JCTC Journal of Chemical Theory and Computation
Article
pubs.acs.org/JCTC

Solvent Binding Analysis and Computational Alanine Scanning of the Bovine Chymosin−Bovine κ-Casein Complex Using Molecular Integral Equation Theory

David S. Palmer,*[†,‡] Jesper Sørensen,[§,∥] Birgit Schiøtt,[∥] and Maxim V. Fedorov*[†,‡]

[†]Department of Physics, University of Strathclyde, John Anderson Building, 107 Rottenrow, Glasgow, Scotland G4 0NG, United Kingdom

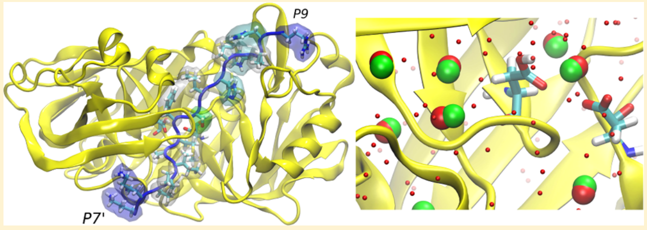
[‡]Max Planck Institute for Mathematics in the Sciences, Inselstrasse 22, DE-04103 Leipzig, Germany

[§]Department of Chemistry and Biochemistry, University of California, San Diego, Urey Hall, 9500 Gilman Drive, La Jolla, California 92093, United States

[∥]The Center for Insoluble Protein Structures (inSPIN) and the Interdisciplinary Nanoscience Center (iNANO), Department of Chemistry, Aarhus University, Langelandsgade 140, DK-8000 Aarhus C, Denmark

Supporting Information

ABSTRACT: We demonstrate that the relative binding thermodynamics of single-point mutants of a model protein−peptide complex (the bovine chymosin−bovine κ-casein complex) can be calculated accurately and efficiently using molecular integral equation theory. The results are shown to be in good overall agreement with those obtained using implicit continuum solvation models. Unlike the implicit continuum models, however, molecular integral equation theory provides useful information about the distribution of solvent density. We find that experimentally observed water-binding sites on the surface of bovine chymosin can be identified quickly and accurately from the density distribution functions computed by molecular integral equation theory. The bovine chymosin−bovine κ-casein complex is of industrial interest because bovine chymosin is widely used to cleave bovine κ-casein and to initiate milk clotting in the manufacturing of processed dairy products. The results are interpreted in light of the recent discovery that camel chymosin is a more efficient clotting agent than bovine chymosin for bovine milk.



INTRODUCTION

The key thermodynamic parameter characterizing the binding of a ligand (L) by a receptor (R) is the binding free energy (ΔG_{bind}) for the process R + L → RL.[1] The binding free energy is strongly influenced by the solvent environment, which modulates effects such as hydrophobicity and competitive solvent binding.[2] Of the large number of different methods that have been proposed to calculate the absolute or relative binding free energies of biomolecular complexes,[3,4] the vast majority have employed explicit or implicit continuum representations of the solvent environment (or both), while other methods for modeling solvent have received considerably less attention.[5−7] The Integral Equation Theory (IET) of Molecular Liquids is a promising theoretical framework for modeling solvent in biomolecular simulations. IET allows calculation of solvent density distributions and solvation thermodynamic parameters at significantly lower computational expense than explicit solvent simulations. The theory may be used to study specific solute−solvent interactions that are not accessible by continuum solvent models. Moreover, the theory is easily generalizable to many different pure and mixed solvent systems. Recent developments in IET based methods have made it possible to make accurate calculations of hydration free energies (HFE) across multiple classes of compounds at relatively low computational expense.[8,9] Furthermore, IET has found an ever-increasing number of successful applications, including computing solubility of druglike molecules,[10] fragment-based drug design,[11] modeling the binding of water[12] and ions[13,14] by proteins, predicting tautomer ratios,[15] interpreting solvent densities around biomacromolecules,[16] and sampling molecular conformations.[17] The attributes of IET make it a useful method to complement traditional implicit continuum or explicit solvent simulations.

The aim of this work is to compare several different IET based approaches for binding free energy calculation, computational alanine scanning, and solvent binding analysis. In the first part of this paper, we compare binding free energies calculated using the well-known Molecular Mechanics Poisson−Boltzmann Surface Area (MM-PBSA) method, and an IET based analogue, Molecular Mechanics Three-Dimensional Reference Interaction Site Model (MM-3DRISM), where the latter calculations were performed using three different hydration free energy functionals (the Gaussian Fluctuations, Kovalenko−

Received: July 12, 2013
Published: October 9, 2013

Hirata, and Universal Correction functionals) within the scope of the Three-Dimensional Reference Interaction Site Model (3DRISM), which is a method based on the integral equation theory of molecular liquids. This work is an extension and further validation of a method previously tested by Genheden et al.[18] In the second part of this paper, we apply these methods to computational alanine scanning of a protein−peptide complex that is important in the food industry. Finally, the 3DRISM distribution functions are used to predict water-binding sites on the surface of the protein, which are shown to be in good agreement with experimental data. The manuscript has two sets of messages: the first set relates to the feasibility of using molecular integral equation theory for modeling protein−ligand complexes; the second set relates to the implications of our results for a biomolecular complex that is important in the food industry.

The test systems used for these calculations are the complexes of bovine chymosin with single-point mutants of a 16-residue fragment of bovine κ-casein. Bovine chymosin is an aspartic protease found in calf stomachs, where it is responsible for selectively cleaving the milk protein κ-casein.[19] The enzyme was first purified industrially in 1874 and remains the most widely used enzyme to initiate milk clotting in cheese manufacturing. There has recently been renewed interest in understanding the biological action of the enzyme as it has been demonstrated that the camel variant of the enzyme has 70% higher clotting activity and only 20% of the unspecific protease activity for bovine milk, which has led to it being successfully marketed as an alternative to the bovine enzyme.[20] The two enzymes have high sequence identity (85%) and high sequence similarity (94%),[21] but the difference in catalytic efficacy is not well understood, due in part to a lack of experimental structural information about the chymosin−κ-casein complexes. Four X-ray crystal structures of apo- or inhibitor-bound bovine chymosin are currently available,[22−25] but there are no experimental structural coordinates for the bovine chymosin−bovine κ-casein complex. The crystal structure of a doubly glycosylated variant of camel chymosin in its apo form has recently been determined.[26]

κ-Casein is a 169 residue protein that is found in bovine milk, where it helps to solubilize α_{s1}-, α_{s2}-, and β-caseins by promoting the formation of aggregates referred to as casein micelles. The catalytic action of bovine chymosin is to cleave κ-casein at the *Phe105*−*Met106* bond, which causes the hydrophilic C-terminal end of κ-casein to dissociate, thereby destabilizing the casein micelle and causing precipitation of the insoluble casein proteins. The amino acid sequences of κ-caseins from different species in the region of the cleavage site are given in Figure 1 (κ-casein residues are given in italics in the text throughout). A Px or Px′ nomenclature is used to denote κ-casein residues on either side of the cleavage site; e.g., *Ser104*, *Phe105*, *Met106*, and *Ala107* are referred to as P2, P1, P1′, and P2′, respectively. Similarly, the regions of chymosin that interact with the P2, P1, P1′, and P2′ residues are denoted S2, S1, S1′, and S2′ pockets, respectively.[27] On the basis of the crystal structure of a chymosin−inhibitor complex (1CZI) and previous molecular modeling studies, κ-casein is thought to bind in an extended secondary structure.[25,28] This is consistent with circular dichroism, solution NMR, and molecular modeling studies of unbound κ-casein, which show an extended structure in the region of the scissile bond.[29,30] On the basis of geometric considerations and mutagenesis studies, it has been proposed that the P8−P7′ residues are located in the chymosin binding cleft during catalysis. Additionally, an arginine residue in the P9 position has been implicated in binding because it is conserved in bovine, camel, pig, buffalo, and goat chymosin.[31] Accordingly, a variant of bovine κ-casein, in which the P9 position is occupied by a histidine, has been shown to be a poor substrate.[32]

Bovine:	R-**H**-P-**H**-P-**H**-**L**-S-F-**M**-A-I-P-P-K-K	(97-112)
Camel:	R-**R**-P-**R**-P-**R**-**P**-S-F-**I**-A-I-P-P-K-K	(89-104)
Goat:	R-**H**-P-**H**-P-**H**-**L**-S-F-**M**-A-I-P-P-K-K	(97-112)
Pig:	R-**R**-P-**R**-P-**H**-**A**-S-F-**I**-A-I-P-P-K-K	(96-111)
Horse:	R-**H**-P-**C**-P-**H**-**P**-S-F-**I**-A-I-P-P-K-K	(97-112)
Sheep:	R-**H**-P-**H**-P-**H**-**L**-S-F-**M**-A-I-P-P-K-K	(97-112)

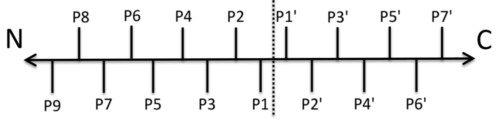


Figure 1. The primary sequence of the chymosin sensitive region of κ-casein in different species aligned. The P*n* and P*n*′ numbering follows the Schechter and Berger nomenclature, where *n* increases with the distance from the scissile bond. The residues that differ between some of the species are highlighted in bold. The residue numbers are shown to the right in parentheses. The complete bovine κ-casein protein contains 169 residues.

