Coarse-grained Modeling with Constant pH of the Protein Complexation Phenomena

Master's Dissertation

student

Sergio Alejandro Poveda Cuevas

supervisor

Fernando Luís Barroso da Silva

University of São Paulo

Interinstitutional Graduate Program in Bioinformatics

Ribeirão Preto – São Paulo

2017

Coarse-grained Modeling with Constant pH of the Protein Complexation Phenomena

Abstract

Theoretical studies of the molecular mechanisms responsible for the formation and stability of protein complexes have gained importance due to their practical applications in the understanding of the molecular basis of several diseases, in protein engineering and biotechnology. The objective of this project is to critically analyze and refine a *coarse-grained* force field for protein-protein interactions based on experimental thermodynamic properties and to apply it to cancer-related S100A4 protein system. Our ultimate goal is to generate knowledge for a better understanding of the physical mechanisms responsible for the association of particular proteins in different environments. We studied the role of short and long-range interactions on the complexation of homo-associations. Furthermore, we analyzed the influence of the pH and its correlation with the charge regulation mechanism. We analyzed and refined the adjustable Lennard-Jones parameter for a mesoscopic model based on experimental second virial data for lysozyme, α-chymotrypsinogen, and ribonuclease A via Monte Carlo simulations. From of that, the S100A3 protein was used to test the new calibrated parameters. Finally, we evaluated the dimerization process of S100A4 proteins, observing the role of physical-chemistry variables involved in the thermodynamical stability of different oligomers.

Keywords: Molecular modeling, computer simulation, Monte Carlo, coarse-grained model, protein complexation, second virial coefficient B_2.

Resumo

Estudos teóricos dos mecanismos moleculares responsáveis pela formação e estabilidade dos complexos de proteínas vêm ganhando importância devido às suas aplicações práticas no entendimento da base molecular de várias doenças, em engenharia de proteínas e biotecnologia. O objetivo deste projeto é analisar criticamente e aperfeiçoar um campo de força de *granulidade grossa* para interação proteína-proteína com base em propriedades termodinâmicas experimentais e aplicá-lo ao sistema proteico S100A4 relacionado com o câncer. Nosso objetivo final é gerar conhecimento para uma melhor compreensão dos mecanismos físicos responsáveis pelas associações de proteínas particulares em diferentes ambientes. Estudamos o papel das interações de curto e longo alcance na complexação de homo-associações. Além disso, analisamos a influência do pH e sua correlação com o mecanismo de regulação de cargas. Por meio de simulações Monte Carlo, analisamos e refinamos o parametro ajustável de Lennard-Jones para um modelo mesoscópico, usando dados experimentais do segundo virial para a lisozima, o α-quimotripsinogênio e a ribonuclease A. A partir disso, a proteína S100A3 foi usada para testar os novos parâmetros calibrados. Finalmente, foi avaliado o processo de dimerização das proteínas S100A4, observando o papel de algumas variáveis físico-químicas envolvidas na estabilidade termondinâmica de diferentes oligómeros.

Palavras chave: Modelagem molecular, simulação computacional, Monte Carlo, modelo de granulidade grossa, complexação de proteina, segundo coeficiente de virial B_2.

Distribution by (name and addres): Sergio Alejandro Poveda Cuevas, Interinstitutional Graduate Program in Bioinformatics, University of São Paulo, Ribeirão Preto, Brazil (2017). E-mail: alejandropc@usp.br.

Dedicado a mi bella familia...

Acknowledgments

Several “thank you” in three different languages (spanish, english, and portugues).

Quiero dar mis más profundos agradecimientos a mis padres Aura Rosa y José Fredy, por su constante apoyo y confianza depositados incluso en las situaciones más complicadas que pudo atravesar nuestra familia. Cada una de sus invaluables enseñanzas han hecho de mi una mejor persona cada día, y con este pequeño trozo de experiencia académica, quiero resaltar que todos sus esfuerzos han valido la alegría (no la pena) de vivir en este corto período de tiempo.

A Adriana, por siempre estar a mi lado, confiar en mis capacidades y brindarme tus más grandes enseñanzas de respeto hacia la humanidad. Por cultivar en mi, una persona más sensible y comprensiva, sin tu presencia sería una piedra tosca y sin sentido en el devenir de la vida.

Agradezco a mi hermana y mis hermanos Laura Catalina, Leonardo Andrés y Freddy Jackson, por darme siempre la mejor magnitud, dirección y sentido; nuestra posición en esta historia y reciprocidad, siempre nos retroalimentarán positivamente como vectores de sabiduría, conocimiento y sensatez. Agradezco a sus esposas Paola Cepeda y Sandra Suárez por siempre estar del lado de las soluciones cuando los problemas se presentan.

