MODELLING OF PROTEIN SOLUTION PROPERTIES

A Thesis submitted to the University of London

for the degree of Doctor of Philosophy

by

Sabine M. Agena, M.Sc., Dipl.-Ing.

Department of Chemical & Biochemical Engineering

University College London

London, UK

November 1997

This thesis is dedicated with much love

to my husband Amit and

to my parents Gabriele and Helmut

Abstract

The research pursued in this work concentrates on modelling two protein solution properties: activity coefficients and solubility. While modelling protein solubility was the prime objective activity coefficients were considered first as deviation from ideal solution behaviour was expected to occur for protein containing systems.

Activity coefficients of protein related compounds, amino acids and peptides, were studied first hypothesising that these compounds represent proteins and because activity coefficient data is documented for those compounds but not for proteins. The predictive UNIFAC model was studied but failed, which led to examination of the related UNIQUAC model. The objective of the work was the creation of a model base, which was achieved. This model base was then utilised for protein containing systems.

Protein activity coefficient data was made available and could be successfully modelled using the established framework. Furthermore, the activity coefficient data was examined over different pH and temperature ranges, salt types and concentrations. A qualitative comparison of the data to protein solubility results of other researchers was pursued and used to confirm the model approach.

Having demonstrated that protein solubility was qualitatively represented through protein activity coefficients a quantitative solubility approach was addressed next. For two protein systems the solubility behaviour was modelled successfully as a function of salt concentration and temperature. Findings of other researchers were integrated into the discussion of the model results while also the calculated protein activity coefficients were examined and a qualitative confirmation of the model was achieved.

The models used in this study for the description of two protein solution properties, when applied to six different proteins over various pH and temperature ranges, salt types and concentrations showed qualitative and quantitative success. They should find application to many other protein systems.

Acknowledgements

I would like to express my sincere gratitude to all those who provided support throughout my work at University College London, UK and at National Aeronautics and Space Administration, USA.

Special thanks go to Dr. D. Bogle at U.C.L. and to Dr. M. Pusey at NASA for supporting me and my work. I would like to thank Dr. D. Bogle for advising me during my Ph.D. program and for arranging a European research project in which I participated for two years. Furthermore, I am very grateful to Dr. M. Pusey who invited me to join his Biophysics research team and supported the final year of this work.

Furthermore, I would like to thank the team members of our European Union Project in the United Kingdom, Denmark, Austria and Spain for their co-operation and hospitality. Moreover, I like to thank Dr. L. Constantinou, Prof. R. Gani and Prof. P.

Rasmussen in Denmark for inviting me to their department and for their support. Special thanks go also to Prof. F. Pessoa from Brazil for many interesting discussions. Last, I like to thank Prof. D. Winzor, Prof. P. Wills, Prof. A. Lenhoff and Prof. H. Blanch who took the time to discuss my work.

Furthermore, I am very grateful to the European Union who granted me a personal research fellowship to pursue part of this work.

Table of Contents

List of Tables

List of Figures

Chapter 1

Modelling of Protein Solution Properties

The motivation for this work is presented and the development and pursuit of the research project is described while the relevant publications are reviewed.

1.1 Introduction

This work concentrates on the modelling of protein solution properties. Proteins were the targeted compounds as they are a main product range in the rapidly developing biochemical industry. From biochemical processes various proteins are produced e.g. enzymes used in washing powders, hormones such as insulin used for medical purposes and growth hormones used for cattle breeding.

Two kinds of solution properties were studied for different proteins: activity coefficients and solubility. The process of selecting these protein properties in order to study and model them are presented in this chapter, chapter one, while also the overall significance of this work is elaborated.

This work was pursued due to the fact that knowledge of protein solution properties, either from experiment or from models, is of importance in a variety of areas such as biochemical process development, protein crystal growth and medical research. The area of optimal process development, where e.g. product quality, processing cost and process controllability determine an optimal process, was the main focus for this work [Bogle et al.,1996] and guided this work. In order to develop optimal biochemical processes, which is becoming increasingly important due to rising industrial competition, the properties of the different compounds, i.e., product and impurities, have to be evaluated. Already the first step towards the optimal process development, which deals with the selection and arrangement of process units, i.e., process synthesis, requires extensive assessment of compound properties. Process synthesis concepts have been discussed and applied with success for chemical engineering purposes [Siirola and Rudd, 1971; Barnicki and James, 1990; Jaksland et al., 1995] but only a few works are aimed at biochemical engineering problems.

Wheelwright (1987) discussed applications of process synthesis concepts for biochemical processes. He introduced heuristics based on physicochemical compound properties. Key properties of the product, e.g. protein, and the impurities have to be screened, and those properties that demonstrate greatest difference for product and impurities determine, which separation and purification processes are chosen. Asenjo (1990) introduced an expert system that allows for computer aided synthesis of biochemical processes. He also introduced a set of heuristic rules to select process units as a function of physicochemical properties of the different compounds or systems, e.g. bacterial cells, occurring during processing. In 1996 Wai et al. studied a related concept using a property ratio matrix to create initial process alternatives that are likely to lead to the optimal process.

All the mentioned synthesis methodologies require a diverse range of compound properties as do any of the other process development stages such as process design and process simulation. The properties needed for these methodologies are most conveniently obtained from models and it is the aim of this research project to study models that estimate compound properties over system conditions and predict properties over different compound types.

Following from the fact that compound properties are needed for the first step, i.e., process synthesis, towards the development of optimal biochemical processes, the major process units and the relating compound properties exploited in these units were examined. This was pursued in an attempt to screen the properties for their importance. Protein production units used in the downstream processing from homogenisation, solubilisation to freeze - dying were viewed with respect to the exploited physicochemical properties and are listed in table 1.1. Depending on the unit operation different properties are exploited e g. an ion exchange unit is applied when differences in the ionic surface charges are found at given system conditions for product and impurities. Likewise, e.g. sedimentation is applied as a separation step if density and size differences are encountered for product and impurities. The main physiochemical properties exploited in the most commonly used biochemical unit operations are: solubility, density, size, molecular weight, partition coefficient, biospecific attraction, ionic surface charge, hydrophopicity, covalent binding, isoelectric point, melting point and glass temperature [Wheelwright, 1987; Burgess, 1988; Asenjo, 1990].

The decision to focus on protein solubility followed various reasons: occurrence of solubility related unit operations in the downstream processing, the availability of experimental data, and the non availability of an estimation or prediction method. The fact that protein solubility has been investigated since the mid 19th century

Table 1.1: Biochemical unit operations and the exploited protein properties

biochemical unit operation	physicochemical property
solubilisation	solubility
sedimentation	density, size
centrifugation	density, size
filtration	size
membrane filtration	size, molecular weight
extraction	partition coefficient
partitioning	partition coefficient
precipitation	solubility
evaporation	vapour pressure*
bio - reactions	biospecific attraction
ion exchange	ionic surface charge
ion exchange chromatography	ionic surface charge
hydrophobic interaction chromatography	hydrobicity
covalent chromatography	covalent binding
affinity chromatography	biospecific attraction
chromatofocusing	isoelectric point
size exclusion chromatography	size
partitioning chromatography	partition coefficient
crystallisation	solubility, melting point
drying by evaporation of water	vapour pressure*
lyophilization	glass temperature, vapour pressure*

* solvent

[Hofmeister, 1888, and references within; Ducruix and Giege, 1992, and references within] until today emphasises its importance. The importance to model a protein's solubility is further demonstrated by the fact that solubility differences of compounds are exploited in 80 % of the applied isolation procedures [Scopes, 1987] and in 57 % of the large scale processes [Bonnerjea et al., 1986]. Furthermore, only recently useful experimental data for protein solubility as a function of salt concentration, temperature and pH [Cacioppo et al., 1991; Ewing et al., 1994] was made available, which therefore allows for comprehensive modelling to be undertaken. Moreover, a model for protein solubility, that is applicable for process development purposes, is not yet available.

This work was motivated by the requirements for optimal biochemical process development where e.g. a solubility model supports the design of a crystallisation unit or can clarify if a solid and liquid phase are encountered for a certain unit operation. However, not only engineering and production purposes link up to this work but also studies aiming at the protein crystal growth process and at medical research. For the protein crystal process a protein's solubility is a major property which once it can be modelled or predicted will support the creation of crystals. To understand a protein's crystallisation process, which is directly related to a protein's solubility behaviour, and to produce protein crystals is a major research objective of various teams [Ducruix and Giege, 1992, and references within]. Crystals are needed to determine a protein's three dimensional structure. Knowledge of a protein's three dimensional structure not only enlightens our understanding of a protein's function and biological purpose but also guides us when designing drugs.

The three dimensional description for human serum albumin [Carter et al., 1989; He and Carter, 1992] was established as a result of such crystallisation work. Human serum albumin carries important drugs such as aspirin in the blood and its structural information will allow for further and enhanced drug creations and especially for those drugs that need to be transported via the blood. Creation of the three dimensional model of canavalin [McPherson and Rich, 1973; McPherson and Spencer, 1975], a protein found in the jack bean and a major source of protein for humans and domestic animals, will allow us to produce plants that are more nutritional and resistant to pest. Furthermore, protein crystal growth research is applied to learn how elastase damages lung tissue while general research towards the medical treatment for cancer, AIDS, diabetes, sickle cell anaemia and rheumatoid arthritis is supported.

However, a precondition for this kind of work is that protein crystals are obtained but the crystal growth process is a limiting step. To obtain a crystal long periods of experimental trials are encountered due to long periods of nucleation and growth [Blundell and Johnson, 1976; McPherson, 1982], which might last over months. For the enzyme lysozyme crystallisation periods of well above a month are encountered when e.g. hexagonal crystals are produced [communication with Pusey, 1997]. A method to model or even predict crystallisation conditions and therefore a model that describes protein solubility, is therefore also of great interest in this research area.

This work aims at a model for protein solubility. Models that describe protein solubility are few. The most well known empirical description of protein solubility with respect to salts as precipitant agents was introduced by Cohn in 1925 and has been extensively used and studied by others till today [Green, 1931; Dixon and Webb, 1961; Foster et al., 1971; Niktari et al., 1989]. In 1977 a theoretically based model followed. Melander and Horvath (1977) described protein solubility as a function of salt concentration by relating the solubility behaviour to hydrophobic effects. But Przybycien and Bailey (1989) showed that this model was not consistent with their experimental solubility results. Further theoretical models were recently presented by Chiew et al. (1995) and Kuehner et al. (1996), who used molecular thermodynamics to describe salt induced protein precipitation. The models incorporate different interaction potentials to represent protein solubility and they succeeded in finding a semi - quantitative agreement with their experimental results.

However, a solubility model is needed that computes quantitatively correctly the protein solid - liquid equilibrium for a variety of protein systems and has a predictive potential. Therefore, a semi - empirical model was approached, which had been used with slight variations for the modelling of the solid - liquid equilibrium of organic [Gmehling et al., 1978; Nass, 1988; Peres and Marcedo, 1994] and inorganic compounds [Nicolaisen et al., 1993] but not for proteins. The theoretical framework adapted for our systems of interest, protein - salt - water, uses the solubility product to represent the equilibrium of liquid and crystal protein, while a protein's deviation from ideal solution behaviour is also accounted for.

Proteins have been demonstrated to show high deviations from ideal solution behaviour [Ross and Minton, 1977], which is expected due to the size difference encountered between e.g. the protein and solvent molecule, i.e., water. Furthermore, various different interactions occur for protein molecules, salt ions and water molecules and create deviation from ideal solution behaviour. Therefore, the modelling of protein activity coefficients representing the deviation from ideal solution behaviour due to size difference and interactions was studied here. The only previous studies aiming at protein activity coefficients are those of Ross and Minton (1977), who studied sickle cell anaemia by means of activity coefficient examination, and Wills et al. (1993), who examined the excluded volume contributions for proteins by interpreting the deduced virial coefficients which also led to activity coefficient studies. Both teams based their work on the virial expansion, which gives protein activity coefficients as a function of system composition at constant temperature and pH. For this work another activity coefficient model was aimed at as not only activity coefficients over system composition but also over system temperature had to be expressed while additionally a predictive potential of the model was an objective.

Activity coefficient models have been derived from Wilsons concept [Wilson, 1964] to describe solution behaviour by means of local composition, which was developed from molecular thermodynamic theories [Smith and van Ness, 1987]. For a solution, local compositions, different from the overall mixture composition, are assumed, accounting for the short - range order and non - random orientation that result from differences in molecular size and intermolecular forces. Three local composition models are notable: the Wilson equation [Wilson, 1964], the Non - Random - Two - Liquid equation (NRTL) [Renon and Prausnitz, 1968] and UNIversal QUAsi - Chemical model (UNIQUAC) [Abrams and Prausnitz, 1975].

The third local composition model, UNIQUAC model, was chosen for this work as it uses a minimal number of parameters compared to the other models and because the

UNIQUAC Functional - group Activity Coefficient model (UNIFAC) [Fredenslund et al., 1975] was developed from it. The UNIFAC model is used to calculate activity coefficients from contributions of defined structural molecular groups, which make up the molecules in solution. This method creates a predictive tool not only over system conditions as given with the UNIQUAC model but also over different but related systems, i.e., compounds. The UNIQUAC and UNIFAC methods were additionally chosen because they describe system composition and temperature. While proteins had never been studied with local composition models, systems, mainly hydrocarbons, that bear similarities with the protein - salt ions - water systems studied here, were examined and have been described with some success using these approaches.

Various small organic compounds, which share some of the polar properties with proteins, were described via local composition models including compounds such as amino acids and peptides [Nass, 1988; Gupta and Heidemann, 1990; Phino et al., 1994; Peres and Marcedo, 1994], which are the structural building blocks of proteins. Furthermore, polymers which share size and hydrophobic features with proteins have been examined with UNIQUAC and UNIFAC related models [Kontogeorgis et al., 1997, and references within], where the description of the solution behaviour focused on athermal approaches. Moreover electrolytes, which have ionic charges in common with proteins, were modelled describing activity coefficients and solubility [Nicolaisen et al., 1993] using additionally the Debye Hückel law [Debye and Hückel, 1923; Hückel, 1925, Debye, 1927] to describe the long range interactions of ions at low salt concentrations. This modelling approach for simple electrolytes is also of interest as salt ions are compounds occurring in the systems studied here. Following these facts the UNIQUAC and UNIFAC models and their different versions were studied here for protein containing systems.

In this work first of all amino acid and peptide systems, which are the building blocks of proteins, were studied. In a systematic manner different UNIFAC and UNIQUAC models were applied to model activity coefficients for amino acids and peptides. This part of the work was used, to clarify if the predictive UNIFAC model is applicable for protein related compounds and to establish which model version would be appropriate if at all for proteins. This first part of this work is presented in chapter three.

In the second part of this work protein activity coefficients were addressed. A method to obtain protein activity coefficients from osmotic pressure measurements was studied and the determined activity coefficient data was modelled using a local composition model. Systems consisting of a protein, salt ions and water were examined. The model performance for protein activity coefficients over salt concentration, pH and temperature was studied, while also a qualitative examination of protein solubility was pursued. This second part of the work is documented in chapter four.

In the third and final part of this work the directly measurable solution property, solubility, was investigated. Protein solubility over salt concentration and temperature was approached and described via the modelling of protein activity coefficients. This part of the research is presented in chapter five.

Results and discussions are presented and summarised for these three parts. The relevant theoretical background for this work is presented in the next chapter, chapter two. Final chapters discuss the overall conclusions and future research projects, chapters six and seven, respectively.

Chapter 2

Theoretical Background

This chapter discusses proteins and their characteristics. Furthermore, it introduces the major definitions and models that were studied in this work.

2.1 Introduction

The modelling of a protein solution property, protein solubility, was aimed at in this work using activity coefficients. Consequently first of all proteins and their characteristics are discussed. This is followed by the definitions for activity coefficients and their two different standard states, the symmetric and unsymmetric convention. Additionally, the activity coefficient definition with respect to the excess

Gibbs energy is introduced leading to the UNIQUAC [Abrams and Prausnitz, 1975] and UNIFAC [Fredenslund et al., 1975] expressions, which are models used to describe activity coefficients and solution behaviour as a function of system composition and temperature. The UNIQUAC and UNIFAC models are introduced in detail, including the different UNIQUAC and UNIFAC versions examined in this work. The presented theories are found to some extent in various references [Prausnitz et al., 1986; Smith and van Ness, 1987; Nicolaisen, 1994]. The discussion relating to proteins is found in references such as Creighton (1984) and Lehninger (1982).

2.2 Characterisation of Proteins

Proteins are macromolecules built from twenty natural occurring amino acids. Table 2.2.1 shows the structures of the different amino acid residues that build a particular amino acid and are also found in proteins as amino acids are the building blocks of proteins. These amino acid residues, R_{aa}, are situated on the following structure to configure an amino acid:

$$H_2N\text{-}CH(R_{aa})\text{-}COOH \qquad (2.2.1)$$

Table 2.2.1: Molecular structure and characterisation of the twenty amino acid residues

Amino acid	Property	Residue	Ionised form
Glycine	non - polar	$-H$	-
Alanine	non - polar	$-CH_3$	-
Valine	non - polar	$-CH(CH_3)_2$	-
Leucine	non - polar	$-CH_2-CH(CH_3)_2$	-
Isoleucine	non - polar	$-CH(CH_3)(CH_2(CH_3))$	-
Phenylalanine	non - polar	$-CH_2-C:-CH:-CH$ $\vert:\ o\ \vert:$ $(CH)_2\ :-CH$	-
Proline	non - polar	**HN-CH-COOH** $\vert\ \vert$ **$(CH_2)_3$**	-
Tryptophan	non - polar	$-CH_2-C=CH-\ NH$ $\vert:\quad\vert:$ $C\ :-\ C$ $\vert:\ o\ \vert:$ $(CH)_2:-(CH)_2$	-
Serine	hydroxyl	$-CH_2-OH$	-
Threonine	hydroxyl	$-CH(OH)(CH_3)$	-
Aspartic acid	acidic	$-CH_2-C(=O)(OH)$	$-CH_2-C^-(:-O)(:-O)$
Glutamic acid	acidic	$-(CH_2)_2-C(=O)(OH)$	$-(CH_2)_2-C^-(:-O)(:-O)$
Asparagine	amido	$-CH_2-C(-O)-NH_2$	-
Glutamine	amido	$-(CH_2)_2-C(=O)-NH_2$	-
Tyrosine	basic	$-CH_2-C:-CH:-CH$ $\vert:\ o\ \vert:$ $(CH)_2\ :-CH-OH$	$-CH_2-\ C:-CH:-CH$ $\vert:\ o\ \vert:$ $(CH)_2:-CH-O^-$
Lysine	basic	$-(CH_2)_4-NH_2$	$-(CH_2)_4-N^+H_3$
Arginine	basic	$-(CH_2)_3-NH-C(=NH)(NH_2)$	$-(CH_2)_3-NH-C^+(:-NH_2)(:-NH_2)$
Histidine	basic	$-CH_2-C=CH-N$ $\vert\quad\vert\vert$ $NH\ -CH$	$-CH_2-C=CH-NH$ $\vert\quad\vert:$ $NH\ :-C^+H$
Cysteine	disulphide bonds	$-CH_2-SH$	$-CH_2-S^-$
Methionine	sulphur	$-CH_2-CH_2-S-CH_3$	-

:- partial double bounds, o aromatic

For proline the whole structure is given in table 2.2.1 because proline is an imino acid as opposed to an amino acid.

A protein is created from a number of amino acids which bind under a condensation reaction resulting in a sequential arrangement:

$$(\text{-NH-CH}(R_{aa})\text{-CO-})_x \qquad (2.2.2)$$

Following these reactions a peptide or protein is created, which consists of the backbone as given above, 2.2.2. This backbone is build of units consisting of a planar peptide bond (-CO-NH-) and a carbon, the α - carbon, from which the particular amino acid residue sets off. Following the condensation reactions a chain molecule is created and its left end is defined as the amino terminus:

$$H_2N\text{-CH}(R_{aa})\text{-CO-}\ldots \qquad (2.2.3)$$

while on the other end the carboxyl terminus is found:

$$\ldots\text{-NH-CH}(R_{aa})\text{-COOH} \qquad (2.2.4)$$

The twenty amino acids that build proteins contribute to the characteristics of proteins. Therefore the twenty amino acids and their characteristics are discussed next. However, amino acids do not describe matters of molecule size which have to be additionally considered for proteins. The amino acids or amino acid residues are grouped according to their polarity, their charge, their steric flexibility etc.

Some amino acids residues have been demonstrated to accept and release protons and the ionised forms of these residues are given in table 2.2.1 while the acid base reaction is documented in table 2.2.2. For six residues ionisation occurs depending on the environment. Aspartic acid and glutamic acid release protons in the acidic pH region and are referred to as acidic. Argine, lysine, tyrosine, cystein and histidine (weaker base) are basic. Moreover, the termini are involved in acid base reactions. The documented ranges for the pK_a values for the different ionised groups indicate in which pH region ionisation occurs e.g. at a pH of five the carboxyl terminus, aspartic acid and glutamic acid are found predominantly in their base form while all other residues are predominantly in their acid form. Following these ionisation reactions proteins are found to be charged molecules, i.e., polyelectrolytes. In strongly acid solution proteins are positively charged and in strongly alkaline solution they are negatively charged. Due to this property proteins migrate in an electric field. At a net charge of zero no migration occurs and the prevailing pH is defined as a protein's characteristic pI (isoelectric point).

Following the formation of ionised groups a strong hydrophilic character is observed while the aliphatic residues participate in hydrophobic interactions. Residues such as glycine, alanine, valine, leucine and isoleucine, table 2.2.1, exhibit a hydrophobic character due to their aliphatic molecule structure. Similar to these, methionine

Table 2.2.2: Ionised amino acid groups and their pK_a and reaction

Group	pK_a*	Reaction: acid → base + H^+
carboxyl terminus	3.5 - 4.3	$-COOH \rightarrow -COO^- + H^+$
Aspartic acid	3.9 - 4.0	$-COOH \rightarrow -COO^- + H^+$
Glutamic acid	4.3 - 4.5	$-COOH \rightarrow -COO^- + H^+$
Histidine	6.0 - 7.0	$-NH_2{:}-C^+{:}-NH_2^- \rightarrow -NH{=}C-NH_2^- + H^+$
amino terminus	6.8 - 8.0	$-H_3N^+ \rightarrow -NH_2 +H^+$
Cystein	9.0 - 9.5	$-SH \rightarrow -S^- + H^+$
Tyrosine	10.0 - 10.3	$-OH \rightarrow -O^- + H^+$
Lysine	10.4 - 11.1	$-H_3N^+ \rightarrow -NH_2 + H^+$
Arginine	12.0	$-C^+({:}-NH_2)({:}-NH_2) \rightarrow -C(-NH_2)(=NH) + H^+$

* Creighton, 1984

exhibits a non - polar and relatively non - reactive character. The amino acid residues serine, threonine and tyrosine have hydroxyl groups and support hydrogen bonds. Asparagine and glutamine, build hydrogen bonds due to their amido groups.

Looking at steric flexibility it is found that glycine due to its small residue is highly flexible and is often found in the bends of proteins while the steric hindered proline limits protein flexibility. Aromatic groups found in phenylalanine, tyrosine and tryptophan contribute to hydrophobic behaviour and steric hindrance. For cystein a very typical characteristic is the creation of disulphide bonds, which are responsible for strong intra- and interstrand interactions.

The interactions, which contribute mainly to the specific characteristics and structure of a protein are the covalent bindings. Covalent linkage of atoms is established by sharing of electrons. With covalent bonds such as the peptide bonds and the disulphide bonds the major characteristics of a protein's structure and therefore characteristics are settled. An example are disulphide bindings, which lead with increasing number to a decrease of a protein's solubility.

London dispersion forces describe short range non - covalent interactions, which also contribute to a protein's properties. Between all atoms weak interactions occur due to the distribution of electronic charges and their location. Their movements create fluctuating dipoles which result in attractive forces between atoms. Repulsion occurs when the electron orbitals are forced to overlap.

Interactions that are closely related to the dispersion forces are those of permanent dipoles. These can lead to hydrogen bonding. Hydrogen bonding occurs for proteins between hydrogen atoms and nitrogen atoms or hydrogen atoms and oxygen atoms. These atoms are involved in dipole moments due to great differences in the electron density of covalent bonds. Where opposite dipole charges are found attraction occurs, otherwise repulsion is found.

The ionic bonds, which are established between differently charged groups tend to bring together parts over large distances while repulsion occurs between ionic groups of same charge. Ionic bonds are the strongest non - covalent forces but do depend on the dielectric constant of the solvent as electrostatic interactions are reduced in solvents with high dielectric constants such as water.

When dealing with proteins another form of interaction has to be considered. Hydrophobic interactions describe the behaviour of groups not being attracted by water or polar substances. While polar molecule groups form hydrogen bonds with water, non - polar groups are not soluble and tend to form hydrophobic clusters minimising the contact area with polar solvents e.g. water, which is the solvent proteins are most commonly exposed to. The hydrophobic parts of a protein are accumulated in the centre while the hydrophilic parts are exposed at the surface to interact with the polar solvent. Therefore, a compact and dense structure is observed for some proteins and those are referred to as globular proteins. In this work globular proteins were studied.

Furthermore, aggregation of protein monomers occurs. Dimers or higher order aggregates are held together by non - covalent forces as described above. During the course of this work two proteins that occur as dimers were examined for the protein activity coefficient and solubility modelling work, β - lactoglobulin and concanavalin A.

The described properties and interactions, which result from the molecular structure, imply the solution behaviour that might occur between proteins and other system compounds such as water or salt ions. Moreover, the interactions within a protein are pictured which lead to a protein's very specific three dimensional structure unlike the random coiling observed for aliphatic polymers.

The described structures, interactions and properties of proteins lead to certain classifications. The most common classification is achieved by means of molecular weights which divides the group of proteins into two groups. The low molecular weight proteins, which are built from only a few amino acids, are referred to as peptides while the higher molecular weight ones are called proteins. These two classes are defined differently with respect to the reference. One definition uses the number of peptide bonds where a polypeptide with more than fifteen peptide bonds is considered a protein [Freifelder, 1986]. Other references are found, which give a definition according to the number of amino acids incorporated or the molecular weight. Typically proteins have molecular weights of 5000 Da to 200000 Da [Lehninger, 1982] and the average molecular weight of the amino acid, which is incorporated into a protein and constitutes the building block of a protein, is about 120 Da [Creighton, 1984].

Furthermore, proteins are classified according to their composition [Elmore, 1986]. Unconjugated and conjugated proteins are referred to. The first class is composed of only amino acids as discussed here, while the second one has additional molecular structures which are not amino acids. The group of conjugate proteins is divided into nucleoproteins, lipoproteins, glycoproteins, chromoproteins in accordance to their prosthetic groups.

Moreover, certain properties are used to define proteins. A protein's solubility behaviour allows for yet another categorisation. Albumins (soluble in water and dilute solutions of salt), globulins (few soluble in water but soluble in dilute salt solutions), prolamines (insoluble in water but soluble in 50 - 90 % aqueous ethanol), glutelins (insoluble in the mentioned solvents but soluble in dilute solutions of acids or bases) and scleroproteins (insoluble in most solvents) were defined over the years.

Another widely applied method of classification was established using a protein's biological function. Such categories are hormones, enzymes, antibodies, structural proteins etc. Since the first creation of recombinant proteins at around 1972 another classification evolved differentiating between recombinant and non - recombinant proteins. Furthermore, classification due to molecular size, amino acid composition, conformation (helix content), origin and many other factors are known.

In this work six different proteins, serum albumin (horse blood), α - chymotrypsin (bovine pancreas), β - lactoglobulin (bovine milk), ovalbumin (chicken egg), lysozyme (chicken egg) and concanavalin A (jack bean), were studied. These proteins originate from a variety of sources as indicated above and have very different biological functions. The protein α - chymotrypsin is a protease and its biological function is to hydrolyse peptide bonds of proteins. The protein lysozyme, which is the smallest protein (14600 Da) studied during the course of this work, cleaves bacterial cell walls. Lysozyme therefore prevents infections while it is also applied on the biochemical processing scale as a biochemical tool to disrupt bacterial cells. The protein serum albumin again has a very different function. It occurs in the blood and regulates its osmotic pressure while it also is a transport vehicle that delivers other compounds and drugs through the blood stream. Not only the biological function of the studied proteins differs but also the examined system conditions differ e.g. pH and temperature ranges. However, all these proteins have in common that they are globular proteins, i.e., of a compact structure. Furthermore, they all originate from natural sources that produce them at high amounts, which makes them available at the quantities and qualities needed for biophysical studies. Therefore, these six proteins could be studied here in the first place. The two proteins, β - lactoglobulin and concanavalin A, were studied to represent proteins that are dimers. The protein concanavalin A is the biggest protein studied here with a molecular weight 102668 Da .

2.3 Activity Coefficients and Standard States

To model a protein's solution properties such as solubility the activity coefficients have to be addressed. A description of a solution property has to consider the deviation from ideal solution behaviour which is expressed through activity coefficients. For protein containing systems deviation from ideal solution behaviour is expected as demonstrated by Ross and Minton (1977) for haemoglobin.

Activity coefficients are used to describe a compound's deviation from ideal behaviour while referring to ideal solution behaviour. Two means of referring to the ideal solution behaviour are known and the fugacity, f, is used to define these. The fugacity of a compound, i, in solution is represented by a product of activity, a, and the reference fugacity:

$$f_i = a_i \cdot f_i^{ref} \qquad (2.3.1)$$

The reference fugacity, f^{ref}, is defined independently for different solution compounds and two reference states are used: Raoult's law behaviour and Henry's law behaviour. For Raoult's law behaviour $f_i^{ref} = f_i(T, P, \text{pure i})$ refers to the fact that the partial pressure, p, is proportional to a compound's vapour pressure, P^S:

$$p_i = x_i \cdot P_i^S \qquad (2.3.2)$$

where x is the mole fraction. The reference fugacity for Henry's law behaviour is $f_i^{ref} = H_i$ and the Henry constant, H, is proportional to the partial pressure:

$$p_i = x_i \cdot H_i \qquad (2.3.3)$$

Both laws refer to ideal solution behaviour and were found to describe the ideal solution behaviour of different systems satisfactorily, e.g. benzene - toluene is described by Raoult's law [Laidler and Meiser, 1982]. However, ideal solution behaviour is only observed for mixtures of similar compounds, which are of the same molecular size and exhibit the same interactions. Other systems exhibit deviation from ideal solution behaviour, which has to be accounted for with reference to ideal behaviour.

For a compound for which the standard state is the pure liquid, the activity, a, in the standard state is equal to unity. In these cases the standard fugacity, f°, is equal to:

$$f_i^\circ = f_i^{ref} = P_i^S \qquad (2.3.4 - 5)$$

It follows for the chemical potential, μ, that:

$$\mu_i(T,P) = \mu_i^\circ(T,P) + R \cdot T \cdot \ln(a_i) \qquad (2.3.6)$$

where $a_i \rightarrow 1$ and $\mu_i \rightarrow \mu_i^\circ$ as $x_i \rightarrow 1$ which refers to Raoult's law and introduces the symmetric reference system. For Henry's law a hypothetical standard state is created where the compound's activity is equal to one in pure solvent, which introduces the unsymmetric reference state with $a_i \rightarrow 1$ as $x_i \rightarrow 0$.

The activity is defined as the product of a compound's concentration and activity coefficient, γ. Different concentration scales lead to different activity and activity coefficient expressions:

$$
\begin{aligned}
a_{i(x)} &= x_i \cdot \gamma_{i(x)} \\
a_{i(m)} &= m_i \cdot \gamma_{i(m)} \qquad (2.3.7 - 9) \\
a_{i(c)} &= c_i \cdot \gamma_{i(c)}
\end{aligned}
$$

where x is the mole fraction, m the molal and c the molar concentration scale. Deviations from ideal solution behaviour are described by the activity coefficients, which are a function of the specified standard state and the concentration scale. The activity coefficients are denoted differently with respect to the standard state chosen. For Henry's law behaviour, i.e., the unsymmetric convention, a special notation is made: *. Otherwise Raoult's law behaviour, i.e., the symmetric convention, is applied. The explicit activity coefficient expressions studied in this work are introduced in the following chapters and the symmetric and unsymmetric activity coefficient equations are presented.

2.4. The Original UNIQUAC Model

The original UNIQUAC model [Abrams and Prausnitz, 1975] links the microscopic and the macroscopic solution scale, creating a thermodynamic framework to describe solution activity coefficients of pure or mixture systems over temperature. The three dimensional arrangements and interactions of molecules as represented by lattice theory (microscopic level) relate to the excess Gibbs energy (macroscopic level). By means of the molar excess Gibbs energy, g^E, deviation from ideal solution behaviour is described:

$$\ln\gamma_i = \left[\frac{\partial(n_{tot}\cdot g^E / R\cdot T)}{\partial n_i}\right]_{P,T,n_j(j\neq i)} \qquad (2.4.1)$$

The activity coefficient, γ, of a compound, i, is obtained by differentiating an expression for the molar excess Gibbs energy with respect to that compound's moles, n, while temperature, T, pressure, P, and the moles of other system compounds, j, are constant.

Exploiting the relationship given in equation 2.4.1, and aiming at the description of phase equilibrium and furthermore equilibrium and compound properties, an excess Gibbs energy expression was introduced by Abrams and Prausnitz (1975). They expressed the molar excess Gibbs energy as two additive terms:

$$\frac{g^E}{R\cdot T} = \frac{g_C^E}{R\cdot T} + \frac{g_R^E}{R\cdot T} \qquad (2.4.2)$$

The two terms contributing to the molar excess Gibbs energy are the combinatorial term, g_C^E, and the residual term, g_R^E.

The combinatorial term or entropy term for multi - component mixtures is given by:

$$\frac{g_C^E}{R\cdot T} = \sum_1^i x_i \cdot \ln\frac{\Phi_i}{x_i} + \frac{z}{2}\cdot\sum_1^i q_i \cdot x_i \cdot \ln\left(\frac{\Phi_i}{\theta_i}\right)^{-1} \qquad (2.4.3)$$

where

$$\Phi_i = \frac{r_i \cdot x_i}{\sum_1^j r_j \cdot x_j} \quad \text{and} \quad \theta_i = \frac{q_i \cdot x_i}{\sum_1^j q_j \cdot x_j} \qquad (2.4.4 - 5)$$

The combinatorial term accounts for deviation from ideal solution behaviour due to differences in size and shape of the mixture compounds. The mole fraction, x, refers to the mixture composition for all compounds, i, while the parameters Φ and θ represent the volume and surface fraction of the different compounds, respectively. The structural parameters r and q refer to the volume and surface area of each compound, i.e., molecule, and z indicates the number of nearest neighbours, i.e., the co - ordination number. The value, z, is generally set equal to 10 but ranges from 6 - 12 representing the unit cell configuration of the lattice, e.g. z = 6 for the regular cubic lattice and z = 12 for the hexagonal lattice [Tanford, 1961]. To compute the combinatorial contribution the structural parameters, r and q, have to be determined for each system compound. These are derived from a compound's van der Waals volume, V_w, and surface area, A_w, while being normalised with respect to the methylene molecular group in polyethylene, which was an arbitrary choice by Abrams and Prausnitz (1975):

$$r = \frac{V_W}{15.17} \qquad (2.4.6)$$

$$q = \frac{A_W}{2.5 10^9} \qquad (2.4.7)$$

To obtain a compound's van der Waals volume and surface area the group contribution method by Bondi (1968) is generally applied. However, depending on the compound Bondi's method might not be adequate and ways to obtain this data for proteins were developed. These are described when relevant.

The residual term or enthalpy term is given by:

$$\frac{g_R^E}{R \cdot T} = -\sum_1^i q_i \cdot x_i \cdot \ln\left(\sum_1^j \theta_j \cdot \psi_{ji}\right) \qquad (2.4.8)$$

$$\psi_{ji} = \exp\left(-\left[\frac{u_{ji} - u_{ii}}{T}\right]\right) \qquad (2.4.9)$$

With the residual term short - range interactions of a centre molecule with its ten (z) surrounding next neighbours are introduced using binary interaction parameters, u. Interaction parameters describe the sum of interactions between a nearest neighbour and a centre molecule over the various binary interactions occurring per compound pair. The interactions between identical and different molecule pairs are described by a number of binary interaction parameters. Three interaction parameters describe for two compounds, i and j, the binary interactions between identical, u_{ii} and u_{jj}, and different, u_{ij} with $u_{ij} = u_{ji}$, compound types. The total number of binary interaction parameters, b, needed per system is a function of the number of system compounds, i, and is given by:

$$b = (i) + (i-1) + ... + (i-i) \qquad (2.4.10)$$

These interaction parameters are available from previous modelling attempts for some compounds. However, they are a function of the studied model version and might therefore not apply. Following these facts the interaction parameters have to be established for some compounds and this was the case for this work as proteins had not been studied before. A programming procedure was designed and FORTRAN programs were developed during the course of this work in order to obtain these parameters. These procedures are explained where relevant, while the programs are printed in the appendices.

Partial molar differentiation of the given g^E expressions leads to the activity coefficient, γ, for a compound, i, over the two possible reference states, the symmetric and unsymmetric reference states:

$$\ln\gamma_i = \ln\gamma_i^C + \ln\gamma_i^R \qquad (2.4.11)$$

$$\ln\gamma_i^* = \ln\gamma_i^{*,C} + \ln\gamma_i^{*,R} \qquad (2.4.12)$$

where γ^C is the combinatorial and γ^R is the residual activity coefficient. In table 2.4.1 the resulting expressions are given with respect to the two terms, combinatorial and residual term, and the two reference states, symmetric and unsymmetric conventions. The unsymmetric activity coefficient, γ*, is obtained from the symmetric activity coefficient, γ, using:

Table 2.4.1: The combinatorial term and residual term expressions for activity coefficient calculations using the original UNIQUAC model

Combinatorial term:

$$\ln\gamma_i^{C} = \ln\left(\frac{\Phi_i}{x_i}\right) + 1 - \frac{\Phi_i}{x_i} - \frac{z}{2}\cdot q_i \cdot \left[\ln\left(\frac{\Phi_i}{\theta_i}\right) + 1 - \frac{\Phi_i}{\theta_i}\right]$$

$$\ln\gamma_i^{*,C} = \ln\left(\frac{\Phi_i}{x_i}\right) - \frac{\Phi_i}{x_i} - \ln\left(\frac{r_i}{r_{Sol.}}\right) + \frac{r_i}{r_{Sol.}}$$
$$-\frac{z}{2}\cdot q\cdot_i \left[\ln\left(\frac{\Phi_i}{\theta_i}\right) - \frac{\Phi_i}{\theta_i} - \ln\left(\frac{r_i\cdot q_{Sol.}}{r_{Sol.}\cdot q_i}\right) + \frac{r_i\cdot q_{Sol.}}{r_{Sol.}\cdot q_i}\right]$$

Residual term:

$$\ln\gamma_i^{R} = q_i \cdot \left[1 - \ln\left(\sum_{1}^{k}\theta_k\cdot\psi_{ki}\right) - \sum_{1}^{k}\frac{\theta_k\cdot\psi_{ik}}{\sum_{1}^{j}\theta_j\cdot\psi_{jk}}\right]$$

$$\ln\gamma_i^{*,R} = q_i \cdot \left[-\ln\left(\sum_{1}^{k}\theta_k\cdot\psi_{ki}\right) - \sum_{1}^{k}\frac{\theta_k\cdot\psi_{ik}}{\sum_{1}^{j}\theta_j\cdot\psi_{jk}} + \ln\psi_{Sol.,i} + \psi_{i,Sol.}\right]$$

Sol.: solvent

$$\gamma^{*} = \frac{\gamma}{\gamma^{\infty}} \qquad (2.4.13)$$

where γ^{∞} defines the infinite dilution activity coefficient of a compound in pure solvent.