Chymosin is a mammalian aspartic protease comprising 323 amino acids, which folds into a bilobal globular structure, with two similar β-barrel domains. Unlike viral aspartic proteases such as HIV protease, chymosin is monomeric, with a 9% sequence identity between N- and C-terminal lobes.[24] The secondary structure is 13% helical (nine helices, 44 residues) and 48% β sheet (29 strands, 158 residues),[33] and these features are organized to give a pseudosymmetry along a single cleft between the N- and C-terminal lobes. In the center of the binding cleft are two catalytic Asp34 and Asp216 residues, which occur in conserved Asp-Thr-Gly motifs, with the Asp side chains orientated toward each other, in an approximately planar geometry. The Asp residues are stabilized by a network of hydrogen bonds, which includes two threonine interactions coined the "fireman's-grip."[24,34,35] A water molecule is observed between the catalytic Asp residues in all crystal structures of apo chymosin. The protein contains three disulfide bridges (Cys47-Cys52, Cys207-Cys211, and Cys250-Cys283) and a cis-proline residue (Pro25) that is conserved in mucorpepsin, endothiapepsin, and porcine pepsin.[22,24]

Specific interactions between the surface of chymosin and water molecules that exhibit high residence times are known to be important for modulating both the structure and dynamics of chymosin and its catalytic behavior. The crystal structures of chymosin and related aspartic proteases contain a large number of structural water molecules. From an analysis of 10 of these enzymes, 17 water molecules have been identified as conserved in the aspartic protease family, with 16 of these present in the crystal structures of bovine chymosin.[36] The conserved waters include the catalytic water molecule, five molecules in the S1′−S7′ half of the binding cleft, and one in the S8−S1 end. A water molecule close to Ser35 is thought to take part in a hydrogen bond chain that helps to stabilize Asp34.[37] The same water molecule may also help to stabilize the open form of chymosin through a hydrogen bond to Tyr77.[37]

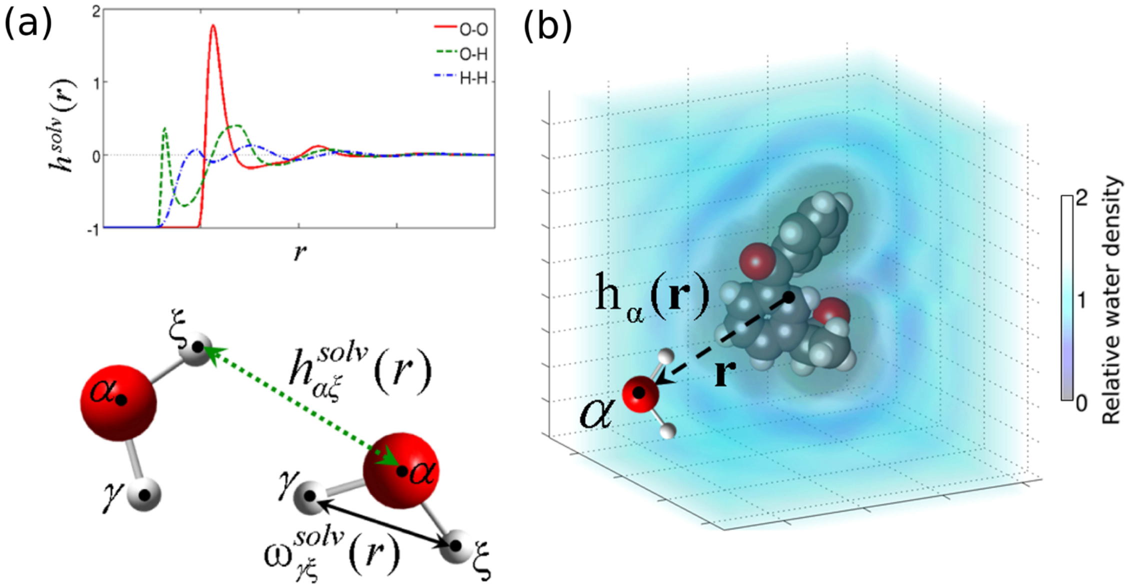


Figure 2. Correlation functions in the 3DRISM approach. (a) Site–site intramolecular ($\omega^{\text{solv}}_{\gamma\xi}(r)$) and intermolecular ($h^{\text{solv}}_{\alpha\xi}(r)$) correlation functions between sites of solvent molecules. The graph shows the radial projections of water solvent site–site density correlation functions: oxygen–oxygen (O–O, red solid), oxygen–hydrogen (O–H, green dashed), and hydrogen–hydrogen (H–H, blue dash-dotted). (b) Three-dimensional intermolecular solute–solvent correlation function $h_\alpha(\mathbf{r})$ around a model solute (diclofenac). This figure is based on Figure 1 from our earlier work.[42]

THEORY

Calculation of ΔG_{bind}. The key thermodynamic parameter characterizing the binding of a ligand (L) by a receptor (R) is the binding free energy (ΔG_{bind}) for the process:

$$\mathrm{R} + \mathrm{L} \rightarrow \mathrm{RL} \tag{1}$$

Several methods are available for calculating the binding free energies of protein–ligand complexes at various degrees of accuracy. Free energy perturbation and thermodynamic integration methods are thermodynamically rigorous and in principle very accurate but require a great deal of simulation time to provide adequate sampling, which makes them less suitable for large-scale studies. Many different end-point techniques have been developed to estimate binding-free energies at lower computational expense, e.g., the linear-interaction-energy (LIE) approach. The molecular mechanics Poisson–Boltzmann surface area (MM-PBSA) and molecular mechanics Generalized-Born surface area (MM-GBSA) methods have been used extensively for binding free energy calculations because they are relatively fast and have been shown to afford reasonably accurate estimates of binding free energy for protein–ligand systems.

In the MM-PBSA method, the binding free energy of an enzyme–substrate complex may be given as

$$\Delta G_{\text{bind}} = G_{\text{solvated}}(\text{complex}) - [G_{\text{solvated}}(\text{enzyme}) + G_{\text{solvated}}(\text{substrate})] \tag{2}$$

The free energy of each species is evaluated as

$$G_{\text{solvated}} = \langle E_{\text{gas}}\rangle + \langle \Delta G_{\text{solvation}}\rangle - TS \tag{3}$$

$$\langle E_{\text{gas}}\rangle = \langle E_{\text{internal}}\rangle + \langle E_{\text{electrostatic}}\rangle + \langle E_{\text{vdW}}\rangle \tag{4}$$

$$\langle E_{\text{internal}}\rangle = \langle E_{\text{bond}}\rangle + \langle E_{\text{angle}}\rangle + \langle E_{\text{torsion}}\rangle \tag{5}$$

where E_{bond}, E_{angle}, and E_{torsion} are contributions to the internal energy E_{internal} caused by the strain from the deviation of the bonds, angles, and torsion angles from their equilibrium values; $E_{\text{electrostatic}}$ is the electrostatic interaction energy; E_{vdW} is the van der Waals interaction energy; E_{gas} is the absolute gas-phase energy; $\Delta G_{\text{solvation}}$ is the solvation free energy; T is the temperature; and S is the entropy. $\Delta G_{\text{solvation}}$ is commonly denoted as ΔG_{hyd} when the solvent is water. The binding free energy of a single complex is calculated as the ensemble average of the binding free energies of a set of different conformers of the protein–ligand complex (as indicated by the broken brackets, ⟨...⟩).

Computational Alanine Scanning. In many practical applications, it is useful to have an estimate of the way that changes in the chemical structure of the host or guest molecule affect the binding free energy. For protein complexes, one of the most widely used methods is computational alanine screening, in which the difference in the binding free energy is calculated between the wild-type complex and a mutant complex, where one or more residues in the mutant complex have been changed to alanine. Although in principle computational alanine scanning could be performed with many different methods for calculating binding free energies, in practice the MM-PBSA method is commonly used because it is not prohibitively computationally expensive.

$$\Delta\Delta G_{\text{bind}} = \Delta G_{\text{bind,mutant}} - \Delta G_{\text{bind,wildtype}} \tag{6}$$

Here, we use alanine scanning to probe the effects of point mutations on binding thermodynamics in the bovine chymosin–bovine κ casein complex.

Calculation of ΔG_{hyd} Using Implicit Continuum Solvent Models. The solvation free energy term, ΔG_{hyd}, is commonly calculated using implicit continuum solvent models based on, for example, the Poisson–Boltzmann (MM-PBSA) equation, in which the total solvation free energy is computed from a polar part (E_{elec}) and a nonpolar part (G_{nonpolar}):

$$\langle \Delta G_{\text{hyd}}\rangle = \langle E_{\text{PB}}\rangle + \langle G_{\text{nonpolar}}\rangle \tag{7}$$

For each calculation of solvation free energy (i.e., for enzyme, substrate, and complex), the nonpolar solvation term describes the process of transferring a nonpolar molecule in the shape of the molecule of interest from a vacuum to water, including the creation of a cavity in water and the van der Waals interactions between the nonpolar molecule and the water molecules. The polar solvation term describes the contribution to the free energy due to polarization of the solvent environment by the solute.

