LBNL-46713
Preprint

# ERNEST ORLANDO LAWRENCE BERKELEY NATIONAL LABORATORY

# Calculation of Phase Diagrams for Aqueous Solutions

R.A. Curtis, H.W. Blanch, and J.M. Prausnitz

**Chemical Sciences Division**

November 2000

Submitted to
*Journal of Physical Chemistry*

REFERENCE COPY
Does Not
Circulate

Library Annex Reference
Lawrence Berkeley National Laboratory

LBNL-46713
Copy 1

**DISCLAIMER**

This document was prepared as an account of work sponsored by the United States Government. While this document is believed to contain correct information, neither the United States Government nor any agency thereof, nor The Regents of the University of California, nor any of their employees, makes any warranty, express or implied, or assumes any legal responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by its trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof, or The Regents of the University of California. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof, or The Regents of the University of California.

Ernest Orlando Lawrence Berkeley National Laboratory
is an equal opportunity employer.

## DISCLAIMER

This document was prepared as an account of work sponsored by the United States Government. While this document is believed to contain correct information, neither the United States Government nor any agency thereof, nor the Regents of the University of California, nor any of their employees, makes any warranty, express or implied, or assumes any legal responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by its trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof, or the Regents of the University of California. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof or the Regents of the University of California.

LBNL-46713

# "CALCULATION OF PHASE DIAGRAMS FOR AQUEOUS SOLUTIONS"

**R. A. Curtis; H. W. Blanch and J.M. Prausnitz**

Department of Chemical Engineering
University of California
and
Chemical Sciences Division
Lawrence Berkeley National Laboratory
University of California
Berkeley, CA 94720, U.S.A.

November 2000

This work was supported by the Director, Office of Science, Office of Basic Energy Sciences, Chemical Sciences Division of the U.S. Department of Energy under Contract Number DE-AC0-76SF00098.

# Calculation of Phase Diagrams for Aqueous Protein Solutions

R.A. Curtis, H.W. Blanch, and J.M. Prausnitz
*Chemical Engineering Department, University of California, Berkeley and Chemical Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, CA, 94720, Fax # (510)-643-1228*

## Abstract

In recent publications, phase diagrams have been generated based on simple models of globular proteins interacting via anisotropic interactions. In these models, protein solubility is determined from the favorable energetic interactions due to the formation of protein-protein contacts in the crystal that overcomes the unfavorable loss in entropy from constraining a protein molecule upon crystallization. In this work, we develop a statistical-mechanical description for protein crystallization where a key part is the quantitative calculation of this entropy loss. We calculate the entropic term from experimental crystallographic data for lysozyme and show that the empirical correlation of the osmotic second virial coefficient with lysozyme solubility corresponds to 6-8 protein contacts per protein crystal. In addition, our model predicts that the two-body potential of mean force between lysozyme molecules is highly anisotropic. This has important implications for determining the position of a fluid-fluid critical point metastable to the fluid-solid equilibrium. That position is important because, as shown previously, crystallization kinetics are maximized at temperatures slightly exceeding the fluid-fluid critical temperature.

## 1. Introduction

Obtaining protein crystals for x-ray diffraction has always been a difficult task. Although there exist empirical rules for determining conditions favorable for protein crystallization, protein crystallization remains predominantly a trial-and-error process. Every protein has different biophysical properties (e.g. surface hydrophobicity, charge distribution, size, and conformational lability) that control the solution behavior of the protein and consequently the protein's ability to crystallize. Finding a rational predictor for conditions favorable to crystallize proteins had remained elusive until George and Wilson[1] proposed that a window in the osmotic second virial coefficient, $B_2$, exists for protein crystallization. $B_2$ provides a direct measure of the protein-protein pair potential. As a necessary condition for protein crystallization, George and Wilson found that in aqueous systems, $B_2$ should be in the region $-2 \times 10^{-4}$ and $-8 \times 10^{-4}$ mLmol/g$^2$, where negative values of $B_2$ indicate attractive interactions. For $B_2$ more positive than $-2 \times 10^{-4}$

mLmol/g$^2$, the protein crystal solubility is sufficiently high that quality crystals cannot be obtained. For solutions where $B_2$ is more negative than -8 x $10^{-4}$ mLmol/g$^2$, amorphous precipitation is likely to occur because protein-protein attractions are so strong that the protein molecules do not have adequate time to orient themselves to form crystals before forming an amorphous agglomerate. Unfortunately, while many protein solutions satisfy the George and Wilson criteria, they do not form crystals over reasonable time scales. Consequently, there remains a need for a better crystallization diagnostic.

The amorphous phase formed by a higher concentration of aqueous protein molecules probably corresponds to the dense liquid phase of a liquid-liquid equilibrium. Furthermore, it was shown by Rosenbaum et al.[2] and by Poon[3] that the range of $B_2$ values in the crystallization window corresponds to conditions for the metastable liquid-liquid critical point in the vicinity of liquid-solid equilibrium. The presence of the metastable critical point is very important for protein crystallization because density fluctuations are enhanced in this region, lowering the free energy for formation of critically-sized nuclei and increasing the rate of crystallization from a characteristic time of months or weeks to hours.[4] Recent work[5,6] has focused on developing more specific crystallization diagnostics through determination of accurate protein phase diagrams. Theoretical studies of molecules interacting in isotropic fluids have shown that the range of the pair potential controls the distance between the critical point for liquid-liquid equilibrium and liquid-solid equilibrium.[7-10] In particular, liquid-liquid coexistence is metastable with respect to solid-liquid equilibrium when the range of the particle attraction is less than one quarter of the particle diameter.

Experimental evidence of the relationship between $B_2$ and the protein phase diagram has been provided by Guo et al.[11] who showed that there is a strong correlation of $B_2$ with protein-crystal solubility for a wide range of solution conditions. This correlation has important implications for understanding fluid-crystal thermodynamic behavior. $B_2$ provides a measure of the dilute two-body protein interaction Boltzmann-averaged over all possible solvent configurations and also over the orientations of the protein molecules. However, protein-crystal solubility is determined primarily by protein-protein interactions in the crystal where the protein molecules are closely packed and where the molecules sample only a very small number of orientations. Because of

the experimentally observed correlation between $B_2$ and protein solubility, it is likely that protein-protein interactions in solution can be extrapolated to describe interactions in the crystal. Further, it is likely that the pair potential in the liquid is highly anisotropic because such interactions are required to fix the orientation of the protein molecule in the crystal.[12,13] Finally, because the protein molecule is fixed translationally and orientationally in the crystal, the potential function for the protein molecules in the crystal depends only on the characteristic energy of the protein-protein "bonds" formed in the crystal. It is this energy that should be used in the calculation of $B_2$.