2.5 The Original UNIFAC Model

Different group contribution models have been developed [Lydersen, 1955; Hoshino et al., 1982; Klincewicz and Reid, 1984; Mavrovouniotis et al., 1988; Mavrovouniotis, 1990; Elbro et al., 1991; Constantinou and Gani, 1994] and demonstrated that compound properties relate not only to the molecule type but also to the molecular groups, which build a molecule. A great variety of compounds, mainly hydrocarbons [Lyman et al., 1982; Reid et al., 1987], were represented by a limited number of molecular groups and these were successfully correlated to a compound's molecular structure and properties, either pure or mixture properties. These model types, group contribution models, are able to describe properties not only of previously modelled compounds or systems but also new and unexamined systems as the molecular groups allow the creation of any kind of compound and therefore system. This makes these types of models more widely applicable than other model types. However, extensive and good experimental data over various conditions and systems has to be available to develop satisfactory group contribution models, which should make the models applicable for systems and system conditions related to the originally used systems, i.e., systems of the database used to train the model and to obtain the needed parameters.

Fredenslund et al. (1975) integrated such a group contribution concept into the previously presented UNIQUAC model, chapter 2.4, and created the UNIFAC model. The group contribution approach of the UNIFAC model uses also molecular groups, which combine to build a compound and lead to that compound's properties using the properties of the molecular groups that build that compound. For calculations with the UNIFAC model, as in the case of the UNIQUAC model, the structural parameters, r and q, of a compound, i, are needed. In the case of the UNIFAC model these are obtained from the structural parameters, R and Q, of the defined molecular groups, p, as a function of their occurrence, v, in that same compound:

$$r_i = \sum_{1}^{p} v_p^i \cdot \mathrm{R}_p \qquad (2.5.1)$$

$$q_i = \sum_{1}^{p} v_p^i \cdot Q_p \qquad (2.5.2)$$

Two sets of structural parameters are introduced. One for the compound, r and q, and one set for the molecular groups, R and Q. Both sets of structural parameters relate to the van der Waals volume and surface area of either a compound or a molecular group as discussed before. By means of these molecular groups and their parameters the group contribution approach was introduced into the UNIFAC model. For the combinatorial term of the UNIFAC model the same equations apply as for the UNIQUAC model, equations 2.4.3 - 2.4.5, and the same theory applies as described before, chapter 2.4.

A different set of equations was derived for the residual term of the UNIFAC model. The interaction parameters of the molecular groups, Γ, were integrated into the residual term, which leads to the residual activity coefficient, γ^R:

$$\ln\gamma_i^R = \sum_1^p v_p^i \cdot \left(\ln\Gamma_p - \ln\Gamma_p^i\right) \qquad (2.5.3)$$

where Γ_p is the activity coefficient of a molecular group and Γ^i_p is the residual activity coefficient of group p in a reference solution containing only molecules of type i while v refers to the number a group occurs per compound. The activity coefficient of a molecular group is given by:

$$\ln\Gamma_p = Q_p \cdot \left[1 - \ln\left(\sum_1^p \Theta_p \cdot \Psi_{pq}\right) - \sum_1^p \frac{\Theta_p \cdot \Psi_{qp}}{\sum_1^l \Theta_l \cdot \Psi_{qp}}\right] \qquad (2.5.4)$$

which also applies for calculations of the reference activity coefficient. The variable Θ is the area fraction of a molecular group and is calculated by:

$$\Theta_p = \frac{Q_p \cdot X_p}{\sum_1^q Q_q \cdot X_q} \qquad (2.5.5)$$

where X is the mole fraction of a molecular group. The interaction parameter, ψ, is given by:

$$\Psi_{pq} = \exp\left(-\left[\frac{U_{pq} - U_{qq}}{T} \right] \right) \qquad (2.5.6)$$

where U is a binary interaction parameter for molecular groups and T is the system temperature. The UNIFAC model takes the same basic approach as the UNIQUAC model and its parameters and variables bear the same physical meaning as presented in the previous chapter 2.4. However, for the UNIFAC model molecular groups, which configure the different system compounds, are used as opposed to a description of the compounds as an entity, which is done with the UNIQUAC model.

The molecular groups that are used to build a compound have been defined and are found in the most extensive and recent publication by Hansen et al. (1991), which also aims at biotechnology compounds and is therefore of particular interest to this work. Hansen et al. (1991) defined the molecular groups and related almost all of these groups to the structural parameters, R and Q, and interaction parameters, U, which are a function of the database, i.e., model training systems, applied by Hansen et al. (1991).

However, reconstruction of a system's compounds using the defined molecular groups allows for activity coefficient predictions applying the determined parameters and introduced UNIFAC equations.

2.6 The Modified Model

In 1987 a modified model [Larsen et al., 1987] was developed. The combinatorial term was changed compared to the original one and was since widely used for polymers [Kontogeorgis et al., 1997, and references within]. Additionally, a temperature dependence was introduced by Larsen et al. (1987) for the residual term and its interaction parameters. These alterations improved excess enthalpy estimations from deviation of around 50 % to around 10 % [Fredenslund, 1989]. Consequently, modelling of systems that span a certain temperature range can be improved using the modified residual term. Therefore, the modified residual term was studied for this work when temperature ranges of up to 30 K had to be described for protein solubility.

For the modified model the combinatorial activity coefficients, γ^{C}, of a compound, i, is described by:

$$\ln \gamma_i^C = \ln\left(\frac{\omega_i}{x_i}\right) + 1 - \frac{\omega_i}{x_i} \qquad (2.6.1)$$

where

$$\omega_i = \frac{x_i \cdot r_i^{\frac{2}{3}}}{\sum_{1}^{j} x_j \cdot r_j^{\frac{2}{3}}} \qquad (2.6.2)$$

The modified combinatorial term is a function of the volume fraction, ω, and the mole fraction, x, and resembles closely the original combinatorial term, equation 2.4.3 and table 2.4.1. However, the second part of the original combinatorial term is eliminated. Due to this the entropy state of solution is described only by means of compound sizes and reflects the originally suggested expression of Flory - Huggins [Larsen et al., 1987].

The volume fraction of the modified model, equation 2.6.2, is determined from the mole fractions of compounds, i and j, and relates closely to the definition given for the original model, equation 2.4.4. However, the volume parameter, r, was assigned a smaller contribution as examined and discussed by Kikic et al. (1980). The introduced exponent, which is generally and for this work set equal to 2/3, lowers the volume contribution. The value of the exponent might vary with the type of molecule. The structural volume parameter, r, is obtained in the manner described before, equations 2.4.6 or 2.5.1, and represents a compound's van der Waals volume as discussed before.

The residual term of the modified model is derived in the same manner as for the original model, chapter 2.4, but a higher order temperature dependence is added for the binary interaction parameters, $u_{i,j}$. This results in up to three parameters per binary interaction parameter instead of one when compared to the original model:

$$u_{ij} = u_{ij}^{\circ} + u_{ij}^{t} \cdot \left(T - T_o\right) + u_{ij}^{t'} \cdot \left(T \cdot \ln\frac{T_o}{T} + T - T_o\right) \qquad (2.6.3)$$

However, in this work the temperature, T, dependence was terminated after the second term in order to introduce as few parameters as necessary into the solubility model. The temperature, T_o, is an arbitrary reference temperature which is fixed to 300 K for this work.

2.7 Appraisal of the Models

The basic model concepts introduced in this work are those of the UNIFAC and UNIQUAC models. Both models refer to the same theoretical approach but vary with the extent of their applicability. The UNIQUAC model allows for predictions over systems conditions such as temperature and composition. The UNIFAC model allows additionally for predictions over different but related compounds as a function of the defined molecular groups.

Both models have been applied to study phase equilibria and compound properties. Pure and mixture compound properties were described for a variety of compounds by various authors [Abrams and Prausnitz, 1975; Fredenslund et al., 1975; Prausnitz et al., 1986; Smith and van Ness, 1987]. The compounds investigated were mainly hydrocarbons and small organic compounds. These relate to some degree to proteins as discussed earlier, chapter 2.2, and therefore these models were approached.

However, some limitations are indicated for the original models as presented in 1975. Compounds that bear an electrolyte character just like proteins, which are polyelectrolytes, were not considered. Likewise, macromolecules such as polymers and proteins were not included in the early examinations. However, attempts were made to overcome these limitations and since then polymers and simple electrolytes were studied with these local composition models while this work is the first dedicated to proteins.

For polymers it was found that a description of the entropy state of solution alone is pursued as hardly any interactions are encountered for polymer mixtures built from aliphatic polymers. Ways to express the entropy state of solution for polymer mixtures were pursued using the Flory - Huggins combinatorial term as introduced with the modified model in the previous chapter. While the original combinatorial term bears also the Flory - Huggins expression, it also carries a Staverman - Guggenheim correction, which in most cases is often quite small [Larsen et al., 1987]. For the modelling of polymer properties also variations for the definition of the volume fraction are noted [Couthinho et al., 1994] and differences in the definition of the volume term and the magnitude of its contribution are introduced. These works are of relevance for protein containing systems due to the fact that they have the macromolecular character in common with polymers. Consequently, they encouraged the study of the original models and their consecutive versions for proteins. However, the original model and its entropy term were addressed first as this term already introduces the size impact of the different compounds into the model as discussed before. While for polymer solutions athermal behaviour is expected the same cannot be assumed for proteins. Having examined the characteristics of proteins in chapter 2.2, it is for certain that interactions occur between proteins and other system compounds. Therefore, the enthalpy state of solution needs to be described for protein containing systems as done in this work.

Referring to the fact that proteins are polyelectrolytes, studies pursued for simple electrolytes were of interest. Various research works have been pursued to deal with electrolyte systems for the UNIQUAC and UNIFAC models since 1986 [Sander et al., 1986, Nicolaisen, 1994]. To represent the long - range interactions that occur between charged groups the Debye Hückel law [Debye and Hückel, 1923; Hückel, 1925, Debye, 1927] was introduced. Interactions between charged groups are observed at low ionic strength in electrolyte solutions. However, high ionic strength solutions do not encounter charge - charge forces as these are screened off. Dealing with protein systems that additionally constitute of buffer and salt ions, high ionic strengths are observed and a screening of the ionic long range interactions is noted [Kuehner et al., 1996]. Therefore, this approach was not adapted for the studied systems and only short - range forces such as dispersion and dipole forces were accounted for.

Furthermore, it was of interest for this work that recently a new database of molecular groups and their parameters [Hansen et al., 1991] was published for the UNIFAC model, which was documented to have been extended to compounds relevant in biotechnology applications. Success of these parameters for the prediction of protein activity coefficients or those of related compounds is certainly of interest for this work. The determination of parameters for the target systems would not be required, which is a work intensive process depending on the number of parameters that need to be determined. For the establishment of e.g. a solubility model about twenty parameters were determined by Nicolaisen (1994) for a system of five compounds. However, following the fact that a new database, which considered also biotechnology compounds, was introduced for the predictive UNIFAC model encouraged the use of the discussed modelling frameworks for biotechnology compounds: proteins.

Consequently, first the UNIFAC model and the new database was examined. The approaches and results of that work are discussed in the next chapter, chapter three. This work is then followed by two further result sections which deal with additional model approach examinations for proteins and lead to the modelling of protein solubility.

2.8 Summary

The characteristics of proteins, major definitions and models for this work were introduced and discussed. However, this introduction was limited to the theory needed throughout this work and for the next chapter, chapter three. Further definitions and models are introduced when relevant.

2.9 Nomenclature

a	activity
A_w	Van der Waals surface area
b	total number of binary interaction parameters
c	molar concentration
Da	dalton [g/mol]
f	fugacity
g_C^E	combinatorial term of the molar excess Gibbs energy
g^E	molar excess Gibbs energy
g_R^E	residual term of the molar excess Gibbs energy
H	Henry's constant
i	compound
m	molal concentration
n	number of moles
p	partial pressure
P	pressure
pI	isoelectric point
pK_a	dissociation constant on the pH scale
P^S	vapour pressure
q	structural surface area parameter of a compound
Q	structural surface area parameter of a group
R	gas constant
r	structural volume parameter of a compound
R	structural volume parameter of a group
R_{aa}	amino acid residue

T	temperature
u	UNIQUAC binary interaction parameters of a compound
U	UNIQUAC binary interaction parameters of a group
UNIFAC model	UNIQUAC Functional Activity Coefficients model
UNIQUAC model	UNIversal QUAsi Chemical model
V_w	Van der Waals volume
x	compound mole fraction
X	group mole fraction
z	lattice co - ordination number

Greek letters

Φ	volume fraction (org.)
Γ	group activity coefficient
Θ	group area fraction
Ψ	group interaction parameter
γ	compound activity coefficient
μ	chemical potential
ν	molecular group occurrence per compound
θ	compound area fraction
ω	volume fraction (mod.)
ψ	molecule interaction parameter

Subscript

^{o}	standard
c	molar concentration
i	compound
j	compound

k	compound
l	molecular group
m	molal concentration
p	molecular group
q	molecular group
tot	total
x	mole fraction

Superscript

∞	infinite dilution
$^\circ$	standard
*	unsymmetric convention
C	combinatorial term
i	compound
R	residual term
ref	reference
t'	temperature (third term)
t	temperature

Chapter 3

Calculation of Amino Acid and Peptide Activity Coefficients

It was the objective of this part to evaluate which model approach describes a solution and mixture property, activity coefficients, of amino acids and peptides.

3.1 Introduction

Aiming at the description of protein solubility by means of protein activity coefficients required first the examination of the various activity coefficient models. The UNIFAC [Fredenslund et al., 1975] and UNIQUAC [Abrams and Prausnitz, 1975] models and their consecutive versions were investigated for their ability to describe activity coefficients using binary systems of amino acids and peptides in water. These compounds were chosen as they are the building blocks of proteins, the target compounds, and it is hypothesised that models performing for these compounds will also perform for proteins due to their molecular similarities. Furthermore, activity coefficient data is available for these systems but not for proteins.

A systematic study of the various model versions was adopted with the aim to establish, which model type is applicable for amino acids and peptides and therefore most likely also for proteins. However, a model that performs for amino acid and peptide activity coefficient calculations would need to be examined for proteins next as molecular size differences of these compound classes are significantly different.

The UNIFAC group definitions, the group parameters, the model approaches and activity coefficient reference states were investigated for amino acids and peptides. Furthermore, for the first time the original UNIFAC model [Fredenslund et al., 1975] with the established molecular groups and parameters of Hansen et al. (1991) was studied. This was of particular interest for this work as biotechnology compounds had been considered by Hansen et al. (1991) making the model likely to succeed for proteins and related compounds, i.e., amino acids and peptides. While first of all the UNIFAC model was approached because of its predictive properties, an examination of the UNIQUAC model was also pursued.

3.2 Activity Coefficient Data

Activity coefficient data of six aqueous amino acid and three aqueous peptide systems at 25 °C was obtained [Edsall and Wyman, 1958]. Hence, the activity coefficient data of nine different compounds was examined in this work: glycine, alanine, valine, serine, threonine, proline, glycylglycine, glycylalanine and triglycine. These compounds reflect the different properties that are found for the different residues of amino acids, peptides or proteins, chapter 2.2. The first three compounds represent non - polar characteristics while the next two, serine and threonine, are polar. The imino acid proline represents a steric hindered residue and the three small peptides refer to a protein's peptide bonding and backbone formation.

The data for these nine systems had to be converted into the thermodynamic concentration scale: mole fractions and activity coefficients referring to the mole fraction scale. To convert the data from the molal scale the following equations were applied [Soehnel and Garside, 1992; Nicolaisen, 1994]:

$$x_i = \frac{m_i \cdot M_{Solvent}}{\sum_1^j m_j \cdot M_{Solvent} + 1} \qquad (3.2.1)$$

$$\gamma_{i(x)} = \gamma_{i(m)} \cdot \left(1 + M_{Solvent} \cdot \sum_1^j m_j\right) \qquad (3.2.2)$$

The mole fraction, x, of compound, i, is obtained from the molal concentration, m, of that compound, the solvent's molecular weight, M, and the molal concentrations of further compounds, j. The activity coefficient on the mole fraction scale, γ_x, is calculated from the molal activity coefficient, γ_m, the solvent's molecular weight and the molal concentrations of further system compounds.

In cases where the data is provided on the molar scale the following conversions were applied:

$$x_i = \frac{c_i \cdot M_{Solvent}}{\rho_{Solution} + \sum_1^j c_j \cdot \left(M_{Solvent} - M_j\right)} \qquad (3.2.3)$$

$$\gamma_{i(x)} = \gamma_{i(c)} \cdot \left(\frac{\rho_{Solution} + \sum_1^j c_j \cdot \left(M_{Solvent} - M_j\right)}{\rho_{Solvent}} \right) \qquad (3.2.4)$$

To obtain a compound's mole fraction and activity coefficient on the mole fraction scale the molar concentrations, c, of all compounds, i and j, are needed. Additionally, the solvent's molecular weight and compounds' molecular weights have to be known while also the solution and solvent density, ρ, are necessary for the calculations. As the solution density was not available for the mixture systems studied the solvent density, i.e., water density, was applied. This was justified following the fact that water is the main compound of the systems where the second compound, either an amino acid or a peptide, occurs at low concentrations, appendix A. Conversions into the thermodynamic scale, x and γ_x, were pursued as part of this work and the obtained data of all nine systems, which were studied here, is given in appendix A.

3.3 Study of the UNIFAC and UNIQUAC Models

The activity coefficient data of nine amino acids and peptides, which build a binary system with water as the solvent, was modelled using the UNIFAC and UNIQUAC models. The models used in this chapter were introduced in detail in the previous chapter, chapter two, and the programs to pursue these investigations were obtained from different sources.

Routines for all UNIFAC models using the symmetric convention were supplied by Prof. R. Gani from the Chemical Engineering Department at the Technical University of Denmark, Lyngby. To use these models slight modifications were added to the programs as part of this work. This made the programs applicable for the systems, i.e., amino acids and peptides, studied here. A routine for the UNIFAC model based on the unsymmetric convention was made available by Prof. F.L.P. Pessoa from the Chemical Engineering Department of the Federal University of Rio de Janeiro in

Brazil. The programming routine for the UNIQUAC model used here was designed and programmed as part of this work and is documented in appendix B.

3.3.1 Study of the UNIFAC Model

In order to apply the UNIFAC model [Fredenslund et al., 1975] the molecular structures of all system compounds were examined and the molecular groups required to assemble the compounds according to Hansen et al. (1991) definitions were determined. In table 3.3.1.1 the structures of the nine compounds [Lehninger, 1982] and the defined UNIFAC molecular groups, i.e., UNIFAC groups [Hansen et al., 1991], are given. Up to five different UNIFAC groups are used for a molecule e.g. threonine. For one case, glycylalanine, the CONHCH group is not available and was therefore substituted by the structurally related CONHCH$_2$ group. This substitution will be discussed when reviewing the results as further approaches of this kind were pursued. For these groups, table 3.3.1.1, parameters and binary interaction parameters as given by Hansen et al. (1991) were applied for the model calculations unless stated differently. For water, the second system compound, the established groups and parameters were applied [Abrams and Prausnitz, 1975].

Table 3.3.1.1: Molecular structures of amino acids and peptides and their UNIFAC groups

Compound	Molecule structure	UNIFAC groups
Glycine	H_2N-CH(H)-C(=O)(OH)	H_2NCH_2, COOH
Alanine	H_2N-CH(CH_3)-C(=O)(OH)	H_2NCH, COOH, CH_3
Valine	H_2N-CH(CH(CH_3)$_2$)-C(=O)(OH)	H_2NCH, COOH, 2 CH_3, CH
Serine	H_2N-CH(CH_2OH)-C(=O)(OH)	H_2NCH, COOH, CH_2, OH
Threonine	H_2N-CH(CH(OH)(CH_3))-C(=O)(OH)	H_2NCH, COOH, CH, CH_3, OH
Proline	CH_2-NH CH_2 | CH_2-C(H)-C(=O)(OH)	HNCH, COOH, 3 CH_2
Glycylglycine	H_2N-CH(H)-C(=O)NH-CH(H)-C(=O)(OH)	H_2NCH_2, COOH, $CONHCH_2$
Glycylalanine	H_2N-CH(H)-C(=O)NH-CH(CH_3)-C(=O)(OH)	H_2NCH_2, COOH, $CONHCH_2$, CH_3
Triglycine	H_2N-CH(H)-C(=O)NH-CH(H)-C(=O)NH-CH(H)-C(=O)(OH)	H_2NCH_2, COOH, 2 $CONHCH_2$

In table 3.3.1.2 the results for the calculations for the nine systems are listed. Four different sets of calculations were pursued with the UNIFAC model studying the group definitions, group parameters, model versions and activity coefficient reference states. The model performances for these calculations are given in table 3.3.1.2 referring to the root mean square deviation, rmsd. The root mean square deviation describes the error between modelled data, cal., and experimentally determined data, exp., over all data points, N. The root mean square deviation is evaluated using the standard square deviation, ssd:

$$ssd = \sum_{1}^{N}\left(\frac{\text{exp.}-cal.}{\text{exp.}}\right)^{2} \qquad (3.3.1.1)$$

and

$$rmsd = 100\% \cdot \left(\frac{1}{N}\cdot ssd\right)^{0.5} \qquad (3.3.1.2)$$

The root mean square deviation is computed as a percentage and a low percentage indicates a good model performance, where a root mean square deviation of 0 % is the best achievable model performance. All model evaluations of this work were scaled and judged using the root mean square deviation.

For this part of the work four consecutive sets of calculations were pursued. The results of these are presented next. Following this, the results of the four different calculations are reviewed together. Moreover the work of Pinho et al. (1994) was

Table 3.3.1.2: Performance of the activity coefficient calculations for amino acids and peptides using various UNIFAC model versions

case	compounds	model type	rmsd [%]
1	*Glycine*	original UNIFAC (symmetric)	*73*
2	Alanine	original UNIFAC (symmetric)	11
3	Valine	original UNIFAC (symmetric)	377
4	Serine	original UNIFAC (symmetric)	14
5	Threonine	original UNIFAC (symmetric)	53
6	Proline	original UNIFAC (symmetric)	395
7	Glycylglycine	original UNIFAC (symmetric)	95
8	Glycylalanine	original UNIFAC (symmetric)	92
9	Triglycine	original UNIFAC (symmetric)	99
10	*Glycine*	original UNIFAC (symmetric) and altered groups	*69*
11	Valine	original UNIFAC (symmetric) and altered groups	882
12	Threonine	original UNIFAC (symmetric) and altered groups	53
13	Proline	original UNIFAC (symmetric) and altered groups	678
14	*Glycine*	modified UNIFAC (symmetric)	*60*
15	Valine	modified UNIFAC (symmetric)	113
16	Threonine	modified UNIFAC (symmetric)	44
17	Proline	modified UNIFAC (symmetric)	124
18	*Glycine*	original UNIFAC (unsymmetric)	*49*
19	Proline	original UNIFAC (unsymmetric)	36

integrated into this review as they had investigated yet another UNIFAC model version. They had defined amino acid and peptide specific groups and determined their parameters, which led to activity coefficient predictions. Their results and those from this work will be discussed and used to direct this work further.

The first set of calculations, table 3.3.1.2, pursued as part of this work used the original UNIFAC model on the symmetric scale to compute activity coefficients for all nine systems (cases 1 - 9). The model gave deviations from 11 - 395 % for the nine systems. The poor performance of the original UNIFAC model on the symmetric scale using the parameters of Hansen et al. (1991) was thought to be mainly due to the fact that some interaction parameters were missing for the defined groups. For the H_2NCH_2 and H_2NCH groups binary interaction parameters with the COOH group are listed by Hansen et al. (1991) as zero. This implies that the sum of interactions occurring between the molecular groups is zero, which is not necessarily true for these groups. If they are in the vicinity of each other interactions are likely to occur due to their dipole moments, chapter 2.2. Therefore these interaction parameters might have led to the unsatisfactory results. Following this hypothesis a second set of calculations was pursued using structurally related groups with interaction parameters.

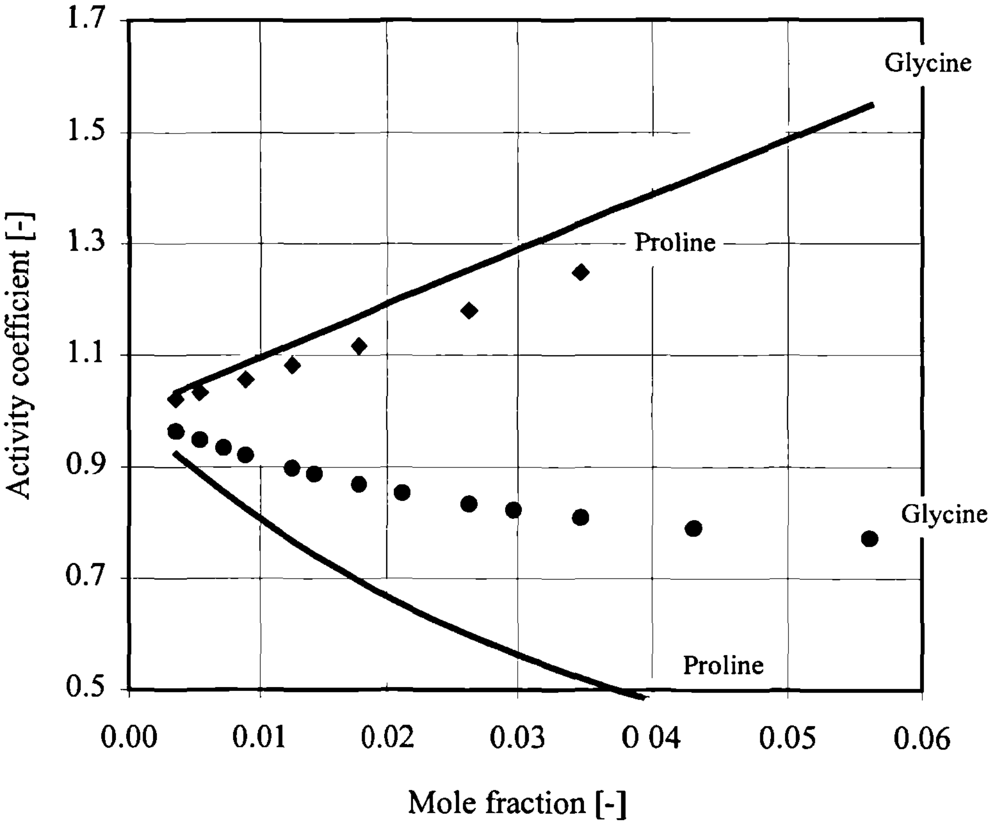
For the second set of calculations (cases 10 - 13), table 3.3.1.2, the lack of interaction parameters was overcome by introduction of different UNIFAC groups. Groups, which have interaction parameters listed, and relate most closely in their molecular structure to the ones used for the first calculations were used. This approach was taken as it had been demonstrated by Fredenslund et al. (1975) and Hansen et al. (1991) that similar molecular structures such as CH, CH_2 and CH_3 bear the same property contribution and therefore were assigned into one group, main group, of same interaction parameters. Furthermore, it can be assumed that similar molecular structures have similar interactions occurring. However, this is only true for some relating molecular structures but not all, as manifested in the group contribution method where a set of different molecular groups is created. Still, the groups that missed interaction parameters, H_2NCH_2, H_2NCH and COOH, were substituted with the CH_2NH, CHNH and COO main groups, respectively. Following this, four selected systems, glycine, valine, threonine and proline, were studied with altered groups. These systems were selected as they showed the lowest and highest deviations in the previous calculations and because they demonstrate different activity coefficient behaviour, i.e., negative and positive deviation from ideal behaviour. Furthermore, they represent the different residue characteristics as discussed in chapter 2.2. Following the hypothesis that the assumed interaction parameters of zero caused the poor results for the original UNIFAC model (cases 1 -

9) the same model type was applied but altered groups, i.e., CH_2NH, CHNH and COO, with their interaction parameters were used. However, no improvement was found for these calculations. For the same model type with a different set of groups the deviations almost doubled when compared to the first calculations. A doubling of the deviation was observed in two cases: valine (case 11 and 3) and proline (case 13 and 6). This result was interpreted as a possible deficiency in the model. Therefore, the original UNIFAC model was substituted with a different model approach, the modified UNIFAC model, to examine this hypothesis.

For the third set of calculations, table 3.3.1.2, the modified model (cases 14 - 17) was studied. The symmetric scale was used and the same set of four systems was studied while the first set of group definitions was applied. For the systems previously performing most poorly, valine and proline, better results were obtained and an improvement by a factor of three is observed compared to the first calculations (cases 3 and 6). Still, deviations of around 100 % were obtained which is not satisfactory. Qualitative examination of the calculated activity coefficients and those from experiment demonstrates also that no improvement was made. The qualitative behaviour of experimentally determined and modelled activity coefficients shows no correlation. At this point of the work it was unclear what caused the high deviations: the group definitions, their parameters or the model approach. Neither a change of the group definitions and parameters nor a change of the model version showed improvement. However, the activity coefficient reference state had not yet been examined. Therefore, the unsymmetric reference state was examined for its impact on the activity coefficient description.

The fourth and last set of calculations, table 3.3.1.2, examined the original UNIFAC model using the unsymmetric scale for the glycine and proline systems (cases 18 - 19). The results of these calculations and the model approach used are directly comparable to the first calculations (cases 1 and 6) as only the applied reference state differs. The unsymmetric scale has a major impact on the model performance. Deviations of modelled and experimentally determined activity coefficients are the lowest so far obtained. For proline a deviation of only 36 % and for glycine a deviation of only 49 % resulted. Noticeable improvement was made by introduction of the unsymmetric scale but still no improvement for the qualitative behaviour is found. Figure 3.3.1.1. demonstrates that the model calculates the opposite behaviour to the behaviour determined from experiment. However, the fact that the unsymmetric scale reduces the deviation of proline by a factor of ten when compared to the symmetric scale leads to the conclusion that amino acids and related compounds, i.e., proteins, follow Henry's law for reference purposes. This was also found by Fraaije et al. (1991) from adsorption experiments with proteins, which are

Figure 3.3.1.1: Experimentally determined and modelled (original UNIFAC unsymmetric) activity coefficients of glycine and proline against the amino acid and peptide mole fraction



Symbols represent experimentally determined data. Lines represent model calculations.

the target compounds of this work and relate to amino acids and peptides. As a result of this work the unsymmetric standard state was applied for the activity coefficient calculations in the following parts.

The results from the four calculation sets demonstrate that the UNIFAC model is not applicable in its present state for the compounds examined. A combination of causes might be responsible for the poor performances. However, the model failure is thought to be mainly due to the fact that the compounds were not fully represented. Model parameters and group definitions seem inappropriate. For glycylalanine the group representing the peptide bond of the molecule is missing. This documents that the UNIFAC group definitions are not appropriate for proteins as the peptide bond is an important molecular group for all proteins. The peptide bond builds a protein's backbone as discussed in chapter 2.2. Additionally, binary interactions of so far three molecular groups are not available. The interactions of the amino terminus with the carboxyl terminus and the carboxyl group of the acidic residues (aspartic acid, glutamic acid) are not described. The missing interactions and the non existence of a peptide group show that similar compounds were not used to establish the model parameters, which most possibly leads to the unsatisfactory quantitative and qualitative performance. However this does not yet invalidate the model as only incorrect group definitions and parameters are possibly responsible for the bad model performances. Additionally, confirmation of the unsymmetric reference scale by means of these UNIFAC model studies allows for the assumption that the modelling base is satisfactory for the studied compounds. However, this is further questioned and discussed next.

The UNIFAC model might perform satisfactorily for the systems of interest hypothesising that the model approach is appropriate. Introduction of the systems into the database, which is used to establish the model parameters should improve the model performance. Pinho et al. (1994) have demonstrated that indeed the UNIFAC model performs better when the model parameters are established from a database that contains amino acids and peptides. They computed activity coefficients with their newly established model parameters and found average deviations of about 0.5 % with a maximum of 1.2 % for proline. Their results implied that the modelling base is appropriate.

However, predictions pursued by Pinho et al. (1994) for related but new systems, that were not previously included in the database used to obtain the model parameters, showed maximum deviations of 17 % (glyclglycine) or even 28 % (glycylalanine) when compared to the experimentally determined data over composition. This predictive performance is already better than the one found from the work pursued here and confirms the impact of the reference database. Still, the performance is not satisfactory and this is most likely still due to the limitations of the database used to obtain the model parameters. The database of Pinho et al. (1994) consisted of only seven systems and was therefore not fully representative for related but new systems. Limited experimental data leads to unsatisfactory group contribution methods and is certainly a matter of concern as only limited experimental property data is available for the targeted compounds, i.e., proteins. This makes the study of the UNIFAC model at this stage unrealistic, where as a next step a correlative approach is focused on as no parameters are available to describe the target compounds. However, the UNIFAC model in its present state with the group definitions and parameters of Hansen et al. (1991) had to be examined for protein related compounds in order to clarify its applicability as it might have been successful as it is successful for many other compounds.

Following this, the related UNIQUAC model was examined next as part of this work. This model allows closer examination of the modelling approach due to the fact that no group contribution approach is added. This reduces the number of parameters required for calculations. Furthermore, the amount of experimental data necessary is reduced while also deviations due to possibly incorrectly defined UNIFAC groups are eliminated. However, the predictive ability over related but different compounds is lost.

3.3.2 Study of the UNIQUAC Model

The UNIQUAC model [Abrams and Prausnitz, 1975] evolved from the same theoretical approach as the UNIFAC model but it represents system compounds differently. Compounds are represented as molecular entities. Due to this kind of compound description less model parameters are needed, which elucidates model calculations and model approach investigations in comparison to the previously examined UNIFAC model. Additionally, less experimental data is needed for the model as less parameters have to be determined and experimental data is limited for the targeted compounds, i.e., proteins. Aiming predominantly at the establishment of a model base and following the fact that only limited reference data is available the original UNIQUAC model was applied here. In any case a correlative approach was to be examined as a next step as no appropriate parameters are available for the target compounds. The unsymmetric scale was used following the presented results with the UNIFAC model, chapter 3.3.1.

Three of the previously studied systems were examined: glycine, proline and glycylglycine. The first two systems were chosen as they have been examined here for all the different UNIFAC model types and because they demonstrate different activity coefficient behaviour over composition. Glycylglycine was chosen to examine a peptide containing system.

In order to apply the UNIQUAC model the binary interaction parameters and structural parameters had to be determined. The structural parameters, r and q, were obtained from the publication of Peres and Marcedo (1994) and are listed in table 3.3.2.1. The interaction parameters, u, for the three systems had to be obtained independently. The procedure for the parameter determination is discussed in the next chapter, chapter four, while the interaction parameters determined as a result of this work are given in table 3.3.2.1. Per compound pair one interaction parameter was determined and for each of the binary systems investigated three interaction parameters resulted. In total seven new parameters were obtained to describe the three systems*. The results of the model calculations are given in table 3.3.2.2. Use of the original UNIQUAC model shows much better results when compared to the UNIFAC model, table 3.3.1.2. With the UNIQUAC model root mean square deviations of 0.1 - 0.3 % were achieved for the three systems when comparing modelled and experimentally determined activity coefficient data. Additionally, the agreement of model calculations and the experimentally determined activity

* and for all three systems an identical water – water interaction parameter was determined.

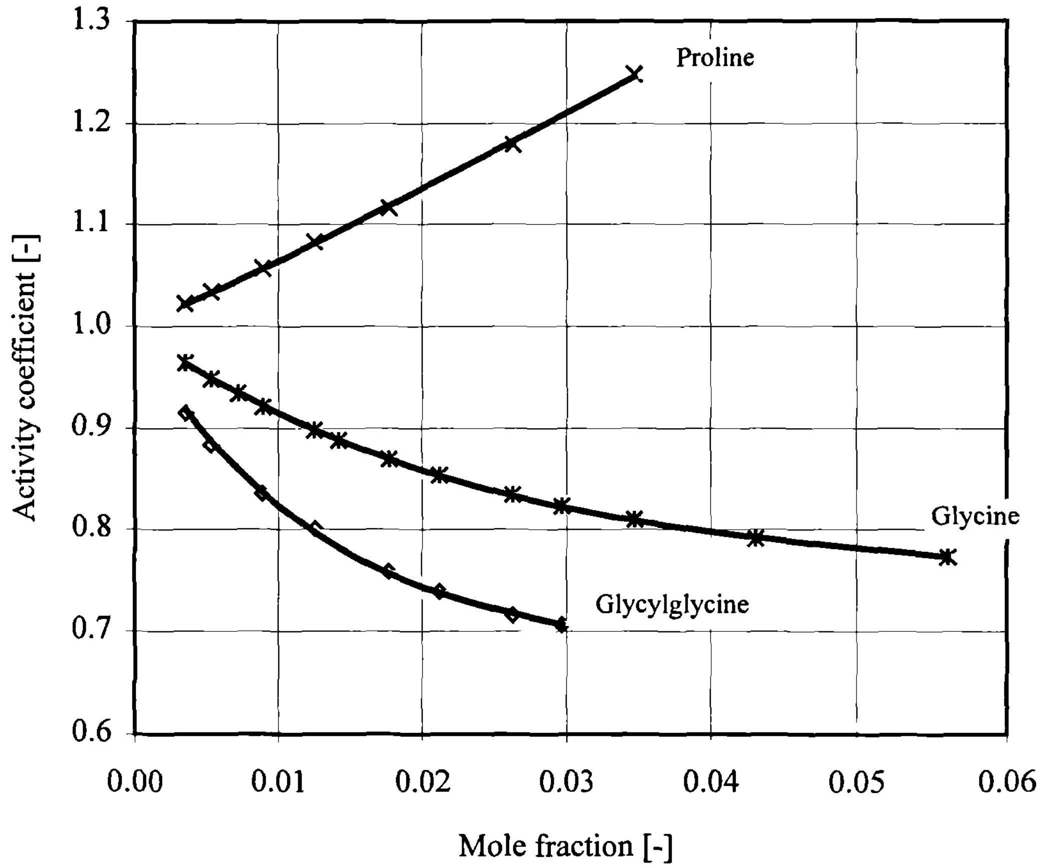
Table 3.3.2.1: Structural parameters and binary interaction parameters for glycine, proline, glycylglycine and water

	H_2O	Glycine	Proline	Glycylglycine	r	q
	[K]	[K]	[K]	[K]	[-]	[-]
H_2O	78.69	- 264.0	- 302.2	- 269.1	1.40	0.92
Glycine		- 801.3			2.67	2.46
Proline			- 806.3		4.30	3.46
Glycylglycine				- 806.6	4.65	4.04

Table 3.3.2.2: Performance of the activity coefficient calculations for glycine, proline and glycylglycine using the original UNIQUAC model (unsymmetric)

case	compounds	model	rmsd [%]
1	Glycine	original UNIQUAC (unsymmetric)	0.1
2	Proline	original UNIQUAC (unsymmetric)	0.2
3	Glycylglycine	original UNIQUAC (unsymmetric)	0.3

Figure 3.3.2.1: Experimentally determined and modelled (original UNIQUAC unsymmetric) activity coefficients of glycine, proline and glycylglycine against the amino acid and peptide mole fraction



Symbols represent experimentally determined data. Lines represent model calculations.

