Parameter Estimation of Thermodynamic Models for Equilibrium of Proteins Solutions with salts


b DOPI/IQ/University of the State of Rio de Janeiro, R. São Francisco Xavier 524 - Rio de Janeiro, Brazil
c DEQ/EQ/Federal University of Rio de Janeiro, Av. Athos da Silveira Ramos 149- B1.E-sl.201, Rio de Janeiro, Brazil

Abstract

The main challenges of biomolecules production on an industrial scale can be divided into challenges in the upstream processes, which involve the product synthesis, and challenges in downstream processes involving the separation and purification of biomolecules. Among the unit operations employed in biotechnological processes, many of them need the knowledge of thermodynamic properties of the solutions, such as solubility and phase equilibrium of the system. In fact, the simulation of unit operations involved in protein purification, and therefore the prediction of the behavior of proteins in each unit operation, requires the understanding of both the phase behavior of systems and the chemical behavior of fluids in the processes. To do this, it is necessary the calculation of phase equilibrium, which can be quite complex, since it is possible to find systems with multiple phases in equilibrium, multiple chemical species distributed in phases, the presence of chemical reactions and the presence of partially or totally dissociated electrolytes. The accuracy of the results depends on model parameters that must be estimated and is fundamental for the correct process design. Usually, the numerical problem resulting of the resolution of the phase equilibrium equations corresponds to a nonlinear programming problem (NLP). In this context, the proposed paper presents an analysis of the utilization of different optimization methods for the problem of parameter estimation for the calculus of the equilibrium of solutions of proteins with salts. This optimization problem may present some particularities which difficult the utilization of traditional algorithms (e.g. multiple local optima). In this work more modern numerical techniques are employed to circumvent these obstacles.

Keywords: proteins, equilibrium, parameter estimation, optimization method

1. Introduction

Among the unit operations employed in biotechnological processes, many of them need the knowledge of thermodynamic properties of the solutions, such as solubility and phase equilibrium of the system. The accuracy of phase equilibrium calculation depends on model parameters that must be estimated and is fundamental for the correct process design. For example, the simulation of the chromatographic profile of proteins solutions involves the description of adsorption equilibrium of the system protein/chromatographic resin. To do that, there is a need for appropriate models to represent the nonideality of the phases in equilibrium, especially the mobile phase. Therefore, there is a need of thermodynamic models able to predict satisfactorily the activity
coefficient of proteins and salts which were not well explored in adsorption modeling yet.

Literature has some attempts to describe biomolecules behavior in solution, although in most cases the works just deals with amino acids and simple peptides. Less numerous are the works modeling protein solutions. Most of it is based on the potential strengths and equations of state with interaction potential of hard spheres. These models are related to complex structures with a lot of parameters to be determined and have been applied only in narrow range of experimental conditions. Gibbs free energy models have been employed only in a few studies (Agena et al., 1997; Agena et al., 1998; Coutinho and Pessoa, 2004).

This paper presents an optimization analysis of the problem of parameter estimation for the calculus of the equilibrium of proteins with salts solutions using an excess Gibbs free energy model. The optimization problem may present some particularities which difficult the utilization of traditional algorithms (e.g. multiple local optima).

2. Thermodynamic Problem

The experimental data analyzed are osmotic pressure measurements of an aqueous solution containing a protein converted to activity coefficient through a virial expansion (Wills et al., 1993), as done by Agena et al. (1997) and Coutinho and Pessoa (2004). The modeling is performed considering a binary system composed of the solute (protein) and a pseudo solvent (saline solution). The thermodynamic model used is the original UNIQUAC model with symmetrical convention for solvent and unsymmetrical convention for proteins.

2.1. System analyzed

The data are available from different system conditions from Moon et al. (2000) resulting in six systems. The measurements are of lysozyme in ammonium sulfate solutions over an ionic strength of 1.0 and 3.0M at pH 4, 7, and 8. Protein concentrations vary in a range from 4 to 20 g/L and temperature is always 25°C.