Calculation of ΔG_{hyd} Using 3DRISM. 3DRISM[11,38−40] is a theoretical method for modeling solution phase systems based on classical statistical mechanics. The 3DRISM equations relate 3D intermolecular *solvent site–solute* total correlation functions ($h_\alpha(\mathbf{r})$) and direct correlation functions ($C_\alpha(\mathbf{r})$; index α corresponds to the solvent sites):[38,40]

$$h_\alpha(\mathbf{r}) = \sum_{\xi=1}^{N_{\mathrm{solvent}}} \int_{R^3} c_\xi(\mathbf{r} - \mathbf{r}')\, \chi_{\xi\alpha}(|\mathbf{r}'|)\, \mathrm{d}\mathbf{r}' \tag{8}$$

where $\chi_{\xi\alpha}(r)$ is the bulk solvent susceptibility function and N_{solvent} is the number of sites in a solvent molecule (see Figure 2). The solvent susceptibility function $\chi_{\xi\alpha}(r)$ describes the mutual correlations of sites ξ and α in solvent molecules in the bulk solvent. It can be obtained from the solvent intramolecular correlation function ($\omega_{\xi\alpha}^{\mathrm{solv}}(r)$), site–site radial total correlation functions ($h_{\xi\alpha}^{\mathrm{solv}}(r)$), and the solvent site number density (ρ_α): $\chi_{\xi\alpha}(r) = \omega_{\xi\alpha}^{\mathrm{solv}}(r) + \rho_\alpha h_{\xi\alpha}^{\mathrm{solv}}(r)$ (from here onward, we imply that each site is unique in the molecule, so that $\rho_\alpha = \rho$ for all α).[40] In this work, these functions were obtained by solution of the RISM equations of the solvent.[40,41]

In order to calculate $h_\alpha(\mathbf{r})$ and $C_\alpha(\mathbf{r})$, N_{solvent} *closure* relations are introduced:

$$h_\alpha(\mathbf{r}) = \exp(-\beta u_\alpha(\mathbf{r}) + h_\alpha(\mathbf{r}) - c_\alpha(\mathbf{r}) + B_\alpha(\mathbf{r})) - 1 \qquad \alpha = 1, ..., N_{\mathrm{solvent}} \tag{9}$$

where $u_\alpha(\mathbf{r})$ is the 3D interaction potential between the solute molecule and α solvent site, $B_\alpha(\mathbf{r})$ represents bridge functionals, $\beta = 1/k_{\mathrm{B}}T$, k_{B} is the Boltzmann constant, and T is the temperature.

In general, the exact bridge functionals $B_\alpha(\mathbf{r})$ in eq 9 are represented as an infinite series of integrals over high order correlation functions and are therefore practically incomputable, which makes it necessary to incorporate some approximations[40,43,44] or to estimate the form of these functionals from molecular simulation.[45] In the current work, we use a closure relationship proposed by Kovalenko and Hirata (the KH closure),[46] which was designed to improve convergence rates and to prevent possible divergence of the numerical solution of the RISM equations:[46]

$$h_\alpha(\mathbf{r}) = \begin{cases} \exp(\Xi_\alpha(\mathbf{r})) - 1 & \text{when } \Xi_\alpha(\mathbf{r}) \le 0 \\ \Xi_\alpha(\mathbf{r}) & \text{when } \Xi_\alpha(\mathbf{r}) > 0 \end{cases} \tag{10}$$

where $\Xi_\alpha(\mathbf{r}) = -\beta u_\alpha(\mathbf{r}) + h_\alpha(\mathbf{r}) - c_\alpha(\mathbf{r})$. We note that in the literature the combination of the KH closure relations and the 3DRISM equations are sometimes referred to as 3DRISM-KH theory, but for succinctness we will use 3DRISM instead, since we use only the KH closure here.

The 3D interaction potential between the solute molecule and α site of solvent ($u_\alpha(\mathbf{r})$, eq 9) is estimated as a superposition of the site–site interaction potentials between solute sites and the particular solvent site, which depend only on the absolute distance between the two sites. We use the common form of the site–site interaction potential represented by the long-range electrostatic interaction term and the short-range term (Lennard-Jones potential).[8]

Within the framework of the RISM theory there exist several approximate functionals that allow one to analytically obtain values of the hydration free energy (HFE) from the total $h_\alpha(\mathbf{r})$ and direct $c_\alpha(\mathbf{r})$ correlation functions.[18,47,48]

Kovalenko–Hirata Free Energy Functional (3DRISM/KH). The Kovalenko–Hirata free energy functional for 3DRISM is given by

$$\Delta G_{\mathrm{hyd}}^{\mathrm{KH}} = k_{\mathrm{B}}T \sum_{\alpha=1}^{N_{\mathrm{solvent}}} \rho_\alpha \int_{R^3} \left[\frac{1}{2} h_\alpha^2(\mathbf{r}) \Theta(-h_\alpha(\mathbf{r})) - \frac{1}{2} h_\alpha(\mathbf{r}) c_\alpha(\mathbf{r}) - c_\alpha(\mathbf{r})\right] \mathrm{d}\mathbf{r} \tag{11}$$

where ρ_α is the number density of solvent sites α, and Θ is the Heaviside step function:

$$\Theta(x) = \begin{Bmatrix} 1 \text{ for } x > 0 \\ 0 \text{ for } x < 0 \end{Bmatrix} \tag{12}$$

Gaussian Fluctuations Free Energy Functional (3DRISM/GF). The Gaussian fluctuations (GF) HFE functional was initially developed by Chandler, Singh, and Richardson, for 1D RISM, and adopted by Kovalenko and Hirata for the 3DRISM case:[40,49]

$$\Delta G_{\mathrm{hyd}}^{\mathrm{GF}} = k_{\mathrm{B}}T \sum_{\alpha=1}^{N_{\mathrm{solvent}}} \rho_\alpha \int_{R^3} \left[-\frac{1}{2} c_\alpha(\mathbf{r}) h_\alpha(\mathbf{r}) - c_\alpha(\mathbf{r})\right] \mathrm{d}\mathbf{r} \tag{13}$$

Universal Correction Free Energy Functional (3DRISM/UC). HFEs calculated using the KH and GF free energy functional have only a *qualitative* agreement with experimental results. The error in hydration free energies calculated by the GF functional in 3DRISM is strongly correlated with the partial molar volume calculated by 3DRISM.[5,9,42] The 3DRISM/UC free energy functional developed from this observation is a linear combination of the $\Delta G_{\mathrm{hyd}}^{\mathrm{GF}}$, the dimensionless partial molar contribution, ρV, and a bias correction, b (intercept):[42]

$$\Delta G_{\mathrm{hyd}}^{\mathrm{3DRISM/UC}} = \Delta G_{\mathrm{hyd}}^{\mathrm{GF}} + a(\rho V) + b \tag{14}$$

where the values of the scaling coefficient a and intercept b were obtained by linear regression against experimental data for a diverse data set of organic molecules. For the combination of methods used here, the coefficients have the values a = −3.2217 kcal/mol and b = 0.5783 kcal/mol. The 3DRISM/UC free energy functional has been shown to give accurate hydration free energies for a wide range of chemically diverse organic molecules.[5,9,42]

We estimate the solute partial molar volume via *solute–solvent site* correlation functions using the standard 3DRISM theory expression:[50]

$$V = k_{\mathrm{B}}T\eta(1 - \rho_\alpha \sum_{\alpha=1}^{N_{\mathrm{solvent}}} \int_{R^3} c_\alpha(\mathbf{r})\, \mathrm{d}\mathbf{r}) \tag{15}$$

where η is the pure solvent isothermal compressibility and ρ_α is the number density of solute sites α.

Analysis of 3DRISM Calculated Solvent Density. The distribution functions ($g(\mathbf{r}) = h(\mathbf{r}) + 1$) calculated by 3DRISM characterize the average density distribution of solvent molecules around solute at thermodynamic equilibrium. It has previously been demonstrated that in favorable circumstances the peaks in these density distribution functions correspond to regions where solvent molecules are experimentally observed to bind to the solute molecule.[51] Several computational algorithms have previously been proposed to identify solvent binding sites from 3DRISM calculated correlation functions. In the current work, we use the *Placevent* algorithm proposed by Sindhikara et al.[51] Since details of this method have already been published, only a brief description of the method will be presented here.

Given a 3DRISM distribution function ($g(\mathbf{r}) = h(\mathbf{r}) + 1$), a discrete distribution of explicit solvent atoms $D = \{\mathbf{r}_1,\mathbf{r}_2,...\mathbf{r}_n\}$ is found, where n is the total number of water oxygen atoms to be identified, and each element $\mathbf{r}_i = \{x_i,y_i,z_i\}$ contains the Cartesian coordinates of a single water oxygen atom. Solvent oxygen atom sites are placed iteratively starting from the highest density region of the distribution function. In practice, this is done by converting the 3DRISM distribution function to a probability function by

$$P(\mathbf{r}) = \rho V_{\mathrm{grid}} g(\mathbf{r}) \tag{16}$$

One explicit atom is placed per iteration of the algorithm. The total population at iteration i is conserved with the conservation law:

$$\int_v P_0(\mathbf{r}) = N_i + \int_v P_i(\mathbf{r}) \tag{17}$$

This is done by finding a sphere with radius δ_i centered at $\mathbf{r}_i$ that contains a population of one unit. The population in this volume is then set to zero for all future iterations of the algorithm.

$$\int_{\mathbf{r}_i}^{\mathbf{r}_i+\delta_i} P_i(\mathbf{r}')\,\mathrm{d}\mathbf{r}' = 1, \int_{\mathbf{r}_i}^{\mathbf{r}_i+\delta_i} P_{i+1}(\mathbf{r}')\,\mathrm{d}\mathbf{r}' = 0 \tag{18}$$

The process of placing one explicit atom and then reducing the local population by one in its vicinity can be iterated until there are a satisfactory number of explicit atoms, or until some cutoff criteria has been met (for example, until the local water population approaches that of bulk solution).

The output of these calculations is a vector of water oxygen atom coordinates $D = \{\mathbf{r}_1,\mathbf{r}_2,...\mathbf{r}_n\}$, each element of which is associated with a local probability density $P = \{P_{\mathbf{r}_1},P_{\mathbf{r}_2},...P_{\mathbf{r}_n}\}$. Water oxygen atom coordinates (D_i) that are associated with high local probability densities (P_i) are predicted to be more probable than those water oxygen atom coordinates that are associated with low local probability densities. Since the elements of P are given in decreasing order, the indices, i, of D_i and P_i give a rank order of calculated binding sites, from 1 (most probable) to n (least probable).

