# Neuro-Fuzzy Modeling of Nucleation Kinetics of Protein

DEVINDER KAUR, RAJENDRAKUMAR GOSAVI
Electrical and Computer Science and Chemical Engineering
University of Toledo
2801, W. Bancroft, Toledo, OH 43606
USA
http://www.eecs.utoledo.edu/~dkaur/

*Abstract:*- The estimation of the nucleation kinetics of protein using a predictive model will greatly help in reducing the number of parameters for protein nucleation studies. In this paper a Fuzzy Inference Model has been developed using the observed data for nuclear kinetics. The model was trained using neuro fuzzy inference system. The results obtained from the fuzzy model were compared with the experimentally observed values from the data collected. Results show that a better fit can be obtained using the fuzzy logic technique and the nucleation kinetics can be very well estimated at the conditions studied.

*Key Words:*- Neuro-Fuzzy Modeling, Nucleation Kinetics, ANFIS

## 1 Introduction

Nucleation is the first step in the formation of solute crystals in a supersaturated solution. One can say that the nucleation process forms the templates of solid phase on which further growth of the solid phase takes place. Nucleation of crystals is first required in solutions with no crystalline phase present. An understanding of the factors that affect nucleation kinetics is important in any crystallization or precipitation process in order to control the outcome of that process (i.e. crystal size distribution). Supersaturation levels, solution hydrodynamics, temperature, solute concentration, nonidealities and impurities in the protein solution can affect the nucleation rates of protein crystals.

The creation of a new phase in another phase that is not at equilibrium starts with nucleation. That is to say, the random formation of clusters of growth units (ions, atoms, molecules or molecular aggregates), that are able to grow further into the new phase. Nucleation is generally classified into two categories: homogeneous nucleation where external surfaces are absent and heterogeneous nucleation where the surfaces in contact with the mother phase play a key role in the formation of a second phase.

While the growth process of globular proteins was studied extensively, knowledge about the nucleation process is still very limited. There are various ways to determine the nucleation rates. One approach is to allow a system to nucleate at a given solution conditions and change the conditions so that nucleation is suppressed and only growth takes place. The number of grown crystals then provides information about nucleation kinetics. An initial rate method is used to obtain the nucleation kinetic rates for the desired conditions.

There are various instances where use of Fuzzy logic in protein science can be observed. Fuzzy logic has been used to predict protein's sub cellular locations from their dipeptide compositions [1]. It has also been successfully used [2] in molecular modeling where a new method which produces fuzzy structural vectors to predict secondary structural class of a protein from its sequence. Also the classification of protein based on the adaptive Neuro-fuzzy inference system [6] enhanced the protein secondary structure prediction. A hybrid GMC-fuzzy algorithm is used [3] in pH control of the enzymatic hydrolysis of cheese whey in a reaction carried out by alcalase immobilized on agarose gel. A strategy has been discussed [4] that uses fuzzy logic-based factors to modify algebraic rate laws and eventually obtain kinetic models that are suitable for metabolic pathway modeling without complete enzyme mechanisms. A fuzzy model of kinetics of enzymic Penicillin-G conversion has been described. [5].

In this paper comparison of the model fit of nucleation data with that of the fuzzy logic method has been done. The nucleation data with certain conditions has been modeled using non linear regression technique [7]. Since the non linear regression and weighted least square regression methods did not give satisfactory fit on the nucleation data, the nucleation data was split into two regions namely homogenous nucleation and heterogenous nucleation. Data fitting was done on both these regions to achieve a better estimation of nucleation kinteci data. The description of use of Anfis for the data fittting to give a better fit is provided. The results indicate that the nucleation rates in the range of the conditions studied could be estimated with a decent 10% error of the actual nucleation data.

## 2 ANFIS:

ANFIS stands for Adaptive Neuro-Fuzzy inference system (Anfis). ANFIS provides an optimization scheme to find the parameters in the fuzzy system that best fit the data. ANFIS applies two techniques in updating the parameters. For premise parameters that define membership functions, ANFIS employs gradient descent to fine tune them. For consequent parameters that define the coefficients of each output equations, ANFIS uses the least squares method to identify them. It learns from the data and adapts itself to the trend in the data. Anfis then fits a model for the data. A schematic of ANFIS is shown in Fig 1. The Fuzzy logic system that has been modeled can be denoted by a schematic as shown in Figure 2.