In this work we calculate aqueous protein phase equilibria for protein molecules modeled as hard spheres with sticky sites on the surface. With important modifications, our calculation follows that of Haas et al.[14] and of Sear.[15] Haas et al. calculated protein solubility from a simple cell model for the protein crystal and an "ideal-gas" (ideal dilute solution) phase for the liquid. The solubility is correlated with $B_2$ by relating the characteristic energy of the protein-protein contacts in the crystal to an expression for $B_2$ that includes an anisotropy factor. The anisotropy factor provides a measure of the fraction of orientations where the protein-protein interaction is attractive. Here, we extend this earlier calculation by using Wertheim's model[16,17] for associating fluids to calculate the protein pair potential in an aqueous medium. In Wertheim's model, the pair potential is given by an angle-average of a set of site-specific square-well potentials. In addition, we also include an entropic term in the phase-equilibrium calculation to account for the decrease in rotational and translational entropies of the protein molecule as it goes from the liquid to the crystal. Sear[15] has shown that this model can adequately predict the phase behavior of protein solutions for a range of parameter space. Here we investigate if the parameter space is consistent with the $B_2$ - solubility correlation and discuss the significance of the rotational entropic term that has been neglected in most previous work for calculating protein phase equilibria. Further, we investigate the location of the liquid-liquid critical point for fluids of molecules with different numbers of sticky sites.

## 2. Fluid-Phase Model

Protein molecules are modeled as hard spheres with sticky square-well sites located on the surfaces. We use the results of Wertheim[16] who has developed an accurate

first-order perturbation theory for calculating the free energy of a fluid of particles containing one or two sticky sites minus that of a reference fluid, where the particles contain zero sites. Jackson et al.[17] extended Wertheim's theory to include fluids where the molecules interact with more than two attractive sites. The hard-sphere fluid is used for the reference system.

The Helmholtz energy per particle, $a_f$, is given by the sum of two contributions:

$$\beta a_f = \beta a_{hs} + \beta a_{assoc} \tag{1}$$

where $a_{hs}$ is the Helmholtz energy per particle for a hard-sphere fluid, $a_{assoc}$ is the contribution to the Helmholtz energy per particle arising from the interacting sites, and $\beta = 1/k_BT$, where $k_B$ is Boltzmann's constant and T is temperature. For phase-equilibrium calculations, we require expressions for the chemical potential of the protein, $\mu$, and the compressibility factor, Z, given by relations similar to Equation (1):

$$\beta\mu_f = \beta\mu_{hs} + \beta\mu_{assoc} \tag{2}$$

and

$$Z_f = Z_{hs} + Z_{assoc} \tag{3}$$

The Helmholtz energy per particle, $a_{hs}$, is obtained from integration of the compressibility factor. The integration constant is determined by the low-density ideal-gas limit as follows

$$\beta a_{hs} - \beta a_{ig} = \int_0^{\rho_f} (Z_{hs} - 1)\frac{d\rho_f}{\rho_f} \tag{4}$$

where $a_{ig}$ is the Helmholtz energy per particle for an ideal-gas fluid at number density, $\rho_f$, and $Z_{hs}$ is the compressibility factor for a hard-sphere fluid. $Z_{hs}$ is given by the equation of state[18]

$$Z_{hs} = \frac{1 + \eta_f + \eta_f^2 - \eta_f^3}{(1 - \eta_f)^3} \tag{5}$$

where $\eta_f$ is the protein packing fraction ($=\pi\sigma^3/\rho_f$) of the fluid and $\sigma$ is the hard-sphere protein diameter.

To evaluate the integral in Equation (4), the ideal-gas contribution to the Helmholtz energy must be determined. Because there are no interactions between particles in an ideal solution, the partition function $Q_{ig}$ corresponding to an ideal gas, is

given by the product of the individual particle partition functions, $q_{ig}$. For the canonical ensemble in classical statistical mechanics, the Helmholtz energy is[19]

$$-\beta A_{ig} = \ln Q_{ig} = \ln \frac{q_{ig}^N}{N!\Lambda^{3N}} \tag{6}$$

where $A_{ig}$ is the Helmholtz energy of an ideal-gas fluid of N particles, and $\Lambda$ is the thermal wavelength of the particle. The particle partition function is obtained by integrating the single-particle configurational integral over all possible translational and orientational coordinates:

$$q_{ig} = \int d\mathbf{r} d\Omega = 8\pi^2 V. \tag{7}$$

The integral over the translational coordinates of a particle, **r**, gives the volume of the fluid, V. The integration over the set of Euler angles specifying the orientation of the molecule, $\Omega$, is included in Equation (7) because the available orientation space available to a protein molecule in the crystal is significantly lower than that for a molecule in the fluid. All possible coordinates are equally likely for a protein molecule in the fluid and the integration over rotational coordinates gives a factor of $8\pi^2$. The Helmholtz energy per particle, $a_{ig}$ is obtained from Equations (6) and (7) and $a_{ig} = A_{ig}/N$

$$\beta a_{ig} = \ln \frac{\rho_f \Lambda^3}{8\pi^2} - 1. \tag{8}$$

Integration of Equation (4) gives the full expression for $\beta a_{hs}$

$$\beta a_{hs} = \ln \frac{\rho_f \Lambda^3}{8\pi^2} - 1 + \frac{4\eta - 3\eta^2}{(1-\eta)^2}. \tag{9}$$

The chemical potential of the hard-sphere reference system can be obtained from

$$\beta \mu_{hs} = \beta a_{hs} + Z_{hs}. \tag{10}$$

Substituting Equations (5) and (9) into Equation (10) gives

$$\beta \mu_{hs} = \ln \frac{\Lambda^3}{8\pi^2} + \ln \rho_f + \frac{8\eta_f - 9\eta_f^2 + 3\eta_f^3}{(1-\eta_f)^3}. \tag{11}$$

The sum of the first two terms on the right side of Equation (11) gives the ideal-gas chemical potential and the last term is the contribution to the chemical potential from "turning on" hard-sphere interactions.

## 3. Wertheim's Theory of Associating Fluids

Figure 1a shows the pair potential of Wertheim's model. In this model, anisotropic "bonds" between site A on molecule 1 and site B on molecule 2 are represented by a site-specific square-well potential given by

$$-\Gamma(r_{12},\Omega_1,\Omega_2) = \varepsilon_f \quad \text{if} \quad r_{ab} < r_c$$

$$-\Gamma(r_{12},\Omega_1,\Omega_2) = 0 \quad \text{if} \quad r_{ab} > r_c \tag{12}$$

where $r_{ab}$ is the distance between sites a and b; $r_c$ is the width of the square-well potential or the site-site cut-off separation; $\Omega_1, \Omega_2$ are the sets of Euler angles denoting the positions of the interaction sites relative to the centers of the protein molecules 1 and 2, respectively; and $\varepsilon_f$ is the interaction energy between the sites. Wertheim's theory gives a simple expression for the Helmholtz energy of a hard-sphere fluid with M attractive sites (where "bonds" can only form between similar sites, i.e. only A-A, B-B bonds are allowed) relative to that for a fluid of hard spheres with no sticky sites.