■ METHODS

Molecular Dynamics Simulations. We have applied the MM-3DRISM binding free energy calculation method to computational alanine scanning of the bovine chymosin−κ-casein complex. Here, we have used 81 snapshots taken every 1 ns from 20 to 100 ns of a 100 ns molecular dynamics simulation. In previous work, we used MM-PBSA with 480 snapshots to investigate the binding thermodynamics of this complex,[52] but it was found to be computationally intractable to use the same number of snapshots with the more computationally expensive MM-3DRISM method tested here. Therefore, to provide a frame-by-frame comparison, the MM-PBSA calculations were repeated here using the same 81 snapshots as the MM-3DRISM calculations, as described below. We note that 81 snapshots is more than twice as many as have been used in previous MM-3DRISM[18] and some MM-PBSA studies[53] of similar complexes. Only a brief summary will be provided here of the molecular dynamics simulation from which the snapshots were taken, since the details have previously been reported.[21,54] In short, the simulation was run using NAMD[55] with the AMBER FF03 force field parameters developed by Duan et al. and the TIP3P water model. Production simulations were performed in the isothermal−isobaric (NPT) ensemble, where the Nose−Hoover Langevin piston pressure control was used to regulate the pressure at 1.01325 bar, and the temperature of the system was maintained at 300 K by means of Langevin dynamics. Periodic boundary conditions were applied, and all electrostatic interactions were calculated using the Particle Mesh Ewald (PME) method. For van der Waals interactions, a cutoff of 10 Å was set, using a switching distance of 9 Å. All of the hydrogen to heteroatom bond distances were held fixed using the SHAKE algorithm. The equations of motion were integrated every 2 fs using the Velocity Verlet algorithm, and snapshots were stored every 2 ps.

We note that since the potential energy surfaces of an arbitrary solute modeled in explicit solvent, PBSA, or 3DRISM solvent are almost certainly not identical, the ensemble of conformers generated by explicit solvent molecular dynamics may not represent the same Boltzmann sampling that would have been obtained had either of the other two solvent models been used.

Calculation of ΔG_{bind} Using MM-3DRISM. The binding free energy of the 16 residue fragment of κ-casein in bovine chymosin was calculated using the MM-3DRISM method,[18,56] which is a variant of the MM-PBSA method.[52,57,58] The calculations were performed using end-point analysis on single trajectories of each complex, which is computationally less demanding and has been shown to provide results that are closer to experimental values due to cancellation of errors.[59] The Amber03 force field was used to compute bonded and nonbonded interactions. All nonbonded interactions were calculated with no cutoff distance. The entropic contributions to the binding free energy were calculated from translational, rotational, and vibrational contributions, where the latter were computed by normal-mode analysis.[60] The MM-PBSA.py program in Amber11 was used to expediate most of the necessary computations,[61] with the exception of the 3DRISM calculations, which were performed separately as described below.

Calculation of ΔG_{hyd} Using 3DRISM. The ΔG_{hyd} term from MM-3DRISM was computed using three different solvation free energy functionals within the scope of 3DRISM (KH, GF, and UC). All 3DRISM calculations were performed using the nonbonded terms of the Amber03 force field.

RISM Calculations. RISM calculations were performed assuming infinitely dilute solution. We used the Lue and Blankschtein version of the SPC/E model of water (MSPC/E).[62] This differs from the original SPC/E water model[63] by the addition of modified Lennard-Jones (LJ) potential parameters for the water hydrogen, which were altered to prevent possible divergence of the algorithm.[64−67] The Lorentz−Berthelot mixing rules were used to generate the solute−water LJ potential parameters.[68] The following LJ parameters (for water hydrogen) were used to calculate the interactions between solute sites and water hydrogens: $\sigma_{H_w}^{LJ}$ = 1.1657 Å and $\varepsilon_{H_w}^{LJ}$ = 0.0155 kcal/mol.

3DRISM Calculations. The 3DRISM calculations were performed using the NAB simulation package[41] in the AmberTools 1.5 set of routines.[69] The 3D-grid around a solute was generated such that the minimum distance between any solute atom and the edge of the solvent box (*buffer* in NAB notation) was equal to 30 Å. The linear grid spacing in each of the three directions was 0.3 Å. We employed the MDIIS iterative scheme,[70] where we used five MDIIS vectors, an MDIIS step size of 0.7, and a residual tolerance of 10^{-10}. The KH closure was used for solution of the 3DRISM equations. Solvent susceptibility functions were taken from the 1D RISM calculations.

Solvent Susceptibility Functions. Solvent susceptibility functions were calculated with the 1D RISM method present in AmberTools 1.5. The dielectrically consistent RISM was employed,[71] using the KH closure. The grid size for 1D functions was 0.025 Å, which gave a total of 16 384 grid points. We employed the MDIIS iterative scheme, where we used 20 MDIIS vectors, an MDIIS step size of 0.3, and a residual tolerance of 10^{-12}. The solvent was considered to be pure water with a number density 0.0333 Å^{-3} and a dielectric constant of 78.497. The final susceptibility solvent site−site functions were stored and then used as input for the 3DRISM calculations. The solvent isothermal compressibility evaluated from the 1D RISM calculation was $k_BT\eta$ = 1.949459 Å^3.

Calculation of ΔG_{hyd} Using PBSA. To provide a comparison to the MM-3DRISM results, the binding free energy and computational alanine scanning results were also evaluated using the MM-PBSA method, which is equivalent to the MM-3DRISM method, except that the change in solvation free energy on binding is calculated using the Poisson−Boltzmann surface area (PBSA) method, rather than 3DRISM. All MM-PBSA calculations were performed using the MM-PBSA.py script in Amber11. The Poisson−Boltzmann equation was solved with the recommended default settings, which have been used in previous work.[52] The dielectric constants of the solute and solvent were set to 1 and 80, respectively. The size of the probe for the PB energy grid was 1.4 Å, and the default grid spacing was 0.5 Å.

Computational Alanine Scanning. Computational alanine scanning calculations were carried out for 15 of the 16 residues of the bovine κ-casein fragment in bovine chymosin (the remaining residue, *AlaP2′*, was naturally not mutated). These calculations were performed using the protocol of Massova and Kollman,[72] which uses the same trajectories as for the MM-PBSA calculations. The mutation is introduced after the simulation has been performed, and it is thereby assumed that the mutation will not have a major effect on the dynamics of the enzyme−substrate complex. This is supported by molecular dynamics simulations of apo chymosin and the chymosin-κ-casein complex, which show that no large structural changes occur on binding. A recent study by Bradshaw et al. has shown that this is a valid assumption for many protein−ligand systems.[73] The method is computationally much less demanding, given the amount of residue we are mutating, and more importantly, use of the same trajectories allows for cancellation of errors, which has proven to provide more accurate results.[72,74] For each setup, the side-chain atoms of the selected amino acid residue were removed, so that only the backbone and the C_β atom remained, to which hydrogens were subsequently added. Owing to the prohibitive computational cost and in accordance with previous studies,[72,74,75] the entropy term was neglected in all alanine scanning calculations.

Since we subtract the wild-type binding free energy from that of the mutant, negative values of $\Delta\Delta G_{bind}$ indicate favorable mutations, and positive values indicate unfavorable mutations. On the basis of the magnitude of $\Delta\Delta G_{bind}$, residues can be classified as so-called warm- (>1 kcal/mol) or hot-spots (>2 kcal/mol) that contribute disproportionately to the binding free energy. In previous work,[52] we performed computational alanine scanning using MM-PBSA on eight residues of bovine κ-casein and some residues of chymosin that were considered to be important for understanding the binding thermodynamics. Here, where our main interest is in comparing different computational methods, we present computational alanine scanning results for all of the residues of the P9−P7′ fragment of bovine κ-casein (excluding AlaP2′ as mentioned previously), including seven residues that were not considered in the earlier work.

Statistical Analysis. To compare free energies calculated by different computational models, a correlation coefficient and the root mean squared deviation (RMSD) were evaluated:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(x^i - y^i)^2}{\sum_{i=1}^{n}(x^i - M(y^i))^2} \tag{19}$$

$$\mathrm{RMSD}(x, y) = \sqrt{\frac{1}{N}\sum_i (x^i - y^i)^2} \tag{20}$$

where index *i* runs through the set of *N* selected molecules, and x^i and y^i are values calculated by different computational methods, for molecule *i* for a given property. The total deviation can be split into two parts: bias (or mean displacement, *M*) and standard deviation (SD), which are calculated by the formulas:

$$\mathrm{bias} = M(x - y) = \frac{1}{N}\sum_{i\in S}(x^i - y^i) \tag{21}$$

$$\sigma(x - y) = \sqrt{\frac{1}{N}\sum_{i\in S}(x^{(i)} - y^{(i)} - M(x - y))^2} \tag{22}$$

The bias gives the systematic error, which can be corrected by a simple constant term. The standard deviation gives the random error that is not explained by the model. One can see the connection between these three formulas:

$$\mathrm{RMSD}(x, y)^2 = M(x, y)^2 + \sigma(x, y)^2 \tag{23}$$

Statistical analyses were carried out in the R Statistical Computing Environment.[76] Python scripts were used to manipulate raw data files.

Computational Details. The MM-3DRISM calculations reported here were performed using a computing cluster at the Max Planck Institute for Mathematics in the Sciences, Leipzig, Germany. The most time-consuming step in these calculations is solving the 3DRISM equations; the remaining steps require minimal computational expense (∼10 to 20 min on a single CPU). The mean time required to solve the 3DRISM equations for a single conformer of apo chymosin was ∼245 min (on an Intel(R) Core (TM)2 Duo CPU E8600 3.33 GHz processor). By their nature, MM-3DRISM calculations are trivially parallel (e.g., one calculation per node), but the time required for a single calculation could be significantly reduced by using advanced numerical algorithms[77,78] or by performing the simulations using parallel computation.[41]

RESULTS

Binding Free Energy Calculations. Four different methods were used to calculate the binding free energy of the bovine chymosin−bovine *κ*-casein complex. The component energy terms for each method are summarized in Table 1.

