Correlation between the osmotic second virial coefficient and solubility for equine serum albumin and ovalbumin. *Acta Crystallogr. D, Biol. Crystallogr.* **58**:1544–1548 (2002). doi:10.1107/S0907444902014385.
7. E. Y. Chi *et al.* Physical stability of proteins in aqueous solution: mechanism and driving forces in nonnative protein aggregation. *Pharm. Res.* **20**:1325–1336 (2003). doi:10.1023/A:1025771421906.
8. B. Guo *et al.* Correlation of second virial coefficients and solubilities useful in protein crystal growth. *J. Cryst. Growth.* **196**:424–433 (1999). doi:10.1016/S0022-0248(98)00842-2.
9. P. M. Tessier *et al.* Self-interaction chromatography: a novel screening method for rational protein crystallization. *Acta Crystallogr. D, Biol. Crystallogr.* **58**:1531–1535 (2002). doi:10.1107/S0907444902012775.
10. J. J. Valente *et al.* Colloidal behavior of proteins: effects of the second virial coefficient on solubility, crystallization and aggregation of proteins in aqueous solution. *Curr. Pharm. Biotechnol.* **6**:427–436 (2005). doi:10.2174/138920105775159313.
11. D. Christendat *et al.* Structural proteomics of an archaeon. *Nat. Struct. Mol. Biol.* **7**:903–909 (2000). doi:10.1038/82823.
12. C. Luan *et al.* High-throughput expression of *c. elegans* proteins. *Genome Res.* **14**:2102–2110 (2004). doi:10.1101/gr.2520504.
13. S. M. Singh, and A.K. Panda. Solubilization and refolding of bacterial inclusion body proteins. *J Biosci Bioeng.* **99**:303–310 (2005). doi:10.1263/jbb.99.303.
14. E. H. Koo, P.T. Lansbury, and J.W. Kelly. Amyloid diseases: Abnormal protein aggregation in neurodegeneration. *Proc. Natl. Acad. Sci. U.S.A.* **96**:9989–9990 (1999). doi:10.1073/pnas.96.18.9989.
15. A. Pande *et al.* Crystal cataracts: Human genetic cataract caused by protein crystallization. *Proc. Natl. Acad. Sci. U.S.A.* **98**:6116–6120 (2001). doi:10.1073/pnas.101124798.
16. J. G. Ho *et al.* The likelihood of aggregation during protein renaturation can be assessed using the second virial coefficient. *Protein Sci.* **12**:708–716 (2003). doi:10.1110/ps.0233703.
17. T. Loftsson, and M.E. Brewster. Physicochemical properties of water and its effect on drug delivery, a commentary. *Int. J. Pharm.* **354**:248–54 (2008). doi:10.1016/j.ijpharm.2007.08.049.
18. E. Kerns, and L. Di. Physicochemical profiling: overview of the screens. *Drug Discov. Today Technol.* **1**:343–348 (2004). doi:10.1016/j.ddtec.2004.08.011.
19. S. Y. Patro, and T. M. Przybycien. Self-interaction chromatography: a tool for the study of protein-protein interactions in bioprocessing environments. *Biotechnol. Bioeng.* **52**:193–203 (1996). doi:10.1002/(SICI)1097-0290(19961020)52:2<193::AID-BIT2>3.0.CO;2-L.
20. P. M. Tessier, A. M. Lenhoff, and S. I. Sandler. Rapid measurement of protein osmotic second virial coefficients by self-interaction chromatography. *Biophys. J.* **82**:1620–31 (2002).
21. J. J. Valente *et al.* Screening for physical stability of a pseudomonas amylase using self-interaction chromatography. *Anal. Biochem.* **357**:35–42 (2006). doi:10.1016/j.ab.2006.06.007.
22. C. W. Carter, and C. W. Carter. Protein crystallization using incomplete factorial experiments. *J. Biol. Chem.* **254**:12219–23 (1979).

23. G. Taguchi, and S. Konishi. *Orthogonal arrays and linear graphs*. American Supplier Institute, Dearborn, 1987.
24. R. L. Kingston, H. M. Baker, and E. N. Baker. Search designs for protein crystallization based on orthogonal arrays. *Acta Crystallogr. D.* **50**:429–440 (1994). doi:10.1107/S0907444993014374.
25. L. J. DeLucas *et al.* Protein crystallization: virtual screening and optimization. *Prog. Biophys. Mol. Biol.* **88**:285–309 (2005). doi:10.1016/j.pbiomolbio.2004.07.008.
26. L. E. S. D. Souza, and S. Canuto. Efficient estimation of second virial coefficients of fused hard-sphere molecules by an artificial neural network. *Phys. Chem. Chem. Phys.* **3**:4762–4768 (2001). doi:10.1039/b105183k.
27. C. M. Bishop. Neural networks and their applications. *Rev. Sci. Instrum.* **65**:1803–1832 (1994). doi:10.1063/1.1144830.
28. N. J. A. Sloane. Orthogonal Arrays. http://www.research.att.com/~njas/oadir/ (accessed 17 September 2006)
29. J. J. Valente *et al.* Second virial coefficient studies of cosolvent-induced protein self-interaction. *Biophys. J.* **89**:4211–4218 (2005). doi:10.1529/biophysj.105.068551.
30. P. De Phillips, and A. M. Lenhoff. Pore size distributions of cation-exchange adsorbents determined by inverse size-exclusion chromatography. *J. Chromatogr. A.* **883**:39–54 (2000). doi:10.1016/S0021-9673(00)00420-9.
31. J. Fanguy *et al.* Scale-down approaches for measuring protein-protein interactions. In N. E. Chayen (ed.), *Protein crystallization strategies for structural genomics*, International University Line, La Jolla, 2007, pp. 127–152.
32. P. Kratochvil. *Classical light scattering from polymer solutions*. Elsevier, New York, 1987.
33. P. Marrone. Joone–Java Object Oriented Neural Engine. http://www.jooneworld.com/ (accessed 26 June 2006)
34. JMP, Version 7. SAS Institute Inc., Cary, NC. (1989–2007).