$$\beta a_{assoc} = M\left[\ln X - \frac{X}{2} + \frac{1}{2}\right] \tag{13}$$

The interaction energy is assumed to be the same for all bonds. Consequently, the fraction of sites on the molecule that are not involved in bonding, X, is independent of the site label. By definition, a site on molecule 1 is bonded if the same site on molecule 2 is located within the cut-off separation. X is given by the equation of mass action

$$X = \frac{-1+(1+4\rho_f K)^{1/2}}{2\rho_f K}. \tag{14}$$

where K, related to the probability of a site forming a bond, is given by

$$K = 4\pi \int_{\sigma}^{\sigma+r_c} g_{hs}(r)\left\langle \exp(\hat{\varepsilon}_f) - 1\right\rangle_{\Omega_1,\Omega_2} r^2 dr. \tag{15}$$

Here, r is center-to-center separation of the two protein molecules, $\hat{\varepsilon}_f$ is the dimensionless energy parameter ($\hat{\varepsilon}_f = \beta\varepsilon_f$), and $g_{hs}$ is the radial distribution function of the hard-sphere reference system. The brackets in Equation (15) denote an unweighted average over all orientations of the two molecules. We assume that $r^2 g_{hs}$ is constant over the range of the attractive part of the square-well potential and is given by $\sigma^2 g_{hs}(\sigma^+)$, where $g_{hs}(\sigma^+)$ is the hard-sphere distribution function evaluated at surface-to-surface

contact. It is evaluated using the result obtained from the contact-value theorem[20] applied to the Carnahan-Starling equation of state [Equation (5)]:

$$g_{hs}\left(\sigma^{+}\right)=\frac{1-\frac{1}{2}\eta}{\left(1-\eta\right)^{3}}. \tag{16}$$

The change in the compressibility factor due to association is obtained from differentiating the Helmholtz energy with respect to packing fraction at constant temperature. The result is

$$\frac{Z_{assoc}}{M}=\eta_f\left[\frac{1}{X}-\frac{1}{2}\right]\left(\frac{\partial X}{\partial \eta_f}\right)_T. \tag{17}$$

The contribution to the chemical potential of the protein from the attractive-site interactions is determined from

$$\beta\mu_{assoc}=\beta a_{assoc}+Z_{assoc}=\ln X-\frac{X}{2}+\frac{1}{2}+\eta\left[\frac{1}{X}-\frac{1}{2}\right]\left(\frac{\partial X}{\partial \eta_f}\right)_T \tag{18}$$

where $\left(\partial X/\partial \eta_f\right)_T$ is calculated from Equations (14), (15), and (16).

The unweighted angle average of the site-site Mayer f-function is given by the product of a geometric factor and the unaveraged value of the function

$$\left\langle f_{mayer}\right\rangle_{\Omega_1,\Omega_2}=\left\langle \exp\left(\hat{\varepsilon}_f\right)-1\right\rangle_{\Omega_1,\Omega_2}=\left[\frac{\left(r_c+\sigma-r\right)^2\left(2r_c-\sigma+r\right)}{6\sigma^2 r}\right]\left(\exp\left(\hat{\varepsilon}_f\right)-1\right). \tag{19}$$

The geometric factor in the brackets of Equation (19) is the fraction of rotational space corresponding to one "bond". If "bonds" only form between similar sites, the second virial coefficient for a protein molecule with M sites of equal energy is given by

$$B_2=B_{2,hs}-\frac{M}{2}\int_{\sigma}^{\sigma+r_c}\left\langle \exp\left(\hat{\varepsilon}_f\right)-1\right\rangle_{\Omega}4\pi r^2 dr \tag{20}$$

where $B_{2,hs}$ is the hard-sphere contribution to $B_2$, equal to $2\pi\sigma^3/3$. Equation (20) is based on the assumption that two simultaneous "bonds" do not form for any orientation of the two protein molecules. Equation (20) can be simplified by performing the integration over volume. Substituting Equation (19) into Equation (20) and solving for $\hat{\varepsilon}_f$ gives

$$\hat{\varepsilon}_f = \ln\left[\frac{2}{V_{bond}(r_c)M}\left(B_2 - B_{2,hs}\right)+1\right] \tag{21}$$

where $V_{bond}$ is the volume available for the formation of one "bond" between two protein molecules. It follows that $MV_{bond}$ is the total volume available for "bonding" between two protein molecules.

As shown in Figure 1b, the condition that two bonds do not form between two molecules simultaneously is satisfied if the minimum distance between sites on the same molecule is given by

$$d_{site,min} = \sigma \cos^{-1}\left(1 - \frac{r_c}{\sigma}\right) \tag{22}$$

This condition is satisfied for all systems studied here.

Wertheim's first-order association theory does not account for possible ring-like clusters of molecules. In addition, the structure of the fluid is approximated by that of the reference system, a fluid containing only monomers. Both approximations are valid in the low-bonding limit where the fluid consists mainly of monomers and where higher-order clusters are not likely. For fluids of molecules with many sites, the approximations break down for strong interaction strengths at moderate densities of molecules where there is a large degree of association. Here, calculation of the liquidus branch of the liquid-solid equilibrium is not expected to be in significant error because the packing fractions of protein molecules are low in the fluid phase. However, calculation of the liquid-liquid critical point only gives quantitative results due to the approximations above and also due to neglecting the strong density fluctuations that occur at and near the critical point.

## 4. The Solid Phase

Protein crystallization is driven by strong anisotropic interactions that overcome the reversible work of fixing the position and orientation of the protein molecule in the crystal. The essential physics can be captured by using the cell model[19] with an extra term to account for the loss in rotational entropy upon transferring a protein molecule from solution to the crystal. In this model, a protein molecule is confined to a cell where it experiences an average field from its z nearest neighbors. Because the field is

averaged, it is independent of the positions of all other particles and the configurational integral of a single protein is uncoupled with respect to the positions and orientations of its nearest neighbors. Consequently the partition function of the crystal is given by the product of single-particle partition functions. The Helmholtz energy per particle for the crystal, $a_s$, is then given by

$$-\beta a_s = \ln \frac{q_s}{\Lambda^3} \tag{23}$$

The factor N! in the analogous equation for an ideal gas [Equation (8)], is omitted because the particles in the crystal can be distinguished by the locations of their sites on the crystal lattice.