The schematics for ANFIS look similar to that of the neural network except that in ANFIS at layer 4 there are external parameters that affect the optimization process.

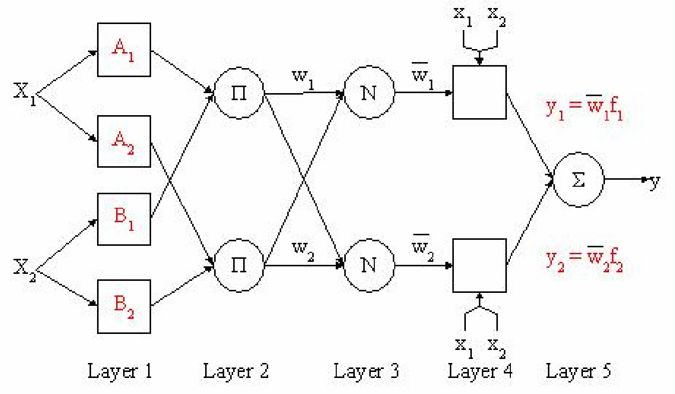

Fig 1: Schematic of ANFIS.

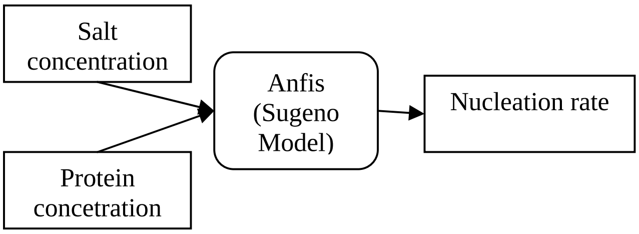

Fig 2: Schematic of the fuzzy inference system.

## 3 What is protein Nucleation?

Crystallization is a rate process similar to chemical reaction. To obtain the rate of reaction, the method of initial rates is often used [Ref Scott Fogler, Livenspeil]. The reaction is started at known initial reactant concentrations and the concentration of the product is measured with respect to time to obtain the reaction rate data. A similar procedure for crystallization requires measurement of the number of nuclei formed with respect to time in a given volume of crystallizing solution during the initial stages of nucleation where the concentration of the solution is known. Since nuclei are very small and by definition a critical nucleus will grow into a crystal of detectable size, counting the number of crystals formed at discreet time intervals should yield the nucleation rate. The assumption underlying this technique is that every nucleus formed will grow into a crystal and can be detected.

This method offers several advantages over the methods discussed above. The only challenge lies in detecting and counting crystals as they are forming and when they are very small, about a few microns in size. Since the amount of growth that takes place is very low, the concentration of the solution does not change significantly during the initial stages of crystallization. No knowledge of growth kinetics or particle size distribution is necessary.

Figure 3 is an example of how a typical nucleation rate data would look like. This nucleation data has been obtained for lysozyme protein for 2%NaCl salt concentration. The slope of the line as shown in the plot gives the nucleation rate for lysozyme at the given conditions of salt concentration and protein concentration. The nucleation data was obtained for various salt concentrations.

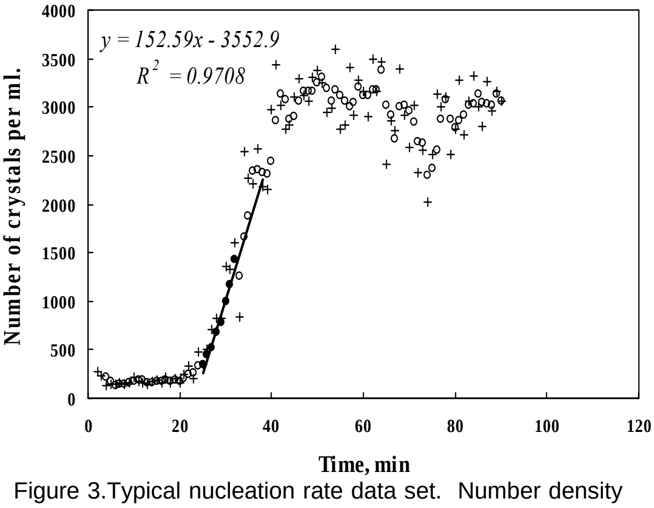

Figure 3.Typical nucleation rate data set. Number density of particles is plotted against time [7].