Table 3.3.2.3.: ~~Experimentation~~ Bench mark error evaluation using two sets of experimentally determined activity coefficient data for glycine, serine and glycylglycine

x	Glycine$^{\#}$	Glycine$^{\$}$	Serine$^{\#}$	Serine$^{\$}$	Glycylglycine$^{\#}$	Glycylglycine$^{\$}$
0.0036	0.9637	0.9644	0.9540	0.9673	0.9145	0.9153
0.0054	0.9531	0.9487	0.9340	0.9492	0.8821	0.8838
0.0089	0.9158	0.9213	0.8951	0.9160	0.8308	0.8354
0.0125	0.9048	0.8981	-	-	0.7917	0.8006
0.0177	0.8910	0.8697	-	-	0.7154	0.7581
0.0263	0.8410	0.8342	-	-	0.7079	0.7155
0.0348	0.8152	0.8098	-	-	-	-
0.0431	0.7962	0.7909	-	-	-	-

Error	Glycine	Serine	Glycylglycine
ssd [-]	0.0008	0.0010	0.0038
rmsd [%]	1.02	1.83	2.53

#: Sober, 1968; $: Edsall and Wyman, 1958 coefficient data is shown in figure 3.3.2.1. Qualitative and quantitative agreement resulted for the three systems when the original UNIQUAC model on the unsymmetric scale was applied, which is partly due to the fact that a correlation is examined. However, the low deviation seems to also confirm the model base.

To judge the model performance a bench mark error ~~an experimentation error~~ was established, * table 3.3.2.3. For glycine, serine and glycylglycine activity coefficient data at same system conditions is available from two different laboratories [Edsall and Wyman, 1958; Sober, 1968]. Comparisons of the two data sets gave root mean square deviations of 1.0 %, 1.8 % and 2.5 % for glycine, serine and glycylglycine, respectively. This indicates an average bench mark ~~experimentation~~ error of 1.8 %. These results demonstrate that the deviations of 0.1 - 0.3 % resulting from the UNIQUAC calculations are well below the bench mark ~~experimentation~~ error, confirming the good model performance. Moreover, these results imply that the model approach is appropriate for amino acid and peptide systems and therefore most likely also for protein containing systems. This is assumed as amino acids and peptides are the molecular building groups of proteins but the next part of this work shall possibly prove this hypothesis. Therefore the same model was approached for the calculations of protein activity coefficients. These studies are presented in the next chapter, chapter four.

* from a range of published values

3.4 Summary

A systematic comparison of different models was followed for the first time to calculate activity coefficients for protein related compounds, amino acids and peptides. This part of the work was pursued in order to evaluate which model would represent an appropriate approach for protein containing systems and would eventually model solution and mixture properties.

Nine amino acid and peptide systems were examined demonstrating that the UNIFAC models and the established parameters are not applicable. Deviations of well above 100 % resulted when comparing experimentally determined and calculated activity coefficients. This is possibly mainly due to missing parameters. Furthermore, molecular groups were not defined according to the needs of the examined systems. Molecular structures such as e.g. peptide bonds were not represented for the studied systems. However, the UNIFAC model calculations demonstrated that the unsymmetric scale should be applied for the calculations of amino acid and peptide activity coefficients. This was indicated by an improvement of the deviations to below 50 %. Further work on the UNIFAC model was not pursued due to the limited amount of experimental data available. Therefore, the UNIQUAC model was approached.

A program of the UNIQUAC model using the unsymmetric activity coefficient scale was developed as part of this work and applied for three systems, two amino acids and one peptide system. The binary interaction parameters for these systems were determined. Deviations of only 0.1 - 0.3 % occurred for the model when calculated and experimentally determined activity coefficients were compared over varying composition. This result is well below the obtained bench mark ~~experimentation~~ error of 1.8 % and indicates that the UNIQUAC model should be examined for protein containing systems and their solution and mixture properties.

3.5 Nomenclature

c	molar concentration
cal.	calculated
exp.	experimental
m	molal concentration
M	molecular weight
N	number of data points
q	structural area parameter of a compound
r	structural volume parameter of a compound
rmsd	root mean square deviation
ssd	standard square deviation
UNIFAC model	UNIQUAC Functional Activity Coefficients model

UNIQUAC model	UNIversal QUAsi Chemical model
x	compound mole fraction

Greek letters

γ	compound activity coefficient
ρ	density

Subscript

c	molar concentration
i	compound
j	compound
m	molal concentration
x	mole fraction

Chapter 4

Calculation of Protein Activity Coefficients

The determination of protein activity coefficients from osmotic pressure measurements and the study of the UNIQUAC model to calculate these protein activity coefficients is documented.

4.1 Introduction

The first part of this work demonstrated that the original UNIQUAC model on the unsymmetric scale successfully calculated activity coefficients for two amino acids and one peptide. As amino acids and peptides are the subunits of proteins it was hypothesised that the same approach is applicable to model activity coefficients of proteins. This hypothesis was examined and the investigation is documented in this chapter, chapter four. These results are also presented and discussed elsewhere [Agena et al., 1996; Agena et al., 1997b; Agena et al., 1998b].

In order to examine the UNIQUAC model for its ability to express protein activity coefficients, activity coefficient data was required. Experimental osmotic pressure data was used to obtain this data. By means of the virial expansion activity coefficients as a function of solution composition were determined using experimental osmotic pressure measurements from various sources. Ten different systems were examined consisting of four different proteins: serum albumin, α – chymotrypsin, β - lactoglobulin and ovalbumin. The determined activity coefficient data was used to study the original UNIQUAC model. The protein activity coefficient behaviour as a function of salt concentration, pH and temperature was examined to validate the model performance while additionally protein solubility was investigated.

4.2 Activity Coefficient Data

Activity coefficient data is not documented for proteins but osmotic pressure measurements were utilised here to obtain protein activity coefficient data. Experimental osmotic pressure data is available from various laboratories at different system conditions [Cohn and Edsall, 1943; Guntelberg and Lindstrom-Lang, 1949; Christensen, 1952; Haynes et al., 1992]. Ten systems and a broad spectrum of system conditions were studied: four globular protein types, about four different salt types, protein concentrations of up to 63 g/L, varying salt concentrations, temperatures around 1 ^{o}C and 25 ^{o}C, and a pH range from 3 - 12. These systems and their conditions are listed in table 4.2.1. while the four proteins and their characteristics have been discussed before, chapter 2.2.

For this work the salt and the solvent, i.e., water, were represented as one pseudo solvent, which is applicable due to the fact that the salt concentration is held constant, table 4.2.1. The systems were regarded as binary systems of a solute, i.e., protein, and a pseudo solvent, PS. The properties of water were assumed for the pseudo solvents as water is its main compound. To confirm this approach the activity coefficient behaviour was examined and discussed in the results section and the results support the concept of the pseudo solvent.

Table 4.2.1: Protein systems and system conditions investigated

System	Protein	Pseudo	Protein conc.	t	pH	Salt, Buffer
		solvent	[g/L]	[°C]	[-]	
S	Serum Albumin	PS 1	10 - 48	1	7.4	Phosphate 0.0667 M
C1	α - Chymotrypsin	PS 2	0.9 - 40	25	3	Potassium Sulphate 0.05 M
C2	α - Chymotrypsin	PS 3	0.9 - 40	25	3	Potassium Sulphate 0.15 M
C3	α - Chymotrypsin	PS 4	0.9 - 40	25	3	Potassium Sulphate 0.3 M
C4	α - Chymotrypsin	PS 5	0.9 - 9	25	5	Potassium Sulphate 0.1 M
C5	α - Chymotrypsin	PS 6	0.9 - 9	25	12	Potassium Sulphate 0.1 M
C6	α - Chymotrypsin	PS 7	0.9 - 9	25	5	Sodium Phosphate 0.1 M
C7	α - Chymotrypsin	PS 8	0.9 - 10	25	8.25	Sodium Phosphate 0.1 M
L	β - Lactoglobulin	PS 9	10 - 47	20	n.k.	Sodium Chloride 1 molal
O	Ovalbumin	PS 10	23 - 63	20	4.85	Sodium Chloride 1 molal

n.k.: not known, conc.: concentration

4.2.1 Virial Expansion

The virial expansion [Wills et al., 1993] was applied for binary systems to derive activity coefficients of proteins. From osmotic pressure, Π, measurements over the molar protein concentration, c, one obtains the virial coefficients, B, using the equation:

$$\frac{\Pi}{R \cdot T} = c \cdot \left(1 + B_2 \cdot c + B_3 \cdot c^2 + \ldots\right) \qquad (4.2.1.1)$$

where T is the temperature and R the gas constant.

Molal activity coefficients of the solute, $\gamma_{(m)}$, with respect to changes in solvent chemical potentials over the molal composition, m, are given by:

$$\ln \gamma_{(m)} = 2 \cdot D_2 \cdot m + \frac{3}{2} \cdot D_3 \cdot m^2 + \ldots \qquad (4.2.1.2)$$

which leads to protein activity coefficients. The coefficients, D, are related to the virial coefficients B by:

$$D_2 = \left(B_2 - \bar{v} \cdot M\right) \cdot \rho \qquad (4.2.1.3)$$

$$D_3 = \left(B_3 - 2 \cdot B_3 \cdot \bar{v} \cdot M + (\bar{v} \cdot M)^2\right) \cdot \rho^2 \qquad (4.2.1.4)$$

where $\bar{v}$ and M represent the partial specific volume and the molecular weight of the solute, respectively, and ρ is the solvent density. The partial specific volumes needed for the activity coefficient determination were obtained from Sober (1968). The protein molecular weights originated from the same publications as the osmotic pressure measurements, chapter 4.2, following the fact that osmotic measurements are used to obtain the molecular weights of proteins.

The conversion between the molal concentration, m, and molar concentration, c, scales [Soehnel and Garside, 1992; Nicolaisen, 1994] were computed using:

$$m_i = \frac{c_i}{\rho_{Solution} - \sum_1^j c_j \cdot M_j} \qquad (4.2.1.5)$$

and

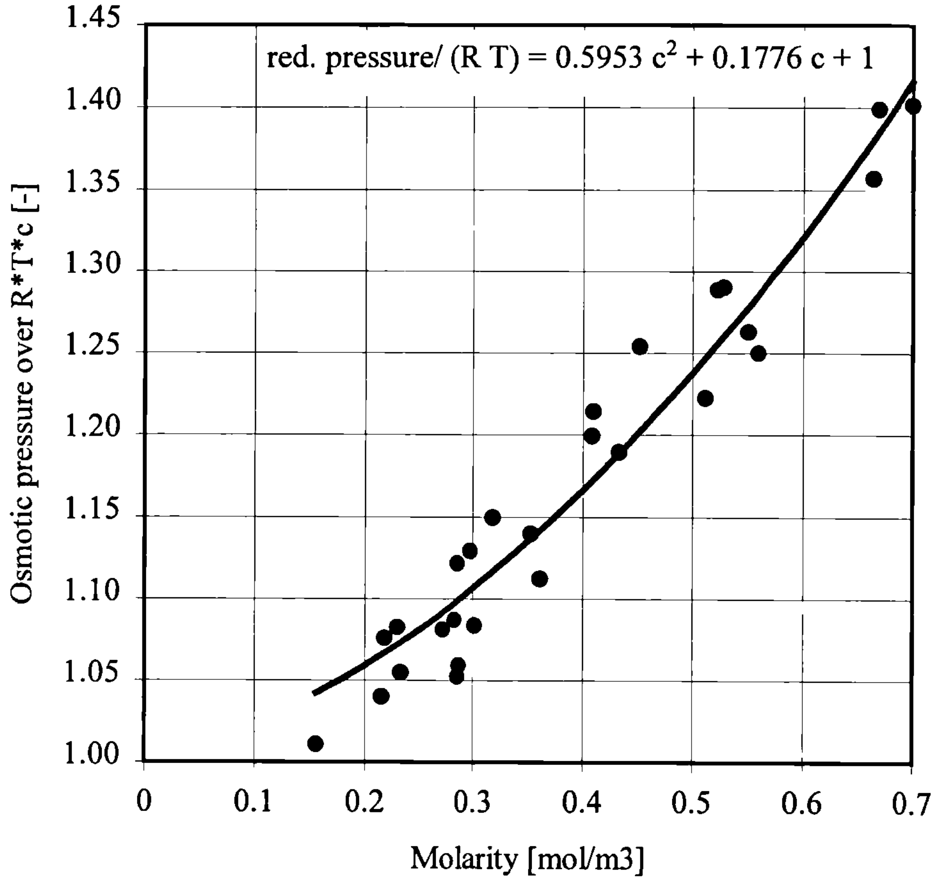
$$c_i = \frac{m_i \cdot \rho_{Solution}}{1 + \sum_1^j m_j \cdot M_j} \qquad (4.2.1.6)$$

The variables are those introduced above while i and j refer to the system compounds.

To obtain activity coefficient data from osmotic pressure measurements, the reduced osmotic pressure, Π/c, was scaled over the molar protein concentration, figure 4.2.1.1. Linear or higher order least square fitting was performed as part of this work in order to determine the virial coefficients, B. For the serum albumin system two coefficients with values of 0.1771 m^3/mol and 0.5953 $(m^3/mol)^2$ resulted. In table

4.2.1.1 the B coefficients for all systems are given and a good fitting performance is

Figure 4.2.1.1: Reduced osmotic pressure of the serum albumin system against protein concentration



Symbols represent experimentally determined data. Lines represent model calculations.

Table 4.2.1.1: Virial expansion data of protein systems

System	Mapparent	spec. vol.	B_2	B_3	B_4	R^2
	[g/mol]	[cm^3/g]	[m^3/mol]	[(m^3/mol)^2]	[(m^3/mol)^3]	[-]
S	68598 *	0.748	0.1776	0.5953	-	1.00
C1	28200	0.736	0.0022	0.0958	-	0.98
C2	26800	0.736	- 0.4776	0.2234	-	0.98
C3	27400	0.736	- 0.649	0.658	- 0.2178	0.98
C4	32200	0.736	0.048	- 0.2745	-	0.91
C5	30900	0.736	- 0.7118	-	-	1.00
C6	30000	0.736	- 0.2737	-	-	0.98
C7	31400	0.736	0.1995	- 0.4547	-	0.94
L	39240	0.732	0.0405	0.0795	-	0.88
O	44990	0.748	0.0576	0.0317	-	0.97

*: crystal molecular weight

indicated by R^2 values close to one. Up to three coefficients were needed to describe protein activity coefficients e.g. the C3 α - chymotrypsin system.

The protein activity coefficient data obtained from experimental osmotic pressure measurements is referred to as the experimentally determined activity coefficient data. The data was converted to the thermodynamic scale, chapter 3.2, for the following examinations and is listed in appendix C. The data obtained as a result of this work was used in the following part for model evaluation purposes.

4.3 Study of the Original UNIQUAC Model

Abrams and Prausnitz (1975) showed that the excess Gibbs energy and therefore activity coefficients, may be represented by a combinatorial term and a residual term using:

$$\ln\gamma_i = \ln\gamma_i^C + \ln\gamma_i^R \qquad (4.3.1)$$

$$\ln\gamma_i^C = f(q_i, r_i, x_i) \qquad (4.3.2)$$

$$\ln\gamma_i^R = f(q_i, x_i, u_{ij}, T) \qquad (4.3.3)$$

Their model, the original UNIQUAC model, was discussed in detail previously, chapter 2.3.1, but is briefly summarised introducing its parameters while also the applied reference states are defined.

The model consists of a combinatorial term, ln γ^C, which describes the molecule orientations in solution through the sizes and shapes of the different compounds, i. The structural parameters q and r represent the size and shape of the compounds and x represent the mole fraction and therefore the system composition. The residual term, ln γ^R, represents the short - range interactions occurring between the compounds by introducing binary interaction parameters, u, while also the system temperature, T, is accounted for. For the activity coefficients of the proteins the unsymmetric convention, $\gamma^*_P \rightarrow 1$ as $x_P \rightarrow 0$, is chosen for reference as the previous examinations, chapter 3.3.1, indicated Henry's law behaviour for amino acids and peptides. For the pseudo solvent the symmetric standard state with $\gamma_{PS} \rightarrow 1$ as $x_{PS} \rightarrow 1$ is chosen. The activity coefficients computed from the UNIQUAC model are referred to as calculated activity coefficients.

To calculate activity coefficients with the UNIQUAC model the structural parameters, q and r, for the proteins had to be determined. As this is the first reported attempt to model protein activity coefficients using the UNIQUAC model detailed examination of the structural parameters for proteins were pursued first.

Creighton (1984) gives a listing of molecular groups that build proteins and reports their van der Waals volumes and surface area contributions, which lead to the molecule volume and area and therefore to the structural parameters of proteins, chapter 2.4. These group properties and the knowledge of the amino acid sequence of proteins were used to calculate and to study protein van der Waals volumes and surface areas aiming at the structural parameters. The results of these computations are documented in table 4.3.1. The amino acid sequences for the five different proteins examined here are given in the cited references: lysozyme [Cranfield and Liu, 1965], serum albumin [Ho et al., 1993], α - chymotrypsin [Matthews et al., 1967], β - lactoglobulin [Pervaiz and Brew, 1985] and ovalbumin [Nisbet et al., 1981]. Twenty amino acid segments, A, occur in proteins and their van der Waals volume and surface area contributions are listed. These segments were as part of this work defined as the particular amino acid residue and the peptide bond plus the α - carbon atom of the protein backbone. Per protein the occurrence of these amino acid segments, ν, was determined from the amino acid sequences. From the number of amino acid segments, ν, per protein and the van der Waals volumes, $V_{W(A)}$, and surface areas, $A_{W(A)}$, of these segments the van der Waals volume, $V_{W(Protein)}$, and surface area, $A_{W(Protein)}$ of a protein were computed:

$$V_{W(Protein)} = \sum_{1}^{A} \nu^{(Protein)} \cdot V_{W(A)} \qquad (4.3.4)$$

$$A_{W(Protein)} = \sum_{1}^{A} \nu^{(Protein)} \cdot A_{W(A)} \qquad (4.3.5)$$

For β - lactoglobulin the amino acid sequence of the monomer was used to determine

Table 4.3.1: Calculated van der Waals volumes and surface areas of proteins

A	$V_{w(A)}$	$A_{w(A)}$	ν	ν	ν	ν	ν
	[Å^3]	[Å^2]	LYS	S	C	L*	O
Glycine	48.4	59.6	12	19	24	4	19
Alanine	65.4	83	12	58	22	15	35
Valine	99.2	127.3	6	33	23	9	31
Leucine	116	148.2	8	69	19	22	32
Isoleucine	122.5	148.2	6	14	10	10	25
Serine	72.5	89.8	10	34	27	7	38
Threonine	89.5	113.2	7	28	22	8	15
Cystein	84.5	104.7	8	35	10	5	6
Methionine	116.9	146.4	2	1	2	4	16
Proline	93.5	112.3	2	30	9	8	14
Aspartic acid	90.7	113.9	8	41	8	10	14
Asparagine	95.6	119.3	13	11	14	5	17
Glutamic acid	107.5	134.8	2	58	4	16	33
Glutamine	112.4	140.2	3	17	10	9	15
Lysine	127.8	159.7	6	62	14	15	20
Arginine	143.3	172.6	11	25	3	3	15
Histidine	109	113.5	1	19	2	2	7
Phenylalanine	136	165.4	3	33	6	4	20
Tyrosine	148.6	184.7	3	18	12	4	10
Tryptophan	166.4	193.7	6	2	0	2	3
$V_{W(Protein)}$ [Å^3]	-	-	12897	62545	22812	16859	39144
$A_{W(Protein)}$ [Å^2]	-	-	15930	77578	28460	20995	48572

A: amino acid segments, LYS: lysozyme, S: serum albumin, C: α - chymotrypsin, L: β - lactogobulin monomer, O: ovalbumin the van der Waals volume and surface area. Since β - lactoglobulin occurs as a dimer of two identical monomers, the van der Waals volume and surface area are 33718 Å^3 and 41990 Å^3, respectively. For lysozyme a van der Waals volume of 12897 Å^3 was computed, table 4.3.1. This result compares well with that of Roth et al. (1996) who used Bondi's method [Bondi, 1968] and calculated 12700 Å^3, which differs by only 1.5 % from the result obtained here. The calculated values were also examined referring to crystallographic results. Based on crystallographic data the surface area of lysozyme was established at 17000 Å^2 [Taratuta et al., 1990, and references within]. The calculated surface area of 15930 Å^2 compares well, differing by about 6 % from those results. This deviation represents the accumulative error of the x-ray measurements, the crystallographic models and the computations pursued here.

The van der Waals volume is related to the partial specific volume and therefore the calculated volumes become also comparable on this scale. Richards (1974) and Chothia (1975) found an average packing density for proteins of 0.75, defining the packing density as the ratio of the van der Waals volume to the actual volume of space, $\bar{v}$, occupied, which leads to:

$$V_W = 0.75 \cdot \bar{v} \qquad (4.3.6)$$

Therefore for lysozyme which has a partial specific volume of 0.703 cm^3/g [Sober, 1968], a van der Waals volume of 7698 cm^3/mol or 12771 Å^3 results. This value compares well to the calculated value of 12897 Å^3, table 4.3.1, and differs by only about 1 %. The same procedure was followed for three other proteins, serum albumin, chymotrypsin and ovalbumin. From the specific volumes van der Waals volumes of 63843 Å^3, 22436 Å^3, and 41881 Å^3 resulted for the three proteins, respectively. Comparison gave deviations of about 2 %, 2 % and 7 % for experimentally determined and computed van der Waals volumes and implies accuracy for the group contribution calculations, table 4.3.1, which show an average deviation of 3.7 %. By these means, it was shown that the correct structural parameters were applied.

This examination was of importance as previous calculations perused as part of this work but not presented here had resulted in model calculations that demonstrated on average a poorer model performance than those finally reported here. For those first calculations higher structural parameters (r higher by 7 % and q higher by 25 %) values were applied as a less accurate method to obtain these parameters was used. The average deviation of the calculations documented here, which were pursued with the confirmed set of parameters, was by a factor ten lower than for those with higher structural parameters. This illustrates the impact of the structural parameters on the model performance and emphasises the importance of these studies.

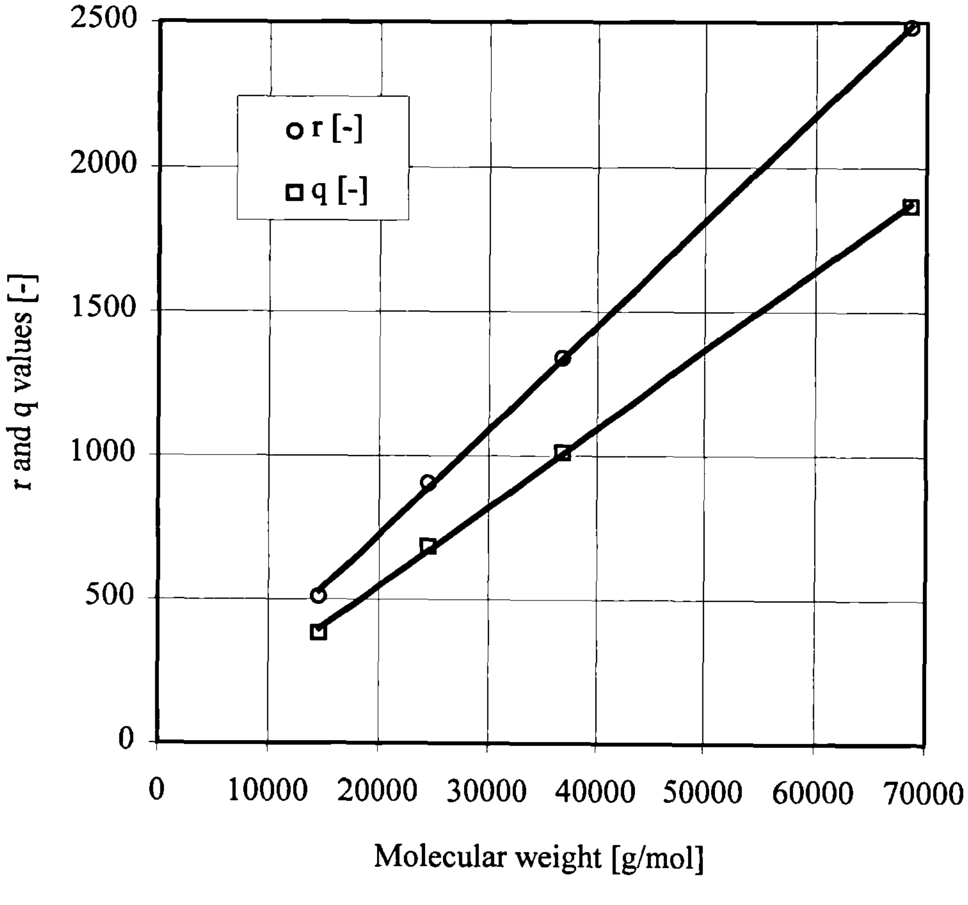
Investigations by Janin (1976) and Teller (1976) showed that volume and surface properties of proteins are proportional to a protein's molecular weight. As molecular weights of proteins are more widely available than their amino acid sequences a prediction method for the structural parameters on the basis of molecular weights was created. The determined van der Waals volumes and surface areas were converted into the structural parameters, equation 2.4.6 - 7, which were correlated to protein molecular weights. For the structural parameters a linear relationship with protein molecular weights, M, resulted:

$$r = 0.0362 \cdot M \qquad (4.3.7)$$

$$q = 0.0273 \cdot M \qquad (4.3.8)$$

For both correlations R^2 of 1 resulted indicating a perfect fit and figure 4.3.1 illustrates this agreement. The protein ovalbumin was not included in the correlation process and it was used to examine the performance of the developed equations. A prediction error of 5 % resulted for each of the two parameters and additionally confirms the new correlations, which also reflect the results of Janin (1976) and Teller (1976). The developed correlations, which were developed as part of this work, simplify the evaluation process for the structural parameters of a protein as now calculations are possible without the knowledge of a protein's amino acid sequence. The structural parameters of proteins were calculated using these newly established correlations and were used for all the following model calculations, table

Figure 4.3.1: The structural parameters of lysozyme, serum albumin, α - chymotrypsin and β - lactoglobulin against their molecular weight



Symbols represent reference data. Lines represent model calculations.

Table 4.3.2: UNIQUAC modelling data of protein systems

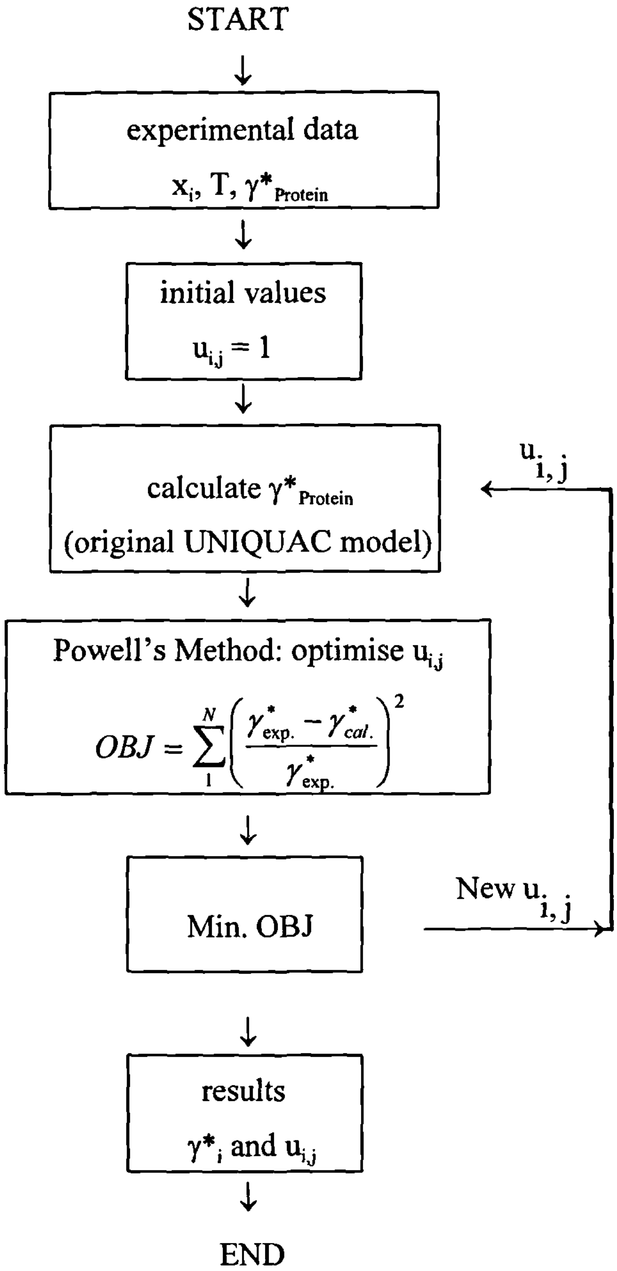
System	PS I	$M_{crystal}$	$q_{Protein}$	$r_{Protein}$	$u_{PS\ i,\ PS\ i}$	$u_{PS\ i,\ Protein}$	$u_{Protein,\ Protein}$
		[g/mol]	[-]	[-]	[K]	[K]	[K]
S	PS 1	68598	1873	2483	-8524.99	-8739.09	-8892.08
C1	PS 2	24500	669	887	670.117	330.113	-102.052
C2	PS 3	24500	669	887	766.012	398.105	-102.052
C3	PS 4	24500	669	887	825.842	439.421	-102.052
C4	PS 5	24500	669	887	379.094	118.679	-127.334
C5	PS 6	24500	669	887	404.658	156.541	-93.0088
C6	PS 7	24500	669	887	394.952	130.132	-127.334
C7	PS 8	24500	669	887	-2717.82	-2820.68	-2488.63
L	PS 9	36800	1005	1332	538.302	241.233	-96.1955
O	PS 10	45000	1229	1629	1181.38	912.220	643.074

$r_{PS\ i} = 0.92$, $q_{PS\ i} = 1.40$

4.3.2. For the pseudo solvent the established structural parameters of water were used in the following work.

In addition to the structural parameters the binary interaction parameters had to be determined to calculate activity coefficients with the UNIQUAC model. For each system up to three binary interaction parameters were determined using a multivariable non - linear optimisation technique of conjugate directions by Powell [Press et al., 1986]. A minimisation of the error between calculated and experimental activity coefficients was performed to guide the optimisation procedure. The structure of the program is displayed in figure 4.3.2. To determine the adjustable parameters initial values for the parameters, i.e., interaction parameters, were applied for the first calculation step. These initial parameters were set equal to one in most cases. Following the programming routine activity coefficients are calculated using equation 4.3.1 - 4.3.3 and the calculated coefficients were compared to the experimentally determined coefficients evaluating the standard square deviation, equation 3.3.1.1. Depending on the computed deviations, i.e., objective function values, new parameters are evaluated by the optimisation routine. This parameter adjusting routine is applied until a minimal deviation of computed and experimentally determined activity coefficients is obtained, which gives optimal parameters and lowest deviation between experimental and computed data. The programming routine developed and applied for this work is given in appendix B.

Figure 4.3.2: Programming structure used to determine interaction parameters


START
experimental data
x_i, T, $\gamma^*_{Protein}$
initial values
$u_{i,j} = 1$
calculate $\gamma^*_{Protein}$
(original UNIQUAC model)
$u_{i,j}$
Powell's Method: optimise $u_{i,j}$
$OBJ = \sum_1^N \left(\frac{\gamma^*_{exp.} - \gamma^*_{cal.}}{\gamma^*_{exp.}} \right)^2$
Min. OBJ
New $u_{i,j}$
results
γ^*_i and $u_{i,j}$
END

x: mole fraction, T: temperature, γ: activity coefficient, u: interaction parameter, i and j: compound, OBJ: objective function, N: number of data points, exp.: experimental, cal.: calculated

From this optimisation routine three model parameters, the interaction parameters u, resulted per system, table 4.3.2. These parameters represent the different interactions occurring between the compounds, i.e., the pseudo solvent and the protein. For different α – chymotrypsin systems the same protein - protein interaction parameter was determined when system pH and temperature are the same, table 4.2.1. This is the case for systems C1, C2 and C3, and likewise for systems C4 and C6. For all other systems, the same does not hold as system conditions differ.

A root mean square deviation of 0.54 % between experimentally determined and calculated activity coefficients is obtained for the model over all ten systems and is documented in table 4.3.3. All systems show good agreement with the experimental behaviour. Higher than average deviation resulted for two systems. One α - chymotrypsin system, C3, shows higher deviation than 0.54 %, which seems due to the more complex solution behaviour of the system as indicated by the number of virial coefficients, table 4.2.1.1, needed to determine the solution behaviour in the first case. The ovalbumin system, O, shows the highest deviation with a root mean square deviation of 2.35 %. This behaviour is possibly due to the high protein concentration of up to 7 % mass fraction, which is highest compared to the other systems examined. All other systems have protein mass fractions of a maximum of 5 %. Therefore, extrapolation to predict activity coefficients above protein mass

Table 4.3.3: Performance of protein activity coefficient calculations using the original UNIQUAC model (unsymmetric)

System	Number of data points	rmsd [%]
S	27	0.54
C1	9	0.08
C2	9	0.31
C3	9	1.65
C4	6	0.07
C5	6	0.14
C6	5	0.05
C7	6	0.19
L	15	0.03
O	7	2.35
Average	9.9	0.54

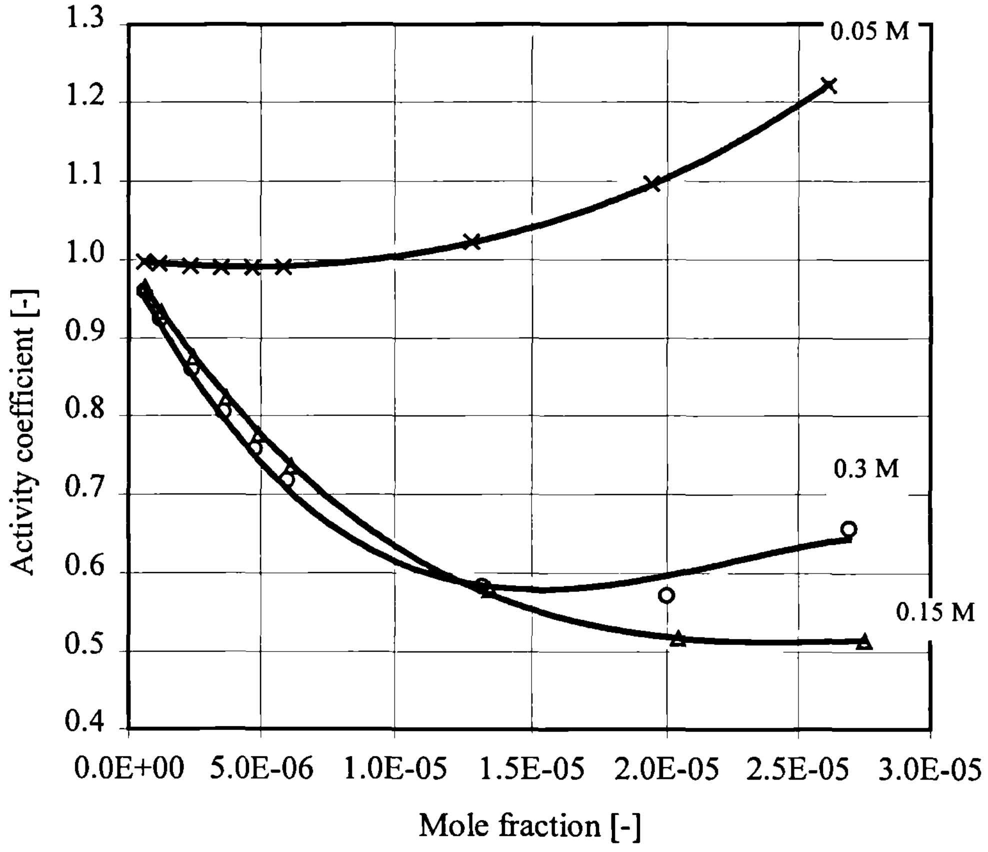
fractions of 5 % is not guaranteed with the determined parameters, which resulted from this work.

To judge the average root mean square deviation of 0.54 %, i.e., the model performance, the previously established bench mark ~~experimentation~~ error, chapter 3.3.2, was reviewed. Activity coefficient data from different laboratories had been compared and three protein related compounds demonstrated an average root mean square deviation of 1.8 %. Comparison of this bench mark ~~experimentation~~ error of 1.8 % and the model deviation of 0.54 % shows that the applied model performs well within the bench mark ~~experimentation~~ error.

For the seven α - chymotrypsin systems, C1 - C7, the experimentally determined activity coefficients and calculated activity coefficients are displayed in figure 4.3.3, 4.3.4 and 4.3.5 demonstrating furthermore the good model performance.