To calculate the phase diagram it is necessary to establish an accurate expression for the single-particle partition function, $q_s$. In protein crystals, the position and orientations of a protein molecule are constrained by strong anisotropic interactions. The individual atoms of the molecule have vibrational amplitudes given by about 0.3 - 0.5 Å and the entire protein molecule has 0.2 - 0.25 Å of translational freedom in principal directions.[21] Because the protein surface area buried by a contact interaction is on the order of 200 - 1200 Å$^2$,[22] we expect that the small vibrations of neighboring protein molecules and their constituent atoms do not alter the field experienced by the constrained particle. We therefore assume that the field is constant for all possible configurations of the protein molecule in the crystal. The single-particle partition function is

$$q_s = \frac{\exp(-\beta u_s) \int d\mathbf{r} \int d\Omega}{\Lambda^3} \tag{24}$$

where $u_s$ is the intermolecular energy per protein molecule in the crystal. In Equation (24), the protein molecule is considered to be rigid such that **r** refers to a 3-dimensional vector that specifies the position of the center of mass of the protein molecule. Integration over **r** yields the volume available to the center of the protein molecule. The integral over the position vector of the protein molecule is given by

$$\int d\mathbf{r} = v_c \tag{25}$$

where $v_c$ is the volume available to the center of mass of the protein molecule, that is, the free volume. Similarly, the integral over the set of Euler angles in Equation (24) yields

the orientation space sampled by the protein molecule in the crystal. This space is denoted by $\Delta\Omega$ ($=\int d\Omega$).

Because we consider only the contribution of one configuration to the internal energy, the energy of this configuration is called the ground-state energy. If the protein forms z contacts of average interaction strength $\varepsilon_s$, the ground-state internal energy per particle is given by

$$u_s = -\left(\frac{z}{2}\right)\varepsilon_s . \tag{26}$$

Substitution of Equation (24) into Equation (23) gives the solid-phase Helmholtz energy per particle

$$\beta a_s = \ln \Lambda^3 - \ln\left(v_c \Delta\Omega\right) - \frac{z}{2}\hat{\varepsilon}_s \tag{27}$$

Equation (27) provides the simplest form of the Helmholtz energy that indicates the different contributions to the free energy of the solid.

The chemical potential is determined from the Helmholtz energy per particle by

$$\beta\mu_s = \beta a_s + Z_s \tag{28}$$

where $Z_s$ is the compressibility factor of the solid. As outlined in the Appendix, $Z_s$ can not be determined accurately for a protein crystal. At the experimental conditions discussed here, $Z_s$ is negligible when compared to $\beta a_s$. Consequently, the chemical potential is equated with the Helmholtz energy per particle.

## 5. Correlation of $B_2$ with Lysozyme Solubility

The empirical correlation of $B_2$ with protein solubility implies that protein-protein interactions in the infinitely dilute region of the fluid can be extrapolated to describe protein-protein interactions in the crystal. Therefore, we set the average interaction strength of the anisotropic "bonds" between protein molecules in the fluid equal to the average interaction strength of the protein-protein contacts in the crystal. The phase-equilibrium criteria are then given by

$$\mu_f\left(\eta_f, \hat{\varepsilon}\right) = \mu_s\left(\eta_s, \hat{\varepsilon}\right) \tag{29}$$

$$P_f\left(\eta_f, \hat{\varepsilon}\right) = P_s\left(\eta_s, \hat{\varepsilon}\right) \tag{30}$$

where $\hat{\varepsilon}$ is the average interaction strength, proportional to the inverse of the reduced temperature $k_BT/\varepsilon$ for the system, and $P_f$ and $P_s$ (see the Appendix) are the pressures and $\eta_f$ and $\eta_s$ are the packing fractions of the fluid and solid phases, respectively. For a system of two phases with one component, there is one degree of freedom. In our calculations, the reduced temperature is set, and Equations (29) and (30) are solved simultaneously for the packing fractions of the equilibrium liquid and crystal phases.

In our formulation presented here, we make a further simplifying assumption that facilitates calculation of the phase diagram. The pressure of the protein crystal cannot be calculated accurately because it is a complex function of the intermolecular forces between protein molecules, consistent with the approximate phase-equilibrium theory discussed in the Appendix. For this reason, the packing fraction of the crystal is set equal to a physically reasonable constant and the effect of pressure is neglected. Since the pressure is a very rapidly changing function of the solid packing, small changes in solid-phase packing fraction alter the pressure without significantly changing the chemical potential.[15] By neglecting the effect of pressure on the solid, we neglect the small displacement in density of the solid required to change the pressure of the solid to match that of the fluid. As a result of neglecting the effect of pressure, the protein solubility is determined only from the condition that the chemical potential of the protein is the same in all phases.

To understand the factors that determine protein solubility, it is helpful to consider the case when the fluid is in the ideal limit where the intermolecular interactions between protein molecules in the fluid are turned off. The chemical potential of the protein in an ideal liquid solution is obtained from differentiating Equation (8) with respect to molecule number, giving

$$\beta\mu_{ig} = \ln\left(\frac{\rho_f \Lambda^3}{8\pi^2}\right). \tag{31}$$

Substituting Equation (31) into the phase-equilibrium criterion of identical chemical potentials [Equation (29)] gives

$$\ln \eta_f = \beta\Delta G_{ent} - \frac{z}{2}\hat{\varepsilon} \tag{32}$$

where

$$\Delta G_{ent} = -\beta^{-1} \ln\left[\frac{3v_c \Delta\Omega}{4\pi^3\sigma^3}\right]. \tag{33}$$

The first term on the right side of Equation (32), $\Delta G_{ent}$ is the entropic contribution to the change in free energy upon transferring a protein molecule from solution to crystal. It is related to the reversible work required to fix the rotation and translation of the protein molecule in its unit cell relative to those in the fluid. Because this process is accompanied by a large decrease in configuration space, the reversible work is positive. The large loss in entropy is compensated by the negative free energy due to the protein-protein contact interactions in the crystal given by the second term on the right side of Equation (32). The magnitude of the entropic term relative to that of the intermolecular-interaction term determines protein solubility. As intermolecular attraction increases, protein solubility is reduced because favorable interactions between protein molecules in the fluid overcome the loss in entropy upon crystallization.

$\Delta G_{ent}$ is determined by the extent of translational and rotational freedom of the protein molecule in the crystal. For lysozyme, $\Delta G_{ent}$ can be estimated from lysozyme crystallographic data. For a reasonable approximation, we use the effective spherical diameter of lysozyme (34.8 Å) obtained from crystal-structure dimensions. In lysozyme crystals, X-ray scattering shows that the entire molecule moves about 0.2-0.25 Å.[21] Consequently, $v_c$ is approximated as $(0.25\ \text{Å})^3$. Calculation of the orientation space is not straightforward because coherent oscillation movements are not detected in X-ray scattering experiments. However, assuming that oscillations generate atomic displacements of the magnitude as above, $\delta\Omega$ can be approximated by

$$\Delta\Omega = \left(\frac{0.25}{\sigma/2}\right)^3 \tag{34}$$

With these reasonable approximations, $\Delta G_{ent}$ is estimated as 18.5 kcal/mol.