Quero expressar um agradecimento especial a meu orientador, professor Fernando Barroso, por todos os ensinamentos, apoio e confiança dados. Obrigado por me apresentar desde uma perspectiva muito mais detalhada, este maravilhoso mundo da simulação computacional em biomoléculas; todas as suas dicas e formas de planejar os problemas, tem me ajudado ver a pesquisa de um deito mais objetivo, crítico e cuidadoso.

Aos membros da banca de qualificação professores Antônio Caliri, Silvana Giuliatti e Norival Alves Santos; e aos membros da banca de defessa professores Leandro Barbosa, Luciana Malavolta, José Cesar Rosa, por suas importantes sugestões neste trabalho.

Aos trabalhos prévios do Laboratório de Biofísica-química Computacional e aos diferentes pesquisadores que contribuiram com o desenvolvimento das diversas

ferramentas computacionais usadas neste trabalho. Sem todas essas importantes contribuições não teria sido possível realizar este estudo.

Quero agradecer aos colegas e amigos da USP Lariani Delboni, Paulo Siani, Rafael Maglia, Yagoub Ibrahim e Junier Marrero, por seus conselhos e interessantes pontos de vista em nossas discussões acadêmicas.

A todos mis amigos que desde lejos y cerca estuvieron adyacentemente conmigo durante mi maestría, brindándome su fuerza, cariño y apoyo. Quiero dar un especial agradecimiento a: Diana Cuevas, Javier Muñoz, Paula Hofmann, David Casilimas, Tatiana Niño, Tatiana Triana, Karen Pulido, Norita Perez, Jennyfer Aldana, Jorge Wilches, Jorge Sosa, Natalia Montellano, David Avellaneda... ¡muchas gracias!

Thank you to professors Luis Gustavo Dias, Catherine Etchebest and Nelson Augusto Alves by their teaching in computational chemistry, molecular modeling and statistical mechanics.

A mis primeros profesores, Edgar Montaño y Carolina Camargo, los cuales me dieron la guía para abrir varias puertas de las áreas: bioinformática, bioquímica y biología molecular.

Ao programa de Bioinformática (USP) e à Faculdade de Ciências Farmacêuticas (USP-RP), em especial às secretarias Patricia Martorelli e Cristiane de Fátima Braulino, responsáveis por todo o suporte burocrático neste mestrado, muito obrigado.

Agradeço ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) e ao governo do Brasil, por o financiamento de meu mestrado.

Thank you so much to all people of different countries (Colombia, Brazil, Peru, Bolivia, Argentina, Mexico, Cuba, Iran, India, Egypt, Sudan, Nigeria, USA, Canada, Spain, France, Germany, Sweden, Slovenia, etc.) that I get to know. They allowed me to break even more all stereotypes that society has implanted. Despite speaking different languages and having different cultures, we laugh in the same way, suffer in the same way and believe in the welfare of a better world.

To everyone who comes to read this work, the efforts have been worth it.

Aquí no tiene nada que ver la cuestión del daño o del no daño. Te ocurrirá lo mismo que al río durante la primavera. El río crece, aumenta su caudal, alimenta la tierra de sus riberas y mantiene su propio curso hasta penetrar en el mar, que lo acoge hospitalariamente por ser su más valioso aliado. "Esfúerzate en comprender hasta ese límite las disposiciones de la Dirección". Franz Kafka - La Gran Muralla China (1919).

Contents

List of Figures

List of Tables

Symbols and Abbreviations

ϵ	Dielectric constant of the medium
ϵ_0	Dielectric constant of the vacuum (in C^2/N m^2)
κ	Inverse Debye screening length (in Å)
μ	Dipole moment (in D)
μ_P	Dipole moment number of the protein (in Å)
ρ_{ij}	Number of contacts between residues i and j
σ	Sum of proteins radii (in Å)
ε_{LJ}	Depth of the Lennard-Jones potential (in k_BT)
A	Free energy of interaction (in k_BT)
B_2	Second virial coefficient B_2 (in mol ml/g^2)
B_{hs}	Hard sphere contribution in the second virial coefficient B_2 (in mol ml/g^2)
B_{q+vdw}	Electrostatic and van der Waals contributions in the second virial coefficient B_2 (in mol ml/g^2)
C,C_P	Charge regulation parameter
c_{axis}	Translation axis of cylinder (in Å)
c_k	Density of the mobile electrolyte specie k