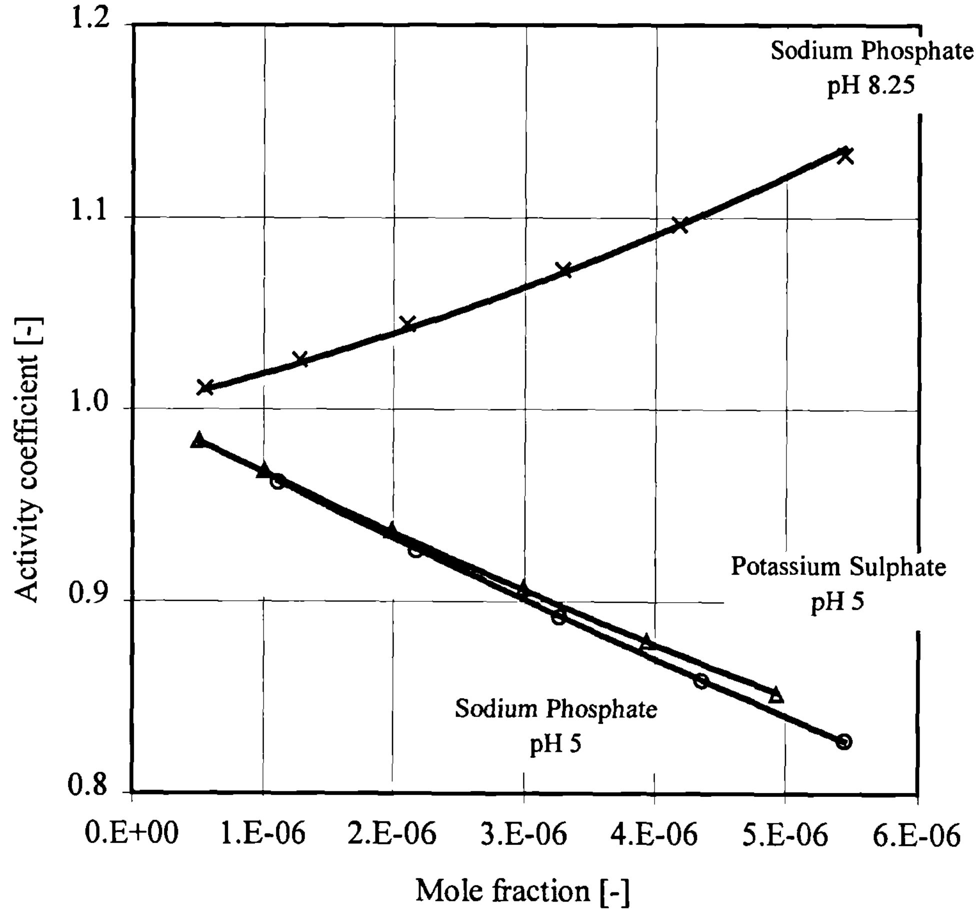
In Figure 4.3.3 the activity coefficients of α – chymotrypsin at different ionic strengths of potassium sulphate are shown. It is observed that with an increase of ionic strength from 0.05 M to 0.15 M a lower activity coefficient results for α – chymotrypsin. Activity coefficients reflect the solubility behaviour since protein solubility is proportional to the inverse of the protein activity coefficient as discussed by Green and Hughes (1955). This implies that the solubility of α – chymotrypsin increases with increasing ionic strength from 0.05 M to 0.15 M, figure 4.3.3. Such a solubility behaviour would occur in the salting - in region, where protein solubility

Figure 4.3.3: Experimentally determined and modelled (original UNIQUAC unsymmetric) activity coefficients of α - chymotrypsin with varying ionic strength of potassium sulphate against protein mole fraction



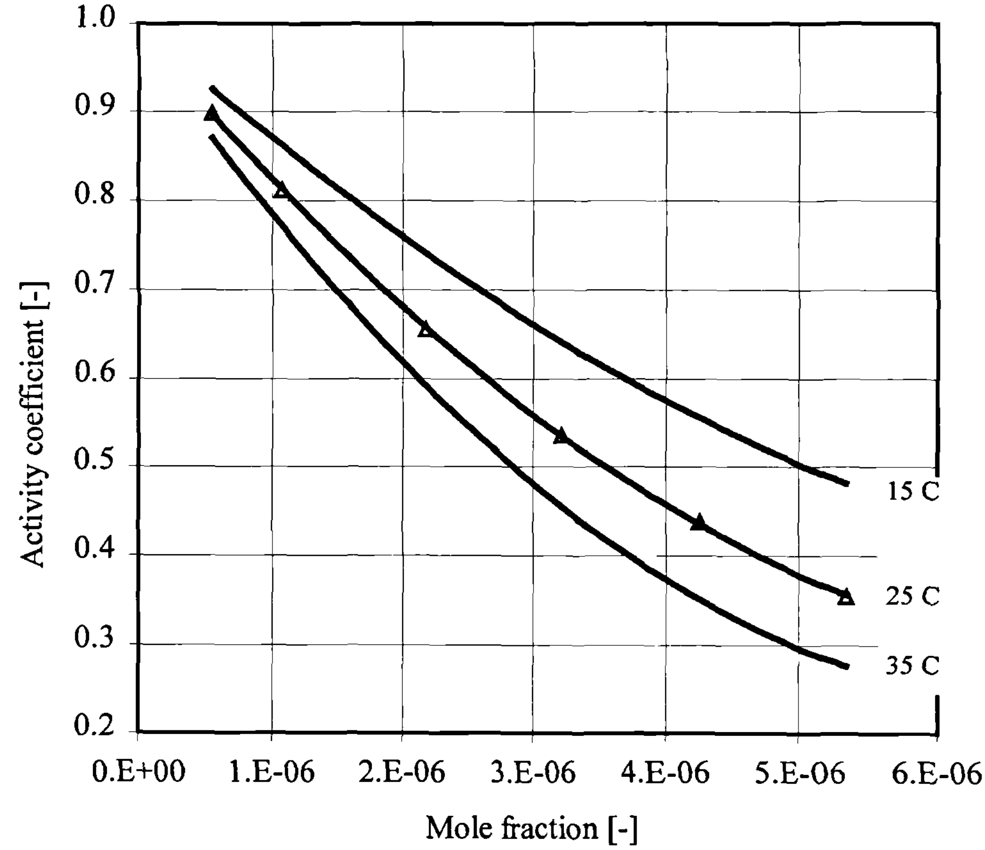
Symbols represent the experimentally determined data. Lines represent model calculations.

Figure 4.3.4: Experimentally determined and modelled (original UNIQUAC unsymmetric) activity coefficients of α - chymotrypsin with varying pH and salt types against protein mole fraction



Symbols represent the experimentally determined data. Lines represent model calculations.

Figure 4.3.5: Experimentally determined and modelled (original UNIQUAC unsymmetric) activity coefficients for α - chymotrypsin at different temperatures against protein mole fraction



Symbols represent the experimentally determined data. Lines represent model calculations.

increases with rising salt concentration until the salting - out region is reached where the opposite solubility behaviour occurs. Ries - Kautt [Ducruix and Giege, 1992] states that generally, at salt concentrations below 0.5 M, the salting - in effect is observed which agrees with the solubility behaviour correlated from the activity coefficients. Comparison at the two higher salt concentrations of 0.15 M and 0.3 M, demonstrates the shift from the salting - in towards the salting - out region where with rising salt concentration a decrease in solubility is observed, figure 4.3.3.

In figure 4.3.4 the activity coefficients and their behaviour with respect to pH and salt types are shown. Comparison of the activity coefficients for the systems with sodium phosphate at different pH demonstrates that at the higher pH higher activity coefficients result than at the lower pH. For the solubility the opposite behaviour is deduced. The solubility is higher at a pH of 5 and lower at a pH of 8.25. It has been established [Bailey and Ollis, 1986] that proteins exhibit their lowest solubility at the isoelectric point. For α – chymotrypsin the isoelectric point is at pH 8.25. The solubility behaviour and activity coefficient behaviour as a function of pH suggests therefore additionally to the examination of salt concentrations that the models are correct.

Figure 4.3.4 also shows the activity coefficients for α – chymotrypsin as a function of two different salt types, potassium sulphate and sodium phosphate. The activity coefficients at a constant pH of 5 are lower for the phosphate anions than for the sulphate anions. A higher solubility for phosphate anions and a lower solubility for sulphate anions results. This implies that the sulphate anion is the more effective precipitation agent of the two anions. This behaviour is in agreement with the results of Hofmeister (1888) on precipitation effectiveness. However, the observed difference is small and compares also well to the result of Shih et al. (1992) who found that the two anions have almost equivalent precipitation effectiveness on a different protein, lysozyme.

For the protein α – chymotrypsin the temperature dependence was predicted with the model parameters established as a result of this work, table 4.3.2. In figure 4.3.5 the experimental data and model performance are given for three temperatures. With increasing temperature the activity coefficients decrease which indicates that the solubility of α – chymotrypsin rises with temperature at given conditions. By lowering the temperature the opposite behaviour is observed. This temperature dependence for solubility is most commonly found from experiments [Ducruix and Giege, 1992] and hence the model prediction tendency observed is confirmed as qualitatively correct.

The experimentally determined activity coefficients were demonstrated to relate to the common solubility behaviour as a function of ionic strength, pH, salt type and temperature. By confirming the experimental data the model was approved, which was shown to follow the experimental data with a deviation of 0.54 %.

4.4 Summary

Protein activity coefficient data has been established here from osmotic pressure measurements for ten protein - salt - water systems. After the development of a prediction method to obtain the structural parameters for globular proteins the determined activity coefficient data was used to obtain a number of new UNIQUAC interaction parameters. These were used to support the estimation and prediction of protein activity coefficients as a function of compound types, composition and temperature. On average a root mean square deviation of 0.54 % resulted when the experimentally determined data and calculated data was compared, showing that the model follows the experimental behaviour. This demonstrated that the original UNIQUAC model, using the interaction parameters generated and the proposed evaluation method for the structural parameters of proteins, is applicable for the systems and compounds of interest.

The systems examined, protein - salt - water systems, cover a wide range of different system conditions and therefore the different effects of these conditions on activity coefficients were studied. Effects of salt concentrations and types, system pH and temperature were examined. The relationship introduced for activity coefficients and protein solubility allowed a qualitative interpretation of protein solubility with respect to these parameters. The fact that the protein solution property, solubility, is described correctly in a qualitative manner through activity coefficients additionally validated the model approach. Having confirmed the model approach and shown correct protein solubility representation through protein activity coefficients encouraged the next part of this work where the quantitative protein solubility description was addressed.

4.5 Nomenclature

A	amino acid segments
A_W	van der Waals area
B	virial coefficient
C	α - chymotrypsin
c	molar concentration
D	virial coefficient
L	β - lactoglobulin

LYS	lysozyme
m	molal concentration
M	molecular weight
N	number of data points
O	ovalbumin
OBJ	objective function
PS	pseudo solvent
q	structural area parameter of a compound
R	gas constant
r	structural volume parameter of a compound
rmsd	root mean square deviation
S	serum albumin
t	temperature
T	temperature
u	interaction parameter
UNIQUAC model	UNIversal QUAsi Chemical model
V_W	van der Waals volume
x	compound mole fraction

Greek letters

Π	osmotic pressure
γ	compound activity coefficient
ν	segment occurrence
$\bar{v}$	partial specific volume
ρ	density

Subscript

A	amino acid segment
cal.	calculated
exp.	experimental
i	compound
j	compound
m	molal concentration
P	protein
PS	pseudo solvent

Superscript

*	unsymmetric
C	combinatorial
R	residual

Chapter 5

Calculation of Protein Solubility

A modelling approach for the description of the solid - liquid equilibrium is introduced and used to represent protein solubility for two different systems as a function of salt concentration and temperature.

5.1 Introduction

It was demonstrated for ten systems that protein activity coefficients are well described with the UNIQUAC model. Furthermore, it was shown that protein solubility can be qualitatively described through protein activity coefficients, chapter four. Therefore, the quantitative modelling of protein solubility was approached next and this part of the work is documented in this chapter. Protein activity coefficients and the solubility product were used to express the solid - liquid equilibrium. Two different globular proteins, lysozyme and concanavalin A, and two salt types, sodium chloride and ammonium sulphate, were examined. Protein solubility for the different systems as a function of the salt type, salt concentration and temperature was investigated using the introduced modelling framework. This work is also presented and discussed elsewhere [Agena et al., 1997a; Agena et al., 1998a].

In the following part of this chapter first the solubility data of the two systems is presented and the protein solubility behaviour of those two systems is examined. The solubility behaviour over composition and temperature is discussed, which introduces the salting - out region, and normal and retrograde solubility behaviour, respectively. Thereafter, the model framework is presented and the results using this modelling approach are discussed. To confirm the model various directions are pursued and various verification approaches are suggested to support the solubility model qualitatively.

5.2 Experimental Solubility Data

Experimental solubility data is necessary in order to model protein solubility but defined and extensive solubility measurements of proteins over various system conditions are rare. Only recently, due to the interest in protein crystallisation, a rapid technique was developed that made solid - liquid equilibrium data for proteins more easily accessible and available.

In 1988 the column solubility method was devised by Pusey and Gernert (1988), which gives solubility measurements from oversaturated and undersaturated protein solutions in the presents of a crystalline phase at constant system conditions within a typical error of 3 % [Pusey and Munson, 1991]. Up to 100 - 200 mg of crystalline protein [communication with Pusey, 1997] are used for a mini - column to pursue the measurements. For two proteins, lysozyme (LYS) and concanavalin A (CON), the solid - liquid equilibrium has been measured using this micro - column technique [Pusey and Munson, 1991].

A selection from these measurements was utilised here for the model studies and is documented in table 5.2.1. The experimental solubility data was screened and converted to the thermodynamic scale, chapter 3.2, and the resulting data is listed in appendix D. Experimental solubility data for aqueous protein systems that are crystallised by monovalent or divalent salts was studied while salt concentrations and temperatures vary. The first system, I, of aqueous lysozyme (M = 14600 g/mol) is a model system and has been studied extensively. The second system, II, introduces concanavalin A with ammonium sulphate as a crystallisation agent, which is the dominant salting - out agent used. With its high molecular weight of 102668 g/mol, concanavalin A is representative of the macromolecular character of proteins. Additionally, the concanavalin A molecule is a dimer allowing for examination of a protein consisting of two subunits, while previously β - lactoglobulin was representatively studied for this characteristic when protein activity coefficients were examined in chapter four.

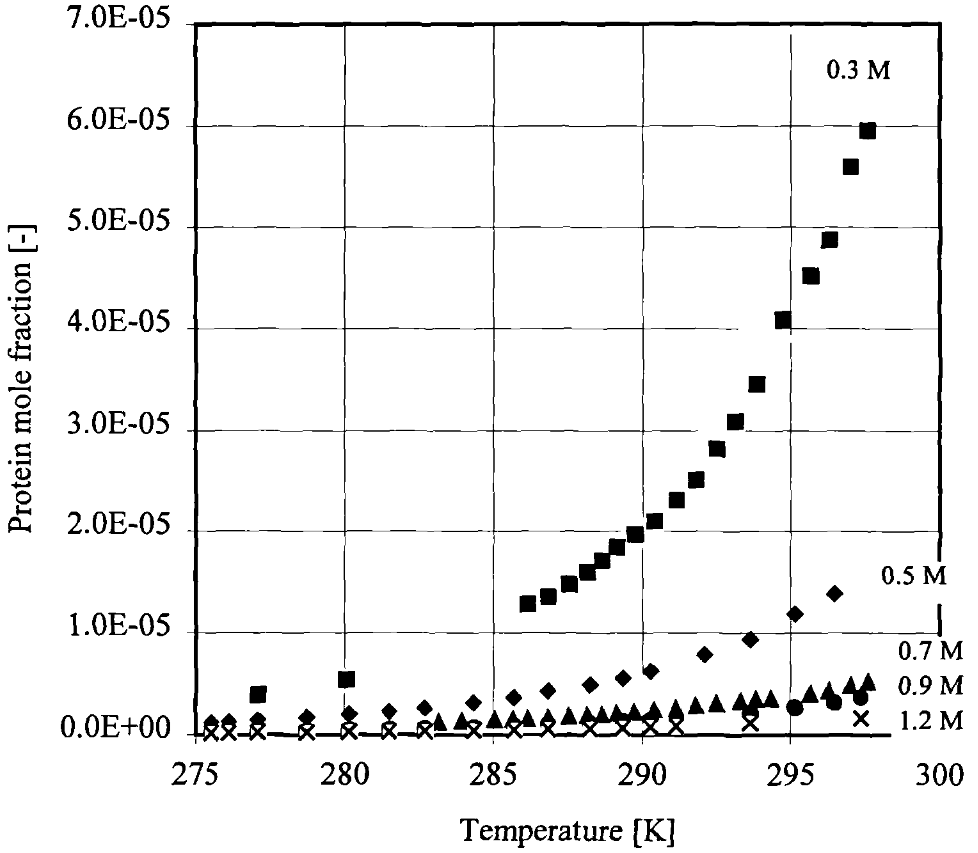
Table 5.2.1: Protein systems and system conditions investigated

System	Protein	M	Salt	Buffer	pH	t
		[g/mol]			[-]	[°C]
I	LYS	14600	NaCl	Sodium acetate	4	2 - 25
	0.2 - 45 g/L		0.3, 0.5, 0.7, 0.9, 1.2 M	0.1 M		
II	CON	102668	$(NH_4)_2SO_4$	tris - acetate	6	18 - 45
	0.3 - 5.5 g/L		0.25, 0.5, 1 M	0.1 M		

The lysozyme (tetragonal crystal form) data had been obtained at a constant pH of 4 in 0.1 M sodium acetate buffer, while temperature and sodium chloride concentration were varied [Cacioppo and Pusey, 1991]. The solubility behaviour as a function of five salt concentrations from 2 % - 7 % (0.3 - 1.2 M) and in a temperature range from about 2 - 25 °C had been obtained. For the second system, concanavalin A (acid treated and recalcified), the solubility had been investigated using ammonium sulphate as a salting - out agent [Cacioppo and Pusey, 1992]. Solubility data had been determined for concanavalin A over three salt concentrations, 0.25, 0.5 and 1 M, and was examined here over the measured temperature range of 18 - 45 °C at pH 6 in 0.1 M tris - acetate buffer.

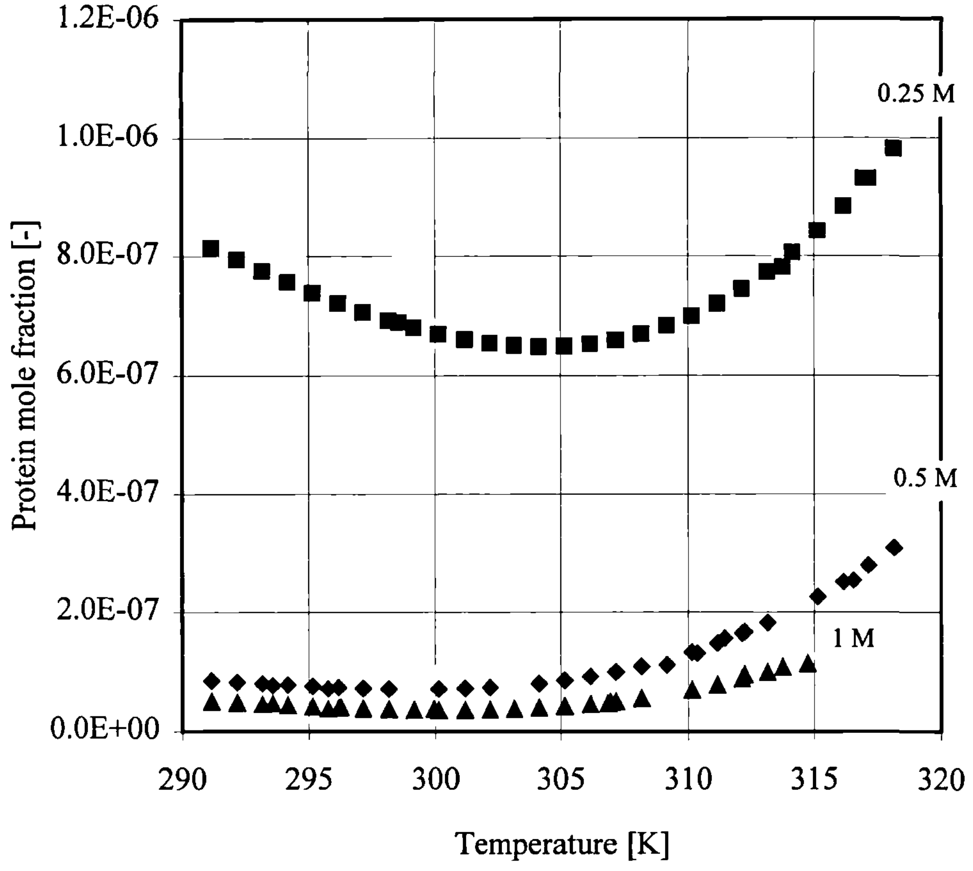
The chosen solubility data represents normal and retrograde solubility behaviour and the salting - out region. In figures 5.2.1 and 5.2.2 the experimental data of the two systems is shown. Figure 5.2.1 shows the solubility of lysozyme as a function of temperature and sodium chloride concentration. A normal solubility, i.e., increasing with temperature, is observed. Concanavalin A, figure 5.2.2, shows the opposite behaviour. At low temperature ranges retrograde solubility, i.e., decreasing with temperature, is observed, which is thought to occur when ammonium sulphate is applied [Jakoby, 1968]. However, at higher temperature ranges normal solubility behaviour is also found for the concanavalin A system. In all systems the salting - out

Figure 5.2.1: Experimental solubility of lysozyme at various sodium chloride concentrations against temperature



Symbols represent the experimentally determined data.

Figure 5.2.2: Experimental solubility of concanavalin A at various ammonium sulphate concentrations against temperature



Symbols represent the experimentally determined data.

region was studied, which is demonstrated by a solubility decrease with increasing salt concentration.

5.3 Solid - Liquid Equilibrium Model

To model protein solubility the solution behaviour and the exchange of liquid and crystal protein between the liquid and solid phase needs to be described. The transfer of protein between the two phases was represented by the solubility product, which is a function of temperature while the solution behaviour, i.e., deviation from ideal solution behaviour, was accounted for and described by the previously examined UNIQUAC model, chapter four. In figure 5.3.1 the solubility model which consists of these two terms is shown. These terms are the phase equilibrium and solution activity coefficient description, which are discussed below. A programming code for the solubility model including the parameter determination routine and solubility calculation routine was developed as part of this work and is given in appendix E.

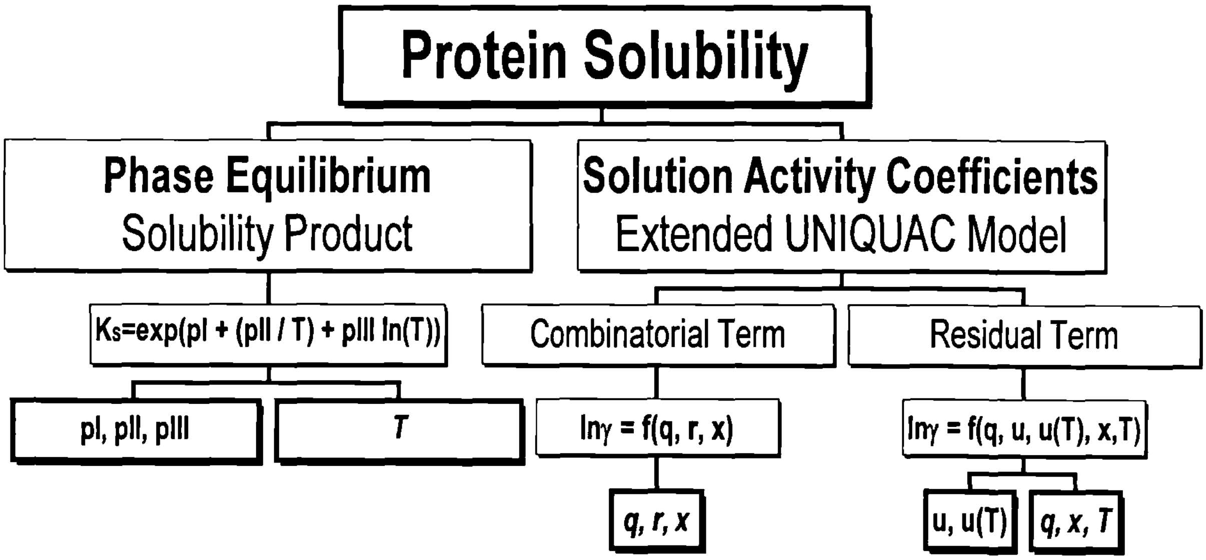
The phase equilibrium term, figure 5.3.1, is introduced to describe the exchange of crystalline and liquid protein between the two phases, i.e., the solid and liquid phase, respectively:

$$\Pr otein^{solid} \overset{Ks}{\rightleftharpoons} \Pr otein^{liquid} \qquad (5.3.1)$$

This exchange of protein between two phases can be represented by the solubility product, Ks, with:

$$Ks = x_p \cdot \gamma_p^* \qquad (5.3.2)$$

Figure 5.3.1: Structure of the solubility model



pI, pII, pIII: solubility product parameters; T: temperature; γ: activity coefficients; q and r: structural parameters; x: mole fraction; u: interaction parameters

This equation, 5.3.2, relates the solubility product to the mole fraction of liquid protein, x_P , which is the protein solubility. Furthermore, the liquid activity coefficient of the protein, γ_P* on the unsymmetric convention $\gamma^* \rightarrow 1$ as $x_P \rightarrow 0$ is introduced in equation 5.3.2. The solubility product, Ks, is represented by the following expression:

$$\ln Ks = pI + \frac{pII}{T} + pIII \cdot \ln(T) \qquad (5.3.3)$$

The solubility product is a function of temperature, T. The adjustable parameters, pI, pII and pIII, describe the solubility product and the given expression is a widely and successfully applied description for the solubility product [Mullin, 1993, and references within].

The solution activity coefficient term, figure 5.3.1, is introduced to describe the protein activity coefficient introduced in equation 5.3.2. For the calculation of activity coefficients the extended UNIQUAC model introduced by Sander et al. (1986) was used. This model relates closely to the original UNIQUAC model, which was shown as part of this work to describe protein activity coefficients, chapter four. In this model the excess Gibbs energy and therefore the activity coefficients are likewise represented by a set of two terms, a combinatorial term and a residual term:

$$\ln\gamma_i = \ln\gamma_i^C + \ln\gamma_i^R \qquad (5.3.4)$$

where

$$\ln\gamma_i^C = f\left(q_i, r_i, x_i\right) \qquad (5.3.5)$$

$$\ln\gamma_i^R = f\left(q_i, x_i, u_{ij}, T\right) \qquad (5.3.6)$$

The combinatorial term, $\ln\gamma^C$, describes the entropy state of solution and is the combinatorial term introduced for the original UNIQUAC model, chapter 2.3.1. The composition of the system is introduced with the mole fraction, x, while the structural parameters, q and r, represent the size and shape of the compounds accounting for the differences of small molecules and macromolecules. The structural parameters of the studied compounds are summarised in table 5.3.1. For the salt ions the structural parameters were obtained from literature [Sander et al., 1986; Pessoa et al., 1992]. In the case of the sodium ion the structural parameters were calculated from the ionic radius as given by Pauling (1960). The structural parameters of the proteins were calculated from the protein molecular weights using the method developed as a result of this work, chapter 4.3. For the water related compounds, i.e., the pseudo solvents, the established structural parameters of water were applied [Abrams and Prausnitz, 1975].

Table 5.3.1: Structural parameters of proteins, salt ions and pseudo solvents

Compound	r [-]	q [-]
Lysozyme	529	399
Concanavalin A	3717	2803
Na^+	0.1425	0.2731
Cl^-	0.9861	0.9917
NH_4^+	0.9097	0.9800
SO_4^{2-}	2.3138	1.3600
PS 1	0.92	1.4
PS 2	0.92	1.4

Two pseudo solvents, PS, were introduced as two different, protein buffers were applied over the two systems: PS 1 and PS 2 referring to system I and II, respectively. Effects of buffer concentration and therefore type have been shown to markedly affect protein solubility [Forsythe and Pusey, 1996] and were hence incorporated by creation of a pseudo solvent. The introduction of pseudo solvents has been established for the previous calculations on protein activity coefficients, chapter four, and was therefore adapted again. This concept was applied as it reduces the number of compounds per system and therefore the number of parameters needed. This is especially of interest with respect to the binary interaction parameters which have to be determined and their number rises with the number of system compounds, equation 2.4.10. Minimisation of system compounds reduces the number of parameters to be determined but it also reduces the predictive power of the model. Introduction of the solvent and buffer as independent compounds would have allowed for predictions over buffer concentration. However, the data used here is constant over the buffer concentration and therefore a prediction over the buffer concentration would not have been advisable even if possible.

While the combinatorial term is that of the original model the residual term, $\ln\gamma^R$, was extended by Sander et al. (1986) and resembles that also introduced with the modified model, equation 2.6.3. However, for this work only the first two temperature terms were introduced for the interaction parameters, u:

$$u_{ij} = u_{ij}^{\circ} + u_{ij}^{\prime} \cdot (T - 300) \qquad (5.3.7)$$

Termination after the second term leads to less secondary interaction parameters per binary interaction parameter, which was an objective as less parameters had to be determined. To describe the behaviour of the studied four compound systems, protein - cation - anion - pseudo solvent, the secondary binary interaction parameters, u° and u^t had to be determined.

5.4 Study of the Solubility Model

In order to model protein solubility for the two systems the values of the solubility product parameters and the interaction parameters had to be determined. The program routine in appendix E was designed for this purpose and performs similar to the one described in chapter 4.3. An Indigo2 Impact (195 MHZ IP28 Processor, CPU: MIPS R10000 Processor Chip Revision: 2.5, Secondary unified instruction/data cache size: 1 Mbyte, Main memory size: 64 Mbytes) performing at 124.4 Whetstone MIPS was used for the parameter determination and solubility calculation. The experimental solubility data was used to guide the determination while the thermodynamic functions, equations 5.3.2, 5.3.3 and 5.3.4, were estimated, minimising the following objective function:

$$ssd = \sum_{1}^{N}\left(\frac{x_{\exp} - x_{cal}\left(K_S, \gamma^*\right)}{x_{\exp.}}\right)^2 \qquad (5.4.1)$$

The multivariable non - linear optimisation technique of conjugate directions by Powell [Press et al., 1986] was adopted to minimise the standard square deviation, ssd. The deviation of experimental solubility, $x_{exp.}$, and modelled solubility, $x_{cal.}$, over all data points, N, was minimised. The solubility is calculated as a function of the solubility product, Ks, and the protein activity coefficients, $\gamma*$, while the solubility product parameters and interaction parameters are optimised.

From the optimisation procedure the parameters, pI, pII and pIII, of the solubility product resulted. The parameters determined as a result of this work are listed in table 5.4.1. For each of the two systems one crystal form was represented, leading to one solubility product per system over the temperature range. In the case of lysozyme it has been established that only the tetragonal crystal form occurs under the studied conditions [Cacioppo and Pusey, 1991; Ewing et al., 1994]. Calculations at higher temperatures would require the solubility product relating to the orthorhombic crystal form of lysozyme. For the other system, concanavalin A, it has not been shown that

Table 5.4.1: Parameters of the thermodynamic solubility product for lysozyme and concanavalin A

Solid phase	pI	pII * 10^{-3}	pIII	temperature range [°C]
Lysozyme	17.9891	-3.966148	-2.3381	2 - 25
Concanavalin A	-2289.08	113.390	333.270	18 - 45

only one crystal form exists under the conditions studied but experimental indication for various crystal forms have not been found during the experimental solubility studies [communication with Pusey, 1996].

In addition to the solubility product parameters, the binary interaction parameters were determined. For the studied systems the interaction parameters are given in tables 5.4.2 and 5.4.3. For the salt total dissociation was described leading to two ions and producing a four compound system. This approach was taken for the salt as previous examinations of the model with a non dissociating salt and therefore a three compound system did not succeed. Deviations were by a two or three - fold higher than found for modelling approaches with a dissociating salt, i.e., four compound system. As ions show different impacts on the crystallisation process as discussed later this concept is supported. Furthermore, representation of the salt as ions is a realistic representation as this occurs in solution. However, the three compound systems was examined in order to reduce the number of interaction parameters as discussed before. Still, as a result of this approach systems consisting of four compounds - protein, cation, anion, pseudo solvent- were modelled and up to twelve binary interaction parameters were determined. Insignificant interaction parameters were indicated during the optimisation process. An independence of the parameter values and the calculated solubility was observed. These interaction parameters did

Table 5.4.2: Binary interaction parameters for the lysozyme system

$u^{\circ}_{i,j}$ [K]	PS 1	Lysozyme	Na^+	Cl^-
PS 1	-3210.69	-3217.27	-3763.78	-3723.26
Lysozyme		0	0	-3543.58
Na^+			0	0
Cl^-				0
$u^{t}_{i,j}$ [-]	PS 1	Lysozyme	Na^+	Cl^-
PS 1	0	0.0123	0	2.1818
Lysozyme		0	0	4.3890
Na^+			0	0
Cl^-				0

Table 5.4.3: Binary interaction parameters for the concanavalin A system

$u^{\circ}_{i,j}$ [K]	PS 2	Concanavalin A	NH_4^+	SO_4^{2-}
PS 2	53.10	10.18	-401.07	21.48
Concanavalin A		1012.89	20.12	619.05
NH_4^+			0	0
SO_4^{2-}				0
$u^{t}_{i,j}$ [-]	PS 2	Concanavalin A	NH_4^+	SO_4^{2-}
PS 2	0	0.3977	-1.2626	4.6523
Concanavalin A		0	5.4760	6.8900
NH_4^+			0	0
SO_4^{2-}				0

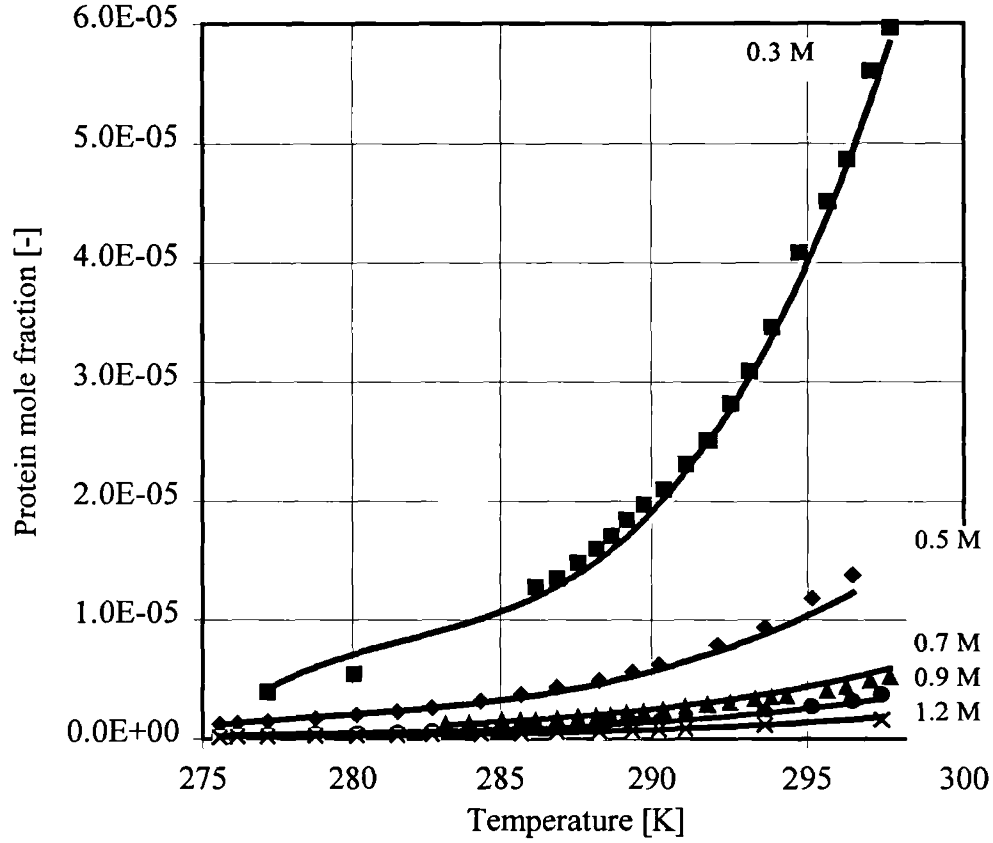
not seem to be of significance for the solubility process and were therefore adjusted to values of zero indicating no interactions. This approach is further discussed and related to the findings of other researchers.

For system I, consisting of lysozyme and sodium chloride, a number of binary interaction parameters are not relevant. The lysozyme - lysozyme interactions and ion - ion interactions are of minor importance. The interactions of the pseudo solvent and chloride with the other compounds are predominant, table 5.4.2. The strong effect of the chloride anion and low impact of the sodium cation reflected in the model has also been noted by Taratuta et al. (1990). They demonstrated experimentally that cation substitution had no effect on the lysozyme coexistence curve while anion substitution had. Likewise, the greater impact of anions over cations on protein solubility was shown by Hofmeister (1888) and recently confirmed by Carbonnaux et al. (1995). While this behaviour was proven through experiment by various researchers it was here for the first time, to the best of my knowledge, shown by theory and the same applies for the next case. For the second system of concanavalin A more than eight interaction parameters had to be determined. While the ion - ion interactions are also irrelevant the cation certainly has an impact and is not to be neglected as in the lysozyme system. This relates to Jakoby's finding (1968), who demonstrated that ammonium sulphate leads to the retrograde solubility behaviour. Furthermore, it implies that ammonium and sulphate ions and not only e.g. the sulphate ions induce the retrograde solubility behaviour.

With the determined parameters protein solubility was calculated as a function of temperature and salt concentration. In figures 5.4.1 and 5.4.2 the modelled and experimental solubility data is shown. These figures demonstrate qualitative agreement of model and experimental behaviour. Correspondingly, the quantitative comparison gives good results. Comparison of the experimental solubility data and the model calculations gives a root mean square deviation of 7 % for lysozyme, table 5.4.4. This is slightly above the deviation range, 3.9 - 5.7 %, reported by Cacioppo and Pusey (1991) for their polynomial correlations over the temperature range alone. The deviation of the solubility model, however, included deviations due to temperature and composition. For the concanavalin A system the deviation of experimental and modelled solubility as a function of salt concentration and temperature is 4.5 %. Overall an average deviation of 5.8 % resulted for the systems when the model calculations and experimental solubility data was compared showing that the model follows well the experimental behaviour of two different systems.

