Artificial neural network prediction of lysozyme solubility for protein crystallization

A S Sokolovskii¹, N A Besedina², Yu Berdnikov¹ and V G Dubrovskii¹

¹ Department of Modern Functional Materials, ITMO University, 197101 St. Petersburg, Russia
² Nanobiotech Laboratory, St. Petersburg Academic University, 194021 St. Petersburg, Russia

Abstract. In current work, we use the method of artificial neural networks to predict the variation of the solubility of lysozyme protein in solution as the complex function of four crystallization parameters: temperature and pH of the bulk solution, the concentration of precipitant and the concentration of buffer solution. The artificial neural network predicts the salting-in effect at small concentrations of sodium chloride. The predicted dependence of the solubility on the concentration of sodium acetate buffer shows the minimum around 0.1M at different temperatures of the bulk solution.

1. Introduction

Understanding of interaction mechanisms and functions of macromolecules in living cells requires the knowledge of the spatial structure of the considered molecules. This information can be accessed by X-ray analysis of crystallized protein structures [1]. Within this approach, the formation of protein crystals is the limiting step in the process of establishing the tertiary structure of the protein.

The counter-diffusion technique is considered as one of the most progressive methods for protein crystallization. Within this technique the results of a single experiment are equivalent to data acquired in a large set of crystallization experiments within standard methods like sitting drop, batch crystallization, etc [2].

In the counter-diffusion process nucleation and growth of protein crystals are controlled by supersaturation of the protein solution, which is defined as the ratio of protein volume concentration to protein solubility. We can calculate the concentrations of protein molecules by modeling their diffusion in capillary and diffusion of precipitant molecules [3]. However, the solubility of protein molecules cannot be accessed directly due to its complex non-linear dependences on the temperature and pH of the bulk solution, the concentration of buffer and precipitant concentration. Within this work we use the approach of artificial neural networks to evaluate the solubility of protein molecules.

2. Method

In this paper, we use the Artificial Neural Network (ANN) with three hidden layers, with 30, 20, 10 hidden neurons in each hidden layer. The four inputs of the ANN are temperature of the bulk solution,
sodium chloride concentration, sodium acetate buffer concentration and pH-level of the bulk solution. The only output gives the solubility of lysozyme protein molecules. Figure 1 summarizes the simplified topology of ANN. The scheme of the ANN with three hidden layers obtained from MATLAB software is shown in figure 2.

![ANN Architecture Topology](image)

**Figure 1.** ANN architecture topology. The ANN use four inputs: temperature and pH of the bulk solution, sodium chloride concentration, sodium acetate buffer concentration. The only output is the solubility of lysozyme protein molecules.

![Trained Neural Network Scheme](image)

**Figure 2.** The scheme of the trained neural network. The ANN contains three hidden layers with 30, 20 and 10 neurons. “w” and “b” denotes the weight and the bias matrixes of each hidden layer, correspondingly.

We use the backpropagation (BP) ANN, with Levenberg-Marquardt (LM), Bayesian regularization (BR) and Quasi-Newton Broyden-Fletcher-Goldfarb-Shanno (BFGS) training algorithms. The performance of ANN is evaluated by mean-squared error (MSE). Modeling of ANN was carried out with Neural Network Toolbox of MATLAB system. We choose and compare three different training algorithms to optimize the evaluation times and value of MSE. The details of the ANN performance are shown in Table 1. BR algorithm provides the lowest MSE value in comparison to other algorithms but requires the longest computation time of 17835 seconds. LM algorithm requires 12% less computation time, but MSE value for LM algorithm is five times greater than for BR algorithm. BFGS algorithm provides the shortest computation times, which is about ten times less than for LM and BR algorithms but the MSE value of 0.0268 is ten times higher than for BR algorithm.
Table 1. Efficiency comparison of training algorithms.

| Name of algorithm | Epochs | Time of evolution, sec | MSE   |
|-------------------|--------|------------------------|-------|
| LM                | 50000  | 15670                  | 0.0098|
| BR                | 50000  | 17835                  | 0.002 |
| BFGS              | 50000  | 1575                   | 0.0268|

The training of the ANN was based on the solubility data published in Refs. [4–7]. All the data correspond to crystallization conditions of lysozyme crystals with tetragonal syngony within the temperature range from 1.6 to 30.7 °C; pH of the solution varies from 4.0 to 5.4; concentration of sodium acetate buffer changes from 0.01M to 0.5M; concentration of sodium chloride remains from 0% to 7% w/v. We use 5193 values of solubility in the set of crystallization conditions for ANN training.

3. Results

The MSE after training during 50000 epochs was estimated to 0.002 for Bayesian regularization algorithm. The obtained MSE value is 2.5 times less than in ref. [8] which can be explained by a larger set of the training data. We also predict the dependence of the solubility on the concentration of sodium acetate buffer, which was not discussed previously. The set of data allow us to analyze the solubility at small sodium chloride concentrations and observe the salting-in effect as shown in figure 3.

![Figure 3](image-url)  

**Figure 3.** The solubility of lysozyme at 18 °C, at pH = 4.5, as a function of NaCl concentration. The solubility increases at small concentrations of sodium chloride which demonstrate the presence of the salting-in effect.

According to ref. [4] the solubility sensitive to the amount of sodium acetate buffer with the minimum near the concentrations of the buffer of 0.1M. The predicted dependence reproduces this minimum for different temperatures and NaCl concentrations. Figure 4 (a,b) shows the obtained predicted dependences of solubility on the concentration of sodium acetate buffer.
Figure 4. Solubility of lysozyme at pH = 4.0, as a function of sodium acetate buffer concentration, for different 2%, 3% and 4% of NaCl concentration: a) at 10 °C; b) at 20 °C

4. Conclusion
We demonstrate that the approach of artificial neural networks can provide an approximation for the solubility as a function of the temperature of the bulk solution, the concentration of buffer and precipitant concentration. In the predicted dependence the salting-in effect is observed at small sodium chloride concentrations. The dependence of the solubility on the concentration of sodium acetate buffer shows the minimum around 0.1M at temperatures of the bulk solution of 10 and 20°C and NaCl concentration of 2%, 3% and 4%. In future, the developed approach of artificial neural networks can be used for prediction of solubility for different protein molecules.

Acknowledgments
The authors gratefully acknowledge the financial support received from the Ministry of Education and Science of the Russian Federation under grant 14.587.21.0040 (project ID RFMEFI5871X0040).

References
[1] Franklin R E and Gosling R G 1953 Nature 172 156
[2] Garcia-Ruiz J M 2003 Methods Enzymol. 368 130–154.
[3] Sokolovskiy A S, Trushin Yu V, Lubov M N, Eliseev I E, Yudenko A N and Dubina M V 2016 J. Phys.: Conf. Ser. 741 012053
[4] Forsythe E L and Pusey M L J. Cryst. Growth 1996 168 112–17
[5] Forsythe E L, Judge R A and Pusey M L J Chem Eng Data 1999 44 637–40
[6] Cacioppo E and Pusey M L J. Cryst. Growth 1991 114 286–92
[7] Otalora F and Garcia-Ruiz J M J. Cryst. Growth 1997 182 141–54
[8] Zhang X, Zhang S and He X J. Cryst. Growth 2004 264 409–16