If we assume that the available bonding volume does not depend on solution conditions, there is a direct correspondence between $B_2$ and the interaction strength parameter, $\hat{\varepsilon}$, according to Equation (21). However, protein-protein interactions consist of complex charge-density-dependent electrostatic interactions, dispersion forces, and hydrophobic forces that have different characteristic decay lengths. Since these forces depend on the intervening aqueous medium (salt concentration, pH, temperature), and

because the correlation between $B_2$ and lysozyme solubility covers a wide range of solution conditions, we expect that the forces that determine $B_2$ also depend on solution conditions.[1] Slightly changing these conditions can result in the formation of different crystal structures that are stabilized by different contact-contact interactions. Because our goal here is to show that a simplified model of these forces can adequately explain the crystallization behavior of protein molecules, the effect of specific forces on crystallization behavior is not considered here. Such effects are not easily generalized for understanding the solution behavior of different proteins.

Although we have a reasonable quantitative estimate of $\Delta G_{ent}$, we choose empirically to fit $\Delta G_{ent}$ to the experimental data relating protein solubility to $B_2$. In this fit, we use Equations (2), (11), and (18) for the chemical potential of the protein in the fluid phase and Equation (28) for that in the solid phase. As $B_2$ decreases, protein solubility falls, not due to small changes in $\mu_f$, but because there are significant changes in $\mu_s$ due to increased attractive interaction of the protein-protein contacts. This result is shown in Figure 2 for different values of the number of attractive sites per protein molecule in the fluid, M, and width of square well, $r_c$. For a pair of proteins in the fluid, parameters M and $r_c$ determine the volume available for "bonding". According to Equation (21), as the available bonding volume decreases, keeping $B_2$ constant requires larger values of $\hat{\varepsilon}$. In determining $\Delta G_{ent}$ by fitting the data, the stronger interaction strengths are compensated by a simultaneous increase in the unfavorable entropic term. Because of this compensation, the fit chemical potential of the solid is essentially independent of parameters M or $r_c$.

The chemical potential of the fluid is only weakly dependent on association-potential parameters because the fluid-phase protein concentrations are low and therefore the fluid behaves almost as an ideal dilute solution. Consequently, the fluid chemical potential is nearly the same function of protein density for all sets of parameters, and the curves in Figure 2 collapse to one curve. The fit values of $\Delta G_{ent}$ corresponding to the parameter sets with 12 or 6 attractive sites and a square-well width of 2 Å provide the most realistic description of the protein-protein interaction in the fluid.

We also fit $\Delta G_{ent}$ to the empirical relation between $B_2$ and protein solubility for different values of z, the number of contacts formed per protein molecule in the crystal.

This fit is for a protein molecule that can form 6 bonds in the fluid with square-well width of 2 Å. According to Equation (32), z is the slope of a plot of the natural log of the protein solubility versus $\hat{\varepsilon}$. Consequently, in a plot of solubility versus $B_2$, the curvatures become steeper as z increases, as indicated in Figure 3 where the curves correspond to z equal to 4, 6, and 12. According to Figure 3, the empirical correlation indicates a lower limit of 4 and an upper limit of 12 contacts per protein molecule. This result is independent of the choice of M or $r_c$ because changing these parameters only serves to change the fit value of $\Delta G_{ent}$, while the shape of the curves remains the same.

Upon comparing the fit values of $\Delta G_{ent}$ with that calculated by Equation (33), it appears that a protein crystal, where the protein molecule forms 6-8 contacts, provides the best description of the experimental correlation. This result is in remarkable agreement with experimental studies showing that protein molecules form 8 - 10 contacts on average.[22]

For comparison, we have fit the empirical correlation of $B_2$ with protein solubility for the case where the protein-protein interaction is given by a uniform square-well potential in the fluid. To calculate the phase equilibrium condition of Equation (29), we model the proteins as hard spheres interacting through a uniform square-well potential. The chemical potential of the hard-sphere reference system is given by Equation (11) and the random-phase approximation (RPA)[23] is used to calculate the contribution to the chemical potential from the attractive interactions in the fluid. The chemical potential of the crystal is calculated as before, where the value of $\hat{\varepsilon}_s$ is equal to the square-well depth of the uniform square-well potential in the fluid. Results are presented in Figure 4 for two different values of z; the width of the square-well potential is 2 Å. The same quality of fit is obtained as in the case where the protein-protein interactions are anisotropic. However, the fit values of $\Delta G_{ent}$ are very small and do not correspond to the physical behavior of protein molecules in crystals. We therefore conclude that the lysozyme molecules interact via highly anisotropic interactions.

## 6. The Liquid-Liquid Critical Point

Determining the location of the liquid-liquid critical point is important for protein crystallization because the rate of nucleation is enhanced near this region, thereby

providing insight for fixing conditions to grow crystals.[4] At the critical point there are no degrees of freedom; the thermodynamic properties must satisfy the critical-point criteria

$$\left(\frac{\partial \mu_f}{\partial \eta_f}\right)_{\hat{\varepsilon}} = 0 \tag{35}$$

and

$$\left(\frac{\partial^2 \mu_f}{\partial \eta_f^2}\right)_{\hat{\varepsilon}} = 0 \tag{36}$$

where the partial derivatives of Equations (35) and (36) are evaluated holding reduced temperature, $\hat{\varepsilon}^{-1}$ constant. $\mu_f$ is calculated from Equations (2), (11), and (18). Equations (35) and (36) can be solved to determine the reduced temperature at the critical point, $\hat{\varepsilon}_c^{-1}$, and the packing fraction at the critical point, $\eta_c$. The results are given in Table 1 for different values of M, with $r_c$ set to 2 Å.

According to the results of Wertheim's theory; as the number of attractive sites per protein molecule increases, the critical-point temperature rises and the top of the liquid-liquid connodal line approaches the liquidus branch of the liquid-solid equilibrium. It is not possible to locate exactly the position of the liquid-liquid critical point because of approximations in Wertheim's theory, as discussed in Section 3. However, based on the calculated results, it appears that the number of attractive sites on the surface of the protein molecule is an important variable that determines the position of the liquid-liquid critical point with respect to the liquid-solid equilibrium. As the number of attractive sites per protein molecule increases, there are a larger number of different orientations available for "bond" formation between protein molecules and the formation of a disordered dense fluid phase becomes thermodynamically more stable. This fluid phase corresponds to the amorphous precipitate. If $B_2$ at the critical temperature is more positive than those values corresponding to the location of the crystallization window, amorphous precipitation may interfere with the crystallization process.

## 7. Conclusions

Following the work of Sear[15] and Haas et al.,[14] it has been shown here that a simple model can explain the phase behavior of aqueous protein solutions; the important

feature of the model follows from anisotropic interactions of protein molecules in the crystal phase. Protein solubility is determined by two competing effects: (a) the unfavorable loss in entropy from constraining a protein molecule upon crystallization and (b) the favorable energetic interactions due to the formation of protein-protein contacts in the crystal. By calculating the entropic term from experimental data, we have shown that the empirical correlation of $B_2$ with protein solubility corresponds to the formation of 6 - 8 protein-protein contacts in the crystal, in good agreement with independent experimental results. Further, our model shows that the location of the liquid-liquid critical point is related to the number of attractive sites per protein molecule. As the number of sites increases, the liquid-liquid connodal line approaches that of the liquid-solid equilibrium. Calculation of the liquid-liquid critical temperature is useful because other studies have shown that crystallization kinetics are most favorable at temperatures very slightly exceeding that critical temperature.