Table 1. Components of the Binding Free Energy of the Bovine Chymosin−Bovine *κ*-Casein Complex, as Calculated Using the MM-PBSA and MM-3DRISM Methodologies[a]

contribution	mean (kcal/mol)	standard error (kcal/mol)
ΔE_{elec}	−1236.23	9.92
ΔE_{vdw}	−143.07	1.29
ΔE_{gas}	−1379.3	10.88
$T\Delta S$	−40.10	1.80
$\Delta G_{\mathrm{hyd}}^{\mathrm{PBSA}}$	1279.45	9.82
$\Delta G_{\mathrm{hyd}}^{\mathrm{3DRISM-KH}}$	1318.82	10.08
$\Delta G_{\mathrm{hyd}}^{\mathrm{3DRISM-GF}}$	1353.7	10.37
$\Delta G_{\mathrm{hyd}}^{\mathrm{3DRISM-UC}}$	1333.21	10.43
$\Delta G_{\mathrm{bind}}^{\mathrm{PBSA}}$	−59.75	1.73
$\Delta G_{\mathrm{bind}}^{\mathrm{3DRISM-KH}}$	−20.38	1.57
$\Delta G_{\mathrm{bind}}^{\mathrm{3DRISM-GF}}$	14.51	1.31
$\Delta G_{\mathrm{bind}}^{\mathrm{3DRISM-UC}}$	−5.98	1.06

[a] All values are give in units of kcal/mol.

For each method, a large favorable change in the electrostatic energy on binding (ΔE_{elec}) is opposed by an unfavorable change in the solvation energy (ΔG_{solv}) of similar magnitude. The magnitude of the van der Waals contribution to the binding free energy (ΔG_{vdw}) is smaller, but this term is significant for each method since it makes complexation more thermodynamically favorable.

The differences in the binding free energies calculated by the four methods tested here result from differences in the way that they compute the change in solvation free energy on binding (since all other terms are identical between these methods). As previously mentioned, all four methods predict the change in solvation free energy to be strongly unfavorable, but they do not agree on the magnitude of the term, with estimates ranging from 1279.45 to 1353.7 kcal/mol. The KH, UC, and GF free energy functionals of 3DRISM compute the changes in solvation free energy on binding to be less favorable than the estimate obtained from the PBSA model by 39.37, 53.76, and 74.25 kcal/mol, respectively. As a consequence, the binding free energies computed by the MM-3DRISM methods are all less thermodynamically favorable than the MM-PBSA value. A similar result has previously been observed for the binding of biotin analogues to avidin.[18] It is not possible to compare our results to experimental ones, unfortunately, since neither hydration free energies or binding free energies have been reported for the molecules or complexes considered here. It is interesting, however, to consider the accuracy of the hydration free energies calculated by the three different 3DRISM free energy functionals compared to the results of the PBSA method and chemical reasoning. While the PBSA method is probably too crude an approach to give accurate hydration free energies for biomacromolecules (e.g., peptides, proteins) and is therefore not an ideal benchmark, it would be expected to give roughly the correct trends. The correlation coefficients between the hydration free energies calculated by the PBSA method and the KH, GF, and UC functionals of 3DRISM for the 81 conformers of bovine *κ*-casein used in the binding free energy calculations are given in Table 2. As can be seen, there is a good correlation between the hydration free energies calculated by all four of the different methods. The same data are plotted for the PBSA and UC methods in Figure 3, which clearly shows that the UC free energy functional predicts the same general trends as the PBSA method (the correlation plots for the KH and GF free energy functionals are provided in the Supporting Information). The hydration free energies computed by these two methods are well correlated over a range of ∼40 kcal/mol. Although the mean and standard deviation of the hydration free energies computed by the UC free energy functional (*mean* = −498.02 kcal/mol, σ = 8.04 kcal/mol) are not in perfect agreement with those computed by the PBSA method (*mean* = −562.62 kcal/mol, σ = 8.29 kcal/mol), both methods predict that the peptide should strongly prefer to be in solution rather than gas phase under conditions considered here. The difference between the three 3DRISM free energy functionals is apparent when the absolute values of the hydration free energies calculated by the KH or GF functionals are considered. The means of the hydration free energies computed by the KH (*mean* = −119.64 kcal/mol, σ = 8.25 kcal/mol) and GF (*mean* = −233.39 kcal/mol, σ = 8.20 kcal/mol) functionals are significantly higher than those computed by either the UC functional or the PBSA method. The systematic error in the hydration free energies computed by the KH or GF free energy functionals is well-known and has previously been reported for both proteins[79] and small organic molecules.[9,42] The error arises from a significant overestimation

Table 2. Correlation Matrix for Hydration Free Energies Computed by the PBSA Method and Three 3DRISM Free Energy Functionals (KH, GF, UC) for 81 Conformers of Bovine *κ*-Casein[a]

	PBSA	KH	GF	UC
PBSA	1.00	0.94	0.94	0.97
KH	0.94	1.00	0.99	0.98
GF	0.94	0.99	1.00	0.99
UC	0.97	0.98	0.99	1.00

[a] Please see the text for details of each method.

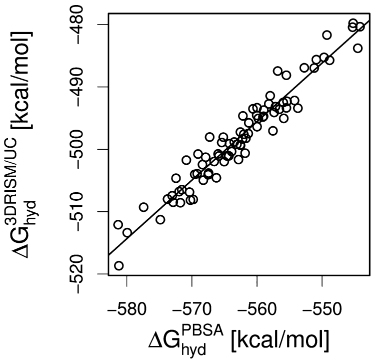


Figure 3. Correlation plot of hydration free energies computed by the PBSA method and the UC free energy functional for 81 conformers of apo bovine chymosin. Please see the text for details of each method.

of excluded volume effects in the standard 3DRISM theory. The results presented here suggest that the UC free energy functional, which includes a partial molar volume correction to account for excluded volume effects, provides more accurate estimates of hydration free energies of biomacromolecules (e.g., proteins) than the KH or GF free energy functionals.

Computational Alanine Scanning. For each residue of the P9−P7′ bovine *κ*-casein peptide complexed to bovine chymosin (excluding *AlaP2′*), computational alanine scanning calculations were performed (Figure 4). These computations allow us to quantify the importance of specific residues to the overall binding free energy. In these results, negative $\Delta\Delta G_{\text{bind}}$ values indicate residues that are favorably mutated to alanine, whereas positive numbers indicate a loss in the binding free energy upon mutation to alanine. The unsigned error of calculations of this nature has previously been reported to be 1 kcal/mol, and values higher than this are said to be significant and termed warm and hot spots with values above 1 and 2 kcal/mol, respectively.

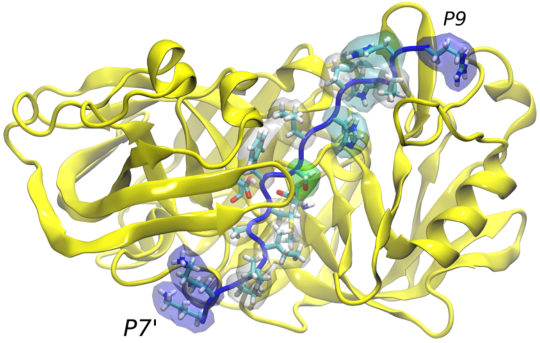


Figure 4. Bovine chymosin, shown in yellow, complexed with bovine *κ*-casein, where the backbone is shown in blue and the residue side chains are shown by sticks. The residues for which alanine scanning calculations have been performed are indicated with a van der Waals surface. The colors of the surfaces are based on the residue type defined by VMD: basic (blue) and nonpolar (white).

The results of the computational alanine scanning calculations for each of the four methods tested here are presented in Figure 5. A good qualitative agreement is observed between the results of each of these methods, but some subtle differences can be observed. The results obtained using the GF and UC free energy functionals are the most highly correlated, which is easily rationalized by the fact that the UC functional is derived from the GF functional (by applying a partial molar volume correction). The $\Delta\Delta G_{\text{bind}}$ values calculated using the KH free energy functional are in many cases a little more positive (0−1 kcal/mol) than those obtained using the GF or UC free energy functionals, while the magnitudes of the MM-PBSA results are in general greater than the corresponding MM-3DRISM results (by 0−3 kcal/mol) per mutated residue. As a consequence of these trends, the MM-PBSA method might in general be expected to identify more warm- or hot-spot residues than the MM-3DRISM methods. Furthermore, the MM-3DRISM-KH approach is more likely to judge mutations to alanine to be unfavorable.