2.2. UNIQUAC original model

The UNIQUAC model, originally developed by Abrams and Prausnitz (1975) is an excess Gibbs free energy model composed of two additive parts: a combinatorial and a residual. The combinatorial part corresponds to an entropic contribution, which depends on the composition, shape and size of the species and only need data of pure species. The residual contribution includes the intermolecular forces responsible for the enthalpy of mixing and depends on adjustable binary interaction parameters ($u_{ij}$). Such parameters are estimated from experimental data and can be found in the literature for some systems. In the case of biomolecules, there are few literature data concerning the binary interaction parameters, and, therefore, these must be estimated. It is important to notice that for different systems the protein/protein interaction parameter is the same when pH and temperature are the same.

2.3. Input data

The application of the virial expansion for binary systems to link the osmotic pressure with virial coefficients is very common in literature. Although it is possible to use any equation of state to perform this relationship, the formulation using the virial coefficients is considered as the one that provides the “experimental” activity coefficient.
Parameter estimation of thermodynamic models for equilibrium of proteins solutions with salts

To calculate the parameters of shape and volume \((r, q)\) of UNIQUAC model, Agena et al. (1997) developed a correlation presented in Equations 5 and 6 where \(M_c\) is the crystal molecular weight of proteins.

\[
q_i = 0.0273M_c \tag{5}
\]
\[
r_i = 0.0362M_c \tag{6}
\]

The employed properties of lysozyme are MW=1700 g/mol; \(V_s=0.734 \text{cm}^3/\text{g}\); \(q=464.1\); and \(r=615.4\). The properties of the pseudosolventes were adopted as the same as pure water (MW=18.015g/mol; \(V_s=1.003 \text{cm}^3/\text{g}\); \(q=1.40\); \(r=0.92\)).

2.4. Optimization

The optimization procedure seeks the minimization of the objective function of Equation 7.

\[
F_{\text{obj}} = \sum_{\text{NDP}} \left( \frac{\gamma_{\text{exp}} - \gamma_{\text{calc}}}{\gamma_{\text{exp}}} \right)^2 \tag{7}
\]

where NDP is the number of data points, and the subscripts \(\text{exp}\) and \(\text{calc}\) refer to the experimental and calculated points, respectively. The estimated parameters (decision variables) are the interaction parameters.

The optimization problem is solved using a set of local optimization methods and a set of stochastic global optimization methods. The first set is composed of the methods Simplex (Press et al., 1986), Subplex (Rowan, 1990), Powell (Press et al., 1986) and BFGS (Press et al., 1986). The second set of methods contains Simulated Annealing, Particle Swarm (Kennedy and Eberhart, 1995) and Genetic Algorithm (Carroll, 1996). The tolerance of almost all methods was \(10^{-8}\). The maximum number of iteration of the methods was 10000. For the simulated annealing the number of moves per temperature level was 200, the temperature cooling parameter was 0.9, the maximum number of moves without acceptance was 20. For the genetic algorithm the number of individuals was 200, the number of chromosomes per individual was 30 and the maximum number of generations was 500. For Particle Swarm the number of particles was 200.

3. Results

3.1. Different optimization methods evaluation

The objective function of the parameter estimation problem generates several local optimum. Optimization methods that are very dependent of initial estimative (usually the deterministic methods like simplex) might be stuck in local optimums and give bad results.

For a specific system with salt concentration equal to 3 M and pH 8, Table 1 presents a comparison between results obtained with different optimization methods and different initial estimative. Due to space limitations in presenting the results, Table 1 does not present all results but just the better for each method.

First, three different initial guesses were used \((u_{ij}=0, u_{ij}=1000, u_{ij}=-1000)\). Such initial guesses were randomly chosen and do not represent any particular physical state. All deterministic methods couldn’t effectively perform a good search and remained stuck in the initial guess since the topologic analysis of the optimization search space shows great extension of many flat areas. The same result was found with the use of Simulated Annealing. Subplex method was the only method dependent of initial estimative that in all cases was able to find optimum results.