## 8. Acknowledgements

For financial support, the authors are grateful to the Office for Basic Energy Sciences of the U.S. Department of Energy and to the National Science Foundation (Grant #CTS-9530793). Robin Curtis thanks Jianzhong Wu and Dusan Bratko for helpful discussions.

## References


(1) George, A.; Wilson, W.W. *Acta Crsytallogr.* **1994**, D51, 361.

(2) Rosenbaum, D.; Zamora, P.C.; Zukoski, C.F. *Phys. Rev. Lett.* **1996**, 76, 150.

(3) Poon, W.C.K. *Phys. Rev. E* **1997**, 55, 3762.

(4) Wolde, P.R.t.; Frenkel, D. *Science* **1997**, 277, 1975.

(5) Rosenbaum, D.F.; Kulkami, A.; Ramakrishnan, S.; Zukoski, C.F. *J. Chem. Phys.* **1999**, 111, 9882.

(6) Agena, S.M.; Brogle, D.L.; Pessoas, F.L.P. *Biotech. Bioeng.* **1997**, 55, 65.

(7) Lekkerkerker, H.; Poon, W.C.K.; Pusey, P.N.; Stroobants, A.; Warren, P.B. *Europhys. Lett.* **1992**, 2, 559.

(8) Gast, A.P.; Russel, C.K.; Russel, W.B. *Faraday Disc. Chem. Soc.* **1983**, 76, 189.

(9) Hagen, M.H.J.; Frenkel, D. *J. Chem. Phys.* **1994**, 101, 4093.

(10) Asherie, N.; Lomakin, A.; Benedek, G.B. *Phys. Rev. Lett.* **1996**, 77, 4832.

(11) Guo, B.; Kao, S.; McDonald, H.; Asanov, A.; Combs, L.L.; Wilson, W.W. *J. Cryst. Growth* **1999**, 196, 424.

(12) Neal, B.L.; Asthagiri, D.; Velev, O.D.; Lenhoff, A.M.; Kaler, E.W. *J. Cryst. Growth* **1999**, 196, 377.

(13) Lomakin, A.; Asherie, N.; Benedek, G.B. *Proc. Natl. Acad. Sci. USA* **1999**, 96, 9465.

(14) Haas, C.; Drenth, J.; Wilson, W.W. *J. Phys. Chem. B* **1999**, 103, 2808.

(15) Sear, R.P. *J. Chem. Phys.* **1999**, 111, 4800.

(16) Wertheim, M.S. *J. Stat. Phys.* **1984**, 35, 19.

(17) Jackson, G.; Chapman, W.G.; Gubbins, K.E. *Mol. Phys.* **1988**, 65, 1.

(18) Carnahan, N.; Starling, K. *J. Chem. Phys.* **1970**, 51, 635.

(19) Hill, T.L. *An Introduction to Statistical Mechanics*; 2nd ed.; Dover Publications Inc.: New York, 1986.

(20) McQuarrie, D. *Statistical Mechanics*; Harper Collins Publishers: New York, 1976.

(21) Finkelstein, A.V.; Janin, J. *Protein Eng.* **1989**, 3, 1.

(22) Janin, J.; Rodier, F. *Proteins: Struct., Func., Gen.* **1995**, 23, 580.

(23) Grimsom, M.J., *J. Chem. Soc. Faraday Trans., II* **1983**, 79, 817.

## Appendix - Pressure of a Protein Crystal

The pressure referred to throughout this paper is an osmotic pressure. Figure 5 illustrates that, at equilibrium, the osmotic pressure of the fluid phase is the same as that of the solid phase.

In this appendix we show that it is difficult to determine this osmotic pressure of a protein crystal. The pressure of the solid can be obtained from the Helmholtz energy:

$$\beta P_s = -\left(\frac{\partial \beta a_s}{\partial 1/\rho_s}\right)_T . \tag{A.1}$$

Equation (27) gives the Helmholtz energy of the solid, where the only term that depends on density is the free volume. Consequently, to evaluate the derivative in Equation (A.1), we must know the relationship between the free volume and the density of the solid. Because we are only concerned with calculating an approximate pressure, we assume that the protein molecule is contained in a cubic lattice where the dimension, L, is given by the hard-sphere diameter of the protein plus the distance available to the center of mass along one direction (~0.25 Å). Since we are approximating the lattice as a cube, the volume of the lattice is not necessarily equal to the inverse of the protein density, $\rho_s$, but instead the volume of the lattice is proportional to $\rho_s^{-1}$

$$L^3 = \frac{\alpha}{\rho_s} \tag{A.2}$$

where the proportionality constant, $\alpha$, is on the order of 1 and is omitted here. The volume available to the protein molecule's center of mass is

$$v_c = (L-\sigma)^3 \quad . \tag{A.3}$$

The pressure of the solid is determined from substituting Equations (A.2) and (A.3) into Equation (A.1):

$$\beta P_s = \rho_s \left(\frac{1}{1-\sigma/L}\right). \tag{A.4}$$

This result is identical to that obtained from the free-volume approximation of a hard-sphere solid. Equation (A.4) gives an approximate pressure of 123 atmospheres. This

should be compared with the pressure of the equilibrated protein solution, $P_f$. This pressure is approximately given by the ideal solution value:

$$\beta P_f = \rho_f . \tag{A.5}$$

For a protein solubility of 60 g/L, the corresponding pressure is 0.1 atm. This should also be the pressure of the solid. The erroneous high pressure predicted from Equation (A.4) is due to neglect of a negative contribution from attractive intermolecular interactions. The interaction field is determined by the nearest-neighbor distances that depend on the density of the protein in the crystal. Here, we have assumed that the field is independent of the position of the molecule in the cell; therefore the attractive pressure is zero. To evaluate the attractive pressure, we require the dependence of the field on the position of the molecule in the cell. Unfortunately that dependence is not easily established because the field is highly anisotropic due to the bonds formed in the crystal; the field depends on the specific form of the sticky-site potential. Because the pressure is the difference between two large numbers, the hard-sphere pressure and the attractive pressure, a reliable result is not possible because both large numbers have an uncertainty that is larger than the difference between the two pressures.

| M | $\eta_c$ | $B_{2,c}$ x $10^4$ (mLmol/g$^2$) | $\hat{\varepsilon}_c$ |
|---|---|---|---|
| 4 | 0.090 | -19.0 | 13.2 |
| 6 | 0.155 | -5.0 | 11.7 |
| 8 | 0.207 | -2.1 | 10.9 |

**Table 1:** Critical-point properties for different number of attractive sites per protein molecule calculated from Wertheim's theory. Here, $\hat{\varepsilon}_c = \varepsilon/k_B T_c$ where $T_c$ is the liquid-liquid critical temperature.