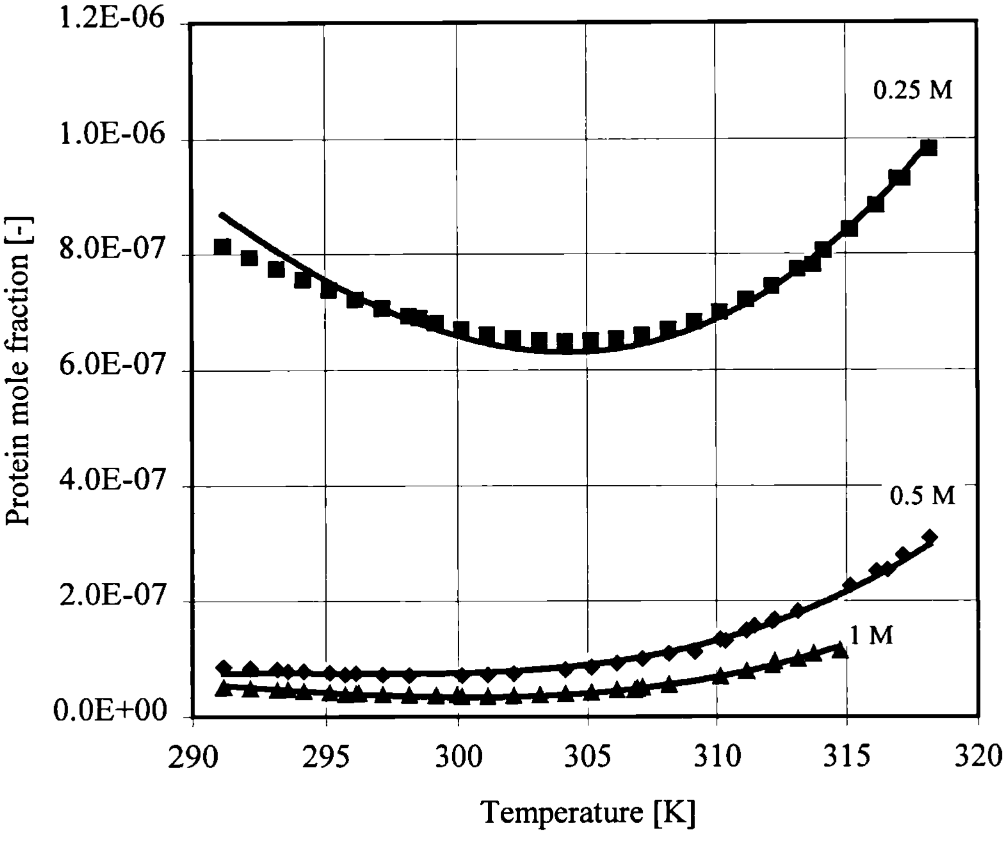
For the two systems the activity coefficient calculations were also examined and used as constraints to direct the applied framework and secure its validity. In figure 5.4.3

Figure 5.4.1: Experimental and modelled solubility of lysozyme at five different sodium chloride concentrations against temperature



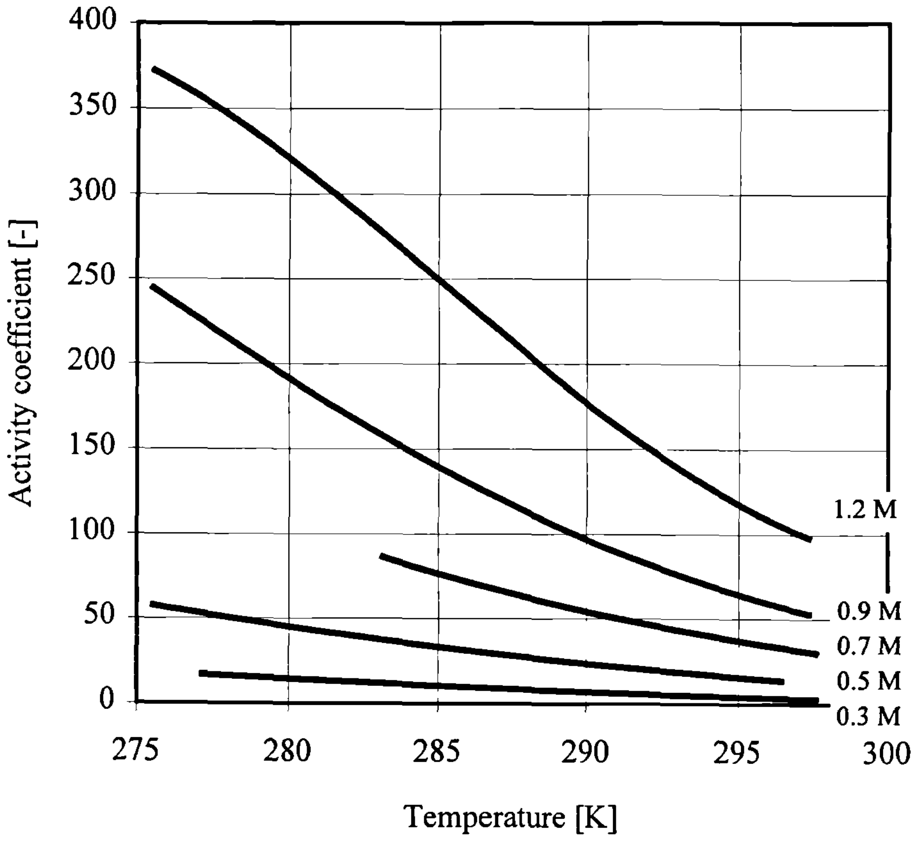
Symbols represent experimentally determined data. Lines represent model calculations.

Figure 5.4.2: Experimental and modelled solubility of concanavalin A at different ammonium sulphate concentrations against temperature



Symbols represent experimentally determined data. Lines represent model calculations.

Figure 5.4.3: Modelled activity coefficients of lysozyme at different sodium chloride concentrations against temperature



Lines represent model calculations.

Table 5.4.4: Performance of solubility calculations for lysozyme and concanavalin A using the solid - liquid equilibrium model

System	N	ssd, objective function value	rmsd [%]
Lysozyme	93	0.45	7.0
Concanavalin A	93	0.19	4.5
Average	93	-	5.8

the calculated activity coefficients of lysozyme against temperature and the salt concentrations are given. The activity coefficients show a 2 fold decrease over the rising temperature range while a 20 fold increase is found over the increasing salt concentration range indicating the impact of these factors on protein solubility for the studied cases.

The qualitative behaviour of the activity coefficients, figure 5.4.3, relates correctly to the experimental solubility behaviour following the fact that activity coefficients are inversely proportional to the solubility at constant temperature, equations 5.3.2 and 5.3.3. The activity coefficients increase with an increase in salt concentration indicating the opposite solubility behaviour, i.e., salting - out behaviour, as found by experiment. The correct solubility behaviour is also deduced with respect to the temperature dependence. This indicates that activity coefficients of proteins represent qualitatively correctly the protein solubility behaviour but also that the model constraint for the activity coefficient calculations was obeyed.

Still, it could not be validated that the activity coefficient results are quantitatively correct. Similar magnitudes for activity coefficients were reported by Ross and Minton (1977) for haemoglobin. They reported values for activity coefficients from about 1 to about 580 over a concentration range of 20 - 400 g/L and comparable quantitative results were obtained for activity coefficients from this model but system conditions are different to the ones studied by Ross and Minton (1977). Therefore, the quantitative results for the protein activity coefficients were not validated while likewise not proven wrong. However, model consistency was additionally shown by means of the qualitative behaviour of the intermediate solution property, protein activity coefficient.

5.5 Summary

It has been demonstrated that the chosen model approach represents protein solubility for two globular proteins of very different size, lysozyme and concanavalin A. Their solubility as a function of added salt concentration and system temperature was studied. The salting - out region, normal and retrograde solubility behaviour of these proteins was modelled. For each of the four compound systems - protein, cation, anion, pseudo solvent - a the minimal number of parameters was determined as part of this work by setting insignificant interaction parameters equal to zero. Using the determined parameters the solubility over the given conditions was calculated and an average deviation of 5.8 % resulted for the proposed model approach when compared to the reference solubility data. To examine the model framework in more detail additionally the calculated protein activity coefficients were viewed and showed to correlate qualitatively correct to the experimental solubility behaviour. Furthermore, some of the determined parameters confirmed the model while likewise for the first time experimental findings were validated using this theoretical approach.

5.6 Nomenclature

CON	concanavalin A
Ks	solubility product
LYS	lysozyme
M	molecular weight
N	number of experimental data points
pI	solubility product parameter
pII	solubility product parameter
pIII	solubility product parameter
PS	pseudo solvent
q	structural area parameter of a compound
r	structural volume parameter of a compound
rmsd	root mean square deviation
ssd	standard square deviation
t	temperature
T	temperature
u	interaction parameter
UNIQUAC model	UNIversal QUAsi Chemical model
x	compound mole fraction

Greek letters

γ	compound activity coefficient

Subscript

cal.	calculated

exp.	experimental
i	compound
j	compound
P	protein

Superscript

*	unsymmetric
C	combinatorial
R	residual
t	temperature

Chapter 6

Conclusions

This chapter summarises the contributions and conclusions of the presented research project.

1. In this work for the first time systematic studies of activity coefficient predictions were pursued and documented for amino acids and peptides using the predictive power of the UNIFAC method and the parameters of Hansen et al. (1991). To study the predictive capacities of this method the UNIFAC groups, model parameters, model versions and activity coefficient reference states were examined. Amino acids and peptides were studied assuming that a model that is successful for these compounds will also be successful for proteins and vice versa. This was assumed as the first set of compounds, amino acids and peptides,

are the building blocks of proteins and therefore resemble many of a protein's characteristics such as interactions but not size.

2. For the first time the most extensive and recent UNIFAC group and parameter database of Hansen et al. (1991) was applied to describe activity coefficients of amino acids and peptides using the UNIFAC model. It was demonstrated with this work that the group definitions are not appropriate for peptides and therefore proteins. A molecular representation of the peptide bond occurring between two different amino acids in an amino acid sequence, i.e., peptide or protein, is missing. A description of a peptide's and protein's backbone is lacking. This was a first indication arising from this work suggesting that the UNIFAC model might be at present limited in its predictive powers when aiming at amino acids, peptides and proteins. Compounds closely related to these were not used to establish the model parameters and therefore a proper description of their properties is less likely to occur.

3. The examination of the database of Hansen et al. (1991) was taken further in this work. It was shown that certain binary interactions, which typically occur for amino acids, peptides and proteins, are not accounted for. Interactions of the amino terminus with the carboxyl terminus and the carboxyl groups of the acidic

residues, aspartic acid and glutamic acid, are not established. However, these molecular groups are very likely to interact when in the vicinity of each other as they carry opposing polarities. Therefore, it was demonstrated with this work that the predictive abilities of the UNIFAC model using the database of Hansen et al. (1991) are limited for the systems aimed at, i.e., protein systems, due to incomplete interaction descriptions.

4. Furthermore, two different UNIFAC models, the original and modified one, were investigated in this work. The ability of these two models to predict activity coefficients of protein related compounds, i.e., amino acids and peptides, was examined. A poor predictive performance resulted for both models. Both the original UNIFAC and the modified UNIFAC model performed unsatisfactorily on the symmetric reference scale. Comparison of the computed activity coefficients and those determined from experiment showed qualitative disagreement and quantitative differences of around a 100 %. This indicated that possibly the model approaches are not appropriate for the studied systems and for related systems such as protein systems. However, no clear conclusion could be drawn as too many additional factors such as possibly incorrect group definitions and missing interaction parameters influenced the model performance. This indicated that a model such as the UNIQUAC model, i.e., a model with no group

contribution method, which therefore uses less parameters was needed to clarify the performance of the model approach.

5. Moreover, for the first time the two activity coefficient reference states, Raoult's law and Henry's law, were studied for amino acid and peptide activity coefficient calculations when using the original UNIFAC model. As a result of this work it was demonstrated that the unsymmetric reference state, i.e., Henry's law, should be used for amino acids and peptides and therefore possibly also for proteins following the previously suggested hypothesis. The conclusion to refer to Henry's law arose not only from this work but was also shown independently when protein adsorption experiments were pursued by Fraaije et al. (1991). They demonstrated that deviation from ideal behaviour has to be expressed with respect to Henry's law for proteins, which agrees with the result of this work and confirms the hypothesis that amino acids, peptides and proteins relate to some extent. Using two completely different approaches the necessary reference state was defined. This part of the work was significant as it demonstrated that the unsymmetric convention is to be applied for further modelling work with proteins and related compounds.

6. Overall an essential contribution was made with this work as it showed that the

UNIFAC model and the most recently established groups and parameters of Hansen et al. (1991) are not applicable to amino acid, peptide and protein systems even though compounds related to biotechnology had been examined by Hansen et al. (1991). The predictive power of the UNIFAC model with the suggested parameters failed for the studied systems, which was demonstrated by inappropriate predictions. However, this part of the work did not rule out the UNIFAC model or its model base for the examined compounds but it demonstrated the inapplicability of the UNIFAC model with the parameters of Hansen et al. (1991).

7. Moreover, it was illustrated in this work that the UNIFAC model is not yet applicable as too little experimental data is available to establish a reliable model. This was discussed using the results of Pinho et al. (1994). Due to the model's group contribution approach more parameters need to be determined which requires more experimental data then generally available for the systems targeted in this work. The related UNIQUAC model overcomes this problem. It uses less parameters and therefore requires less experimental data which is a limiting factor and therefore of importance for this work.

8. Furthermore, it was argued in this work that the UNIQUAC model is the

appropriate model to study the model approach. It was pointed out in this work that the UNIQUAC model is more easily interpreted for its performance then the UNIFAC model as fewer parameters are required and no group information is introduced. Therefore, sources for model deviations are reduced for the UNIQUAC model when compared to the UNIFAC model. Following this, closer model approach examinations are possible using the UNIQUAC model as deviation due to those other factors are screened out. Hence, further research work focused on the UNIQUAC model as correlations using the UNIFAC model seemed less favourable even though the predictive power for different but related compounds is lost using the UNIQUAC model.

9. Further model investigations pursued as part of this work focused on the UNIQUAC model as model approach examinations were a primary objective. As the UNIQUAC model and the UNIFAC model use the same model approaches it can be assumed that any confirmation of the UNIQUAC modelling approach for protein systems is also extendible to the UNIFAC model. However, to transfer any successful model approach from the UNIQUAC model to the UNIFAC model the additional group contribution method would need to be addressed. Still, as a first step it was necessary to confirm the model approach and this was achieved with the examinations pursued in this work. In order to study the

UNIQUAC model interaction parameters were determined to describe the activity coefficients of glycine, proline and glycylglycine in water. Seven new parameters resulted from this work and therefore allowed for activity coefficient calculations.

10. Following the establishment of these parameters the original UNIQUAC model was applied on the unsymmetric scale. The unsymmetric scale was applied due to the findings that resulted from this work. Computation of activity coefficients for two amino acids, glycine and proline, and one peptide, glycylglycine, were successful when using the newly determined interaction parameters and the UNIQUAC model. With this work it was shown that on average deviations as low as 0.2 % (in the worst case up to 0.3 %) would result for the examined systems and the model approach used when comparing calculated activity coefficients and experimentally determined ones. This implied that the modelling approach is correct and that the same one should be successful for the UNIFAC model as both models have a common theoretical base. Furthermore, this part of the work implied that the model should also be applicable for proteins, the target compounds, as these are closely related to the examined amino acids and peptide. With this part of the work a possible protein modelling approach was established, which was a first significant step towards the modelling of protein properties.

11. Additionally, ~~an experimentation~~ a bench mark error was established from this work for activity coefficients in order to create an indicator for the model performance of activity coefficient models. An average deviation of 1.8 %, i.e., ~~experimentation~~ bench mark error, resulted for activity coefficients from two different sources when compared over the composition for three different protein related compounds. This error of 1.8 % for two different sources, i.e., laboratories, gave a measure for model performances, i.e., the model deviations. With ~~an experimentation~~ a bench mark error of 1.8 % for activity coefficients over composition it was established that similar errors should result from the models indicating that the models perform properly and follow the experimentally determined behaviour.

12. Comparison of the ~~experimentation~~ bench mark error of 1.8 % and the average UNIQUAC model deviation of 0.2 %, demonstrated that the model performed very well for amino acid and peptide activity coefficients as the deviation is well below the ~~experimentation~~ bench mark error. This work showed that the model described the experimentally determined activity coefficient behaviour very well for the studied systems and compounds. As these compounds relate to proteins a model performance comparable to this one is expected for proteins.

13. For the first time for ten different systems and four different proteins, serum

albumin, α - chymotrypsin, β - lactoglobulin and ovalbumin, activity coefficient data was made available to pursue this research work further. The data was obtained from osmotic pressure measurements documented in the literature. Protein activity coefficients were made available over varying system composition and over various system conditions such as compound types, temperature and pH using virial expansion. Protein property data over a variety of conditions resulted from this work. To obtain this data was a significant contribution as no such comprehensive data was previously available for proteins. This data allowed for a detailed and systematic examination of protein property models.

14. As a next step in the pursuit of this work the structural parameters for proteins had to be obtained before any UNIQUAC model calculations could be conducted. A new method to derive these parameters was developed as part of this work. Protein molecular weights were introduced to determine the structural parameters of proteins and relate to the research results of Janin (1976) and Teller (1976). The new method was approved by comparison to experimental results. It allows for the calculation of the structural parameters for proteins without the precise knowledge of a protein's molecular structure which is a major improvement in comparison to the previous method. Furthermore, the new method is less time

consuming and erroneous. The new method gave for the prediction of protein structural parameters a deviation of 5 %. Moreover, the method confirmed that pure compound properties can be predicted for some proteins, which is a result of significance and encouragement when aiming at the prediction of protein mixture and solution properties. Furthermore, the developed method is an important contribution as it makes the structural parameters quickly and easily available just as needed for e.g. engineering purposes.

15. As a further result of this work it was presented that the UNIQUAC model performs less well for the calculation of protein activity coefficients when the structural parameters deviate from those determined with the new method. Protein structural parameters that were by up to 25 % higher than the ones computed from the new method gave a ten - fold higher deviation for activity coefficient calculations. This showed how important the detailed examination of the structural parameters was, which was carried out as part of this work.

16. To pursue the protein activity coefficient calculations it was essential to establish interaction parameters for protein containing systems. This was achieved in this work for ten different systems and per system three parameters resulted. For systems with serum albumin, α - chymotrypsin, β - lactoglobulin and ovalbumin

parameters were made available for the first time. For some of the systems the same protein - protein interactions were applied as the same system conditions prevailed, i.e., same pH and temperature. The determined parameters allow now for model calculations and model evaluations, which were the first ones of their kind to be pursued for proteins.

17. With the newly established protein structural parameters and determined interaction parameters the protein activity coefficients were calculated and compared to the previously determined ones. On average a deviation of 0.54 % resulted, which is well below the ~~experimentation~~ bench mark error of 1.8 %. The low deviation indicates that protein activity coefficients are correctly described. It was shown with this part of the work that the original UNIQUAC model on the unsymmetric scale over varying composition performed very well when protein mass fractions are below 5 %. This result was of importance as it pointed out that first of all the assumption that models which perform well for amino acids and peptides do so also for proteins was correct in the systems studied as average errors of similar magnitude resulted. Moreover, it was demonstrated that the model approach used is accurate and describes protein solution behaviour for the ten systems studied but would most likely also perform well for many other protein systems. A significant contribution was made with this part of the work as

a modelling base for protein mixture and solution properties was found.

18. As part of this work the protein activity coefficient behaviour was examined and was verified as correct over changes in salt concentration, pH and salt type when referring to the common solubility behaviour as documented by various other researchers [Bailey and Ollis, 1986; Ducruix and Giege, 1992]. Model predictions over system temperature correlated likewise qualitatively correctly to the common protein solubility behaviour and confirmed the modelling approach. With this part of the work it is was shown that model calculations and predictions are quantitatively and qualitatively correct. A main contribution was made by demonstrating that protein activity coefficients for ten different protein containing systems are successfully described over a wide range of system conditions. It showed that not only a wide variety of protein containing systems but also a wide variety of system conditions are modelled correctly using the introduced approaches. This implies that possibly also different protein containing systems over various system conditions can be modelled.

19. Furthermore, for the first time it was discussed and demonstrated that protein activity coefficients bear solubility information and are able to indicate the protein solubility behaviour with respect to changes in system composition, pH

and temperature. This kind of information is highly valuable and useful for both optimal process development and protein crystal growth where either a property of different compounds at same system conditions is compared or a property of one compound under different system conditions is compared for evaluation purposes. These comparisons can be pursued on a quantitative scale using a secondary property that relates directly to the needed property as demonstrated with protein activity coefficients and solubility. Furthermore, this work demonstrated that solubility can be approached using a different set of experimental measurements, which is valuable as possibly crystallisation conditions might not be known for a protein and therefore solubility measurements cannot be performed in the near future. Here osmotic pressure measurements can possibly be pursued instead of solubility measurements and analysis of the data as presented in this work, should lead to the solubility behaviour. This approach is of relevance as osmotic pressure measurements are in general easier and quicker obtainable than solubility measurements.

20. With this work it was indicated that solubility is qualitatively correct described through activity coefficients and that the examined models obeyed this law. As a consequence the modelling approach was extended aiming at a quantitative solubility description. The protein solubility of lysozyme and concanavalin A as a

function of salt concentration and temperature was modelled. The solubility product parameters and interaction parameters of these systems were therefore determined as part of this work to allow for the quantitative modelling of lysozyme and concanvalin A solubility. Parameters for both systems were obtained for the first time.

21. As a further result of this work it was indicated from the model that for the lysozyme system anion interactions with the protein and solvent are dominant over those of the cation with these compounds as reflected in the interaction parameters. This modelling result was consistent with experimental coexistent measurements by Taratuta et al. (1990) and was also confirmed with respect to Hofmeister's results (1888), which had recently been verified by Carbonnaux et al. (1995). In this work a very different approach was used and for the first time, to the best of my knowledge, a theoretical approach confirmed those experimental findings.

22. Moreover, this work was able to propose for the concanavalin A system that the cation and anion interactions were both important as opposed to the results for the lysozyme system. The result for the concanavalin A system related well to the experimental findings of Jacoby (1968). The model result indicated that both the

ammonium and sulphate ion induce the retrograde solubility behaviour as opposed to only one of the two ions. Again the theoretical approach taken here verified the experimental findings of Jacoby and vice versa. Furthermore, the theoretical result of this work clarified the role of the sulphate ion and ammonium ion. A significant contribution was made with this work as Jacoby's work did not demonstrate that both ions, the ammonium and sulphate ion, were necessary to induce the retrograde behaviour.

23. This work demonstrated for the first time that protein solubility can be modelled accurately using the semi - empirical approach. The solubility of lysozyme and concanavalin A was modelled with an average deviation of 5.8 % over salt concentration and temperature using the solubility model. This model deviation compares very well to that of Cacioppo et al. (1991), who found deviations of up to 5.7 % from polynomial descriptions over temperature alone for the same data. Therefore, this work showed that protein solubility can be successfully modelled for two very different and complex systems. The lysozyme system was modelled over five different salt concentrations and not only the normal but also the retrograde solubility behaviour was described for the concanavalin A system. Furthermore, a rather big protein of around 100,000 g/mol was examined with concanavalin A, which also represents dimers. All, this leads to the conclusion

that also other globular proteins can be successfully described with the solubility model proposed in this work when the salting - out range and normal and retrograde solubility behaviour are sought. Furthermore, this result confirmed the previous statement that arose from the protein activity coefficient modelling work, which stated that other proteins and properties should be likewise successfully described. Just this was shown here by modelling not only activity coefficients but also solubility for proteins and is another significant contribution of this research project.

24. Furthermore, it was shown as a result of this work that protein solubility needs to be represented using a dissociating salt, i.e., ions, as otherwise deviations of a three or four - fold higher result than for the model, which was finally presented. This coincides with the general approach taken for simple electrolyte systems where the salts are represented by the ionic compounds they are build off. However, the approach to represent the salt as an entity had never been examined but failed for the solubility modelling attempts pursued here. Therefore, a four compound system - protein, anion, cation, pseudo solvent - was described instead of a three compound system - protein, salt and pseudo solvent. This certainly more closely describes the systems real behaviour and made additionally the cation and anion investigations possible which were addressed before leading to

significant findings. However, on the other hand more interaction parameters had to be determined for the model.

25. Furthermore, this work demonstrated and discussed for the first time correct solubility model performance by means of the protein activity coefficient behaviour. A qualitatively correct protein activity coefficient behaviour resulted from the solubility model with the newly determined parameters and additionally approved the solubility model. Again the qualitative relationship of protein activity coefficients and solubility was examined and confirmed. Similar approaches were never applied and documented before but are highly recommended in order to confirm correct solubility model performance.

26. Overall this work demonstrated for twelve different protein containing systems and six different proteins that protein solution and mixture properties can be modelled using the proposed approaches. This result is significant as solution properties and mixture properties are the most difficult properties to model. Furthermore, this implies that pure compound properties can be modelled, which was shown in this work. For the structural parameters of proteins prediction methods were created. This and the fact that solution and mixture properties for various proteins over a wide range of different system conditions were accurately

modelled and predicted in this work is encouraging as not only other proteins but also other protein properties should therefore be describable in a manner similar to those suggested here. Still, the specific conditions of the systems studied here need to be considered when the same approaches are transferred to new systems.

Chapter 7

Future Research Objectives

Research projects are suggested which might be pursued in future as a continuation of this work.

This work established that protein solution and mixture properties and in particular protein activity coefficients and protein solubility can be modelled qualitatively and quantitatively correctly. Success of the introduced modelling frameworks was shown for six different proteins over changes in system composition, temperature and pH. Multicompound systems consisting of protein, salt, buffer and water were studied.

Normal and retrograde solubility behaviour in the salting - out region was examined and modelled successfully. The salting - in region, where at low salt concentrations an increase of protein solubility is observed, was not investigated but modelling of this region is as important as modelling the salting - out region. The salting - in region is of direct importance to engineering purposes where not only the crystallisation of proteins but also the solubilisation of proteins is exploited in a processing environment. Still, until today the experimental solubility data mainly originates in the crystallisation community which focuses its interest on the salting - out region. Therefore, little or no experimental data in the salting - in region is available. Following this, any research project aiming at modelling the salting - in region would require a combination of modelling work and experimental measurements.

Further valuable experimental measurements include those of protein solubility of salt and / or buffer free systems. These examinations should refer to the systems and conditions which were studied here. Measurements applying the same system conditions, i.e., temperature and pH but without salt and / or buffer are important reference measurements. These measurements refer to the solubility product for a protein - solvent system over any salt and / or buffer type when reviewing the introduced solubility model and comparing it to Green's studies (1932). Application of these measurements within the modelling framework used here and comparison to the results obtained here could valuably be pursued. However, attempts to obtain these measurements are experimentally demanding as systems without any salt and buffer would need to be produced. This is a complicated process as the protein purification methods apply salts and buffers to produce proteins and therefore introduce these as contaminants. Minimisation of the contamination content leads to protein material losses while also preparative difficulties such as high osmotic pressures when applying dialysis might interfere [communication with Pusey, 1997]. These measurements and the experimental approach to obtain them are valuable and should furthermore lead to model advancements.

The presented and studied solubility model describes the solubility quantitatively correctly as a function of added salt concentration and system temperature but pH and buffer concentration are further factors that might need to be included in the quantitative modelling approach. It is commonly known that the pH has an impact on protein solubility [Bailey and Ollis, 1986] but it is rather new knowledge that the buffer concentration has likewise a strong impact [Forsythe and Pusey, 1996]. Integration of these factors into the studied modelling framework creates further research projects. Data for the lysozyme system studied here is available from the Biophysics laboratory at NASA. However, at least one second system should be studied to confirm the developed model approach. The most appropriate system would be the concanavalin A system which has already been examined in this work.

Still, experimental data as a function of pH, buffer type and concentration is needed for the concanavalin A system. This requires additional experimental research work. Another system of interest is the glucoseisomerase system which demonstrates uncommon solubility behaviour with respect to pH [communication with Judge, 1997]. Furthermore, model and system examinations over different salt types and different crystal forms are likewise of future interest and could lead to further research projects.

The modelling approaches developed here were the first of this kind for proteins and led to a quantitatively correct description of the equilibrium reached after an undefined amount of time. It might be a future research objective to develop dynamic models which indicate at what point in time a certain property value is obtained. This is an issue for optimal process development where time might need to be considered. Dynamic modelling of the solid - liquid equilibrium would enhance the presented model. Studies of nucleation and growth rates have been pursued for the lysozyme system and are available for model development purposes. Experimental studies of nucleation and growth rates under varying conditions are available from the Biophysics group at NASA.

Furthermore, the models can be developed with respect to their description of compound interactions. At present the sum of interactions between the compounds are described. This is done by parameters which are obtained by guidance from experimental data. However, the occurring explicit interactions can be described [Israelachvili, 1985] and the interaction models used can be integrated into the modelling approaches used here. In order to pursue such work rather simple and low compound systems should be approach. The protein systems examined here for their activity coefficient behaviour seem to be a good set of data for this kind of work. For these systems only three different interaction parameters were needed while for the later two systems up to twelve parameters were necessary. However, interactions such as protein - protein and protein - solvent interactions have to be examined and additional experimental measurements might be required to obtain the data needed to describe certain interactions.

In this work the solubility model evaluations for protein containing systems were directed by the experimental solubility data alone. To improve the model evaluation process additional experimental data is useful. Solubility measurements with no salt as discussed previously, are an option as these represent the solubility product. Furthermore, osmotic pressure measurements at the same system conditions can be obtained and integrated as demonstrated in this work. This kind of data can additionally guide the solubility model or confirm the calculated activity coefficients arising from the solubility model. Furthermore, experimental coexistence data for the studied systems under same conditions can be obtained under the precondition that a liquid - liquid separation occurs. Combination of these three equilibrium phenomena, the solubility, the coexistence (liquid - liquid equilibrium) and the osmotic pressure behaviour, leads to a complete phase diagram for a protein containing system and provides further sources to evaluate and confirm the model approaches from a variety of directions, i.e., experimental data sets. These sets of data can be used to either direct the model or to independently confirm the model. However, to obtain a complete protein phase diagram is another challenging research objective while an integrated modelling approach for a whole phase diagram is a further one.

In this work it has been demonstrated that the predictive UNIFAC model with its group contribution approach is not applicable for the systems aimed at. However, other group contribution methods were studied here and confirmed the group contribution approach for proteins. The van der Waals volume and surface area calculations were successful and used a group contribution method. Still, the UNIFAC model was shown to have inappropriate group definitions for proteins and related compounds, which led to the bad performance. The peptide bond between two different amino acids in a sequence is not described and therefore the backbone of a peptide or protein is not represented. A research objective should be to produce group definitions that are applicable for proteins. Here schemes that use amino acids as the basic groups are envisioned unlike the group definitions for hydrocarbons that define their groups on a lower level using up to five atoms per defined group. Furthermore, group definitions that resemble amino acid sequence patterns seem a sensible approach. It has been shown that certain amino acid sequence patterns lead to specific secondary structures e.g. indicating α helixes or β sheets [Taylor, 1990]. Following this, it is assumed that certain amino acid sequence patterns are related to the physicochemical properties of a protein. Investigations into this field are necessary before any group contribution model is approached and are future research projects arising from this work.

The development of a group contribution method that will allow property predictions for proteins or protein containing systems needs a number of systems to train these methods. A set of different systems over a variety of system conditions is necessary which exhibit all the known property behaviour possible. For proteins these amounts of data are generally not available yet and the data needs to be produced in order to do this work. In order to e.g. produce solubility data as documented by Cacioppo and Pusey (1991) not only time and special equipment but also certain amounts of protein are necessary. It is estimated that a maximum of up to 10 g of lysozyme were used [communication with Pusey, 1997] to obtain the solubility data [Cacioppo and Pusey, 1991] which does not reflect pre - trials that might be necessary for proteins that have not been as widely studied as lysozyme. The cost per gram of purified commercial proteins such as β - lactoglobulin, concanavalin A or green fluorescent protein are listed at 35 $ (Sigma company), 150 $ (Sigma company) and 1580000 $ (Clontech company), respectively. It becomes clear that measurements might be unaffordable. Still, the proteins can be produced in the laboratory which lowers costs.

In the cases of some proteins natural sources supply high levels of these proteins. For lysozyme and ovalbumin eggs are a natural source and provide enough protein for examinations. Lysozyme is obtained from eggs and about 1 g of lysozyme results from thirty eggs while for ovalbumin only about two eggs are needed to produce 1 g of ovalbumin [communication with Judge, 1997]. However, this is not given for all proteins. Many proteins are only produced in minimal amounts in organisms while even lower amounts are retrieved. For green fluorescent protein e.g. 50000 jellyfish are needed to produce 200 mg of green fluorescent protein where 40 μg of the protein are available per jellyfish [Perozzo, 1997b].

To obtain proteins from their natural sources in order to study their properties is one of the possibilities but the recent advancements in genetic engineering are an additional one. Furthermore, proteins that occur at only low levels in their natural sources, which are the majority, might be produced at higher levels using artificial expression systems. For a protein such as green fluorescent protein where only very low levels are found per organism genetic engineering will allow for the production of high amounts of this protein. At present e.g. about 10 mg of green fluorescent protein are purified from one litre of fermentation broth using E.coli as a host strain [communication with Perozzo, 1997a; Deschamps et al., 1995]. These artificial systems allow us to produce a variety of proteins at the levels needed. Following this, a variety of new proteins become available for physicochemical property studies of proteins. This kind of work, high level protein expression, is certainly needed to pursue the modelling work further and even more so when aiming at predictive models, which require an experimental database consisting of as many different systems as possible.

Except the fact that genetic engineering allows us to produce proteins at high levels of expression it also allows us to introduce changes in the molecular structure of proteins. This can be exploited to understand which molecular groups or amino acids contribute to a specific property. Moreover, these studies can be combined with the examinations towards the group definitions which were previously suggested with the perspective of developing a group contribution method. A molecular structure - property research project arises from this. However, to do this kind of work first of all a high level expression system is needed and furthermore an effective production and protein purification method has to be applied.

The possibilities introduced by genetic engineering are broad and to use these techniques is necessary in order to proceed with the modelling work of protein properties any further. With respect to genetic engineering many protein systems should be examined in future and plenty of data should be produced as a consequence. This data is applicable for further modelling work and in particular for the extension of the established modelling approaches using a group contribution method to create prediction methods. To derive the group related data from a rather big set of data will introduce a high mathematical load which has to be dealt with. Already the last model of this work, the solubility model, introduced noticeable difficulties. In this case a high performance computer was available and used to determine the needed parameters. While in future certainly a high performance computer is needed when aiming at the development of a group contribution method additional studies towards the improvement of the parameter determination routine might be needed. This would in particular aim at the improvement of the optimisation routine. This is certainly a research project of interest. However, already any extension of the present model approach as suggested here, by either introducing more reference data to direct the model or by extending the modelling framework to describe further system compounds and conditions, increases the mathematical load and requires a solution for this problem.

With respect to the research work pursued here and the projects that were suggested and might arise in future it has to be kept in mind that a description of protein properties for a vast amount of different systems under different system conditions is aimed at, reflecting the needs of a model for engineering and crystal growth purposes. A general modelling base is needed as applied in this work as opposed to a specific modelling approach which will only describe a very specific set of systems due to the specific input data needed e.g. requiring the amount of water found in a specific protein crystal. However, on the other hand a model that bears a theoretical base is needed as opposed to one that is purely empirical as a theoretical model base will ensure that the basic system behaviour is accounted for and that therefore many systems will be represented with the model. Furthermore, a model is aimed at that requires as little input data as possible and moreover input data that is easily and readily available. Such input data are e.g. in the case of the UNIQUAC model the structural parameters and interaction parameters. Easily and widely applicable models on a theoretical base are needed which compute properties accurately. These requirements reflect the compromises that need to be made with respect to this work which had to be kept in mind to secure the success of this work.

With respect to the success of this research project and original objective of producing protein property prediction methods, already significant achievements resulted from this work. It has been demonstrated that protein properties can be described for twelve different systems consisting of protein, salt, buffer and water over various system conditions such as composition, temperature and pH. This indicates that other proteins and properties and furthermore mixture and solution properties of proteins can be described likewise. Moreover, it has been demonstrated in this work that not only model calculations but also prediction methods can be successfully developed for proteins. All this demonstrates that the long term objective of developing predictive models for various protein properties is achievable.

Appendix A

Activity Coefficient Data of Amino Acids and Peptides

The activity coefficient data, which was converted to the thermodynamic scale and was used for the studies in chapter three is listed.

Table A.1: Experimentally determined activity coefficients of amino acids as a function of composition

x	γ_{Glycine}	γ_{Alanine}	$\gamma_{\text{Threonine}}$	γ_{Valine}	γ_{Proline}	γ_{Serine}
0.0036	0.9644	1.0082	0.9921	1.0341	1.0223	0.9673
0.0054	0.9487	1.0124	0.9893	1.0504	1.0336	0.9492
0.0072	0.9344	1.0165		1.0669		0.9160
0.0089	0.9213	1.0207	0.9838	1.0862	1.0566	
0.0125	0.8981		0.9782		1.0825	
0.0142	0.8880	1.0333				
0.0177	0.8697	1.0417	0.9767		1.1162	
0.0212	0.8537	1.0502	0.9756			
0.0263	0.8342	1.0631	0.9763		1.1792	
0.0297	0.8232	1.0718				
0.0331		1.0830				
0.0348	0.8098		0.9781		1.2485	
0.0431	0.7909					
0.0561	0.7728					

Table A.2: Experimentally determined activity coefficients of peptides as a function of composition

x	$\gamma_{Glycylglycine}$	$\gamma_{Glycylalanine}$	$\gamma_{Triglycine}$
0.0036	0.9153	0.9388	0.8542
0.0054	0.8838	0.9169	0.8079
0.0072			
0.0089	0.8354	0.8910	
0.0125	0.8006	0.8799	
0.0142			
0.0177	0.7581	0.8705	
0.0212	0.7384	0.8636	
0.0263	0.7155		
0.0297	0.7065		

Appendix B

FORTRAN Code of the UNIQUAC Model (unsymmetric)

The FORTRAN code for the UNIQUAC model that was created as part of this work for UNIQUAC parameter determinations and activity coefficient calculations is documented.