For the empirical modeling of the nucleation rate kinetics, the classical nucleation theory was used [7]. The classical nucleation theory expresses the homogeneous nucleation rate *J* for a spherical nucleus as

$$J = \frac{2v\sqrt{kT\sigma}}{h} N_1 \exp\left(-\frac{\Delta G_a}{kT}\right) \exp\left(-\frac{16\pi\sigma^3 v^2}{3k^3T^3(\ln S)^2}\right) \qquad (1)$$

where *v* is the molecular volume, *k* is the Boltzmann constant, *T* is absolute temperature, *σ* is the surface energy per unit area of the nuclei, *h* is the Planck's constant, $N_1$ is the number density of monomeric species in the solution, $\Delta G_a$ is the energy barrier to diffusion from bulk solution to the cluster and S is the supersaturation, often expressed as *(C/C*)*, the ratio of bulk concentration of solute to equilibrium solubility. It should be noted that the activity coefficients at bulk and equilibrium concentrations are assumed to be approximately equal by considering *(C/C*)* as the driving force for phase change instead of the ratio of activities. One can lump the parameters in the above equation and transform it into the two-parameter empirical expression below:

$$J = A\ C \exp\left\{\frac{-B}{\left[\ln\left(C/C^*\right)\right]^2}\right\} \qquad (2)$$

## 4. Model Fit Using Adaptive Neuro Fuzzy Logic

The data set that was studied for anfis are the experimental protein nucleation rates at 2%, 2.5%, 3% and 4% NaCl concentrations at pH of 4.5 and temperature 4°C. The input parameters were normalized. The fuzzy system generated a set of if-then rules after evaluating the data.

These rules are of the following format:

If (Salt is Low) and (Protein is Low) then (Nucleation Rate is Value1)

From the data first the fuzzy inference system was created using the genfis1 command. Then the fismat created was trained with anfis command. Genfis1 in combination with anfis learns from the data and then fits the data to the best fit possible hence this procedure is termed as the adaptive neuro fuzzy system of inference.

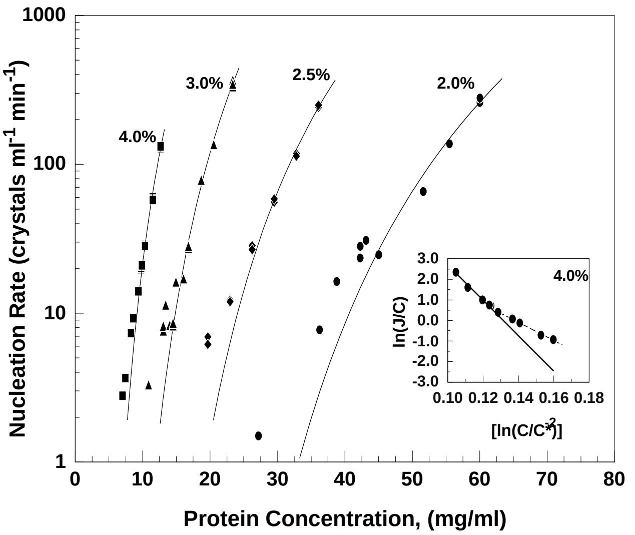

Figure 4: The symbols represent individual nucleation rate measurements for 2% to 4% NaCl concentration as indicated on the graph. The lines represent the nucleation rate model Deviation of model predictions towards the lower protein concentration region is evident when the model is linearized and plotted as in the inset for 4% NaCl nucleation rate data[7].

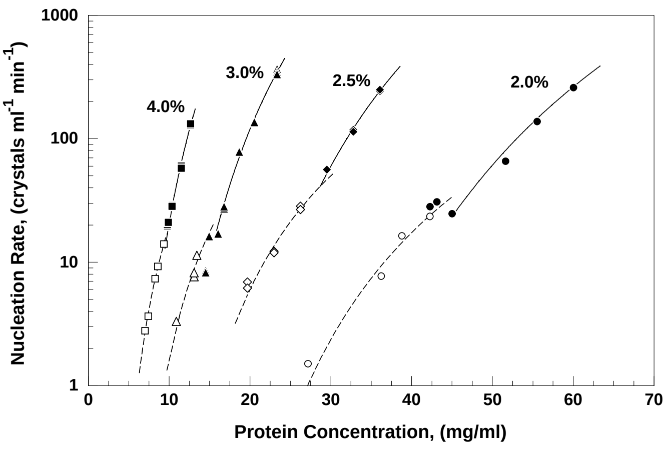

Figure 5: Model fit to data split into two regimes for 2 to 4% NaCl concentration. Solid symbols and lines indicate data points in the homogeneous (high protein concentration) range and open symbols with dashed lines belong to the heterogeneous (low protein concentration) range[7].