Water-Binding Sites. Specific interactions between the surface of chymosin and water molecules that exhibit high residence times are believed to be important for modulating both the structure and dynamics of chymosin and its catalytic behavior.[36,80] Here, we have used the *Placevent* algorithm proposed by Sindhikara et al.[51] to identify specific water-binding sites from the 3DRISM distribution functions. A total of 1532 water binding sites with local solvent density greater than 1.5 times the bulk solvent density were identified and analyzed. The residue numbers of the conserved water molecules identified by experimental structural analysis by Prasad and Suguna[36] are given in Table 3 along with the index, *i*, of the nearest water molecule predicted by 3DRISM. It should be noted that, although the data set of Prasad and Suguna originates from a statistical analysis of many different aspartic proteases, which improves its reliability, the underlying

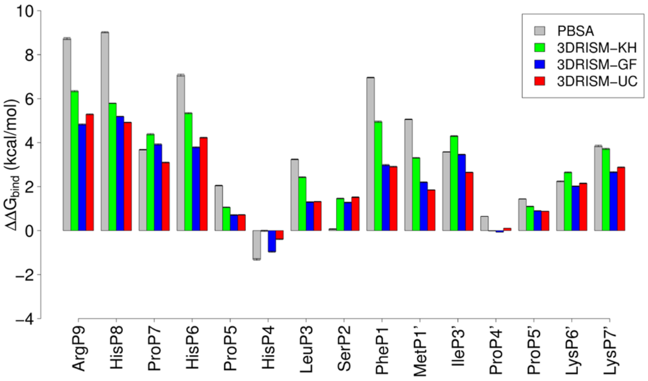


Figure 5. Computational alanine scanning results for the bovine chymosin−bovine *κ*-casein complex, where the *x*-axis labels indicate the residue of bovine *κ*-casein that has been mutated to alanine, and the *y* axis gives the value of $\Delta\Delta G_{\text{bind}} = \Delta G_{\text{bind}}^{\text{mutant}} - \Delta G_{\text{bind}}^{\text{wildtype}}$. The legend indicates that four different methods were tested for the calculation of $\Delta G_{\text{bind,solv}}$. Standard errors were calculated by bootstrap sampling on 10 000 samples. The *P2′* residue is alanine in wildtype bovine *κ*-casein and was therefore not included in the analysis.

Table 3. Comparison between Experimental and Calculated Water Oxygen Atom Positions in Bovine Chymosin[a]

exptl. water (ID)	nearest calcd. water (ID)	*R* (separation, Å)	∑ *R* < 2 Å	$g(\mathbf{r})_0$
1113	1	0.62	1	17.27
598	2	0.31	1	17.07
645	10	0.22	1	14.96
1007	13	0.28	1	14.50
1116	20	0.41	1	13.71
1125	25	0.7	1	13.05
1117	30	5	0	12.53
1083	39	0.48	1	12.09
1006	47	0.77	1	11.77
1027	49	2.4	0	11.67
1031	64	0.7	1	10.79
1045	106	0.33	1	9.41
1005	126	0.45	1	8.78
1020	286	0.39	1	6.61
1035	607	1.12	1	4.53
1053	706	1.31	1	4.07

[a] Columns: (1) residue numbers (in RSCB PDB entry 3CMS) of 16 water molecules that are conserved in mammalian aspartic proteases;[36] (2) nearest water oxygen atom position calculated by 3DRISM given in order of solvent density; (3) distance between the water molecules given in the first two columns; (4) the number of calculated water oxygen atom positions within 2 Å of the given experimental oxygen atom position; (5) the value of the distribution function in the vicinity of the predicted water binding site ($g(\mathbf{r})_0$).

data are taken from X-ray crystallographic experiments, in which it is often difficult to accurately locate water molecules. Imprecise determination of phase or diffraction intensities in X-ray crystallography can result in artifacts in the measured electron-density map, which may erroneously be interpreted as water molecules.[81] One of the 17 water molecules that was identified by Prasad and Suguna as conserved in most aspartic proteases is not observed in any of the crystal structures of bovine chymosin and therefore was omitted from the analysis. All but two of the experimental water binding sites are within 2 Å of a calculated water binding site. The two remaining experimental water molecules are residue numbers 1117 and 1027 in RCSB PDB entry 3CMS, for which the nearest computed water binding sites are at distances of 2.4 and 5 Å, respectively.

Seven of the conserved water binding sites identified by experimental results are within the binding cleft of chymosin (residue numbers 1005, 1113, 1116, 1006, 1125, and 1007 from RCSB PDB entry 3CMS and residue number 645, which was imported from entry 1CMS; see Figure 6). Six of these molecules are found within the first 50 computed water binding sites (Table 3). The seventh solvent molecule is the catalytic water molecule, which is observed in experiments to lie between Asp34 and Asp216 (residue number 1005 in RSCB PDB entry 3CMS). The nearest calculated water binding site is number 126.

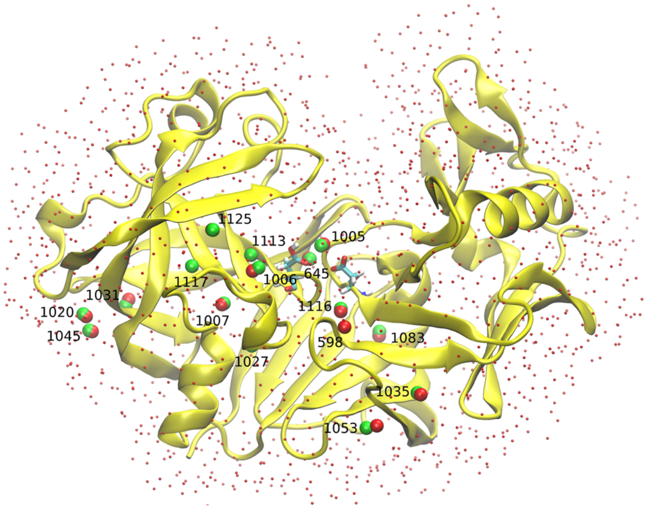


Figure 6. Water binding sites in apo bovine chymosin from experimental results (green) and 3DRISM computation (red). The predicted water binding site closest to each experimentally observed water binding site is illustrated as a large red sphere, while all other predicted water binding sites are shown as small red spheres. Only water oxygen atom sites are illustrated. Experimental data were taken from Prasad and Suguna.[36] The numeric labels give the residue numbers (in RSCB PDB entry 3CMS) of 16 conserved water molecules identified by Prasad and Suguna;[36] these numbers correspond to column 1 in Table 3.

The first water binding site identified from the 3DRISM distribution functions corresponds to an experimentally observed water molecule (residue number 1113 in RCSB PDB entry 3CMS) that is known to be important for stabilizing the catalytic site in an active configuration. The residues in the protein that are close to this water molecule are illustrated in Figure 7. The orientation of the water molecule with respect to the neighboring residues is consistent with a hydrogen bonding pattern suggested previously on the basis of structural analysis[36] and molecular dynamics simulations.[21,54]

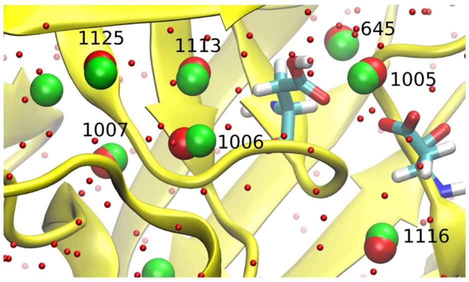


Figure 7. Water binding sites in the binding cleft of bovine chymosin from experimental results (green) and computation (red). Only water oxygen atom sites are illustrated. The numeric labels give the residue numbers (in RSCB PDB entry 3CMS) of seven conserved water molecules in the chymosin binding site; these numbers correspond to column 1 in Table 3.

■ DISCUSSION

The MM-PBSA method for calculating absolute and relative binding free energies has been widely used in the recent literature, since it is thought to provide a reasonable balance between computational expense and expected accuracy. Here, we have tested three different free energy functionals within the scope of the MM-3DRISM method proposed by Genheden et al.[18] All of these 3DRISM free energy functionals (GF, KH, and UC) allow similar estimates of relative binding free energies ($\Delta\Delta G_{\mathrm{bind}}$), which are in reasonably good agreement with the results of MM-PBSA calculations. Although the $\Delta\Delta G_{\mathrm{bind}}$ values calculated using the KH free energy functional are in many cases a little more positive (0−1 kcal/mol) than those obtained using the GF or UC free energy functionals, all three 3DRISM functionals predict values that are on average lower in magnitude (by 0−3 kcal/mol) per mutated residue than the predictions made by MM-PBSA. On the basis of the computational alanine scanning calculations alone, the GF, KH, and UC free energy functionals can all be recommended equally for use with the MM-3DRISM methodology. This is not surprising as the solvation component of the MM-3DRISM binding free energy is computed as the *change* in solvation free energy on binding, and the hydration free energies calculated by these free energy functionals for conformers of bovine κ-casein (or bovine chymosin or the chymosin−κ-casein complex) are observed to be well correlated. However, although it is the relative accuracy of the solvation free energy calculations for receptor, ligand, and complex that are most relevant, it is disconcerting that the absolute hydration free energies calculated using the GF and KH functionals are significantly more positive than those calculated by the PBSA method. By contrast, the hydration free energies computed using the UC functional are in better agreement with those calculated using the PBSA implicit solvent model. It should be noted that the PBSA results probably deviate significantly from reality, but since the hydration free energies of proteins are nearly impossible to measure, this can not easily be directly proven.

As well as being used to calculate solvation thermodynamics, the 3DRISM distribution functions provide information about the average solvent density around the solute of interest, which can be used to identify water-binding sites on the surface of the solute. Peaks in these distribution functions correspond to regions of high solvent density, where solvent molecules might be expected to bind to the surface of the solute. Here, we have used the *Placevent* algorithm proposed by Sindhikara et al.[51] to identify specific water-binding sites on the surface of bovine chymosin and have compared these results to the binding sites identified from experiment by Prasad and Suguna.[36] It is worth noting that these calculations could not be performed using an implicit continuum model (such as the PBSA method). The results of the water-binding calculations are encouraging. All but two of the experimental water binding sites were found within 2 Å of a predicted binding site. The two highest ranked water-binding sites identified by 3DRISM corresponded to experimentally observed binding sites. Moreover, 10 of the 16 conserved water molecules were identified within the 50 highest ranked predicted binding sites. Since the calculations are relatively inexpensive compared to explicit solvent simulation approaches, these results support earlier studies that concluded that 3DRISM might be a useful method for modeling explicit water molecules in computational docking studies, or corroborating the positions of water molecules predicted from experimental crystallography. Some caution is clearly required when interpreting the results; however, as a large number of false positives were also identified among the highest ranked predicted binding sites.