Program routine:

```
C**********************************************************************
C  AUTHOR:   SABINE M. AGENA
C
C
C  VERSION: 1
C
C
C  OBJECTIVE: THIS PROGRAM IS FOR PROTEIN ACTIVITY COEFFICIENTS
C
C
C**********************************************************************
C
      PROGRAM PRO
C
       IMPLICIT REAL*8 (A-H,O-Z)
C
       CHARACTER*30 INEXP
       CHARACTER*30 INEXCA
       CHARACTER*30 INPROG
       CHARACTER*30 INCOMP
       CHARACTER*30 OUT
       CHARACTER*30 OUTT
C
       COMMON / IRUN/ IRUN
       COMMON / NEXP/ NEXP
C
       WRITE(*,*)'FILE NAME OF PROGRAM SPECIFIC DATA:'
       READ(*,'(A)') INPROG
       WRITE(*,*)'FILE NAME OF COMPONENT SPECIFIC DATA:'
       READ(*,'(A)') INCOMP
       WRITE(*,*)'FILE NAME FOR PROGRAM OUTPUT:'
       READ(*,'(A)') OUT
       WRITE(*,*)'FILE NAME FOR RESULT OUTPUT:'
       READ(*,'(A)') OUTT
C
       OPEN(20,FILE=INPROG)
       OPEN(30,FILE=INCOMP)
       OPEN(40,FILE=OUT)
       OPEN(21,FILE=OUTT)
C
       CALL RDINCO
```

```
C
      CALL RDINPR
C
      WRITE(*,*)'DATA WAS READ'
C
      IF (IRUN.EQ.0) THEN
        WRITE(*,*)'PARAMETERS WILL BE DETERMINED'
C
        WRITE(*,*)'FILE NAME OF REFERENCE DATA:'
        READ(*,'(A)') INEXP
C
        OPEN(10,FILE=INEXP)
C
        CALL RDINEX(NEXP)
C
        CALL PDET
C
      ELSE
        WRITE(*,*)'PROTEIN ACTIVITY COEFFICIENTS WILL BE CALCULATED'
C
        WRITE(*,*)'FILE NAME OF REFERENCE DATA:'
C
        READ(*,'(A)') INEXP
C
        OPEN(10,FILE=INEXP)
C
        CALL RDINEX(NEXP)
C
        CALL PCAL
      ENDIF
C
      CLOSE(10)
      CLOSE(15)
      CLOSE(20)
      CLOSE(30)
      CLOSE(40)
      CLOSE(21)
C
      STOP
      END
C
C*********************************************************************
C
C
C TITLE:    SUBROUTINE RDINCO
C
C
C OBJECTIVE:  READS IN THE COMPONENT DATA FROM FILE
```

```
C
C
C********************************************************************
C
      SUBROUTINE RDINCO
C
      IMPLICIT REAL*8 (A-H,O-Z)
C
      COMMON / COMW/ COMW (12)
      COMMON / CORQ/ COR (12), COQ (12)
      COMMON / ZCOOR/ ZCOOR (12)
      COMMON / U/ U(12,12)
C
      READ(30,*) (COMW(I),I=1,12)
      READ(30,*) (COR(I),I=1,12)
      READ(30,*) (COQ(I),I=1,12)
      READ(30,*) (ZCOOR(I),I=1,12)
C
      DO 100 I=1,12
          READ(30,*) (U(I,J),J=I,12)
100   CONTINUE
C
      DO 110 I=1,12
        DO 110 J-1,12
          U(J,I)=U(I,J)
110   CONTINUE
C
      RETURN
      END
C
C********************************************************************
C
C
C  TITLE:     SUBROUTINE RDINPR
C
C
C  OBJECTIVE:  READS PROGRAM SPECIFIC DATA FROM FILE
C
C
C********************************************************************
C
      SUBROUTINE RDINPR
C
      IMPLICIT REAL*8 (A-H,O-Z)
C
      COMMON/ IRUN/ IRUN
      COMMON/ NEXP/ NEXP
      COMMON/ EST/ IUEST(12,12)
      COMMON/ EST1/ MAXFUN, IPRINT,NPEN
      COMMON/ EST2/ ESCALE, EPS, SSQEPS, PMAX, WP
C
      READ(20,*) IRUN
      READ(20,*) NEXP
C
      DO 100 I=1,12
          READ(20,*) (IUEST(I,J),J=I,12)
100   CONTINUE
C
      DO 110 I=1,12
        DO 110 J=1,12
          IUEST(J,I)=IUEST(I,J)
110   CONTINUE
C
      READ(20,*) MAXFUN, IPRINT,NPEN
      READ(20,*) ESCALE, EPS, SSQEPS, PMAX, WP
C
      RETURN
      END
C
C
C*******************************************************************
C
C
C  TITLE:     SUBROUTINE RDINEX
C
C
C  OBJECTIVE:  READS EXPERIMENTAL DATA FROM FILE
C
C
C*******************************************************************
C
      SUBROUTINE RDINEX(NEXP)
C
      IMPLICIT REAL*8 (A-H,O-Z)
C
      COMMON/ TEMP/ TEMPEX(1000)
      COMMON/ GAMPR/ GAMPR(1000)
      COMMON/ XEX/ XPREX(1000), XWAEX(1000)
C
      DO 140 J=1,NEXP
          READ(10,*)TEMPEX(J),XWAEX(J),XPREX(J),GAMPR(J)
          TEMPEX(J)=TEMPEX(J)+273.15
140   CONTINUE
C
      RETURN
      END
```

```
C
C*******************************************************************
C
C
C  TITLE:     SUBROUTINE PCAL
C
C
C  OBJECTIVE:  CALCULATES ACTIVITY COEFFICIENTDS AT GIVEN T
C
C
C*******************************************************************
C
      SUBROUTINE PCAL
C
      IMPLICIT REAL*8 (A-H,O-Z)
C
      COMMON/ NEXP/ NEXP
C
      CALL PUNIQUAC(NEXP)
      CALL PACT(NEXP)
      CALL OUT(NEXP)
C
      RETURN
      END
C
C*******************************************************************
C
C
C  TITLE:     SUBROUTINE PDET
C
C
C  OBJECTIVE: EVALUATES THE PARAMETERS U
C
C
C*******************************************************************
C
      SUBROUTINE PDET
C
      IMPLICIT REAL*8(A-H,O-Z)
C
      COMMON/ NEXP/ NEXP
      COMMON/ CORQ/ COR(12), COQ(12)
      COMMON/ EST/ IUEST(12,12)
      COMMON/ U/ U(12,12)
      COMMON/ EST1/ MAXFUN, IPRINT,NPEN
      COMMON/ EST2/ ESCALE, EPS, SSQEPS, PMAX, WP
C
      COMMON/ FE / F(2000), E(40), X(40)
```

```
C
      II-0
      DO 100 N=1,12
        DO 100 M=N,12
          IF (IUEST(N,M).NE.0) THEN
            II=II+1
            X(II)=U(N,M)
          END IF
100     CONTINUE
C
      N=II
      M=NEXP
      DO 150 I=1,40
         E(I)=EPS
150     CONTINUE
C
      CALL MINF2(M,N)
C
      CALL PACTCAL(N,F,X)
C
      RETURN
      END
C
C***********************************************************************
C
C
C     SUBROUTINE MINF2
C
C
C     OBJECTIVE: THIS SUBROUTINE FINDS THE MINIMUM OF A SUM OF SQUARES
C     OF 'M' GIVEN FUNCTIONS OF N VARIABLES. IT USES A METHOD GIVEN IN
C     PRESS ET AL. (1986) NUMERICAL RECIPES IN FORTRAN, CAMBRIDGE
C
C
C***********************************************************************
C
      SUBROUTINE MINF2(M,N)
C
      IMPLICIT REAL*8(A-H,O-Z)
C
      COMMON /CWORK  /WORK(5000)
      COMMON /ICALC  /ICALC
      COMMON /FE     /F(2000), E(40), X(40)
      COMMON /EST1   /MAXFUN,IPRINT,NPEN
      COMMON /EST2   /ESCALE,EPS,SSQEPS,PMAX,WP
      COMMON /EST3   /WI(8)
C
      ICALC=1
```

```
C
      IER    = 0
      FFOLD  = 0.0D0
      MPLUSN = M+N
      KST    = N+MPLUSN
      NPLUS  = N+1
      KINV   = NPLUS*(MPLUSN+1)
      KSTORE = KINV-MPLUSN-1
C
C       WRITE(6,*) ' INTO CALFUN'
      CALL PACTCAL(N,F,X)
C
      IF (ICALC.EQ.0) GO TO 999
      NN = N+N
      K  = NN
C
      DO 1 I=1,M
        K       = K+1
        WORK(K) = F(I)
 1    CONTINUE
C
      IINV = 2
      K    = KST
      I    = 1
C
 2    CONTINUE
      X(I) = X(I) + E(I)
C       WRITE(6,*) ' INTO CALFUN'
      CALL PACTCAL(N,F,X)
      X(I)=X(I)-E(I)
C
      DO 3 J=1,N
        K       = K+1
        WORK(K) = 0.0D0
        WORK(J) = 0.0D0
 3    CONTINUE
C
      SUM = 0.0D0
      KK  = NN
      DO 4 J=1,M
        KK   = KK+1
        F(J) = F(J)-WORK(KK)
        SUM  = SUM+F(J)*F(J)
 4    CONTINUE
C
      IF (SUM) 5,5,6
 5    CONTINUE
      WRITE(6,7) I
      WRITE(21,7) I
 7    FORMAT(/,1X,'MESSAGE FROM MINF2 : E(',I3,')'
     *      ,' IS UNREASONABLY SMALL')
      DO 8 J=1,M
        NN  = NN+1
        F(J) = WORK(NN)
 8    CONTINUE
      GOTO 10
 6    CONTINUE
      SUM=1.0D0/DSQRT(SUM)
      J=K-N+I
      WORK(J)=E(I)*SUM
      DO 9 J=1,M
        K       = K+1
        WORK(K) = F(J)*SUM
        KK      = NN+J
        DO 11 II=1,I
          KK       = KK+MPLUSN
          WORK(II) = WORK(II)+WORK(KK)*WORK(K)
 11     CONTINUE
 9    CONTINUE
      ILESS  = I-1
      IGAMAX = N+I-1
      INCINV = N-ILESS
      INCINP = INCINV+1
      IF (ILESS) 13,13,14
 13   CONTINUE
      WORK(KINV) = 1.0D0
      GO TO 15
 14   CONTINUE
      B=1.0D0
      DO 16 J=NPLUS,IGAMAX
        WORK(J) = 0.0D0
 16   CONTINUE
      KK=KINV
      DO 17 II=1,ILESS
        IIP      = II+N
        WORK(IIP) = WORK(IIP)+WORK(KK)*WORK(II)
        JL       = II+1
        JLAUX    = JL - ILESS
        IF (JLAUX) 18,18,19
 18       CONTINUE
          DO 20 JJ=JL,ILESS
            KK      = KK+1
            JJP     = JJ+N
            WORK(IIP) = WORK(IIP)+WORK(KK)*WORK(JJ)
            WORK(JJP) = WORK(JJP)+WORK(KK)*WORK(II)
 20       CONTINUE
```

```
19        CONTINUE
        B  = B-WORK(II)*WORK(IIP)
        KK = KK+INCINP
17    CONTINUE
      B  = 1.0D0/B
      KK = KINV
      DO 21 II=NPLUS,IGAMAX
        BB = -B*WORK(II)
        DO 22 JJ=II,IGAMAX
          WORK(KK) = WORK(KK)-BB*WORK(JJ)
          KK      = KK+1
22      CONTINUE
        WORK(KK) = BB
        KK       = KK+INCINV
21    CONTINUE
      WORK(KK) = B
15    CONTINUE
      GO TO (27,24),IINV
24    CONTINUE
      I = I+1
      IF (I-N) 2,2,25
25    CONTINUE
      IINV - 1
      FF  = 0.0D0
      KL  = NN
      DO 26 I=1,M
        KL  = KL+1
        F(I) = WORK(KL)
        FF  = FF+F(I)*F(I)
26    CONTINUE
      ICONT=1
      ISS=1
      MC=N+1
      IPP=IABS(IPRINT)*(IABS(IPRINT)-1)
      ITC=0
      IPS=1
      IPC=0
27    CONTINUE
      FFD = DABS(FF-FFOLD)
C     WRITE(6,*) 'FF,FFOLD,FFD',FF,FFOLD,FFD
      IF (FFD.LT.SSQEPS) IER = 2
      FFOLD = FF
      IF (IER.EQ.2) GOTO 10
      IPC=IPC-IABS(IPRINT)
      IF (IPC) 28,29,29
28    CONTINUE
      CALL FITINF(X,N,FF,ITC,MC,0,IER,MAXFUN)
      IPC=IPP
```

```
29  CONTINUE
    GO TO (34,35),ICONT
35  CONTINUE
    CHANG1 = CHANGE - 1.0D0
    IF (CHANG1.GT.0.D0) GOTO 36
10  CONTINUE
    CALL FITINF(X,N,FF,ITC,MC,1,IER,MAXFUN)
    GOTO 999
36  CONTINUE
    ICONT=1
34  CONTINUE
    ITC=ITC+1
    K=N
    KK=KST
    DO 39 I=1,N
      K=K+1
      WORK(K)=0.0D0
      KK=KK+N
      WORK(I)=0.0D0
      DO 40 J=1,M
        KK=KK+1
        WORK(I)=WORK(I)+WORK(KK)*F(J)
40    CONTINUE
39  CONTINUE
    DM=0.0D0
    K=KINV
    DO 41 II=1,N
      IIP=II+N
      WORK(IIP)=WORK(IIP)+WORK(K)*WORK(II)
      JL=II+1
      IF (JL-N) 42,42,43
42    CONTINUE
      DO 44 JJ=JL,N
        JJP=JJ+N
        K=K+1
        WORK(IIP)=WORK(IIP)+WORK(K)*WORK(JJ)
        WORK(JJP)=WORK(JJP)+WORK(K)*WORK(II)
44    CONTINUE
      K=K+1
43    CONTINUE
      IF(DM-DABS(WORK(II)*WORK(IIP))) 45,41,41
45    CONTINUE
      DM=DABS(WORK(II)*WORK(IIP))
      KL=II
41  CONTINUE
    II=N+MPLUSN*KL
    CHANGE=0.0D0
    DO 46 I=1,N
```

```
       JL=N+I
       WORK(I)=0.0D0
       DO 47 J=NPLUS,NN
         JL=JL+MPLUSN
         WORK(I)=WORK(I)+WORK(J)*WORK(JL)
47      CONTINUE
       II=II+1
       WORK(II)=WORK(JL)
       WORK(JL)=X(I)
       IF(DABS(E(I)*CHANGE)-DABS(WORK(I))) 48,48,46
48      CONTINUE
       CHANGE=DABS(WORK(I)/E(I))
46    CONTINUE
     DO 49 I=1,M
       II=II+1
       JL=JL+1
       WORK(II)=WORK(JL)
       WORK(JL)=F(I)
49    CONTINUE
     FC=FF
     ACC=0.1D0/CHANGE
     IT=3
     XC=0.0D0
     XL=0.0D0
     IS=3
     XSTEP=-DMIN1(0.5D00,ESCALE/CHANGE)
     IF (CHANGE-1.D0) 50,50,51
50    CONTINUE
     ICONT=2
51    CONTINUE
     CALL MINLIN(IT,XC,FC,6,ACC,0.1D0,XSTEP)
     GO TO (52,53,53,53),IT
52    CONTINUE
     MC=MC+1
     IF (MC-MAXFUN) 54,54,55
55    CONTINUE
     IER = 1
     ISS=2
     GOTO 53
54    CONTINUE
     XL=XC-XL
     DO 57 J=1,N
       X(J)=X(J)+XL*WORK(J)
57    CONTINUE
     XL=XC
C      WRITE(6,*) ' INTO CALFUN'
     CALL PACTCAL(N,F,X)
     FC=0.0D0
```

```
   DO 58 J=1,M
     FC=FC+F(J)*F(J)
58  CONTINUE
   GO TO (59,59,60),IS
60  CONTINUE
   K=N
   IF (FC-FF) 61,51,62
61  CONTINUE
   IS=2
   FMIN=FC
   FSEC=FF
   GO TO 63
62  CONTINUE
   IS=1
   FMIN=FF
   FSEC=FC
   GO TO 63
59  CONTINUE
   IF (FC-FSEC) 64,51,51
64  CONTINUE
   K=KSTORE
   GO TO (75,74),IS
75  CONTINUE
   K=N
74  CONTINUE
   IF (FC-FMIN) 65,51,66
66  CONTINUE
   FSEC=FC
   GO TO 63
65  CONTINUE
   IS=3-IS
   FSEC=FMIN
   FMIN=FC
63  CONTINUE
   DO 67 J=1,N
     K=K+1
     WORK(K)=X(J)
67  CONTINUE
   DO 68 J=1,M
     K=K+1
     WORK(K)=F(J)
68  CONTINUE
   GO TO 51
53  CONTINUE
   K=KSTORE
   KK=N
   GO TO (69,70,69),IS
70  CONTINUE
```

```
    K=N
    KK=KSTORE
69  CONTINUE
    SUM=0.0D0
    DM-0.0D0
    JJ=KSTORE
    DO 71 J=1,N
      K=K+1
      KK=KK+1
      JJ=JJ+1
      X(J)=WORK(K)
      WORK(JJ)=WORK(K)-WORK(KK)
71  CONTINUE
    DO 72 J=1,M
      K=K+1
      KK=KK+1
      JJ=JJ+1
      F(J)=WORK(K)
      WORK(JJ)=WORK(K)-WORK(KK)
      SUM=SUM+WORK(JJ)*WORK(JJ)
      DM=DM+F(J)*WORK(JJ)
72  CONTINUE
    GO TO (73,10),ISS
73  CONTINUE
    J=KINV
    KK=NPLUS-KL
    DO 76 I=1,KL
      K=J+KL-I
      J=K+KK
      WORK(I)=WORK(K)
      WORK(K)=WORK(J-1)
76  CONTINUE
    IF (KL-N) 77,78,78
77  CONTINUE
    KL=KL+1
    JJ=K
    DO 79 I-KL,N
      K=K+1
      J=J+NPLUS-I
      WORK(I)=WORK(K)
      WORK(K)=WORK(J-1)
79  CONTINUE
    WORK(JJ)=WORK(K)
    B=1.0D0/WORK(KL-1)
    WORK(KL-1)=WORK(N)
    GO TO 88
78  CONTINUE
    B=1.0D0/WORK(N)
```

```
88  CONTINUE
    K=KINV
    DO 80 I=1,ILESS
      BB=B*WORK(I)
      DO 81 J=I,ILESS
        WORK(K)=WORK(K)-BB*WORK(J)
       K=K+1
81      CONTINUE
      K=K+1
80  CONTINUE
    IF (FMIN-FF) 82,83,83
83  CONTINUE
    CHANGE=0.0D0
    GO TO 84
82  CONTINUE
    FF=FMIN
    CHANGE=DABS(XC)*CHANGE
84  CONTINUE
    XL=-DM/FMIN
    SUM=1.0D0/DSQRT(SUM+DM*XL)
    K=KSTORE
    DO 85 I=1,N
      K=K+1
      WORK(K)=SUM*WORK(K)
      WORK(I)=0.0D0
85  CONTINUE
    DO 86 I=1,M
      K=K+1
      WORK(K)=SUM*(WORK(K)+XL*F(I))
      KK=NN+I
      DO 87 J=1,N
        KK=KK+MPLUSN
        WORK(J)=WORK(J)+WORK(KK)*WORK(K)
87      CONTINUE
86  CONTINUE
    GOTO 14
999  CONTINUE
    RETURN
    END
C
C
C***************************************************************************
C
C
C   SUBROUTINE FITINF
C
C
C   OBJECTIVE: DOCUMENTS THE SSQ AND PARAMETERS DURING THE
```

```
C   DETERMINATION PROCESS
C
C
C************************************************************************
C
      SUBROUTINE FITINF(X,NTPAR,SSQ,ITER,NCAL,NLAST,IER,MFUN)
      IMPLICIT REAL*8 (A-H,O-Z)
C
C
      DIMENSION X(NTPAR)
C
     IF (NLAST.EQ.1.OR.IER.EQ.2) GOTO 10
     WRITE(21,1002) ITER,SSQ
     WRITE(21,1031) NCAL
     WRITE(21,1001)
     WRITE(21,1030) (I,X(I),I=1,NTPAR)
     WRITE(21,1032)
     WRITE(6,1002) ITER,SSQ
     GOTO 20
   10 CONTINUE
     IF (IER.EQ.1) GOTO 30
     WRITE(21,2002) ITER,SSQ
     WRITE(21,1031) NCAL
     WRITE(21,2001)
     WRITE(21,1030) (I,X(I),I=1,NTPAR)
     IF (IER.EQ.2) WRITE (21,2020)
     IF (IER.EQ.0) WRITE (21,2021)
     WRITE(6,2002) ITER,SSQ
     WRITE(6,1031) NCAL
     WRITE(6,2001)
     WRITE(6,1030) (I,X(I),I=1,NTPAR)
     IF (IER.EQ.2) WRITE (6,2020)
     IF (IER.EQ.0) WRITE (6,2021)
     GOTO 40
  30 CONTINUE
     WRITE(21,1035) MFUN
     WRITE(21,1002) ITER,SSQ
     WRITE(21,1001)
     WRITE(21,1030) (I,X(I),I=1,NTPAR)
     WRITE(6,1035) MFUN
     WRITE(6,1002) ITER,SSQ
     WRITE(6,1001)
     WRITE(6,1030) (I,X(I),I=1,NTPAR)
  40 CONTINUE
     WRITE(21,1032)
     WRITE(6,1032)
  20 CONTINUE
C
```

```
C FORMAT DECLA.
C
 1002 FORMAT(/,1X,'FOR ITERATION NUMBER ',I5,//,1X,'SSQ=',D13.6)
 1001 FORMAT(//,1X,'THE ACTUAL VALUES OF THE PARAMETERS ARE :',/)
 1030 FORMAT(1X,'X(',I2,')=',D13.6)
 1031 FORMAT(/,1X,'CALFUN HAS BEEN CALLED',I5,' TIMES')
 1032 FORMAT(/,70('*'),/)
 1035 FORMAT(/,'******WARNING******',/,
     *    1X,'THE NUMBER OF CALLS TO CALFUN HAS EXCEEDE',I5)
 2002 FORMAT(/,1X,'THE ESTIMATION HAS TERMINATED NORMALLY AFTER'
     *   ,I5,' ITERATIONS',//,
     *    1X,'SSQ=',D13.6)
 2001 FORMAT(//,1X,'THE FINAL VALUES OF THE PARAMETERS ARE :',/)
 2020 FORMAT(//,1X,'THE DESIRED ACCURACY IN SSQ WAS ACHIEVED')
 2021 FORMAT(//,1X,'THE DESIRED ACCURACY IN THE PARAMETERS WAS ACHIEVED'
     *)
      RETURN
      END
C
C************************************************************************
C
C
C    SUBROUTINE MINLIN
C
C
C    OBJECTIVE: THIS SUBROUTINE FINDS THE MINIMUM OF AF FUNCTION
C    OF ONE VARIABLE
C
C
C************************************************************************
C
      SUBROUTINE MINLIN(ITEST,X,F,MAXFUN,ABSACC,RELACC,XSTEP)
C
      COMMON /AUXMIN1/ IS,MC,DB,FB,FA,DA,D,IINC
C
C  MODIFICATION BY S.AGENA: AUXMIN1 AND AUXMIN2 WERE CREATED FROM
C  AUXMIN AS THE ALIGNEMENT CREATED A WARNING WHEN COMPLIED WITH THE
C  F77 (NASA, SEP. 96)
C
      COMMON /AUXMIN2/ FC,DC,XINC
      DOUBLE PRECISION ABSACC,D,DA,DB,DC,F,FA,FB,FC,RELACC,X,XINC,XSTEP
      GO TO (1,2,2),ITEST
    2 IS=6-ITEST
      ITEST=1
      IINC=1
      XINC=XSTEP+XSTEP
      MC=IS-3
      IF (MC) 4,4,15
 3 MC=MC+1
  IF (MAXFUN-MC) 12,15,15
12 ITEST=4
43 X=DB
  F=FB
  IF (FB-FC) 15,15,44
44 X=DC
  F-FC
15 RETURN
 1 GO TO (5,6,7,8),IS
 8 IS=3
 4 DC=X
  FC=F
  X=X+XSTEP
  GO TO 3
 7 IF (FC-F) 9,10,11
10 X=X+XINC
  XINC=XINC+XINC
  GO TO 3
 9 DB=X
  FB=F
  XINC=-XINC
  GO TO 13
11 DB=DC
  FB=FC
  DC=X
  FC=F
13 X=DC+DC-DB
  IS=2
  GO TO 3
 6 DA=DB
  DB=DC
  FA=FB
  FB=FC
32 DC=X
  FC=F
  GO TO 14
 5 IF (FB-FC) 16,17,17
17 IF (F-FB) 18,32,32
18 FA=FB
  DA=DB
19 FB=F
  DB=X
  GO TO 14
16 IF (FA-FC) 21,21,20
20 XINC=FA
  FA=FC
  FC=XINC
```

```
   XINC=DA
   DA=DC
   DC=XINC
 21 XINC=DC
   IF ((D-DB)*(D-DC)) 32,22,22
 22 IF (F-FA) 23,24,24
 23 FC=FB
   DC=DB
   GO TO 19
 24 FA=F
   DA=X
 14 IF (FB-FC) 25,25,29
 25 IINC=2
   XINC=DC
   IF (FB-FC) 29,45,29
 29 D=(FA-FB)/(DA-DB)-(FA-FC)/(DA-DC)
   IF(D*(DB-DC))33,33,37
 37 D=0.5D0*(DB+DC-(FB-FC)/D)
   IF(DABS(D-X)-DABS(ABSACC))34,34,35
 35 IF(DABS(D-X)-DABS(D*RELACC))34,34,36
 34 ITEST=2
   GO TO 43
 36 IS=1
   X=D
   IF ((DA-DC)*(DC-D)) 3,26,38
 38 IS=2
   GO TO (39,40),IINC
 39 IF(DABS(XINC)-DABS(DC-D))41,3,3
 33 IS=2
   GO TO (41,42),IINC
 41 X=DC
   GO TO 10
 40 IF(DABS(XINC-X)-DABS(X-DC))42,42,3
 42 X=0.5D0*(XINC+DC)
   IF ((XINC-X)*(X-DC)) 26,26,3
 45 X=0.5D0*(DB+DC)
   IF ((DB-X)*(X-DC)) 26,26,3
 26 ITEST=3
   GO TO 43
   END
C
C*************************************************************************
C
C
C TITLE:    SUBROUTINE OUT
C
C
C OBJECTIVE: CREATION OF OUTPUT FILE
```

```
C
C
C***********************************************************************
C
      SUBROUTINE OUT(NEXP)
C
      IMPLICIT REAL*8(A-H,O-Z)
C
       COMMON / COMW/ COMW (12)
       COMMON / CORQ/ COR (12), COQ (12)
       COMMON / U/ U(12,12)
       COMMON/ XEX/ XPREX(1000), XWAEX(1000)
       COMMON/ GAMPR/ GAMPR(1000)
       COMMON/ GAMMA/ GAMMA(12,1000)
       COMMON/ TEMP/ TEMPEX(1000)
       COMMON/ ERROR/ DEL1(1000),DEL2(1000),DEL3(1000),SSD, RMSD
       COMMON/ GAMM/ GAMCT(12,1000), GAMRS(12,1000)
C
      WRITE(40,1000)
1000  FORMAT('ACTIVITY COEFFICIENT MODEL',///)
C
      WRITE(40,1100)
1100  FORMAT(/,'COMPONENT DATA',//,'NO.',4(2X),'MW',4(2X),'R',4(2X)
     *    ,'Q'//)
      DO 100 I=1,12
        WRITE(40,1200) I,COMW(I),COR(I),COQ(I)
100   CONTINUE
1200  FORMAT(I4,3(F12.4))
C
      WRITE(40,1500)
1500  FORMAT(///,'UNIQUAC PARAMETERS',//,'U(I,J)',//)
      DO 200 I=1,12
        WRITE(40,1600) I,(U(I,J),J=1,12)
200   CONTINUE
1600  FORMAT(I2,12(F12.4))
C
      WRITE(40,1700)
1700  FORMAT(///,'ACTIVITY COEFFICIENT RESULTS',//,
     *    'CAL',4(2X),'EXP',4(2X),'DELT',4(2X),'d X/X'//)
      DO 300 I=1,NEXP
        WRITE(40,1800) GAMMA(4,I),GAMPR(I),DEL1(I),DEL2(I)
300   CONTINUE
1800  FORMAT(4(D12.4))
C
      WRITE(40,1900)
1900  FORMAT(/,'ERRORS',//,'SSD',4(2X),'RMSD',//)
        WRITE(40,2000) SSD,RMSD
2000  FORMAT(2(D12.4))
```

```
C
      WRITE(40,2500)
2500  FORMAT(///,'GAMMA CAL.',//,
     *      'LN CT',4(2X),'LN RS',4(2X),'GAMMA'//)
      DO 500 J=1,NEXP
C        DO 500 I-1,4
         WRITE(40,2600) J,4,GAMCT(4,J),GAMRS(4,J),GAMMA(4,J)
500   CONTINUE
2600  FORMAT(2(I2),3(D12.4))
C
      WRITE(40,2700)
2700  FORMAT(///,'GAMMA CAL.',//,'H2O',4(2X),'PROT.',4(2X)//)
      DO 800 J=1,NEXP
       WRITE(40,2800) J,GAMMA(1,J),GAMMA(4,J)
800   CONTINUE
2800  FORMAT(1(I2),2(D12.4))
C
      WRITE(40,2900)
2900  FORMAT(///,'EXP. DATA',//,'TEMP.',4(2X),
     *      'X H2O',4(2X),'X PROT.',4(2X),'GAMMA PROT.'//)
      DO 900 J=1,NEXP
       WRITE(40,3000) J,TEMPEX(J),XWAEX(J),XPREX(J),GAMPR(J)
900   CONTINUE
3000  FORMAT(1(I2),4(D12.4))
      RETURN
      END
C
C*****************************************************************
C
C
C TITLE:    SUBROUTINE PACTCAL
C
C
C OBJECTIVE: CALCULATION OF ACTIVITY COEFFICIENTS AND
C OBJECTIVE FUNCTION
C
C
C*****************************************************************
C
      SUBROUTINE PACTCAL(N,F,X)
C
      IMPLICIT REAL*8(A-H,O-Z)
C
      COMMON/ NEXP/ NEXP
      COMMON/ CORQ/ COR(12), COQ(12)
      COMMON/ EST/ IUEST(12,12)
      COMMON/ U/ U(12,12)
C
```

```
      DIMENSION F(2000), X(40)
C
      L=0
      DO 100 K=1,12
       DO 100 KK=K,12
         IF (IUEST(K,KK).NE.0) THEN
           L=L+1
           U(K,KK)=X(L)
         END IF
100   CONTINUE
C
      DO 200 I=1,12
        DO 200 J=1,12
          U(J,I)=U(I,J)
200   CONTINUE
C
      CALL PUNIQUAC(NEXP)
      CALL PACT(NEXP)
      CALL POBJ(NEXP,F)
      CALL OUT(NEXP)
C
      RETURN
      END
C
C*******************************************************************
C
C
C  TITLE:    SUBROUTINE PACT
C
C
C  OBJECTIVE: CALCULATION OF ACTIVITY COEFFICIENTS
C
C
C*******************************************************************
C
      SUBROUTINE PACT(NEXP)
C
      IMPLICIT REAL*8(A-H,O-Z)
C
      COMMON/ XEX/ XPREX(1000), XWAEX(1000)
      COMMON/ GAMPR/ GAMPR(1000)
      COMMON/ GAMMA/ GAMMA(12,1000)
      COMMON/ TEMP/ TEMPEX(1000)
      COMMON/ ERROR/ DEL1(1000),DEL2(1000),DEL3(1000),SSD,RMSD
C
      SUM1=0.D0
      SUM2=0.D0
      SUM3=0.D0
```

```
    SUM4=0.D0
    DO 100 J=1,NEXP
       DEL1(J)=DABS(GAMPR(J)-GAMMA(4,J))
       SUM1=SUM1+DEL1(J)
       DEL2(J)–DABS((GAMPR(J)-GAMMA(4,J))/GAMPR(J))
       SUM2-SUM2+DEL2(J)
       DEL3(J)=((GAMPR(J)-GAMMA(4,J))/GAMPR(J))
       SUM3=SUM3+DEL3(J)
       SUM4=SUM4+((DEL3(J))**2)
100  CONTINUE
C
    SSD=SUM4
    RMSD=SQRT(SSD/NEXP)
C
    RETURN
    END
C
C*******************************************************************
C
C
C  TITLE:    SUBROUTINE POBJ
C
C
C  OBJECTIVE: CALCULATES THE OBJECTIVE FUNCTION VALUE, OBJ [-]
C
C
C*******************************************************************
C
    SUBROUTINE POBJ(NEXP,F)
C
    IMPLICIT REAL*8(A-H,O-Z)
C
    COMMON/ XEX/ XPREX(1000), XWAEX(1000)
    COMMON/ GAMPR/ GAMPR(1000)
    COMMON/ GAMMA/ GAMMA(12,1000)
C
    DIMENSION F(2000)
C
    FB=0.D0
C
    DO 200 J=1,NEXP
       FB=((GAMPR(J)-GAMMA(4,J))/GAMPR(J))
       F(J)=FB
200  CONTINUE
C
    RETURN
    END
C
```

```
C*****************************************************************
C
C
C  TITLE:    SUBROUTINE PUNIQ
C
C
C  OBJECTIVE: CALCULATION OF ACTIVITY COEFFICIENTS WITH THE
C  UNIQUAC MODEL (UNSYMMETRIC)
C
C
C*****************************************************************
C
   SUBROUTINE PUNIQUAC(NEXP)
C
   IMPLICIT REAL*8(A-H,O-Z)
C
   COMMON / RQ / RR(12),QQ(12)
   COMMON / CORQ / COR(12), COQ(12)
   COMMON / XTR / XTR(12)
   COMMON / XEX / XPREX(1000), XWAEX(1000)
   COMMON / PT / T
   COMMON / TEMP / TEMPEX(1000)
   COMMON / PM / PM(12)
   COMMON / COMW / COMW(12)
   COMMON / U / U(12,12)
C
   COMMON / GAM / GAM(12)
   COMMON /GCT / GCT(12)
   COMMON /GRS / GRS(12)
   COMMON / GAMLN / GAMLN(12)
   COMMON / GAMMLN / GAMMLN(12,1000)
   COMMON / GAMMA / GAMMA(12,1000)
   COMMON / GAMM / GAMCT(12,1000), GAMRS(12,1000)
C
   DO 200 I=1,12
       RR(I)=COR(I)
       QQ(I)=COQ(I)
C
       PM(I)=COMW(I)
C
200  CONTINUE
C
   DO 300 J=1,NEXP
C
       T=TEMPEX(J)
C
       XTR(1)=XWAEX(J)
       XTR(4)=XPREX(J)
```

```
C
          CALL ACTCOF
C
          DO 320 I=1,12
               GAMMA(I,J)=GAM(I)
               GAMMLN(I,J)= GAMLN(I)
               GAMCT(I,J)  GCT(I)
               GAMRS(I,J)-GRS(I)
320         CONTINUE
C
300  CONTINUE
C
     RETURN
     END
C
C*********************************************************************
C
C
C    SUBROUTINE ACTCOF
C
C
C    OBJECTIVE: THIS SUBROUTINE CALCULATES THE ACTIVITY COEFFICIENTS
C    USING UNIQUAC MODEL
C
C
C*********************************************************************
C
     SUBROUTINE ACTCOF
C
     IMPLICIT REAL*8(A-H,O-Z)
C
     COMMON / GAM / GAM(12)
     COMMON / GCT / GCT(12)
     COMMON / GRS / GRS(12)
     COMMON / GAMLN / GAMLN(12)
C
     DO 2000 I=1,12
       GAM(I) = 0.D0
       GRS(I) = 0.D0
       GCT(I) = 0.D0
2000 CONTINUE
C
       CALL GAMACT
       CALL GAMARS
       DO 1000 I=1,12
         GAMLN(I)=GCT(I)+GRS(I)
         CON=GAMLN(I)
         IF (CON.GE.174.673D0) THEN
```

```
          GAM(I)=7.23D75
        ELSE
          IF (CON.LE.-174.673D0) THEN
            GAM(I)=0.0D0
          ELSE
            GAM(I) = DEXP(CON)
          ENDIF
        ENDIF
1000   CONTINUE
C
    RETURN
    END
C
C********************************************************************
C
C
C   SUBROUTINE GAMACT
C
C
C   OBJECTIVE: THIS SUBROUTINE CALCULATES THE COMBINATORIAL TERM
C
C
C********************************************************************
C
    SUBROUTINE GAMACT
C
    IMPLICIT REAL*8(A-H,O-Z)
C
    COMMON / XTR / XTR(12)
    COMMON / RQ / RR(12),QQ(12)
    COMMON / PTH / PH(12),TH(12),THE(12)
    COMMON / GCT / GCT(12)
    COMMON / ZCOOR / ZCOOR(12)
C
    SR = 0.D0
    SQ = 0.D0
    DO 1000 I=1,12
      SR = SR + XTR(I)*RR(I)
      SQ = SQ + XTR(I)*QQ(I)
1000 CONTINUE
C
    DO 1010 I=1,12
      PH(I)  = RR(I) SR
      TH(I)  = QQ(I)/SQ
      THE(I) = XTR(I)*TH(I)
1010 CONTINUE
C
    DO 1020 I=1,12
```

```
        PHI   = PH(I)
        PHT   = PHI/TH(I)
        GCI   = DLOG(PHI) - PHI
        GCT(I) = GCI-1.D0/2.D0*ZCOOR(I)*QQ(I)*(DLOG(PHT)-PHT)
1020  CONTINUE
C
      DO 1030 I=1,12
        IF (I.EQ.1) THEN
          AUX = 1.D0/2.D0*ZCOOR(I)*QQ(I)
          GCT(I) = GCT(I) + 1.D0 - AUX
        ENDIF
1030  CONTINUE
C
      DO 1040 I-1,12
        IF (I.NE.1) THEN
          RAZ1  = RR(I)/RR(1)
          RAZ2  = QQ(I)/QQ(1)
          RAZ   = RAZ1/RAZ2
          AUX   = 1.D0/2.D0*ZCOOR(I)*QQ(I)
          GCT(I) = GCT(I) - DLOG(RAZ1) + RAZ1 - AUX*(RAZ - DLOG(RAZ))
        ENDIF
1040  CONTINUE
C
      RETURN
      END
C
C********************************************************************
C
C
C   SUBROUTINE GAMARS
C
C
C   OBJECTIVE: THIS SUBROUTINE CALCULATES THE RESIDUAL TERM
C
C
C********************************************************************
C
      SUBROUTINE GAMARS
C
      IMPLICIT REAL*8(A-H,O-Z)
C
      COMMON / PT / T
      COMMON / GRS / GRS(12)
      COMMON / CAUX / US(12),S(12),A(12),C(12,12)
      COMMON / RQ / RR(12),QQ(12)
C
      CALL GARS01
      CALL GARS03
      CALL GARS04
C
      DO 2110 I=1,12
        IF (I.EQ.1) THEN
          GRS(I) = QQ(I)*(1.D0 - DLOG(S(I)) -A(I))
        ENDIF
2110  CONTINUE
C
      DO 2120 I=1,12
        IF (I.NE.1) THEN
          GRS(I) = QQ(I)*(-DLOG(S(I))-A(I)+US(I))
        ENDIF
2120  CONTINUE
C
      RETURN
      END
C
C**********************************************************************
C
C
C     SUBROUTINE GARS01
C
C
C     OBJECTIVE: THIS SUBROUTINE CALCULATES THE RESIDUAL TERM
C
C
C**********************************************************************
C
      SUBROUTINE GARS01
C
      IMPLICIT REAL*8(A-H,O-Z)
C
      COMMON / PT / T
      COMMON / UM / UM(12,12)
      COMMON / U / U(12,12)
      COMMON / CAUX / US(12),S(12),A(12),C(12,12)
C
      DO 1030 I=1,12
        DO 1030 J=I,12
          UM(I,J)=U(I,J)
          UM(J,I)=UM(I,J)
1030    CONTINUE
C
      DO 2000 I=1,12
        IF (I.NE.1) THEN
          US(I) = -(UM(1,I)-UM(I,I))/T+DEXP(-(UM(I,1)-UM(1,1))/T)
        ENDIF
2000  CONTINUE
```

```
C
   RETURN
   END
C
C*********************************************************************
C
C
C   SUBROUTINE GARS03
C
C
C   OBJECTIVE: THIS SUBROUTINE CALCULATES THE RESIDUAL TERM
C
C
C*********************************************************************
C
   SUBROUTINE GARS03
C
   IMPLICIT REAL*8(A-H,O-Z)
C
   COMMON / PT / T
   COMMON / UM / UM(12,12)
   COMMON / PME / PME(12,12)
C
C  exp fortran limitation
C
   DO 2020 I=1,12
     DO 2010 J=1,12
       CONST = - (UM(I,J)-UM(J,J))/T
       IF (CONST.GE.174.673D0) THEN
         PME(I,J) = 7.23D75
       ELSE
         IF (CONST.LE.-174.673D0) THEN
           PME(I,J) = 0.0D0
         ELSE
           PME(I,J) = DEXP(CONST)
         ENDIF
       ENDIF
2010   CONTINUE
2020 CONTINUE
C
   RETURN
   END
C
C*********************************************************************
C
C
C  SUBROUTINE GARS04
C
```

```
C
C   OBJECTIVE: THIS SUBROUTINE CALCULATES THE RESIDUAL TERM
C
C
C*********************************************************************
C
      SUBROUTINE GARS04
C
      IMPLICIT REAL*8(A-H,O-Z)
C
      COMMON / PTH / PH(12),TH(12),THE(12)
      COMMON / PME / PME(12,12)
      COMMON / CAUX / US(12),S(12),A(12),C(12,12)
C
      DO 2040 N=1,12
        SN = 0.D0
        DO 2030 K=1,12
          SN = SN + THE(K)*PME(K,N)
2030    CONTINUE
        S(N) = SN
2040  CONTINUE
C
      DO 2060 N=1,12
        AN = 0.D0
        DO 2050 K=1,12
          C(N,K) = PME(N,K)/S(K)
          AN = AN + THE(K)*C(N,K)
2050    CONTINUE
        A(N) = AN
2060  CONTINUE
C
      RETURN
      END
```

Appendix C

Activity Coefficient Data of Proteins

The activity coefficient data on the thermodynamic scale, which was determined from osmotic pressure measurements and was used for the studies in chapter four is listed.