Then, an initial guess near an optimum found by subplex method was given as initial parameters for all methods. In such scenario, BFGS and Simulated Annealing still didn’t present any improved results but Powell’s method and simplex method were able
to reach an optimum (presented in Table 1). Wiling that BFGS and Simulated Annealing gave a better result an exhausting search for initial estimative had to be performed and is not presented here.

For the stochastic methods, Particle Swarm (PSO) and Genetic Algorithm (GA), the routine was run three times and just the best result is presented.

Table 1 – Results of different optimization methods

<table>
<thead>
<tr>
<th>Method</th>
<th>Initial estimative</th>
<th>Final results</th>
<th>RMSD (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( U_{\text{ps/ps}} ) (cal/mol)</td>
<td>( U_{\text{ps/prot}} ) (cal/mol)</td>
<td>( U_{\text{prot/prot}} ) (cal/mol)</td>
</tr>
<tr>
<td>Subplex</td>
<td>1000 1000 1000</td>
<td>1093.26 800.36 952.17</td>
<td>0.006</td>
</tr>
<tr>
<td>Subplex</td>
<td>-1000 -1000 -1000</td>
<td>-906.73 -1199.64 -1047.82</td>
<td>0.006</td>
</tr>
<tr>
<td>Subplex</td>
<td>0 0 0</td>
<td>93.26 -199.64 -47.82</td>
<td>0.006</td>
</tr>
<tr>
<td>Powell</td>
<td>1000 800 900</td>
<td>1138.32 7424.60 5739.56</td>
<td>0.43</td>
</tr>
<tr>
<td>Simplex</td>
<td>1000 800 900</td>
<td>1471.35 1177.08 1324.22</td>
<td>0.008</td>
</tr>
<tr>
<td>GA</td>
<td>- - -</td>
<td>3095.24 2936.51 3730.16</td>
<td>13.16</td>
</tr>
<tr>
<td>PSO</td>
<td>- - -</td>
<td>1822.99 1415.46 1234.90</td>
<td>0.25</td>
</tr>
</tbody>
</table>

It can be observed that the best results are obtained with the subplex method. The simplex method is extremely dependent from the initial estimative but is more robust than conjugated direction methods and quasi-newton methods when the initial estimative is near an optimum. Between the stochastic methods, particle swarm shows to be better with results similar to subplex. Figure 1 shows graphically the best results of each method and the experimental data of this system.

Figure 1 – Best results of activity coefficient of the system analysed for each method where the dot marks represent the experimental data and the lines represent the calculated results

3.2. Different systems results with the same optimization methods parameters

In this second part, it is presented an analysis of all six systems (Moon et al., 200) comparing subplex and PSO results which was the two method that presented better
results in the first section of this analysis. The subplex initial estimate was \( u_{ij} = 0 \). It can be seen that subplex method has a better performance in general.

Table 2 – Root mean square deviation for the six systems studied (%)

<table>
<thead>
<tr>
<th>System</th>
<th>Subplex</th>
<th>PSO</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1</td>
<td>0.62</td>
<td>0.70</td>
</tr>
<tr>
<td>S2</td>
<td>0.004</td>
<td>0.82</td>
</tr>
<tr>
<td>S3</td>
<td>0.005</td>
<td>0.07</td>
</tr>
<tr>
<td>S4</td>
<td>0.003</td>
<td>0.48</td>
</tr>
<tr>
<td>S5</td>
<td>1.07</td>
<td>1.07</td>
</tr>
<tr>
<td>S6</td>
<td>0.006</td>
<td>0.17</td>
</tr>
</tbody>
</table>

4. Conclusions

The numerical problem resulting from the resolution of the phase equilibrium equations for biosystems corresponds to a nonlinear programming problem (NLP). This work analyzed the use of different optimization methods for the problem of parameter estimation for the calculus of the equilibrium of solutions of proteins with salts. It could be seen that the subplex method has a better performance than the others for the optimization of the system. Although it is a method that depend on initial estimative it is very robust and not so dependent. Future investigations should extend the procedure for analyzing ternary and quaternary systems without the consideration of pseudo solvents.

5. Acknowledgments

The authors thank CNPq for the research founding.

References


The main text can start here. Next paragraphs should start with heading “Elstorderheading”