Since it was first purified in 1874, bovine chymosin has been the most widely used enzyme to initiate milk clotting in cheese manufacturing. The recent discovery that camel chymosin is a superior catalyst to bovine chymosin for proteolysis of bovine κ-casein (70% higher clotting activity and only 20% of the unspecific protease activity) has led to renewed interest in engineering new chymosin mutants for the food industry. We have calculated the Gibbs free energy of binding of a bovine κ-casein fragment in native bovine chymosin. Furthermore, we have investigated the influence of point mutations on the binding free energy by way of computational alanine scanning calculations. It is important to note that the binding free energy is only part of the enzymatic process, which also depends on other factors such as association/dissociaton kinetics. Experimentally, only milk clotting rates, K_M and k_{cat}, have been measured for chymosin complexes, which cannot be directly compared to binding free energies without making undue assumptions. Of the 15 point mutations to κ-casein that were tested during the alanine scanning calculations, only the HisP4Ala mutation was found to be favorable. In the bovine chymosin−bovine κ-casein complex, the HisP4 residue interacts with the side chain of Lys221 in the S4 pocket. From experiments, it is known that a HisP4Lys mutant of bovine κ-casein is a particularly disfavored substrate for bovine chymosin.[82] Furthermore, there is a Lys221Val mutation in the binding pocket of camel chymosin compared to bovine chymosin, which has previously been reported to be important for catalysis.[20,54] It is therefore, reasonable to propose that the S4 pocket in bovine chymosin should be targeted by experimentalists wishing to engineer mutants of bovine chymosin with improved milk clotting activity. This suggestion is further supported by the observation that there are many mutations in the P4 site and S4 pocket between the four different complexes of bovine/camel chymosin and bovine/camel κ-casein. The P4 residue is histidine in bovine κ-casein, while it is arginine in camel κ-casein. Moreover, in bovine chymosin the S4 pocket contains residue Lys221, which is mutated to Val221 in camel chymosin, as previously mentioned. The remainder of the mutations tested in the computational alanine scanning calculations were found to be unfavorable (Figure 5). Several of these results can be correlated with experimental observations. It has previously been proposed that ArgP9 is important for binding because it is conserved in bovine, camel, pig, buffalo, and goat chymosin,[31] which agrees with the observation made here that the mutation ArgP9Ala is strongly disfavored. Our observation that the HisP8-ProP7-HisP6 residues contribute a large amount to the binding free energy would support the suggestion that this so-called His-Pro cluster plays a key role in positioning bovine κ-casein in the bovine chymosin binding cleft.[83] However, our results suggest that the HisP4 residue does not contribute favorably to this interaction, as discussed above. Other factors in the catalytic activity of bovine chymosin, like the self-inhibiting mechanism, recognition and the encounter complex, and a pH switch observed in homologous proteins are worth investigating, and studies are underway in our lab to understand these, as are experimental studies evaluating binding properties of the herein predicted mutants for improved binding of κ-casein to chymosin.

■ CONCLUSIONS

We have reported on the results of solvent binding analysis and computational alanine scanning of the bovine chymosin−κ-casein complex performed using Molecular Integral Equation Theory. The bovine chymosin−bovine κ-casein complex is of industrial interest because bovine chymosin is widely used to cleave bovine κ-casein and to initiate milk clotting in the manufacturing of processed dairy products. There has been renewed interest in protein engineering of bovine chymosin since the recent discoveries that camel chymosin is a more efficient clotting agent than bovine chymosin for bovine milk, while bovine chymosin does not clot camel milk. The results show that the MM-3DRISM method of molecular integral equation theory is a practical alternative to MM-PBSA for computational alanine scanning of protein−peptide complexes. For the 15 single point mutants tested here, the results obtained using MM-3DRISM are in good overall agreement with those obtained using MM-PBSA. Unlike PBSA, however, 3DRISM provides useful information about the distribution of solvent density. We find that experimentally observed water-binding sites on the surface of bovine chymosin can be identified quickly and accurately from the density distribution functions computed by molecular integral equation theory. Since MM-3DRISM provides a realistic model of local solvent density, it is not necessary to include bound water molecules explicitly when calculating binding free energies, which is a major advantage compared to the MM-PBSA method.

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

The complete data sets including all calculated data, correlation plots of hydration free energies, and illustration of the change in local solvation density for 16 single-point mutants. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Authors

*E-mail: david.palmer@strath.ac.uk.
*E-mail: maxim.fedorov@strath.ac.uk.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

D.S.P. is grateful for funding from the European Commission through a Marie Curie Intra-European Fellowship within the seventh European Community Framework Programme (FP7-PEOPLE-2010-IEF). D.S.P. and M.V.F. thank the Scottish Universities Physics Alliance (SUPA) for funding through MVFs SUPA2 start-up funds. D.S.P. and M.V.F. are also grateful for partial financial support from the Deutsche Forschungsgemeinschaft (DFG)−German Research Foundation, Research Grant FE 1156/2-1, and for the use of the computing facilities at the John von Neumann-Institut für Computing (NIC), Juelich Supercomputing Centre (JSC), Forschungszentrum Juelich GmbH, Germany, project IDs HLZ18 and HLZ16. Additional results were obtained using the EPSRC funded ARCHIE-WeSt High Performance Computer (www.archie-west.ac.uk, EPSRC grant no. EP/K000586/1) and the Danish Center for Scientific Computing. B.S. thanks Danish National Research Foundation (DNRF59) and FTP (Danish Research Council, technology and production) for research funding. J.S. is grateful for postdoctoral funding from the Alfred Benzon Foundation.

■ REFERENCES

(1) Jorgensen, W. L. *Science* **2004**, *303*, 1813−1818.

(2) Baron, R.; Setny, P.; McCammon, J. A. *J. Am. Chem. Soc.* **2010**, *132*, 12091−12097.

(3) Karplus, M.; Petsko, G. A. *Nature* **1990**, *347*, 631−639.

(4) Kollman, P. *Chem. Rev.* **1993**, *93*, 2395−2417.

(5) Palmer, D. S.; Chuev, G. N.; Ratkova, E. L.; Fedorov, M. V. *Curr. Pharm. Des.* **2011**, *17*, 1695−1708.

(6) Karino, Y.; Fedorov, M. V.; Matubayasi, N. *Chem. Phys. Lett.* **2010**, *496*, 351−355.

(7) Matubayasi, N.; Shinoda, W.; Nakahara, M. *J. Chem. Phys.* **2008**, *128*, 195107.

(8) Palmer, D. S.; Sergiievskyi, V. P.; Jensen, F.; Fedorov, M. V. *J. Chem. Phys.* **2010**, *133*, 044104.

(9) Palmer, D. S.; Frolov, A. I.; Ratkova, E. L.; Fedorov, M. V. *Mol. Pharmaceutics* **2011**, *8*, 1423−1429.

(10) Palmer, D. S.; McDonagh, J. L.; Mitchell, J. B. O.; van Mourik, T.; Fedorov, M. V. *J. Chem. Theory Comput.* **2012**, *8*, 3322−3337.

(11) Imai, T.; Oda, K.; Kovalenko, A.; Hirata, F.; Kidera, A. *J. Am. Chem. Soc.* **2009**, *131*, 12430−12440.

(12) Imai, T.; Hiraoka, R.; Kovalenko, A.; Hirata, F. *J. Am. Chem. Soc.* **2005**, *127*, 15334−15335.

(13) Imai, T.; Kinoshita, M.; Hirata, F. *Bull. Chem. Soc. Jpn.* **2000**, *73*, 1113−1122.

(14) Yoshida, N.; Phongphanphanee, S.; Maruyama, Y.; Imai, T.; Hirata, F. *J. Am. Chem. Soc.* **2006**, *128*, 12042−12043.

(15) Kast, S. M.; Heil, J.; Guessregen, S.; Schmidt, K. F. *J. Comput.-Aided Mol. Des.* **2010**, *24*, 343−353.

(16) Stumpe, M. C.; Blinov, N.; Wishart, D.; Kovalenko, A.; Pande, V. S. *J. Phys. Chem. B* **2011**, *115*, 319−328.

(17) Miyata, T.; Hirata, F. *J. Comput. Chem.* **2008**, *29*, 871−882.

(18) Genheden, S.; Luchko, T.; Gusarov, S.; Kovalenko, A.; Ryde, U. *J. Phys. Chem. B* **2010**, *114*, 8505−8516.

(19) Foltmann, B.; Pedersen, V. B.; Kauffman, D.; Wybrandt, G. *J. Biol. Chem.* **1979**, *254*, 8447−8456.

(20) Kappeler, S. R.; van den Brink, H. J.; Rahbek-Nielsen, H.; Farah, Z.; Puhan, Z.; Hansen, E. B.; Johansen, E. *Biochem. Biophys. Res. Commun.* **2006**, *342*, 647−654.

(21) Sørensen, J.; Palmer, D. S.; Qvist, K. B.; Schiøtt, B. *J. Agric. Food. Chem.* **2011**, *59*, 5636−5647.

(22) Gilliland, G. L.; Winborne, E. L.; Nachman, J.; Wlodawer, A. *Proteins* **1990**, *8*, 82−101.

(23) Strop, P.; Sedlacek, J.; Stys, J.; Kaderabkova, Z.; Blaha, I.; Pavlickova, L.; Pohl, J.; Fabry, M.; Kostka, V.; Newman, M. *Biochemistry* **1990**, *29*, 9863−9871.