Table C.1: Experimentally determined activity coefficients of serum albumin, S, as a function of composition

x	γ_x
2.8367E-06	1.0495
3.9399E-06	1.0817
3.9924E-06	1.0834
4.2026E-06	1.0905
4.2551E-06	1.0924
4.9643E-06	1.1189
5.1744E-06	1.1275
5.2270E-06	1.1297
5.2270E-06	1.1297
5.2532E-06	1.1308
5.4633E-06	1.1398
5.5421E-06	1.1433
5.8573E-06	1.1576
6.5140E-06	1.1902
6.6716E-06	1.1985
7.5646E-06	1.2500
7.5909E-06	1.2516
8.0374E-06	1.2803
8.4051E-06	1.3055
9.5608E-06	1.3943
9.7709E-06	1.4121
9.8760E-06	1.4213
1.0296E-05	1.4593
1.0480E-05	1.4766
1.2529E-05	1.7052
1.2634E-05	1.7189
1.3238E-05	1.8016

Table C.2: Experimentally determined activity coefficients of α - chymotrypsin, C1, as a function of composition

x	γ_x
5.7972E-07	0.9978
1.1605E-06	0.9960
2.3246E-06	0.9931
3.4935E-06	0.9915
4.6668E-06	0.9911
5.8443E-06	0.9919
1.2816E-05	1.0223
1.9416E-05	1.0948
2.6157E-05	1.2223

Table C.3: Experimentally determined activity coefficients of α - chymotrypsin, C2, as a function of composition

x	γ_x
6.1001E-07	0.9663
1.2211E-06	0.9345
2.4460E-06	0.8763
3.6760E-06	0.8244
4.9106E-06	0.7783
6.1496E-06	0.7373
1.3486E-05	0.5767
2.0430E-05	0.5144
2.7523E-05	0.5155

Table C.4: Experimentally determined activity coefficients of α - chymotrypsin, C3, as a function of composition

x	γ_x
5.9401E-07	0.9570
1.1825E-06	0.9182
2.3719E-06	0.8501
3.5695E-06	0.7934
4.7736E-06	0.7463
5.9639E-06	0.7079
1.3190E-05	0.5919
1.9982E-05	0.5909
2.6920E-05	0.6452

Table C.5: Experimentally determined activity coefficients of α - chymotrypsin, C4, as a function of composition

x	γ_x
5.0717E-07	0.9830
1.0074E-06	0.9666
1.9907E-06	0.9357
2.9882E-06	0.9060
3.9369E-06	0.8792
4.9220E-06	0.8527

Table C.6: Experimentally determined activity coefficients of α - chymotrypsin, C5, as a function of composition

x	γ_x
5.5252E-07	0.8983
1.0756E-06	0.8116
2.1775E-06	0.6553
3.2197E-06	0.5352
4.2532E-06	0.4379
5.3356E-06	0.3550

Table C.7: Experimentally determined activity coefficients of α - chymotrypsin, C6, as a function of composition

x	γ_x
1.1054E-06	0.9623
2.1781E-06	0.9270
3.2680E-06	0.8925
4.3583E-06	0.8592
5.4490E-06	0.8272

Table C.8: Experimentally determined activity coefficients of α - chymotrypsin, C7, as a function of composition

x	γ_x
5.5122E-07	1.0100
1.2740E-06	1.0242
2.1005E-06	1.0422
3.2854E-06	1.0714
4.1785E-06	1.0962
5.4406E-06	1.1355

Table C.9: Experimentally determined activity coefficients of β - lactoglobulin, L, as a function of composition

x	γ_x
5.0868E-06	1.001917
5.4403E-06	1.002743
8.1031E-06	1.011905
9.4391E-06	1.018486
9.6503E-06	1.01965
1.2015E-05	1.035035
1.3608E-05	1.047903
1.4631E-05	1.057275
1.4921E-05	1.060083
1.6096E-05	1.07223
1.7698E-05	1.090759
1.8841E-05	1.10542
2.2399E-05	1.159271
2.2399E-05	1.159271
2.2482E-05	1.160678

Table C.10: Experimentally determined activity coefficients of ovalbumin, O, as a function of composition

x	γ_x
9.3339E-06	1.0088
1.0327E-05	1.0111
1.5076E-05	1.0258
1.7803E-05	1.0371
2.0073E-05	1.0482
2.5631E-05	1.0819
2.7044E-05	1.0920

Appendix D

Solubility Data of Proteins

The solubility data on the thermodynamic scale, which was used for the studies in chapter five is listed.

Table D.1: Experimental solubility data of lysozyme, LYS, as a function of composition and temperature

x_{LYS}	x_{Salt}	Temperature [K]
3.9182E-06	6.2715E-03	277.15
5.4245E-06	6.2796E-03	280.05
1.2833E-05	6.3205E-03	286.15
1.3481E-05	6.3243E-03	286.85
1.4757E-05	6.3314E-03	287.55
1.6040E-05	6.3385E-03	288.15
1.7128E-05	6.3445E-03	288.65
1.8388E-05	6.3513E-03	289.15
1.9697E-05	6.3586E-03	289.75
2.1036E-05	6.3661E-03	290.4
2.3130E-05	6.3776E-03	291.15
2.5112E-05	6.3885E-03	291.85
2.8163E-05	6.4048E-03	292.55
3.0933E-05	6.4196E-03	293.15
3.4617E-05	6.4392E-03	293.85
4.0896E-05	6.4724E-03	294.75
4.5230E-05	6.4957E-03	295.65
4.8669E-05	6.5140E-03	296.25
5.6013E-05	6.5525E-03	297
5.9555E-05	6.5714E-03	297.6
1.2254E-06	9.4532E-03	275.55
1.2879E-06	9.4536E-03	276.15
1.4432E-06	9.4548E-03	277.15
1.6839E-06	9.4569E-03	278.75
1.9950E-06	9.4598E-03	280.15
2.2784E-06	9.4627E-03	281.5
2.5766E-06	9.4658E-03	282.7
3.1684E-06	9.4718E-03	284.35
3.6849E-06	9.4772E-03	285.7
4.2794E-06	9.4832E-03	286.85
4.9118E-06	9.4899E-03	288.25
5.5837E-06	9.4967E-03	289.35

Table D.1: con't

x_{LYS}	x_{Salt}	Temperature [K]
6.2234E-06	9.5031E-03	290.25
7.8269E-06	9.5187E-03	292.15
9.3617E-06	9.5334E-03	293.65
1.1760E-05	9.5549E-03	295.15
1.3802E-05	9.5735E-03	296.45
1.2059E-06	1.2697E-02	283.15
1.3096E-06	1.2699E-02	283.95
1.4841E-06	1.2703E-02	285.05
1.5103E-06	1.2704E-02	285.65
1.5251E-06	1.2705E-02	286.15
1.6188E-06	1.2707E-02	286.85
1.7514E-06	1.2710E-02	287.55
1.8732E-06	1.2712E-02	288.15
1.9613E-06	1.2714E-02	288.65
2.0933E-06	1.2716E-02	289.15
2.2121E-06	1.2719E-02	289.75
2.3935E-06	1.2722E-02	290.4
2.6056E-06	1.2726E-02	291.15
2.8236E-06	1.2730E-02	291.85
3.0341E-06	1.2734E-02	292.55
3.2691E-06	1.2738E-02	293.35
3.4466E-06	1.2742E-02	293.85
3.5524E-06	1.2744E-02	294.35
3.9944E-06	1.2752E-02	295.65
4.2793E-06	1.2757E-02	296.25
4.8115E-06	1.2765E-02	297
5.1496E-06	1.2770E-02	297.6
2.8445E-07	1.5969E-02	275.55
3.1450E-07	1.5969E-02	276.15
3.5798E-07	1.5970E-02	277.15
4.2898E-07	1.5971E-02	278.75
4.8465E-07	1.5973E-02	280.15
5.5442E-07	1.5975E-02	281.5
6.5173E-07	1.5978E-02	282.7

Table D.1: con't

x_{LYS}	x_{Salt}	Temperature [K]
7.7536E-07	1.5982E-02	284.35
9.2468E-07	1.5986E-02	285.7
1.0664E-06	1.5990E-02	286.85
1.2121E-06	1.5996E-02	288.25
1.3604E-06	1.6000E-02	289.35
1.5184E-06	1.6005E-02	290.25
1.6604E-06	1.6009E-02	291.15
2.2400E-06	1.6025E-02	293.65
2.7199E-06	1.6037E-02	295.15
3.2268E-06	1.6048E-02	296.45
3.7386E-06	1.6058E-02	297.35
1.7833E-07	2.2680E-02	275.55
1.9778E-07	2.2680E-02	276.15
2.2827E-07	2.2680E-02	277.15
2.5357E-07	2.2681E-02	278.75
2.9316E-07	2.2683E-02	280.15
3.1329E-07	2.2686E-02	281.5
3.7172E-07	2.2689E-02	282.7
4.3669E-07	2.2694E-02	284.35
4.9908E-07	2.2698E-02	285.7
5.7838E-07	2.2703E-02	286.85
6.4217E-07	2.2709E-02	288.25
7.2416E-07	2.2715E-02	289.35
8.0032E-07	2.2720E-02	290.25
8.6352E-07	2.2725E-02	291.15
1.1952E-06	2.2743E-02	293.65
1.6146E-06	2.2770E-02	297.35

Table D.2: Experimental solubility data of concanavalin A, CON, as a function of composition and temperature

x_{CON}	x_{Salt}	Temperature [K]
8.1443E-07	4.6642E-03	291.15
7.9430E-07	4.6646E-03	292.15
7.7476E-07	4.6650E-03	293.15
7.5599E-07	4.6656E-03	294.15
7.3816E-07	4.6661E-03	295.15
7.2144E-07	4.6668E-03	296.15
7.0599E-07	4.6675E-03	297.15
6.9199E-07	4.6684E-03	298.15
6.8940E-07	4.6688E-03	298.55
6.7961E-07	4.6693E-03	299.15
6.6902E-07	4.6703E-03	300.15
6.6039E-07	4.6714E-03	301.15
6.5389E-07	4.6726E-03	302.15
6.4969E-07	4.6740E-03	303.15
6.4796E-07	4.6754E-03	304.15
6.4889E-07	4.6769E-03	305.15
6.5263E-07	4.6786E-03	306.15
6.5937E-07	4.6804E-03	307.15
6.6929E-07	4.6823E-03	308.15
6.8254E-07	4.6843E-03	309.15
6.9932E-07	4.6865E-03	310.15
7.1980E-07	4.6888E-03	311.15
7.4416E-07	4.6913E-03	312.15
7.7258E-07	4.6939E-03	313.15
7.8111E-07	4.6953E-03	313.75
8.0524E-07	4.6967E-03	314.15
8.4231E-07	4.6996E-03	315.15
8.8400E-07	4.7026E-03	316.15
9.3131E-07	4.7055E-03	316.95
9.3047E-07	4.7059E-03	317.15
9.8193E-07	4.7093E-03	318.15
8.5232E-08	9.5714E-03	291.15

Table D.2: con't

x_{CON}	x_{Salt}	Temperature [K]
8.2844E-08	9.5732E-03	292.15
8.0420E-08	9.5751E-03	293.15
7.7413E-08	9.5757E-03	293.55
7.8066E-08	9.5771E-03	294.15
7.5882E-08	9.5792E-03	295.15
7.1852E-08	9.5804E-03	295.75
7.3970E-08	9.5814E-03	296.15
7.2432E-08	9.5838E-03	297.15
7.1368E-08	9.5863E-03	298.15
7.1076E-08	9.5916E-03	300.15
7.2051E-08	9.5945E-03	301.15
7.3911E-08	9.5976E-03	302.15
8.0695E-08	9.6041E-03	304.15
8.5825E-08	9.6076E-03	305.15
9.2252E-08	9.6113E-03	306.15
1.0008E-07	9.6151E-03	307.15
1.0941E-07	9.6191E-03	308.15
1.1172E-07	9.6228E-03	309.15
1.3301E-07	9.6276E-03	310.15
1.3129E-07	9.6283E-03	310.35
1.4748E-07	9.6322E-03	311.15
1.5576E-07	9.6338E-03	311.45
1.6388E-07	9.6369E-03	312.15
1.6709E-07	9.6375E-03	312.25
1.8231E-07	9.6418E-03	313.15
2.2569E-07	9.6523E-03	315.15
2.5085E-07	9.6578E-03	316.15
2.5403E-07	9.6597E-03	316.55
2.7848E-07	9.6636E-03	317.15
3.0867E-07	9.6695E-03	318.15
4.9661E-08	2.0374E-02	291.15
4.7435E-08	2.0378E-02	292.15
4.5204E-08	2.0382E-02	293.15
4.6658E-08	2.0384E-02	293.55

Table D.2: con't

x_{CON}	x_{Salt}	Temperature [K]
4.3033E-08	2.0387E-02	294.15
4.0985E-08	2.0391E-02	295.15
3.7742E-08	2.0394E-02	295.75
3.9124E-08	2.0397E-02	296.15
3.9734E-08	2.0397E-02	296.25
3.7514E-08	2.0402E-02	297.15
3.6219E-08	2.0407E-02	298.15
3.5303E-08	2.0413E-02	299.15
3.6793E-08	2.0418E-02	299.95
3.4830E-08	2.0420E-02	300.15
3.4865E-08	2.0426E-02	301.15
3.5473E-08	2.0433E-02	302.15
3.6716E-08	2.0440E-02	303.15
3.8662E-08	2.0447E-02	304.15
4.1373E-08	2.0455E-02	305.15
4.4915E-08	2.0463E-02	306.15
4.5853E-08	2.0468E-02	306.85
4.7849E-08	2.0469E-02	306.95
4.9354E-08	2.0471E-02	307.15
5.4755E-08	2.0479E-02	308.15
6.8704E-08	2.0497E-02	310.15
7.7384E-08	2.0507E-02	311.15
8.7290E-08	2.0517E-02	312.15
9.3930E-08	2.0518E-02	312.25
9.8488E-08	2.0527E-02	313.15
1.0650E-07	2.0533E-02	313.75
1.1305E-07	2.0543E-02	314.75

Appendix E

FORTRAN Code of the Solubility Model

The FORTRAN code for the Solubility model, which was created as part of this work for parameter determinations and solubility and activity coefficient calculations is documented.

Program routine:

```
C************************************************************************
C
C
C AUTHOR:   SABINE M. AGENA
C
C
C VERSION: 1
C
C
C OBJECTIVE: THIS PROGRAM IS FOR PROTEIN SOLUBILITY
C
C
C************************************************************************
C
      PROGRAM PRO
C
      IMPLICIT REAL*8 (A-H,O-Z)
C
      CHARACTER*30 INEXP
      CHARACTER*30 INEXCA
      CHARACTER*30 INPROG
      CHARACTER*30 INCOMP
      CHARACTER*30 OUT
      CHARACTER*30 OUTT
C
      COMMON / IRUN/ IRUN
      COMMON / NEXP/ NEXP
C
      WRITE(*,*)'FILE NAME OF PROGRAM SPECIFIC DATA:'
      READ(*,'(A)') INPROG
      WRITE(*,*)'FILE NAME OF COMPONENT SPECIFIC DATA:'
      READ(*,'(A)') INCOMP
      WRITE(*,*)'FILE NAME FOR PROGRAM OUTPUT:'
      READ(*,'(A)') OUT
      WRITE(*,*)'FILE NAME FOR RESULT OUTPUT:'
      READ(*,'(A)') OUTT
C
      OPEN(20,FILE=INPROG)
      OPEN(30,FILE=INCOMP)
      OPEN(40,FILE=OUT)
      OPEN(21,FILE=OUTT)
```

```
C
      CALL RDINCO
C
      CALL RDINPR
C
      WRITE(*,*)'DATA WAS READ'
C
      IF (IRUN.EQ.0) THEN
        WRITE(*,*)'PARAMETERS WILL BE DETERMINED'
C
        WRITE(*,*)'FILE NAME OF REFERENCE DATA:'
        READ(*,'(A)') INEXP
C
        OPEN(10,FILE=INEXP)
C
        CALL RDINEX(NEXP)
C
        CALL PDET
C
      ELSE
        WRITE(*,*)'PROTEIN SOLUBILITY WILL BE CALCULATED'
C
        WRITE(*,*)'FILE NAME OF REFERENCE DATA:'
C
        READ(*,'(A)') INEXP
C
        OPEN(10,FILE=INEXP)
C
        CALL RDINEX(NEXP)
C
        CALL PCAL
      ENDIF
C
      CLOSE(10)
      CLOSE(15)
      CLOSE(20)
      CLOSE(30)
      CLOSE(40)
      CLOSE(21)
C
      STOP
      END
C
C********************************************************************
C
C
C TITLE:    SUBROUTINE RDINCO
C
```

```
C
C OBJECTIVE:  READS IN THE COMPONENT DATA FROM FILE
C
C
C**********************************************************************
C
    SUBROUTINE RDINCO
C
     IMPLICIT REAL*8 (A-H,O-Z)
C
     COMMON / COMW/ COMW (12)
     COMMON / CORQ/ COR (12), COQ (12)
     COMMON / Z/ Z(12)
     COMMON / ZCOOR/ ZCOOR (12)
     COMMON / ABC/ A,B,C
     COMMON / U/ U(12,12), UT(12,12)
C
     READ(30,*) (COMW(I),I=1,12)
     READ(30,*) (COR(I),I=1,12)
     READ(30,*) (COQ(I),I=1,12)
     READ(30,*) (Z(I),I=1,12)
     READ(30,*) (ZCOOR(I),I=1,12)
     READ(30,*) A,B,C
C
     DO 100 I=1,12
         READ(30,*) (U(I,J),J-I,12)
100  CONTINUE
C
     DO 150 I=1,12
       READ(30,*) (UT(I,J),J=I,12)
150  CONTINUE
C
     DO 200 I-1,12
       DO 200 J=1,12
         U(J,I)=U(I,J)
         UT(J,I)=UT(I,J)
200  CONTINUE
C
     RETURN
     END
C
C**********************************************************************
C
C
C TITLE:    SUBROUTINE RDINPR
C
C
C OBJECTIVE:  READS PROGRAM SPECIFIC DATA FROM FILE
```

```
C
C
C*******************************************************************
C
      SUBROUTINE RDINPR
C
      IMPLICIT REAL*8 (A-H,O-Z)
C
      COMMON/ IRUN/ IRUN
      COMMON/ NEXP/ NEXP
      COMMON/ IABC/ IA,IB,IC
      COMMON/ EST/ IUEST(12,12), IUTEST(12,12)
      COMMON/ EST1/ MAXFUN, IPRINT,NPEN
      COMMON/ EST2/ ESCALE, EPS, SSQEPS, PMAX, WP
C
      READ(20,*) IRUN
      READ(20,*) NEXP
      READ(20,*) IA,IB,IC
C
      DO 100 I=1,12
          READ(20,*) (IUEST(I,J),J=I,12)
100   CONTINUE
C
      DO 150 I=1,12
        READ(20,*) (IUTEST(I,J),J=I,12)
150   CONTINUE
C
      DO 200 I=1,12
        DO 200 J=1,12
          IUEST(J,I)=IUEST(I,J)
          IUTEST(J,I)=IUTEST(I,J)
200   CONTINUE
C
      READ(20,*) MAXFUN, IPRINT,NPEN
      READ(20,*) ESCALE, EPS, SSQEPS, PMAX, WP
C
      RETURN
      END
C
C
C*******************************************************************
C
C
C TITLE:    SUBROUTINE RDINEX
C
C
C OBJECTIVE: READS EXPERIMENTAL DATA FROM FILE
C
```

```
C
C*******************************************************************
C
      SUBROUTINE RDINEX(NEXP)
C
      IMPLICIT REAL*8 (A-H,O-Z)
C
      COMMON/ TEMP/ TEMPEX(1000)
      COMMON/ XEX/ XPREX(1000), XWAEX(1000), XSAEX(1000)
      COMMON/ Z/ Z(12)
      COMMON/ X/ X(1000,12)
      COMMON/ XUX/ XUX(1000)
C
      DO 140 J=1,NEXP
          READ(10,*)TEMPEX(J),XSAEX(J),XWAEX(J),XPREX(J)
C
      ZZ=Z(5)+Z(6)
      XUX(J)=XWAEX(J)+XPREX(J)+ZZ*XSAEX(J)
      X(J,1)=XWAEX(J)/XUX(J)
      X(J,4)=XPREX(J)/XUX(J)
      X(J,5)=Z(6)*XSAEX(J)/XUX(J)
      X(J,6)=Z(5)*XSAEX(J)/XUX(J)
      SUMX=X(J,1)+X(J,4)+X(J,5)+X(J,6)
C
140   CONTINUE
C
      RETURN
      END
C
C*******************************************************************
C
C
C TITLE:     SUBROUTINE PCAL
C
C
C OBJECTIVE:  CALCULATES SOLUBILITY AT GIVEN T
C
C
C*******************************************************************
C
      SUBROUTINE PCAL
C
      IMPLICIT REAL*8 (A-H,O-Z)
C
      COMMON/ NEXP/ NEXP
C
      CALL PUNIQUAC(NEXP)
      CALL PSOL(NEXP)
      CALL OUT(NEXP)
C
      RETURN
      END
C
C******************************************************************
C
C
C  TITLE:    SUBROUTINE PDET
C
C
C  OBJECTIVE: EVALUATES THE PARAMETERS U AND UT
C
C
C******************************************************************
C
      SUBROUTINE PDET
C
      IMPLICIT REAL*8(A-H,O-Z)
C
      COMMON/ NEXP/ NEXP
      COMMON/ CORQ/ COR(12), COQ(12)
      COMMON/ ABC/ A,B,C
      COMMON/ IABC/ IA,IB,IC
      COMMON/ EST/ IUEST(12,12), IUTEST(12,12)
      COMMON/ U/ U(12,12), UT(12,12)
      COMMON/ EST1/ MAXFUN, IPRINT,NPEN
      COMMON/ EST2/ ESCALE, EPS, SSQEPS, PMAX, WP
C
      COMMON/ FE / F(2000), E(40), X(40)
C
      II=0
C
      IF(IA.NE.0) THEN
        II=II+1
        X(II)=A
      END IF
C
     IF(IB.NE.0) THEN
       II=II+1
       X(II)=B
     END IF
C
     IF(IC.NE.0) THEN
       II=II+1
       X(II)=C
     END IF
C
```

```
      DO 100 N=1,12
        DO 100 M=N,12
          IF (IUEST(N,M).NE.0) THEN
            II=II+1
            X(II)=U(N,M)
          END IF
100     CONTINUE
C
      DO 200 N=1,12
        DO 200 M=N,12
          IF (IUTEST(N,M).NE.0) THEN
            II=II+1
            X(II)=UT(N,M)
          END IF
200     CONTINUE
C
      N=II
      M=NEXP
      DO 150 I=1,40
          E(I)=EPS
150     CONTINUE
C
      CALL MINF2(M,N)
C
      CALL PSOLCAL(N,F,X)
C
      RETURN
      END
C
C**************************************************************************
C
C
C     SUBROUTINE MINF2
C
C
C     OBJECTIVE: THIS SUBROUTINE FINDS THE MINIMUM OF A SUM OF SQUARES
C     OF 'M' GIVEN FUNCTIONS OF N VARIABLES. IT USES A METHOD GIVEN IN
C     PRESS ET AL. (1986) NUMERICAL RECIPES IN FORTRAN, CAMBRIDGE
C
C
C**************************************************************************
C
      SUBROUTINE MINF2(M,N)
C
      IMPLICIT REAL*8(A-H,O-Z)
C
      COMMON /CWORK   /WORK(5000)
      COMMON /ICALC   /ICALC
      COMMON /FE     /F(2000), E(40), X(40)
      COMMON /EST1   /MAXFUN,IPRINT,NPEN
      COMMON /EST2   /ESCALE,EPS,SSQEPS,PMAX,WP
      COMMON /EST3   /WI(8)
C
      ICALC=1
C
      IER   = 0
      FFOLD = 0.0D0
      MPLUSN = M+N
      KST   = N+MPLUSN
      NPLUS  = N+1
      KINV   = NPLUS*(MPLUSN+1)
      KSTORE = KINV-MPLUSN-1
C
C       WRITE(6,*) ' INTO CALFUN'
      CALL PSOLCAL(N,F,X)
C
      IF (ICALC.EQ.0) GO TO 999
      NN = N+N
      K  = NN
C
      DO 1 I=1,M
        K      = K+1
        WORK(K) = F(I)
1     CONTINUE
C
      IINV = 2
      K   = KST
      I   = 1
C
2     CONTINUE
      X(I) = X(I) + E(I)
C       WRITE(6,*) ' INTO CALFUN'
      CALL PSOLCAL(N,F,X)
      X(I)=X(I)-E(I)
C
      DO 3 J=1,N
        K      = K+1
        WORK(K) = 0.0D0
        WORK(J) = 0.0D0
3     CONTINUE
C
      SUM = 0.0D0
      KK  = NN
      DO 4 J=1,M
        KK   = KK+1
        F(J) = F(J)-WORK(KK)
      SUM  = SUM+F(J)*F(J)
4    CONTINUE
C
     IF (SUM) 5,5,6
5    CONTINUE
     WRITE(6,7) I
     WRITE(21,7) I
7    FORMAT(/,1X,'MESSAGE FROM MINF2 : E(',I3,')'
   *      ,' IS UNREASONABLY SMALL')
     DO 8 J=1,M
       NN   = NN+1
       F(J) = WORK(NN)
8    CONTINUE
     GOTO 10
6    CONTINUE
     SUM=1.0D0/DSQRT(SUM)
     J=K-N+I
     WORK(J)=E(I)*SUM
     DO 9 J=1,M
       K       = K+1
       WORK(K) = F(J)*SUM
       KK      = NN+J
       DO 11 II=1,I
         KK       = KK+MPLUSN
         WORK(II) = WORK(II)+WORK(KK)*WORK(K)
11     CONTINUE
9    CONTINUE
     ILESS  = I-1
     IGAMAX = N+I-1
     INCINV = N-ILESS
     INCINP = INCINV+1
     IF (ILESS) 13,13,14
13   CONTINUE
     WORK(KINV) = 1.0D0
     GO TO 15
14   CONTINUE
     B=1.0D0
     DO 16 J=NPLUS,IGAMAX
       WORK(J) = 0.0D0
16   CONTINUE
     KK=KINV
     DO 17 II=1,ILESS
       IIP       = II+N
       WORK(IIP) = WORK(IIP)+WORK(KK)*WORK(II)
       JL        = II+1
       JLAUX     = JL - ILESS
       IF (JLAUX) 18,18,19
18       CONTINUE
```

```
          DO 20 JJ=JL,ILESS
            KK     = KK+1
            JJP    = JJ+N
            WORK(IIP) = WORK(IIP)+WORK(KK)*WORK(JJ)
            WORK(JJP) = WORK(JJP)+WORK(KK)*WORK(II)
20          CONTINUE
19          CONTINUE
        B  = B-WORK(II)*WORK(IIP)
        KK = KK+INCINP
17    CONTINUE
      B  = 1.0D0/B
      KK = KINV
      DO 21 II=NPLUS,IGAMAX
        BB = -B*WORK(II)
        DO 22 JJ=II,IGAMAX
          WORK(KK) = WORK(KK)-BB*WORK(JJ)
          KK     = KK+1
22        CONTINUE
        WORK(KK) = BB
        KK      = KK+INCINV
21    CONTINUE
      WORK(KK) = B
15    CONTINUE
      GO TO (27,24),IINV
24    CONTINUE
      I = I+1
      IF (I-N) 2,2,25
25    CONTINUE
      IINV = 1
      FF  = 0.0D0
      KL  = NN
      DO 26 I=1,M
        KL  = KL+1
        F(I) = WORK(KL)
        FF  = FF+F(I)*F(I)
26    CONTINUE
      ICONT=1
      ISS=1
      MC=N+1
      IPP=IABS(IPRINT)*(IABS(IPRINT)-1)
      ITC=0
      IPS=1
      IPC=0
27    CONTINUE
      FFD = DABS(FF-FFOLD)
C     WRITE(6,*) 'FF,FFOLD,FFD',FF,FFOLD,FFD
      IF (FFD.LT.SSQEPS) IER = 2
      FFOLD = FF
```

```
    IF (IER.EQ.2) GOTO 10
    IPC=IPC-IABS(IPRINT)
    IF (IPC) 28,29,29
28  CONTINUE
    CALL FITINF(X,N,FF,ITC,MC,0,IER,MAXFUN)
    IPC=IPP
29  CONTINUE
    GO TO (34,35),ICONT
35  CONTINUE
    CHANG1 = CHANGE - 1.0D0
    IF (CHANG1.GT.0.D0) GOTO 36
10  CONTINUE
    CALL FITINF(X,N,FF,ITC,MC,1,IER,MAXFUN)
    GOTO 999
36  CONTINUE
    ICONT=1
34  CONTINUE
    ITC=ITC+1
    K=N
    KK=KST
    DO 39 I=1,N
      K=K+1
      WORK(K)=0.0D0
      KK=KK+N
      WORK(I)=0.0D0
      DO 40 J=1,M
        KK=KK+1
        WORK(I)=WORK(I)+WORK(KK)*F(J)
40    CONTINUE
39  CONTINUE
    DM=0.0D0
    K=KINV
    DO 41 II=1,N
      IIP=II+N
      WORK(IIP)=WORK(IIP)+WORK(K)*WORK(II)
      JL=II+1
      IF (JL-N) 42,42,43
42    CONTINUE
      DO 44 JJ=JL,N
        JJP=JJ+N
        K=K+1
        WORK(IIP)=WORK(IIP)+WORK(K)*WORK(JJ)
        WORK(JJP)=WORK(JJP)+WORK(K)*WORK(II)
44    CONTINUE
      K=K+1
43    CONTINUE
      IF(DM-DABS(WORK(II)*WORK(IIP))) 45,41,41
45    CONTINUE
```

```
        DM=DABS(WORK(II)*WORK(IIP))
        KL=II
41     CONTINUE
      II=N+MPLUSN*KL
      CHANGE=0.0D0
      DO 46 I=1,N
        JL=N+I
        WORK(I)=0.0D0
        DO 47 J=NPLUS,NN
          JL=JL+MPLUSN
          WORK(I)-WORK(I)+WORK(J)*WORK(JL)
47       CONTINUE
        II=II+1
        WORK(II)-WORK(JL)
        WORK(JL)=X(I)
        IF(DABS(E(I)*CHANGE)-DABS(WORK(I))) 48,48,46
48       CONTINUE
        CHANGE=DABS(WORK(I)/E(I))
46     CONTINUE
      DO 49 I=1,M
        II=II+1
        JL=JL+1
        WORK(II)=WORK(JL)
        WORK(JL)=F(I)
49     CONTINUE
      FC=FF
      ACC=0.1D0/CHANGE
      IT=3
      XC=0.0D0
      XL=0.0D0
      IS=3
      XSTEP=-DMIN1(0.5D00,ESCALE/CHANGE)
      IF (CHANGE-1.D0) 50,50,51
50     CONTINUE
      ICONT=2
51     CONTINUE
      CALL MINLIN(IT,XC,FC,6,ACC,0.1D0,XSTEP)
      GO TO (52,53,53,53),IT
52     CONTINUE
      MC=MC+1
      IF (MC-MAXFUN) 54,54,55
55     CONTINUE
      IER = 1
      ISS=2
      GOTO 53
54     CONTINUE
      XL=XC-XL
      DO 57 J=1,N
```

```
         X(J)=X(J)+XL*WORK(J)
57     CONTINUE
       XL=XC
C        WRITE(6,*) ' INTO CALFUN'
       CALL PSOLCAL(N,F,X)
       FC=0.0D0
       DO 58 J=1,M
         FC–FC+F(J)*F(J)
58     CONTINUE
       GO TO (59,59,60),IS
60     CONTINUE
       K=N
       IF (FC-FF) 61,51,62
61     CONTINUE
       IS=2
       FMIN=FC
       FSEC=FF
       GO TO 63
62     CONTINUE
       IS=1
       FMIN=FF
       FSEC=FC
       GO TO 63
59     CONTINUE
       IF (FC-FSEC) 64,51,51
64     CONTINUE
       K=KSTORE
       GO TO (75,74),IS
75     CONTINUE
       K=N
74     CONTINUE
       IF (FC-FMIN) 65,51,66
66     CONTINUE
       FSEC=FC
       GO TO 63
65     CONTINUE
       IS=3-IS
       FSEC=FMIN
       FMIN=FC
63     CONTINUE
       DO 67 J=1,N
         K=K+1
         WORK(K)=X(J)
67     CONTINUE
       DO 68 J=1,M
         K=K+1
         WORK(K)=F(J)
68     CONTINUE
```

```
    GO TO 51
53  CONTINUE
    K=KSTORE
    KK=N
    GO TO (69,70,69),IS
70  CONTINUE
    K=N
    KK=KSTORE
69  CONTINUE
    SUM=0.0D0
    DM-0.0D0
    JJ=KSTORE
    DO 71 J=1,N
      K=K+1
      KK=KK+1
      JJ=JJ+1
      X(J)=WORK(K)
      WORK(JJ)=WORK(K)-WORK(KK)
71  CONTINUE
    DO 72 J=1,M
      K=K+1
      KK=KK+1
      JJ=JJ+1
      F(J)=WORK(K)
      WORK(JJ)=WORK(K)-WORK(KK)
      SUM=SUM+WORK(JJ)*WORK(JJ)
      DM=DM+F(J)*WORK(JJ)
72  CONTINUE
    GO TO (73,10),ISS
73  CONTINUE
    J=KINV
    KK=NPLUS-KL
    DO 76 I=1,KL
      K=J+KL-I
      J-K+KK
      WORK(I)=WORK(K)
      WORK(K)=WORK(J-1)
76  CONTINUE
    IF (KL-N) 77,78,78
77  CONTINUE
    KL=KL+1
    JJ=K
    DO 79 I=KL,N
      K=K+1
      J=J+NPLUS-I
      WORK(I)=WORK(K)
      WORK(K)=WORK(J-1)
79  CONTINUE
```

```
   WORK(JJ)=WORK(K)
   B=1.0D0/WORK(KL-1)
   WORK(KL-1)=WORK(N)
   GO TO 88
78  CONTINUE
   B=1.0D0/WORK(N)
88  CONTINUE
   K=KINV
   DO 80 I=1,ILESS
     BB=B*WORK(I)
     DO 81 J=I,ILESS
       WORK(K)=WORK(K)-BB*WORK(J)
      K=K+1
81    CONTINUE
     K=K+1
80  CONTINUE
   IF (FMIN-FF) 82,83,83
83  CONTINUE
   CHANGE=0.0D0
   GO TO 84
82  CONTINUE
   FF=FMIN
   CHANGE=DABS(XC)*CHANGE
84  CONTINUE
   XL=-DM/FMIN
   SUM=1.0D0/DSQRT(SUM+DM*XL)
   K=KSTORE
   DO 85 I=1,N
     K=K+1
     WORK(K)=SUM*WORK(K)
     WORK(I)=0.0D0
85  CONTINUE
   DO 86 I=1,M
     K=K+1
     WORK(K)=SUM*(WORK(K)+XL*F(I))
     KK=NN+I
     DO 87 J=1,N
       KK=KK+MPLUSN
       WORK(J)=WORK(J)+WORK(KK)*WORK(K)
87    CONTINUE
86  CONTINUE
   GOTO 14
999 CONTINUE
   RETURN
   END
C
C
C**********************************************************************
```

```
C
C
C    SUBROUTINE FITINF
C
C
C    OBJECTIVE: DOCUMENTS THE SSQ AND PARAMETERS DURING THE
C    DETERMINATION PROCESS
C
C
C************************************************************************
C
      SUBROUTINE FITINF(X,NTPAR,SSQ,ITER,NCAL,NLAST,IER,MFUN)
      IMPLICIT REAL*8 (A-H,O-Z)
C
C
      DIMENSION X(NTPAR)
C
      IF (NLAST.EQ.1.OR.IER.EQ.2) GOTO 10
      WRITE(21,1002) ITER,SSQ
      WRITE(21,1031) NCAL
      WRITE(21,1001)
      WRITE(21,1030) (I,X(I),I=1,NTPAR)
      WRITE(21,1032)
      WRITE(6,1002) ITER,SSQ
      GOTO 20
   10 CONTINUE
      IF (IER.EQ.1) GOTO 30
      WRITE(21,2002) ITER,SSQ
      WRITE(21,1031) NCAL
      WRITE(21,2001)
      WRITE(21,1030) (I,X(I),I=1,NTPAR)
      IF (IER.EQ.2) WRITE (21,2020)
      IF (IER.EQ.0) WRITE (21,2021)
      WRITE(6,2002) ITER,SSQ
      WRITE(6,1031) NCAL
      WRITE(6,2001)
      WRITE(6,1030) (I,X(I),I=1,NTPAR)
      IF (IER.EQ.2) WRITE (6,2020)
      IF (IER.EQ.0) WRITE (6,2021)
      GOTO 40
   30 CONTINUE
      WRITE(21,1035) MFUN
      WRITE(21,1002) ITER,SSQ
      WRITE(21,1001)
      WRITE(21,1030) (I,X(I),I=1,NTPAR)
      WRITE(6,1035) MFUN
      WRITE(6,1002) ITER,SSQ
      WRITE(6,1001)
      WRITE(6,1030) (I,X(I),I=1,NTPAR)
   40 CONTINUE
      WRITE(21,1032)
      WRITE(6,1032)
   20 CONTINUE
C
C FORMAT DECLA.
C
 1002 FORMAT(/,1X,'FOR ITERATION NUMBER ',I5,//,1X,'SSQ=',D13.6)
 1001 FORMAT(//,1X,'THE ACTUAL VALUES OF THE PARAMETERS ARE :',/)
 1030 FORMAT(1X,'X(',I2,')=',D13.6)
 1031 FORMAT(/,1X,'CALFUN HAS BEEN CALLED',I5,' TIMES')
 1032 FORMAT(/,70('*'),/)
 1035 FORMAT(/,'******WARNING******',/,
     *      1X,'THE NUMBER OF CALLS TO CALFUN HAS EXCEEDE',I5)
 2002 FORMAT(/,1X,'THE ESTIMATION HAS TERMINATED NORMALLY AFTER'
     *     ,I5,' ITERATIONS',//,
     *     1X,'SSQ=',D13.6)
 2001 FORMAT(//,1X,'THE FINAL VALUES OF THE PARAMETERS ARE :',/)
 2020 FORMAT(//,1X,'THE DESIRED ACCURACY IN SSQ WAS ACHIEVED')
 2021 FORMAT(//,1X,'THE DESIRED ACCURACY IN THE PARAMETERS WAS ACHIEVED'
     *)
      RETURN
      END
C
C***********************************************************************
C
C
C   SUBROUTINE MINLIN
C
C
C   OBJECTIVE: THIS SUBROUTINE FINDS THE MINIMUM OF AF FUNCTION
C   OF ONE VARIABLE
C
C
C***********************************************************************
C
      SUBROUTINE MINLIN(ITEST,X,F,MAXFUN,ABSACC,RELACC,XSTEP)
C
      COMMON /AUXMIN1/ IS,MC,DB,FB,FA,DA,D,IINC
C
C  MODIFICATION BY S.AGENA: AUXMIN1 AND AUXMIN2 WERE CREATED FROM
C  AUXMIN AS THE ALIGNEMENT CREATED A WARNING WHEN COMPLIED WITH THE
C  F77 (NASA, SEP. 96)
C
      COMMON /AUXMIN2/ FC,DC,XINC
      DOUBLE PRECISION ABSACC,D,DA,DB,DC,F,FA,FB,FC,RELACC,X,XINC,XSTEP
      GO TO (1,2,2),ITEST
 2 IS=6-ITEST
   ITEST=1
   IINC=1
   XINC=XSTEP+XSTEP
   MC=IS-3
   IF (MC) 4,4,15
 3 MC–MC+1
   IF (MAXFUN-MC) 12,15,15
12 ITEST=4
43 X=DB
   F=FB
   IF (FB-FC) 15,15,44
44 X=DC
   F=FC
15 RETURN
 1 GO TO (5,6,7,8),IS
 8 IS=3
 4 DC=X
   FC=F
   X=X+XSTEP
   GO TO 3
 7 IF (FC-F) 9,10,11
10 X=X+XINC
   XINC=XINC+XINC
   GO TO 3
 9 DB=X
   FB=F
   XINC=-XINC
   GO TO 13
11 DB=DC
   FB=FC
   DC=X
   FC=F
13 X=DC+DC-DB
   IS=2
   GO TO 3
 6 DA=DB
   DB=DC
   FA=FB
   FB=FC
32 DC=X
   FC=F
   GO TO 14
 5 IF (FB-FC) 16,17,17
17 IF (F-FB) 18,32,32
18 FA=FB
   DA=DB
19 FB=F
```

```
   DB=X
   GO TO 14
 16 IF (FA-FC) 21,21,20
 20 XINC=FA
   FA−FC
   FC=XINC
   XINC=DA
   DA=DC
   DC=XINC
 21 XINC=DC
   IF ((D-DB)*(D-DC)) 32,22,22
 22 IF (F-FA) 23,24,24
 23 FC=FB
   DC=DB
   GO TO 19
 24 FA=F
   DA=X
 14 IF (FB-FC) 25,25,29
 25 IINC=2
   XINC=DC
   IF (FB-FC) 29,45,29
 29 D–(FA-FB)/(DA-DB)-(FA-FC)/(DA-DC)
   IF(D*(DB-DC))33,33,37
 37 D=0.5D0*(DB+DC-(FB-FC)/D)
   IF(DABS(D-X)-DABS(ABSACC))34,34,35
 35 IF(DABS(D-X)-DABS(D*RELACC))34,34,36
 34 ITEST=2
   GO TO 43
 36 IS=1
   X=D
   IF ((DA-DC)*(DC-D)) 3,26,38
 38 IS=2
   GO TO (39,40),IINC
 39 IF(DABS(XINC)-DABS(DC-D))41,3,3
 33 IS=2
   GO TO (41,42),IINC
 41 X=DC
   GO TO 10
 40 IF(DABS(XINC-X)-DABS(X-DC))42,42,3
 42 X=0.5D0*(XINC+DC)
   IF ((XINC-X)*(X-DC)) 26,26,3
 45 X=0.5D0*(DB+DC)
   IF ((DB-X)*(X-DC)) 26,26,3
 26 ITEST=3
   GO TO 43
   END
C
C**********************************************************************
```

```
C
C
C TITLE:    SUBROUTINE OUT
C
C
C OBJECTIVE: CREATION OF OUTPUT FILE
C
C
C*********************************************************************
C
   SUBROUTINE OUT(NEXP)
C
   IMPLICIT REAL*8(A-H,O-Z)
C
    COMMON / COMW/ COMW (12)
    COMMON / CORQ/ COR (12), COQ (12)
    COMMON / U/ U(12,12), UT(12,12)
    COMMON/ XEX/ XPREX(1000), XWAEX(1000), XSAEX(1000)
    COMMON/ X/ X(1000,12)
    COMMON/ SOL/ SOLPR(1000)
    COMMON/ GAMMA/ GAMMA(12,1000)
    COMMON/ TEMP/ TEMPEX(1000)
    COMMON/ ERROR/ DEL1(1000),DEL2(1000),DEL3(1000),SSD, RMSD
    COMMON/ GAMM/ GAMCT(12,1000), GAMRS(12,1000)
C
   WRITE(40,1000)
1000 FORMAT('SOLUBILITY MODEL',///)
C
   WRITE(40,1100)
1100 FORMAT(/,'COMPONENT DATA',//,'NO.',4(2X),'MW',4(2X),'R',4(2X)
   *   ,'Q'//)
   DO 100 I=1,12
     WRITE(40,1200) I,COMW(I),COR(I),COQ(I)
100  CONTINUE
1200 FORMAT(I4,3(F12 4))
C
   WRITE(40,1500)
1500 FORMAT(///,'UNIQUAC PARAMETERS',//,'U(I,J)',//)
   DO 200 I=1,12
     WRITE(40,1600) I,(U(I,J),J=1,12)
200  CONTINUE
1600 FORMAT(I2,12(F12.4))
C
   WRITE(40,3500)
3500 FORMAT(///,'UNIQUAC PARAMETERS', /,'UT(I,J)',//)
   DO 3200 I=1,12
     WRITE(40,3600) I,(UT(I,J),J=1,12)
3200  CONTINUE
```

```
3600 FORMAT(I2,12(F12.4))
C
   WRITE(40,1700)
1700 FORMAT(///,'SOLUBILITY RESULTS',//,
   *     'CAL',4(2X),'EXP',4(2X),'DELT',4(2X),'d X/X'//)
   DO 300 I=1,NEXP
     WRITE(40,1800) SOLPR(I),XPREX(I),DEL1(I),DEL2(I)
300  CONTINUE
1800 FORMAT(4(D12.4))
C
   WRITE(40,1900)
1900 FORMAT(/,'ERRORS',//,'SSD',4(2X),'RMSD',//)
     WRITE(40,2000) SSD,RMSD
2000 FORMAT(2(D12.4))
C
   WRITE(40,2500)
2500 FORMAT(///,'PROTEIN GAMMA CAL.',//,
   *     'LN CT',4(2X),'LN RS',4(2X),'GAMMA'//)
   DO 500 J=1,NEXP
C     DO 500 I=1,4
      WRITE(40,2600) J,4,GAMCT(4,J),GAMRS(4,J),GAMMA(4,J)
500  CONTINUE
2600 FORMAT(2(I2),3(D12.4))
C
   WRITE(40,2700)
2700 FORMAT(///,'GAMMA CAL.',//,'H2O',4(2X),'PROT.',4(2X)//)
   DO 800 J=1,NEXP
    WRITE(40,2800) J,GAMMA(1,J),GAMMA(4,J)
800  CONTINUE
2800 FORMAT(1(I2),2(D12.4))
C
   WRITE(40,2900)
2900 FORMAT(///,'EXP. DATA',//,'TEMP.',4(2X),
   *     'X H2O',4(2X),'X PROT.',4(2X),'X SALT'//)
   DO 900 J=1,NEXP
    WRITE(40,3000) J,TEMPEX(J),XWAEX(J),XPREX(J),XSAEX(J)
900  CONTINUE
3000 FORMAT(1(I2),4(D12.4))
C
   RETURN
   END
C
C*******************************************************************
C
C
C TITLE:   SUBROUTINE PSOLCAL
C
C
```

```
C OBJECTIVE: CALCULATION OF SOLUBILITY AND OBJECTIVE FUNCTION
C
C
C********************************************************************
C
      SUBROUTINE PSOLCAL(N,F,X)
C
      IMPLICIT REAL*8(A-H,O-Z)
C
      COMMON/ NEXP/ NEXP
      COMMON/ CORQ/ COR(12), COQ(12)
      COMMON/ ABC/ A,B,C
      COMMON/ IABC/ IA,IB,IC
      COMMON/ EST/ IUEST(12,12), IUTEST(12,12)
      COMMON/ U/ U(12,12), UT(12,12)
C
      DIMENSION F(2000), X(40)
C
      L=0
C
      IF (IA.NE.0) THEN
        L=L+1
        A=X(L)
      END IF
C
      IF (IB.NE.0) THEN
        L=L+1
        B=X(L)
      END IF
C
      IF (IC.NE.0) THEN
        L=L+1
        C=X(L)
      END IF
C
      DO 100 K=1,12
       DO 100 KK=K,12
         IF (IUEST(K,KK).NE.0) THEN
           L=L+1
           U(K,KK)=X(L)
         END IF
100    CONTINUE
C
      DO 200 K=1,12
       DO 200 KK=K,12
         IF (IUTEST(K,KK).NE.0) THEN
           L=L+1
           UT(K,KK)=X(L)
      END IF
200   CONTINUE
C
    DO 300 I=1,12
      DO 300 J=1,12
        U(J,I)=U(I,J)
        UT(J,I)=UT(I,J)
300   CONTINUE
C
    CALL PUNIQUAC(NEXP)
    CALL PSOL(NEXP)
    CALL POBJ(NEXP,F)
    CALL OUT(NEXP)
C
    RETURN
    END
C
C****************************************************************
C
C
C  TITLE:     SUBROUTINE PSOL
C
C
C  OBJECTIVE: CALCULATION OF SOLUBILITY
C
C
C****************************************************************
C
    SUBROUTINE PSOL(NEXP)
C
    IMPLICIT REAL*8(A-H,O-Z)
C
    COMMON/ ABC/ A,B,C
    COMMON/ SOL/ SOLPR(1000)
    COMMON/ XEX/ XPREX(1000), XWAEX(1000), XSAEX(1000)
    COMMON/ XUX/ XUX(1000)
    COMMON/ GAMMA/ GAMMA(12,1000)
    COMMON/ TEMP/ TEMPEX(1000)
    COMMON/ ERROR/ DEL1(1000),DEL2(1000),DEL3(1000),SSD,RMSD
C
    SUM1=0.D0
    SUM2=0.D0
    SUM3=0.D0
    SUM4-0.D0
C
    DO 100 J=1,NEXP
       S=0
       S=DEXP(A+(B/TEMPEX(J))+C*(DLOG(TEMPEX(J))))
```

```
         SOLPR(J)-XUX(J)*S/GAMMA(4,J)
C
         DEL1(J)=DABS(XPREX(J)-SOLPR(J))
         SUM1=SUM1+DEL1(J)
         DEL2(J)=DABS((XPREX(J)-SOLPR(J))/XPREX(J))
         SUM2=SUM2+DEL2(J)
         DEL3(J)=((XPREX(J)-SOLPR(J))/XPREX(J))
         SUM3=SUM3+DEL3(J)
         SUM4=SUM4+((DEL3(J))**2)
100   CONTINUE
C
      SSD-SUM4
      RMSD=SQRT(SSD/NEXP)
C
      RETURN
      END
C
C*****************************************************************
C
C
C TITLE:     SUBROUTINE POBJ
C
C
C OBJECTIVE: CALCULATES THE OBJECTIVE FUNCTION VALUE, OBJ [-]
C
C
C*****************************************************************
C
      SUBROUTINE POBJ(NEXP,F)
C
      IMPLICIT REAL*8(A-H,O-Z)
C
      COMMON/ XEX/ XPREX(1000), XWAEX(1000), XSAEX(1000)
      COMMON/ SOL/ SOLPR(1000)
C
      DIMENSION F(2000)
C
      FB=0.D0
C
      DO 200 J=1,NEXP
         FB=((XPREX(J)-SOLPR(J))/XPREX(J))
         F(J)=FB
200   CONTINUE
C
      RETURN
      END
C
C*****************************************************************
```

```
C
C
C TITLE:    SUBROUTINE PUNIQUAC
C
C
C OBJECTIVE: CALCULATION OF ACTIVITY COEFFICIENTS WITH THE
C UNIQUAC MODEL (UNSYMMETRIC)
C
C
C*******************************************************************
C
   SUBROUTINE PUNIQUAC(NEXP)
C
   IMPLICIT REAL*8(A-H,O-Z)
C
   COMMON / RQ / RR(12),QQ(12)
   COMMON / CORQ / COR(12), COQ(12)
   COMMON / XTR / XTR(12)
   COMMON / X / X(1000,12)
   COMMON / PT / T
   COMMON / TEMP / TEMPEX(1000)
   COMMON / PM / PM(12)
   COMMON / COMW / COMW(12)
   COMMON / U / U(12,12),UT(12,12)
C
   COMMON / GAM / GAM(12)
   COMMON /GCT / GCT(12)
   COMMON /GRS / GRS(12)
   COMMON / GAMLN / GAMLN(12)
   COMMON / GAMMLN / GAMMLN(12,1000)
   COMMON / GAMMA / GAMMA(12,1000)
   COMMON / GAMM / GAMCT(12,1000), GAMRS(12,1000)
C
   DO 200 I=1,12
       RR(I)=COR(I)
       QQ(I)=COQ(I)
C
       PM(I)=COMW(I)
C
200  CONTINUE
C
   DO 300 J=1,NEXP
C
       T=TEMPEX(J)
C
       XTR(1)=X(J,1)
       XTR(4)=X(J,4)
       XTR(5)=X(J,5)
```

```
        XTR(6)=X(J,6)
C
        CALL ACTCOF
C
        DO 320 I=1,12
            GAMMA(I,J)=GAM(I)
            GAMMLN(I,J)= GAMLN(I)
            GAMCT(I,J)=GCT(I)
            GAMRS(I,J)=GRS(I)
320       CONTINUE
C
300  CONTINUE
C
    RETURN
    END
C
C********************************************************************
C
C
C   SUBROUTINE ACTCOF
C
C
C   OBJECTIVE: THIS SUBROUTINE CALCULATES THE ACTIVITY COEFFICIENTS
C   USING UNIQUAC MODEL
C
C
C********************************************************************
C
    SUBROUTINE ACTCOF
C
    IMPLICIT REAL*8(A-H,O-Z)
C
    COMMON / GAM / GAM(12)
    COMMON / GCT / GCT(12)
    COMMON / GRS / GRS(12)
    COMMON / GAMLN / GAMLN(12)
C
    DO 2000 I=1,12
      GAM(I) = 0.D0
      GRS(I) = 0.D0
      GCT(I) = 0.D0
2000  CONTINUE
C
      CALL GAMACT
      CALL GAMARS
      DO 1000 I=1,12
        GAMLN(I)=GCT(I)+GRS(I)
        CON=GAMLN(I)
        IF (CON.GE.174.673D0) THEN
          GAM(I)=7.23D75
        ELSE
          IF (CON.LE.-174.673D0) THEN
            GAM(I)=0.0D0
          ELSE
            GAM(I) = DEXP(CON)
          ENDIF
        ENDIF
1000    CONTINUE
C
    RETURN
    END
C
C*********************************************************************
C
C
C   SUBROUTINE GAMACT
C
C
C   OBJECTIVE: THIS SUBROUTINE CALCULATES THE COMBINATORIAL TERM
C
C
C*********************************************************************
C
    SUBROUTINE GAMACT
C
    IMPLICIT REAL*8(A-H,O-Z)
C
    COMMON / XTR / XTR(12)
    COMMON / RQ / RR(12),QQ(12)
    COMMON / PTH / PH(12),TH(12),THE(12)
    COMMON / GCT / GCT(12)
    COMMON / ZCOOR / ZCOOR(12)
C
    SR = 0.D0
    SQ = 0.D0
    DO 1000 I=1,12
      SR = SR + XTR(I)*RR(I)
      SQ = SQ + XTR(I)*QQ(I)
1000  CONTINUE
C
    DO 1010 I=1,12
      PH(I)  = RR(I)/SR
      TH(I)  = QQ(I) SQ
      THE(I) = XTR(I)*TH(I)
1010  CONTINUE
C
```

```
  DO 1020 I=1,12
    PHI   = PH(I)
    PHT   = PHI/TH(I)
    GCI   = DLOG(PHI) - PHI
    GCT(I) = GCI-1.D0/2.D0*ZCOOR(I)*QQ(I)*(DLOG(PHT)-PHT)
1020 CONTINUE
C
  DO 1030 I=1,12
    IF (I.EQ.1) THEN
      AUX = 1.D0/2.D0*ZCOOR(I)*QQ(I)
      GCT(I) = GCT(I) + 1.D0 - AUX
    ENDIF
1030 CONTINUE
C
  DO 1040 I=1,12
    IF (I.NE.1) THEN
      RAZ1  = RR(I)/RR(1)
      RAZ2  = QQ(I)/QQ(1)
      RAZ   = RAZ1/RAZ2
      AUX   = 1.D0/2.D0*ZCOOR(I)*QQ(I)
      GCT(I) = GCT(I) - DLOG(RAZ1) + RAZ1 - AUX*(RAZ - DLOG(RAZ))
    ENDIF
1040 CONTINUE
C
  RETURN
  END
C
C***********************************************************************
C
C
C  SUBROUTINE GAMARS
C
C
C  OBJECTIVE: THIS SUBROUTINE CALCULATES THE RESIDUAL TERM
C
C
C***********************************************************************
C
  SUBROUTINE GAMARS
C
  IMPLICIT REAL*8(A-H,O-Z)
C
  COMMON / PT / T
  COMMON / GRS / GRS(12)
  COMMON / CAUX / US(12),S(12),A(12),C(12,12)
  COMMON / RQ / RR(12),QQ(12)
C
  CALL GARS01
```

```
      CALL GARS03
      CALL GARS04
C
      DO 2110 I=1,12
        IF (I.EQ.1) THEN
          GRS(I) = QQ(I)*(1.D0 - DLOG(S(I)) -A(I))
        ENDIF
2110  CONTINUE
C
      DO 2120 I=1,12
        IF (I.NE.1) THEN
          GRS(I) = QQ(I)*(-DLOG(S(I))-A(I)+US(I))
        ENDIF
2120  CONTINUE
C
      RETURN
      END
C
C**********************************************************************
C
C
C     SUBROUTINE GARS01
C
C
C     OBJECTIVE: THIS SUBROUTINE CALCULATES THE RESIDUAL TERM
C
C
C**********************************************************************
C
      SUBROUTINE GARS01
C
      IMPLICIT REAL*8(A-H,O-Z)
C
      COMMON / PT / T
      COMMON / UM / UM(12,12)
      COMMON / U / U(12,12),UT(12,12)
      COMMON / CAUX / US(12),S(12),A(12),C(12,12)
C
      DO 1030 I=1,12
        DO 1030 J=I,12
          UM(I,J)=U(I,J)+(UT(I,J)*(T-300))
          UM(J,I)=UM(I,J)
1030    CONTINUE
C
      DO 2000 I=1,12
        IF (I.NE.1) THEN
          US(I) = -(UM(1,I)-UM(I,I))/T+DEXP(-(UM(I,1)-UM(1,1))/T)
        ENDIF
```

```
2000 CONTINUE
C
   RETURN
   END
C
C*********************************************************************
C
C
C   SUBROUTINE GARS03
C
C
C   OBJECTIVE: THIS SUBROUTINE CALCULATES THE RESIDUAL TERM
C
C
C*********************************************************************
C
   SUBROUTINE GARS03
C
   IMPLICIT REAL*8(A-H,O-Z)
C
   COMMON / PT / T
   COMMON / UM / UM(12,12)
   COMMON / PME / PME(12,12)
C
C  exp fortran limitation
C
   DO 2020 I=1,12
     DO 2010 J=1,12
       CONST = - (UM(I,J)-UM(J,J))/T
       IF (CONST.GE.174.673D0) THEN
         PME(I,J) = 7.23D75
       ELSE
         IF (CONST.LE.-174.673D0) THEN
           PME(I,J) = 0.0D0
         ELSE
           PME(I,J) = DEXP(CONST)
         ENDIF
       ENDIF
2010   CONTINUE
2020 CONTINUE
C
   RETURN
   END
C
C*********************************************************************
C
C
C   SUBROUTINE GARS04
```

```
C
C
C   OBJECTIVE: THIS SUBROUTINE CALCULATES THE RESIDUAL TERM
C
C
C**************************************************************
C
   SUBROUTINE GARS04
C
   IMPLICIT REAL*8(A-H,O-Z)
C
   COMMON / PTH / PH(12),TH(12),THE(12)
   COMMON / PME / PME(12,12)
   COMMON / CAUX / US(12),S(12),A(12),C(12,12)
C
   DO 2040 N=1,12
     SN = 0.D0
     DO 2030 K=1,12
       SN = SN + THE(K)*PME(K,N)
2030    CONTINUE
     S(N) = SN
2040  CONTINUE
C
   DO 2060 N=1,12
     AN = 0.D0
     DO 2050 K=1,12
       C(N,K) = PME(N,K)/S(K)
       AN = AN + THE(K)*C(N,K)
2050    CONTINUE
     A(N) = AN
2060  CONTINUE
C
   RETURN
   END
```