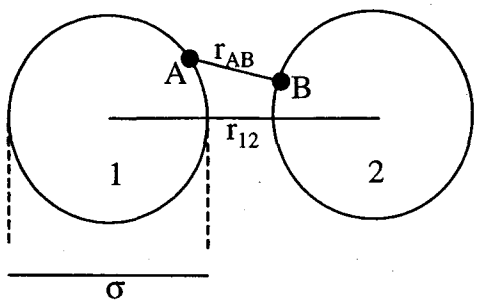


**Figure 1a :** Schematic representation of Wertheim's attractive-site interaction where $r_{AB}$ is the separation between sites A and B. The interaction is given by a square well potential of width $r_c$ and depth ε. The center-to-center separation is $r_{12}$ and the protein diameter is σ.

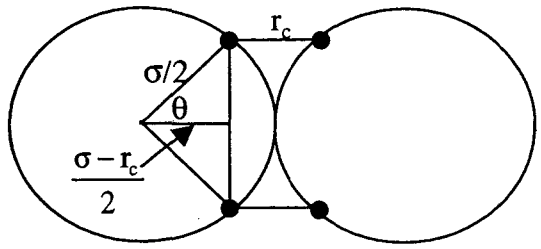


**Figure 1b :** Schematic representation of two-body configuration for protein molecules with sticky sites, which gives the minimum distance between sites on the same molecule such that two bonds are not formed simultaneously. This minimum distance is given by $\theta\sigma$, where $\theta = \cos^{-1}\left(1 - \frac{r_c}{\sigma}\right)$

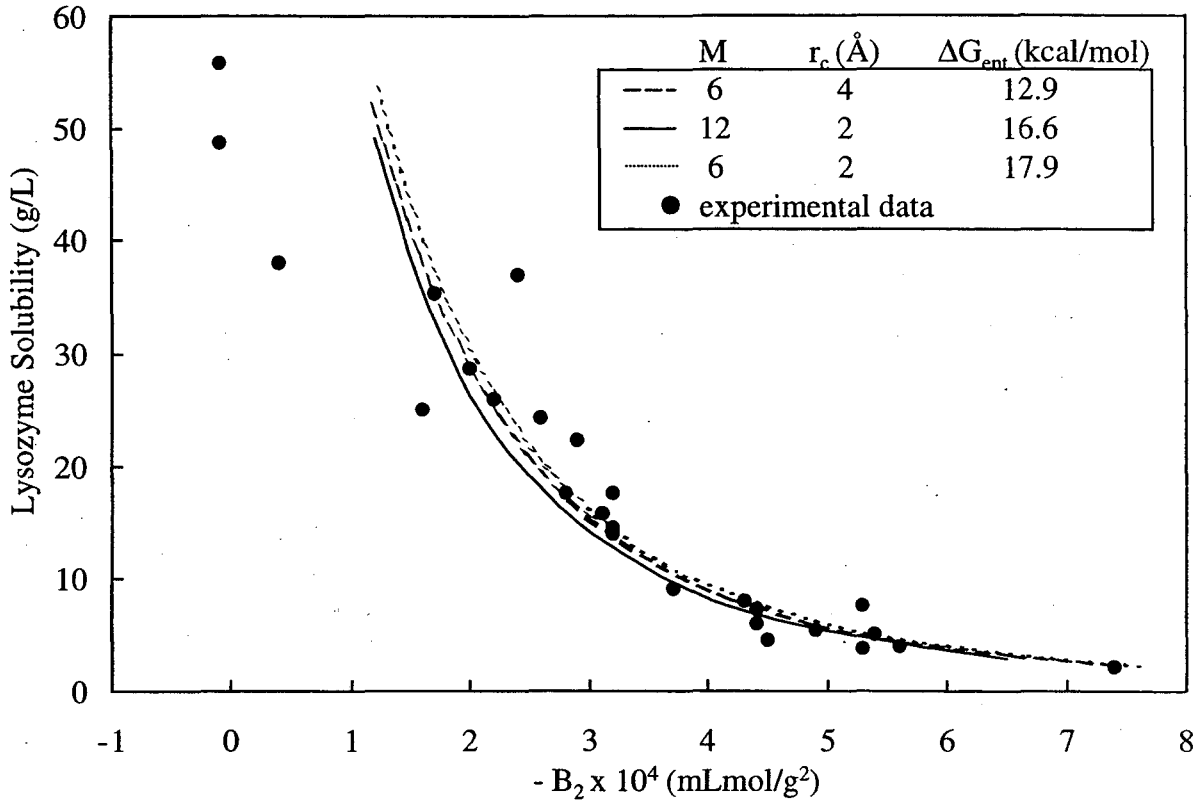


**Figure 2 :** Correlation of lysozyme solubility with osmotic second virial coefficient, $B_2$. The only fit parameter in the calculation is $\Delta G_{ent}$. The number of contacts per protein molecule in the crystal, z, is 6. The number of attractive sites, M, and the square-well potential width, $r_c$, have only a small effect on the fit curve. The value of the interaction parameter, $\hat{\varepsilon}$ ,depends on M and $r_c$. However, the fit value of $\Delta G_{ent}$ compensates for changes in $\hat{\varepsilon}$.