After treating with genfis1 it was observed that the model gave a reasonable fit with very small RMSE values on training data. Figures 6.a and 6.b show how well the model fits using genfis 1. In figure 6.b it can be seen that the model fit follows the data points with a reasonable accuracy.

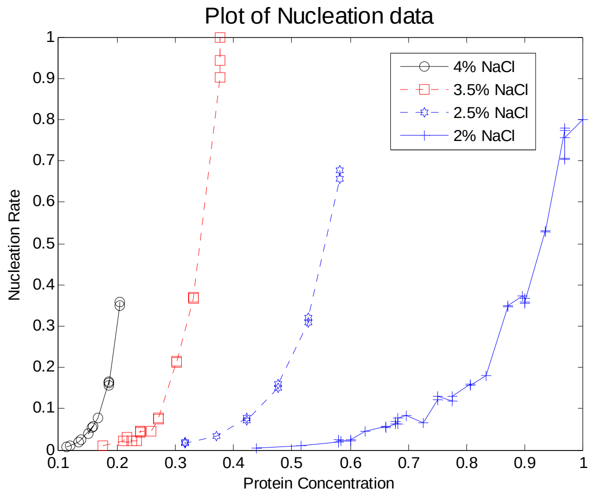

Figure 6: a) Nucleation Rate before using anfis

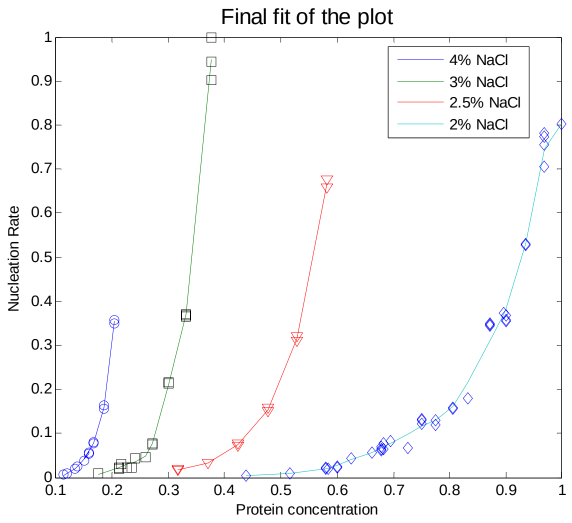

Figure 6: b) Nucleation with model fit using anfis

The error on the training data was obtained in figure 10. This graph can be used to determine the number of epochs required to obtain the minimum error. In this case the anfis used 100 epochs.

The model fit obtained using various epochs lead to nucleation estimate that are not much different than each other.

| Observed J | Model J [7] | Model J, FL* |
|---|---|---|
| 2.79 | 2.71 | 1.81 |
| 3.66 | 3.90 | 3.35 |
| 7.35 | 7.28 | 7.41 |
| 9.25 | 9.10 | 9.51 |
| 14.04 | 14.15 | 20.23 |
| 19.58 | 16.93 | 20.23 |
| 20.17 | 16.93 | 21.67 |
| 21.01 | 18.67 | 15.70 |
| 28.76 | 27.53 | 28.56 |
| 29.01 | 27.53 | 28.56 |
| 28.32 | 27.53 | 28.56 |
| 60.52 | 63.79 | 58.70 |
| 60.17 | 63.79 | 58.70 |
| 57.64 | 63.79 | 58.70 |
| 128.83 | 128.00 | 130.74 |
| 132.11 | 128.00 | 130.74 |
| **SSE**** | **106.54** | **82.76** |
| **$R^2$** | **0.9956** | **0.9966** |