References

Abrams, D.S., Prausnitz, J.M. (1975) Statistical thermodynamics of liquid mixtures: a new expression for the excess Gibbs energy of partly or completely miscible systems. AIChE J., 21 (1), 116.

Agena, S.M., Wai, P. and Bogle, I.D.L. (1996) Property prediction methods for biochemical synthesis and design. Research Event of the Institution of Chemical Engineers, Leeds, UK

Agena, S.M., Pusey, M.L. and Bogle, I.D.L. (1997a) Solid - liquid equilibrium model for biopolymers. Conference of the American Institute of Chemical Engineers, Spring Meeting, Houston, USA

Agena, S.M., Bogle, I.D.L. and Pessoa, F.L.P. (1997b) An activity coefficient model for proteins. Biotech. Bioeng., 55 (1) 65 - 71

Agena S.M., Pusey M., Bogle I.D.L. (1998a) A model for protein solubility. Submitted.

Agena S.M., Bogle I.D.L., Pusey M. (1998b) Studies of protein solution properties using osmotic pressure measurements. Submitted.

Asenjo, J.A. (ed. Plye, D.L.) (1990) Separations for biotechnology. Elsevier.

Bailey, J.E., Ollis, D.F. (1986). Biochemical engineering fundamentals. McGraw-Hill.

Barnicki S.D., James R.F. (1990) Separation system synthesis: a knowledge-based approach. 1. Liquid mixture separation. Ind. Eng. Chem. Res., 29, 421.

Blundell T.L., Johnson L.N. (1976) Protein crystallography. Academic Press.

Bogle, I., Wai, P., Agena, S., Gani, R., Bagherpour, K., Nielsen, J., Carlsen, M., Aracil, J., Martinez, M., Bautista, F., Soriano, R., Coteron, A., Norodoslawsky, M., Altenburger, J., Schichl, M. and Halasz, L. (1996) Process synthesis, design and simulation of integrated biochemical processes. Fifth World Congress of Chemical

Engineering, San Diego, USA

Bondi, A. (1968) Physical properties of molecular crystals, liquids and glasses. John Wiley & Sons.

Bonnerjea, J., Oh, S., Hoare, M., Dunnill, P. (1986) Protein purification: the right step at the right time. Bio/Tech., 4, 954.

Burgess, R.R. (eds. Oxender, D.L., Fox, C.F.) (1988) Protein engineering. Alan R Liss, Inc.

Cacioppo, E., Pursey, M.L. (1991) The solubility of the tetragonal form of hen egg white lysozyme from pH 4.0 to 5.4. J. Crystal Growth, 114, 286.

Cacioppo, E., Pusey, M.L. (1992) The effect of acid treatment and calcium ions on the solubility of concanavalin A. J. Crystal Growth, 122, 208.

Carbonnaux C., Ries-Kautt M., Ducruix A. (1995) Relative effectiveness of various anions on the solubility of acidic hypoderma lineatum collagenase at pH 7.2. Protein Science, 4, 2123.

Carter D.C., He X.-M., Munson S.H., Twigg P.D., Gernert K.M., Broom M.B., Miller T.Y. (1989) Three-dimensional structure of human serum albumin. Science, 244, 1195.

Chiew, Y.C., Kuehner, D., Blanch, H.W., Prausnitz, J.M. (1995) Molecular thermodynamics for salt - induced protein precipitation. AIChE J., 41, 2150.

Chothia, C. (1975) Structural invariants in protein folding. Nature, 254, 304.

Christensen, L.K. (1952) Denaturation and enzymatic hydrolysis of lactoglobulin. Comp.-rend. Lab. Carlsberg, 28, 37.

Cohn, E.J (1925) The physical chemistry of the proteins. Physiol. Rev., 5, 349.

Cohn, E.J., Edsall, J.T. (1943) Proteins, amino acids and peptides. Reinhold Publishing Corporation.

Constantinou L., Gani R. (1994) A new group-contribution method for the estimation of properties of pure compounds. AIChE J., 40, 1697.

Coutinho J.A.P., Andersen S.I., Stenby E.H. (1994) *Evaluation of activity coefficient models in prediction of alkane solid-liquid equilibria.* IVC-SEP report, Department of Chemical Engineering, Technical University of Denmark (DTU), Lyngby, Denmark.

Cranfield, R.E., Liu, A.K. (1965) The disulfide bonds of egg white lysozyme. J. Bio. Chem., 240 (5), 1997.

Creighton, T.E. (1984) Proteins: structures and molecular properties. W.H. Freeman and Company.

Debye P., Hückel E. (1923) Zur Theory der Elektrolyte. I. Gefrierpunktserniedrigung und verwandte Erscheinungen. Phy. Zeitsch., 24 (9), 185.

Debye P. (1927) Das electrische Ionenfeld und das Aussalzen. Z. physik. Chem., 130, 56.

Deschamps J.R., Miller C.E., Ward K.B. (1995) Rapid purification of recombinant green fluorescent protein using the hydrophobic properties of an HPLC size - exclusion column. Protein Expres. Purif., 6, 555.

Dixon M., Webb E.C. (1961) Enzyme fractionation by salting out: A theoretical note. Advan. Protein Chem., 16, 197.

Ducruix, A., Giege, R. (1992) Crystallization of nucleic acids and proteins. A practical approach. IRL Press.

Edsall, J.T., Wyman, J. (1958) Biophysical chemistry. Academic Press Inc.

Elbro H.S., Fredenslund Aa., Rasmussen P. (1991) Group contribution method for the prediction of liquid densities as a function of temperature of solvents, oligomers, and polymers. Ind. Eng. Chem. Res., 30 (12), 2577.

Elmore D.T. (1968) Peptides and proteins. Cambridge University Press.

Ewing, F., Forsythe, E., Pusey, M. (1994) Orthorombic lysozyme solubility. Acta Cryst., D50, 424.

Forsythe, E.L., Pusey, M.L. (1996) The effects of acetate buffer concentration on lysozyme solubility. J. Crystal Growth, 168, 112.

Foster P.R., Dunnill P., Lilly M.D. (1971) Salting-out of enzymes with ammonium sulfate. Biotech. Bioeng., XIII, 713.

Fraaije, J.G.E.M., Norde, W., Lyklema, J. (1991) Interfacial thermodynamics of protein adsorption, ion co - adsorption and ion binding in solution. II. Model interpretation of ion exchange in lysozyme chromatography. Biophy. Chem., 40, 317.

Fredenslund, Aa., Jones, R.L., Prausnitz, J.M. (1975) Group-contribution estimation of activity coefficients in nonideal liquid mixtures. AIChE J., 21, 1086.

Fredenslund Aa. (1989) UNIFAC and related group contribution models for phase equilibria. Fluid Phase Equil., 52, 135.

Freifelder D. (1986) Molecular biology. Jones and Barlett Publishers.

Gmehling, J.G., Anderson, T.F., Prausnitz, J.M. (1978) Solid - liquid equilibria using UNIFAC. Ind. Eng. Chem. Fundam., 17 (4), 269.

Green A.A. (1931) Studies in the physical chemistry of the proteins. VIII. The solubility of hemoglobin in concentrated salt solutions. A study of the salting out of proteins. J. Biol. Chem., 93, 495.

Green A.A. (1932) Studies in the physical chemistry of the proteins. X. The solubility of hemoglobin in solutions of chlorides and sulfates of varying concentration. J. Biol. Chem., 95, 47.

Green A.A., Hughes W.L. (1955) Protein fractionation on the basis of solubility in aqueous solutions of salts and organic solvents. Methods in Enzymology, I, 67.

Guntelberg, A.V., Linderstrom-Lang, K. (1949) Osmotic pressure of plakalbumin and ovalbumin solutions. Compt.-rend. Lab. Carlsberg, 27, 1.

Gupta, R.B., Heidemann, R.A. (1990) Solubility models for amino acids and antibotics. AIChE J., 36 (3), 333.

Hansen, H.K., Rasmussen, P., Fredenslund, Aa., Schiller, M., Gmehling, J. (1991) Vapor-liquid equilibria by UNIFAC group contribution. 5. Revision and extension. Ind. Eng. Chem. Res., 30, 2352.

Haynes, C.A., Tamura, K., Korfer, H.R., Blanch, H.W., Prausnitz, J.M. (1992)

Thermodynamic properties of aqueous α-chymotrypsin solutions from membrane osmometry measurements. J. Phys. Chem., 96, 905.

He X.-M., Carter D.C. (1992) Atomic structure and chemistry of human serum albumin. Nature, 358, 209.

Ho, J.X., Holowachuk, E.W., Norton, E.J., Twigg, P.D., Carter, D.C. (1993) X-ray and primary structure of horse serum albumin (Equus caballus) at 0.27-nm resolution. Eur. J. Biochem., 215, 205.

Hofmeister, F. (1888) Zur Lehre von der Wirkung der Salze. Arch. Exp. Pathol. Pharmakol., 24, 247.

Hoshino D., Nagahama K., Hirata M. (1982) Prediction of acentric factor of alkanes by the group contribution method. J. Chem. Eng. Japan, 15 (2), 153.

Hückel E. (1925) Zur Theorie konzentrierterer wässeriger Lösungen starker Elektrolyte. Phy. Zeitsch., 26 (2), 93.

Israelachvili J.N. (1985) Intermolecular and surface forces. Academic Press.

Jakoby, W.B. (1968) A technique for the crystallization of proteins. Analytical Biochem., 26, 295.

Jaksland C.A., Gani R., Lien K.M. (1995) Separation process design and synthesis based on thermodynamic insight. Chem. Eng. Science, 50 (3), 511.

Janin, J. (1976) Surface area of globular proteins. J. Mol. Biol., 105, 15.

Judge R. (1997) Personal communication.

Kikic I., Alessi P., Rasmussen P., Fredenslund Aa. (1980) On the combinatorial part of the UNIFAC and UNIQUAC models. Can. J. Chem. Eng., 58, 253.

Klincewicz K.M., Reid R.C. (1984) Estimation of critical properties with group contribution methods. AIChE J., 30 (1), 137.

Kontogeorgis, G.M., Nikolopoulos, G.I., Fredenslund, Aa., Tassios, D.P. (1997) Improved models for the prediction of activity coefficients in nearly athermal mixtures. Part II. A theoretically-based GE-model based on the van der Waals partition function. Fluid Phase Equil., 127, 103.

Kuehner, D.E., Blanch, H.W., Prausnitz, J.M. (1996) Salt - induced protein precipitation; Phase equilibria from an equation of state. Fluid Phase Equil., 116, 140.

Laidler, K.J., Meiser, J.H. (1982) Physical chemistry. The Benjamin/Cummings Publishing Company, Inc.

Larsen, B.L., Rasmussen, P., Fredenslund, Aa. (1987) A modified UNIFAC group-contribution model for prediction of phase equilibria and heat of mixing. Ind. Eng. Chem. Res., 26, 2274.

Lehninger, A.L. (1982) Principles of biochemistry. Worth Publishers, Inc.

Lydersen A.L. (1955) *Estimation on critical properties of organic compounds by the method of group contribution.* Report no.3, College of Engineering, University of Wisconsin.

Lyman, W.J., Reehl, W.F., Rosenblatt, D.H. (1982) Handbook of chemical property estimation methods. Mc Graw - Hill.

Matthews, B.W., Sigler, P.B., Henderson, R., Blow, D.M. (1967) Three-dimensional structure of tosyl - α - chymotrypsin. Nature, 214, 652.

Mavrovouniotis M.L., Bayol P., Lam T.K.M., Stephanopoulos G., Stephanopoulos G. (1988) A group contribution method for the estimation constants for biochemical reactions. Biotech. Techniques, 2 (1), 23-28.

Mavrovouniotis M.L. (1990) Group contributions for estimating standard Gibbs energies of formation of biochemical compounds in aqueous solution. Biotech. Bioeng., 36, 1070-1082.

McPherson A., Rich A. (1973) X-ray crystallographic study of the quaternary structure of canavalin. J. Biochem., 74, 155.

McPherson A., Spencer R. (1975) Preliminary structure analysis of canavalin from Jack Bean. Arch. Biochem. Biophys. Res. Commun., 57, 494.

McPherson A. (1982) Preparation and analysis of protein crystals. John Wiley & Sons.

Melander, W., Horvath, C. (1977) Salt effects on hydrophobic interactions in precipitation and chromatography of proteins: An interpretation of the lyotropic series. Archives Biochem. Biophy., 183, 200.

Mullin J.W. (1993) Crystallization. Butterworth-Heinemann.

Nass, K.K. (1988) Representation of the solubility behavior of amino acids in water. AIChE J., 34 (8), 1257.

Nicolaisen, H., Rasmussen, P., Sorensen, J.M. (1993) Correlation and prediction of mineral solubilities in the reciprocal salt system (Na^+, K^+) (Cl^-, SO_4^{2-})-H_2O AT 0-100°C. Chem. Eng. Sci., 48 (18), 3149.

Nicolaisen, H. (1994) *Phase equilibria in aqueous electrolyte solutions.* Ph.D. thesis, Department of Chemical Engineering, Technical University of Denmark (DTU), Lyngby, Denmark.

Niktari M., Richardson P., Ravenhall P., Flanagan M.T., Molloy J., Hoare M. (1989) The modelling and control of fractionation processes for enzyme and protein purification. Proceedings Am. Control Conf., 3, 2436.

Nisbet, A.D., Saundry, R.H., Moir, A.J.G., Fothergill, L.A., Fothergill, J.E. (1981) The complete amino-acid sequence of hen ovalbumin. Eur. J. Biochem., 115, 335.

Pauling L.C. (1960) The nature of the chemical bond. Cornell University Press.

Peres, A.M., Macedo, E.A. (1994) Representation of solubilities of amino acids using the UNIQUAC model for electrolytes. Chem. Eng. Science, 49 (22), 3803.

Perozzo M.A. (1997a) Personal communication.

Perozzo M.A. (1997b) *X-ray crystallographic structural analysis of aequorea victoria green fluorescent protein.* Ph.D. thesis, Department of Chemistry, The Catholic University of America, Washington, D.C., USA.

Pervaiz, S., Brew, K. (1985) Homology of β - lactoglobulin, serum retinol-binding protein, and protein HC. Science, 228, 335.

Pessoa, F.L.P., Rasmussen, P., Fredenslund, Aa. (1992) Calculation of vapor-liquid equilibria in water - sulfuric acid - sulfate salt systems using a revised extended UNIQUAC equation. Latin American Applied Research, 22, 195.

Pinho, S.P., Silva, C.M., Macedo, E.A. (1994) Solubility of amino acids: a group-contribution model involving phase and chemical equilibria. Ind. Eng. Chem. Res., 33, 1341.

Prausnitz, J.M., Lichtenthaler, R.N., Gomes de Azevedo, E. (1986) Molecular thermodynamics of fluid-phase equilibria. Prentice-Hall.

Press, W.H., Teukolsky, S.A., Vetterling, W.T., Flannery, B.P. (1986) Numerical recipes in Fortran. Cambridge University Press.

Przybycien, T.M., Bailey, J.E. (1989) Solubility - activity relationship in the inorganic salt - induced precipitation of α - chymotrypsin. Enz. Microb. Technol., 11, 264.

Pusey, M.L., Gernert, K. (1988) A method for rapid liquid-solid phase solubility measurements using the protein lysozyme. J. of Crystal Growth, 88, 419.

Pusey, M.L., Munson, S. (1991) Micro - apparatus for rapid determination of protein solubilities. J. Crystal Growth, 113, 385.

Pusey, M.L. (1996) Personal communication.

Pusey, M.L. (1997) Personal communication.

Reid, R.C., Prausnitz, J.M., Poling, B.E. (1987) The properties of gases & liquids. McGraw - Hill.

Renon, H., Prausnitz, J.M. (1968) Local composition in thermodynamic excess for liquid mixtures. AIChE J., 14 (1), 135.

Richards, F.M. (1974) The interpretation of protein structures: total volume, group volume distributions and packing density. J. Mol. Biol., 82, 1.

Ross, P.D., Minton, A.P. (1977) Analysis of non-ideal behavior in concentrated hemoglobin solutions. , J. Mol. Biol., 112, 437.

Roth, C.M., Neal, B.L., Lenhoff, A.M. (1996) Van der Waals interactions involving proteins. Biophy. J., 70, 977.

Sander, B., Fredenslund, Aa., Rasmussen, P. (1986) Calculation of vapour-liquid equilibria in mixed solvent/salt systems using an extended UNIQUAC equation. Chem. Eng. Science, 41 (5), 1171.

Scopes, R. K. (1987) Protein purification. Springer Verlag.

Shih, Y.-C., Prausnitz, J. M., Blanch, H.W. (1992) Some Characteristics of protein precipitation by salts. Biotech. Bioeng., 40, 1155.

Siirola J.J., Rudd D.F. (1971) Computer-aided synthesis of chemical process designs. Ind. Eng. Chem. Fundam., 10 (3), 353.

Smith, J.M., van Ness, H.C. (1987) Introduction to chemical engineering thermodynamics. McGraw-Hill.

Sober, H.A. (ed.) (1968) Handbook of biochemistry. CRC Press.

Soehnel O., Garside J. (1992) Precipitation. Butterworth-Heinemann.

Tanford, C. (1961) Physical chemistry of macromolecules. John Wiley & Sons.

Taratuta, V.C., Holschbach, A., Thurston, G.M., Blankschtein, D., Benedek, G.B. (1990) Liquid-liquid phase separation of aqueous lysozyme solution: effects of pH and salt identity. J. Phy. Chem., 94 (5), 2140.

Taylor, G.L. (ed. Crabbe M.J.C.) (1990) Structure prediction, enzyme biotechnology. Ellis -Horwood.

Teller, D.C. (1976) Accessible area, packing volumes and interaction surfaces of globular proteins. Nature, 260, 729.

Wai, P.P.C., Bogle I.D.L., Bagherpour K., Gani R. (1996) Process synthesis and simulation strategies for integrated biochemical process design. Computers. Chem. Eng., 20, S357.

Wheelwright, S.M. (1987) Designing downstream processes for large-scale protein purification. Bio/Tech., 5, 789.

Wheelwright S.M. (1991) Protein purification. Carl Hansen Verlag.

Wills, P.R., Comper, W.D., Winzor, D.J. (1993) Thermodynamic nonideality in macromolecular solutions: interpretation of virial coefficients. Arch. Biochem. Biophy. 300, 206.

Wilson, G.M. (1964) Vapour-liquid equilibrium XI. A new expression for the excess free energy of mixing, J. Am. Chem. Soc., 86, 127.