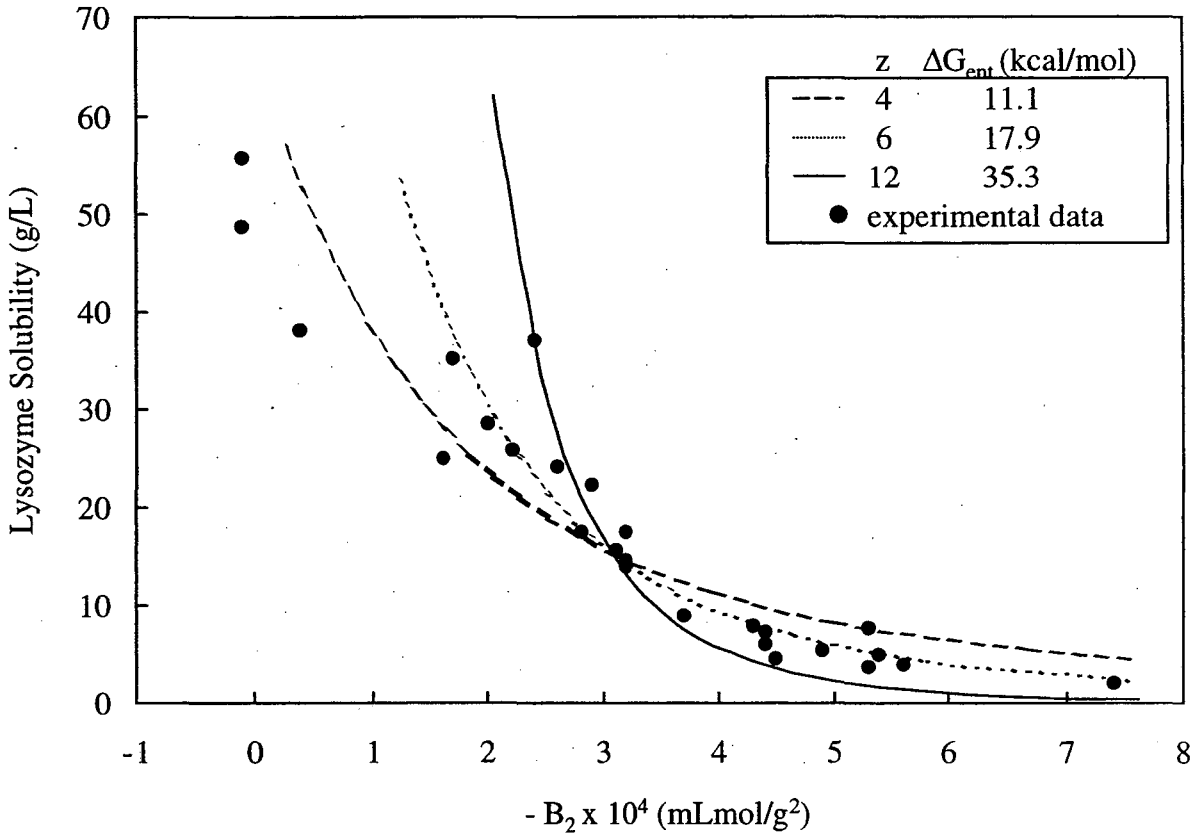


**Figure 3 :** Correlation of lysozyme solubility with osmotic second virial coefficient, $B_2$. The number of attractive sites per protein molecule, M, is 6 and the cut-off separation between sites, $r_c$, is 2 Å. The correlation is fit by varying $\Delta G_{ent}$ for different values of z, the number of contacts per protein molecule in the crystal. The best fit to the data is obtained when z = 6.

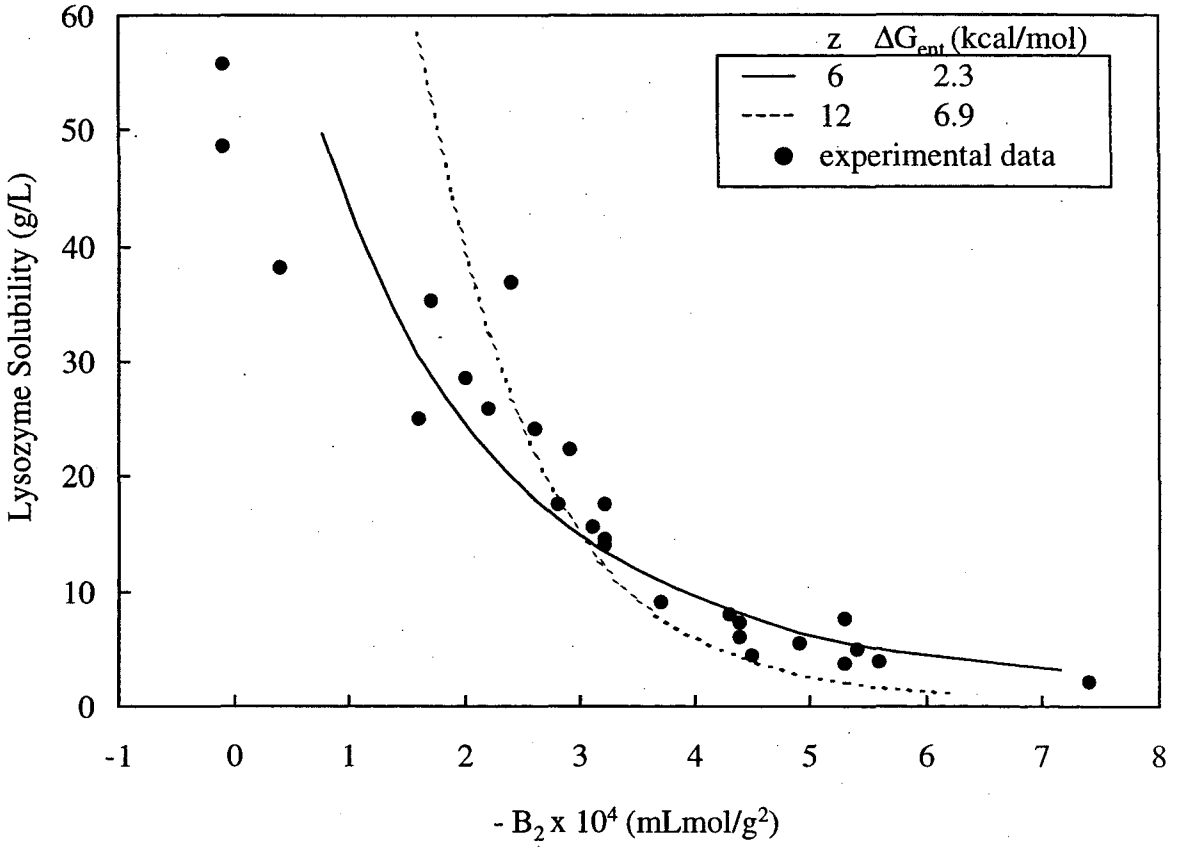


**Figure 4 :** Correlation of lysozyme solubility with osmotic second virial coefficient, $B_2$. The calculations are performed for a uniform square-well potential with a width of 2 Å using the RPA. The correlation is fit with two values of z by varying $\Delta G_{ent}$. Both curves provide an accurate fit. However, the values of $\Delta G_{ent}$ are physically unrealistic.

| Solid phase (s) | Outside phase (o) | Fluid phase (f) |
|---|---|---|
| protein and solvent | solvent alone | protein and solvent |
| $P_{s,eq} = P_o + P_s$ | $P_o$ | $P_{f,eq} = P_o + P_f$ |

**Figure 5 :** In McMillan-Mayer solution theory, the solid and fluid phases are in a hypothetical equilibrium with an outside solution that contains solvent at the same chemical potential. The pressures referred to in the text are osmotic pressures of the solid and the fluid, $P_s$ and $P_f$. The osmotic pressure is defined as the pressure necessary to increase the solvent chemical potential of the fluid phase( or the solid phase) to account for the lower number density in that phase. The equilibrium condition that the pressure of the solid is equal to that of the fluid ($P_{s,eq} = P_{f,eq}$) means that the osmotic pressure of the solid is equal to that of the fluid at equilibrium. This result can also be obtained from the equivalent equilibrium condition $\mu_s = \mu_o = \mu_f$ where $\mu$ is the chemical potential of the solvent.

Ernest Orlando Lawrence Berkeley National Laboratory
One Cyclotron Road | Berkeley, California 94720

Prepared for the U.S. Department of Energy under Contract No. DE-AC03-76SF00098