Table 1.1: Model estimates for 4% NaCl

| Observed J | Model J [7] | Model J, FL |
|---|---|---|
| 3.32 | 2.95 | 2.47 |
| 7.62 | 8.74 | 8.62 |
| 8.21 | 8.74 | 8.62 |
| 11.37 | 10.03 | 9.51 |
| 8.33 | 5.46 | 10.87 |
| 8.25 | 7.81 | 11.90 |
| 8.6 | 7.81 | 11.90 |
| 16.36 | 10.30 | 12.75 |
| 16.09 | 10.30 | 12.75 |
| 16.26 | 10.30 | 12.75 |
| 17.04 | 20.11 | 18.09 |
| 27.16 | 29.84 | 28.82 |
| 27.45 | 29.84 | 28.82 |
| 28.28 | 29.84 | 28.82 |
| 79.53 | 71.07 | 77.49 |
| 78.48 | 71.07 | 77.49 |
| 78.58 | 71.07 | 77.49 |
| 134.91 | 145.20 | 136.30 |
| 135.71 | 145.20 | 136.30 |
| 136.46 | 145.20 | 136.30 |
| 332.65 | 347.29 | 349.47 |
| 368.48 | 347.29 | 349.47 |
| 347.45 | 347.29 | 349.47 |
| **SSE** | **1262.09** | **735.58** |
| **$R^2$** | **0.9956** | **0.9975** |

Table 1.2: Model estimates for 3% NaCl

| Observed J | Model J [7] | Model J, FL |
|---|---|---|
| 6.92 | 5.44 | 6.60 |
| 6.18 | 5.44 | 6.60 |
| 12.25 | 13.55 | 12.01 |
| 11.96 | 13.55 | 12.01 |
| 28.42 | 27.06 | 27.75 |
| 26.75 | 27.06 | 27.75 |
| 55.629 | 52.39 | 56.78 |
| 58.56 | 52.39 | 56.78 |
| 118.02 | 121.60 | 116.33 |
| 113.89 | 121.60 | 116.33 |
| 242.4 | 244.35 | 245.92 |
| 249.8 | 244.35 | 245.92 |
| **SSE** | **163.28** | **42.52** |
| **$R^2$** | **0.9981** | **0.9995** |

Table 1.3: Model estimates for 2.5% NaCl

| Observed J | Model J [7] | Model J, FL |
|---|---|---|
| 1.51 | 1.02 | 1.51 |
| 3.42 | 3.91 | 3.72 |
| 3.74 | 3.91 | 3.72 |
| 7.44 | 8.98 | 6.89 |
| 8.66 | 8.98 | 6.89 |
| 7.71 | 9.38 | 7.52 |
| 8.4 | 11.12 | 10.13 |
| 9.2 | 11.12 | 10.13 |
| 9.47 | 11.12 | 10.13 |
| 16.34 | 14.47 | 14.63 |
| 20.824 | 20.14 | 21.04 |
| 20.789 | 20.14 | 21.04 |
| 20.373 | 20.14 | 21.04 |
| 22.97 | 23.10 | 24.17 |
| 25.74 | 23.10 | 24.17 |
| 23.54 | 23.10 | 24.17 |
| 23.49 | 23.99 | 25.13 |
| 28.13 | 12.04 | 25.13 |
| 30.86 | 14.71 | 28.19 |
| 24.74 | 22.45 | 36.26 |
| 44.69 | 30.58 | 42.82 |
| 47.72 | 30.58 | 42.82 |
| 48.33 | 30.58 | 42.82 |
| 43.76 | 40.95 | 47.72 |
| 47.67 | 40.95 | 47.72 |
| 47.85 | 40.95 | 47.72 |
| 58.48 | 58.96 | 59.66 |
| 57.82 | 58.96 | 59.66 |
| 65.76 | 77.38 | 79.48 |
| 127.42 | 112.40 | 113.93 |
| 127.25 | 112.40 | 113.93 |
| 128.87 | 112.40 | 113.93 |
| 137.65 | 140.39 | 136.85 |
| 135.03 | 146.08 | 141.42 |
| 131.08 | 146.08 | 141.42 |
| 131.48 | 146.08 | 141.42 |
| 195.75 | 196.80 | 194.89 |
| 194.71 | 196.80 | 194.89 |
| 194.52 | 196.80 | 194.89 |
| 285.46 | 254.10 | 273.74 |
| 287.6 | 254.10 | 273.74 |
| 260.06 | 254.72 | 274.44 |
| 259.6 | 254.72 | 274.44 |
| 278.71 | 254.72 | 274.44 |
| 295.18 | 322.65 | 295.30 |
| **SSE** | **6385.59** | **2039.03** |
| **$R^2$** | **0.9833** | **0.9947** |

Table 1.4: Model estimates for 2% NaCl

* Fuzzy Logic ** Sum of Square of Errors

| # of Epochs | Training data RMSE |
|---|---|
| 50 | 0.01479 |
| 70 | 0.01476 |
| 75 | 0.01483 |
| 80 | 0.01476 |
| 85 | 0.01480 |
| 100 | 0.01474 |
| 150 | 0.01472 |
| 200 | 0.01470 |
| 300 | 0.01190 |
| 500 | 0.01189 |

Table 2: Comparison of RMSE values for various epochs.