(24) Newman, M.; Safro, M.; Frazao, C.; Khan, G.; Zdanov, A.; Tickle, I. J.; Blundell, T. L.; Andreeva, N. *J. Mol. Biol.* **1991**, *221*, 1295−1309.

(25) Groves, M. R.; Dhanaraj, V.; Badasso, M.; Nugent, P.; Pitts, J. E.; Hoover, D. J.; Blundell, T. L. *Protein Eng. Des. Sel.* **1998**, *11*, 833−840.

(26) Langholm Jensen, J.; Mølgaard, A.; Navarro Poulsen, J.-C.; Harboe, M. K.; Simonsen, J. B.; Lorentzen, A. M.; Hjernø, K.; van den Brink, J. M.; Qvist, K. B.; Larsen, S. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2013**, *69*, 901−913.

(27) Schechter, I.; Berger, A. *Biochem. Biophys. Res. Commun.* **1967**, *27*, 157−162.

(28) Plowman, J. E.; Creamer, L. K. *J. Dairy Res.* **1995**, *62*, 451−467.

(29) Plowman, J. E.; Creamer, L. K.; Smith, M. H.; Hill, J. P. *J. Dairy Res.* **1997**, *64*, 299−304.

(30) Plowman, J. E.; Smith, M. H.; Creamer, L. K.; Liddell, M. J.; Coddington, J. M.; Gibson, J. J.; Engelbretsen, D. R. *Magn. Reson. Chem.* **1994**, *32*, 458−464.

(31) Bairoch, A.; Apweiler, R.; Wu, C. H.; Barker, W. C.; Boeckmann, B.; Ferro, S.; Gasteiger, E.; Huang, H.; Lopez, R.; Magrane, M.; Martin, M. J.; Natale, D. A.; O'Donovan, C.; Redaschi, N.; Yeh, L.-S. L. *Nucleic Acids Res.* **2005**, *33*, D154−159.

(32) Macheboef, D.; Coulon, J. B.; D'Hour, P. *J. Dairy Res.* **1993**, *9*, 373−374.

(33) Kabsch, W.; Sander, C. *Biopolymers* **1983**, *22*, 2577−2637.

(34) Chitpinityol, S.; Crabbe, M. J. C. *Food Chem.* **1998**, *61*, 395−418.

(35) Pearl, L.; Blundell, T. *FEBS Lett.* **1984**, *174*, 96−101.

(36) Prasad, B. V.; Suguna, K. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2002**, *58*, 250−259.

(37) Andreeva, N.; Dill, J.; Gilliland, G. L. *Biochem. Biophys. Res. Commun.* **1992**, *184*, 1074−1081.

(38) Beglov, D.; Roux, B. *J. Phys. Chem.* **1997**, *101*, 7821−7826.

(39) Du, Q. H.; Beglov, D.; Roux, B. *J. Phys. Chem. B* **2000**, *104*, 796−805.

(40) Hirata, F. *Molecular Theory of Solvation*; Kluwer Academic Publishers: Dordrecht, The Netherlands, 2003.

(41) Luchko, T.; Gusarov, S.; Roe, D. R.; Simmerling, C.; Case, D. A.; Tuszynski, J.; Kovalenko, A. *J. Chem. Theory Comput.* **2010**, *6*, 607−624.

(42) Palmer, D. S.; Frolov, A. I.; Ratkova, E. L.; Fedorov, M. V. *J. Phys. Condens. Matter* **2010**, *22*, 492101.

(43) Hansen, J.-P.; McDonald, I. R. *Theory of Simple Liquids*, 4th ed; Academic Press: New York, 2000.

(44) Duh, D. M.; Haymet, A. D. J. *J. Chem. Phys.* **1995**, *103*, 2625−2633.

(45) Chuev, G. N.; Vyalov, I.; Georgi, N. *Chem. Phys. Lett.* **2013**, *561−562*, 175−178.

(46) Kovalenko, A.; Hirata, F. *J. Phys. Chem. B* **1999**, *103*, 7942−7957.

(47) Ratkova, E. L.; Chuev, G. N.; Sergiievskyi, V. P.; Fedorov, M. V. *J. Phys. Chem. B* **2010**, *114*, 12068−12079.

(48) Ten-no, S.; Jung, J.; Chuman, H.; Kawashima, Y. *Mol. Phys.* **2010**, *108*, 327−332.

(49) Chandler, D.; Singh, Y.; Richardson, D. M. *J. Chem. Phys.* **1984**, *81*, 1975−1982.

(50) Imai, T.; Harano, Y.; Kovalenko, A.; Hirata, F. *Biopolymers* **2001**, *59*, 512−519.

(51) Sindhikara, D. J.; Yoshida, N.; Hirata, F. *J. Comput. Chem.* **2012**, *33*, 1536−1543.

(52) Sørensen, J.; Palmer, D. S.; Schiøtt, B. *J. Agric. Food Chem.* **2013**, *61*, 7949−7959.

(53) Kuhn, B.; Gerber, P.; Schulz-Gasch, T.; Stahl, M. *J. Med. Chem.* **2005**, *48*, 4040−4048.

(54) Palmer, D. S.; Christensen, A. U.; Sørensen, J.; Celik, L.; Qvist, K. B.; Schiøtt, B. *Biochemistry* **2010**, *49*, 2563−2573.
(55) Phillips, J. C.; Braun, R.; Wang, W.; Gumbart, J.; Tajkhorshid, E.; Villa, E.; Chipot, C.; Skeel, R. D.; Kale, L.; Schulten, K. *J. Comput. Chem.* **2005**, *26*, 1781−1802.
(56) Blinov, N.; Dorosh, L.; Wishart, D.; Kovalenko, A. *Biophys. J.* **2010**, *98*, 282−296.
(57) Srinivasan, J.; Cheatham, T. E.; Cieplak, P.; Kollman, P. A.; Case, D. A. *J. Am. Chem. Soc.* **1998**, *120*, 9401−9409.
(58) Kollman, P. A.; Massova, I.; Reyes, C.; Kuhn, B.; Huo, S. H.; Chong, L.; Lee, M.; Lee, T.; Duan, Y.; Wang, W.; Donini, O.; Cieplak, P.; Srinivasan, J.; Case, D. A.; Cheatham, T. E. *Acc. Chem. Res.* **2000**, *33*, 889−897.
(59) Gohlke, H.; Case, D. A. *J. Comput. Chem.* **2004**, *25*, 238−250.
(60) Jensen, F.; Palmer, D. S. *J. Chem. Theory Comput.* **2011**, *7*, 223−230.
(61) Miller, B. R.; McGee, T. D.; Swails, J. M.; Homeyer, N.; Gohlke, H.; Roitberg, A. E. *J. Chem. Theory Comput.* **2012**, *8*, 3314−3321.
(62) Lue, L.; Blankschtein, D. *J. Phys. Chem.* **1992**, *96*, 8582−8594.
(63) Berendsen, H. J. C.; Grigera, J. R.; Straatsma, T. P. *J. Phys. Chem.* **1987**, *91*, 6269−6271.
(64) Hirata, F.; Rossky, P. J. *Chem. Phys. Lett.* **1981**, *83*, 329−334.
(65) Lee, P. H.; Maggiora, G. M. *J. Phys. Chem.* **1993**, *97*, 10175−10185.
(66) Kovalenko, A.; Hirata, F. *J. Chem. Phys.* **2000**, *113*, 2793−2805.
(67) Chuev, G.; Fedorov, M.; Crain, J. *Chem. Phys. Lett.* **2007**, *448*, 198−202.
(68) Allen, M. P., Tildesley, D. J. *Computer Simulation of Liquids*; Clarendon Press: Oxford, 1987.
(69) Case, D. A.; Cheatham, T. E.; Darden, T.; Gohlke, H.; Luo, R.; Merz, K. M.; Onufriev, A.; Simmerling, C.; Wang, B.; Woods, R. J. *J. Comput. Chem.* **2005**, *26*, 1668−1688.
(70) Kovalenko, A.; Ten-No, S.; Hirata, F. *J. Comput. Chem.* **1999**, *20*, 928−936.
(71) Perkyns, J. S.; Pettitt, B. M. *Chem. Phys. Lett.* **1992**, *190*, 626−630.
(72) Massova, I.; Kollman, P. A. *J. Am. Chem. Soc.* **1999**, *121*, 8133−8143.
(73) Bradshaw, R. T.; Patel, B.; Tate, E.; Leatherbarrow, R.; Gould, I. *Protein Eng. Des. Sel.* **2011**, *24*, 197−207.
(74) Huo, S.; Massova, I.; Kollman, P. *J. Comput. Chem.* **2002**, *23*, 15−27.
(75) Moreira, I.; Fernandes, P.; Ramos, M. *J. Comput. Chem.* **2007**, *28*, 644−654.
(76) *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2012. ISBN 3-900051-07-0.
(77) Sergiievskyi, V. P.; Hackbusch, W.; Fedorov, M. V. *J. Comput. Chem.* **2011**, *32*, 1982−1992.
(78) Sergiievskyi, V.; Fedorov, M. *J. Chem. Theory Comput.* **2012**, *8*, 2062−2070.
(79) Imai, T.; Kovalenko, A.; Hirata, F. *Chem. Phys. Lett.* **2004**, *395*, 1−6.
(80) Dunn, B. M. *Chem. Rev.* **2002**, *102*, 4431−4458.
(81) Carugo, O.; Bordo, D. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **1999**, *55*, 479−483.
(82) Visser, S.; Slangen, C. J.; van Rooijen, P. J. *Biochem. J.* **1987**, *244*, 553−558.
(83) Gustchina, E.; Rumsh, L.; Ginodman, L.; Majer, P.; Andreeva, N. *FEBS Lett.* **1996**, *379*, 60−62.