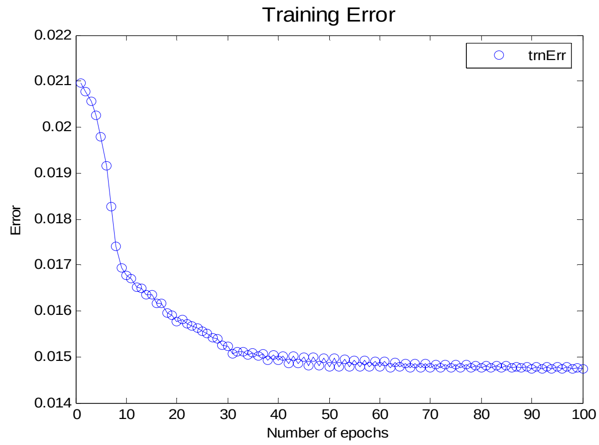

Figure 7: Training Data Error vs. Number of Epochs

The Root Mean Square Error on the fit plotted in Fig. 7 was found to be 0.01474.

Table 2 denotes the training data RMSE for various epochs studied.

The regressed values obtained from the earlier work [7] and those obtained from the fuzzy logic technique were used to compare the fit using $R^2$ values. The comparison is provided in table 1.1, 1.2, 1.3 and 1.4.

## 5 Conclusions:

A fuzzy model for the protein nucleation was created from the data and then the model learnt using adaptive neuro fuzzy modeling generating a set of rules, with the help of which it can estimate the nucleation rate at any conditions of NaCl concentration or protein concentration in the range studied. Thus for any salt concentration in the range of 2% to 4%; for any protein concentration at that salt concentration below 62 mg/ml the nucleation kinetics rate can be estimated. . This implies that the fuzzy logic can be confidently used to estimate the nucleation rate at the conditions studied.

If more data are available at various salt concentrations like 4% to 10% and below 1% apart from what has been used in this work, this fuzzy logic exercise could be extensively used to estimate the nucleation rate of protein at various salt concentrations. The future work in this area will comprise of involving more parameters such as temperature, pH and additive concentrations to estimate the nucleation kinetics of the protein.

*References:*

1. Y. Huang, Y. Li, “Prediction of protein subcellular locations using fuzzy k-NN method”, *Bioinformatics,* 20(1), 21-28, 2004.

2. J. Boberg, T. Salakoski, M. Vihinen. “Accurate prediction of protein secondary structural class with fuzzy structural vectors”, *Protein Engineering,* 8(6), 505-12, 1995.

3. R. Sousa, G. P. Lopes, G. A. Pinto, P. I. F. Almeida, R. C. Giordano. “GMC-fuzzy control of pH during enzymatic hydrolysis of cheese whey proteins”, *Computers & Chemical Engineering,* 28(9), 1661-1672, 2004.

4. B. Lee, J. Yen, L. Yang, J. C. Liao. “Incorporating qualitative knowledge in enzyme kinetic models using fuzzy logic”, *Biotechnology and Bioengineering,* 62(6), 722-729, 1999.

5. R. Babuska, H. J. L. Van Can, H. B. Verbruggen. “Fuzzy modeling of enzymic penicillin-G conversion”, *Proceedings of the World Congress, International Federation of Automatic Control, 13th, San Francisco*, June 30-July 5, 1996.

6. J. A. Hering, P. R. Innocent, P. I. Haris,. “Neuro-fuzzy structural classification of proteins for improved protein secondary structure prediction”, *Proteomics*, 3(8), 1464-1475, 2003.

7. V. Bhamidi, S. Varanasi, C. A. Schall. “Measurement and Modeling of Protein Crystal Nucleation Kinetics”, *Crystal Growth & Design,* 2(5), 395-400, 2002.

8. D. M. Himmelblau. “*Process Analysis by Statistical Methods*”, Wiley: New York, 1970.