dr	Displacement distance (in Å)
e	Elementar charge (equivalent to 1.602×10^{-19} C)
F	Work done by force to a given distance (in N)
$f\,(r)$	Number of contacts between all residues found at a distance r
$g\,(r)$	Radial distribution function
k_B	*Boltzmann*'s constant (equivalent to 1.3807×10^{-23} J mol^{-1} K^{-1})
l_B	*Bjerrum*'s length (in Å)
l_{cyl}	Cylinder height (in Å)
M	Molecular weight of protein (in g/mol)
Mw	Molecular weight of a particular amino acid (g/mol)
N	Number of aminoacids
N_A	Avogadro number (equivalent to 6.02×10^{23} mol^{-1})
pK_a	Dissociation constant of a particular amino acid
r	Separation distance (in Å)
R_{ai}	Radius of Lennard-Jones sphere of the amino acid i (in Å)
r_{cyl}	Cylinder radius (in Å)
r_k	Amino acid coordinates
s	Number of simulations
T	Temperature of the system (in *Kelvin*)
U	Internal energy (in k_BT)
u^{el}	Coulombic potential energy

u^{vdw}	Lennard-Jonnes potential energy
$w(r)$	Potential of mean force (in k_BT)
z	Charge or valence of a given chemical species
Z,Z_P	Charge number or valence
A3	Protein S100A3
A4	Protein S100A4
$A4^{d}$	Dimer of S100A4
$A4^{m}$	Monomer of S100A4
adv	Approach of different value
auv	Approach of unique value
auv^{h}	Approach of unique value at high salt concentration
BOL	*Born-Oppenhimer* level
CG	Coarse-grained models
CHY	Protein α-chymotrypsinogen
DLVO	Derjaguin, Landau, Verwey and Overbeek theory
LF	Protein lactoferrin
LJ	Lennard-Jones
LYS	Protein lysozyme
MC	Monte Carlo
MD	Molecular dynamics
MML	*McMillan-Mayer* level

MO	Membrane osmometry
NTD	N-terminal domain of the spidroin protein
PBE	Poisson-Boltzmann equation
R51A	Mutant type for S100A3
RIB	Protein Ribonuclease
RMSD	Root mean square deviation
$RMSD_{BB}$	Root mean square deviation between backbones
SANS	Small-angle neutron scattering
SAXS	Small-angle x-ray scattering
SEC	Size exclusion chromatography
SIC	Self-interaction chromatography
SL	*Schrödinger* level
SLS	Static light scattering
TK	Tandford-Kirkwood model
vdW	van der Waals forces
WT	Wild type

Chapter 1

Introduction

The biological and chemical systems are found in aqueous solutions that often contain ions. This inherently produces a mutual and complex interaction that allows to observe different physical and chemical phenomena. Macromolecules such as proteins are highly dependent on characteristics and composition of the micro-environment where they are present.[1, 2] Proteins are bodies constituted by amino acids (*aa*) covalently bonded through the so-called peptide bonds. These apparently similar molecular units have a singular lateral chain which confers each one, a particular physical-chemistry property giving to proteins specific features in relation to its topology, geometry, size and charge, to name a few aspects. Proteins are macromolecules that can be described as thermodynamic systems, thus, they obey the acid-base equilibrium, which involves the exchange of ions on both directions (*i.e.* from protein to the medium or vice versa).[3, 4, 5] This behavior is possible, because proteins have seven ionized residues which have the ability to donate or receive protons: aspartic acid (Asp), glutamic acid (Glu), cysteine (Cys) not involved in SS bonds, tyrosine (Tyr), arginine (Arg), histidine (His), lysine (Lys), and its carboxyl (CTR) and amine (NTR) terminals.[6] Therefore, following a logical order of ideas, salt or pH concentration in a determinate solution produces naturally physical effects in the electrostatic nature of proteins creating a medium suitable to observe various processes (*e.g.* protein *folding*, or *protein-protein* interaction).[7, 8]

With the development of quantum mechanics it became possible to understand the nature of the molecular interactions. Intermolecular forces are electrostatics in origin, and in consequence they can be effectively described by the theory of classical electrostatics; this alternative view is consistently interesting in terms of viability, since the solution of the Schrödinger equation is not easy, even for systems as "simple" as two hydrogen atoms.[9, 10] For this reason, it was convenient to establish a practical and effective classification of intermolecular interactions taking

into account that all they have the same fundamental origin. In this way, they have been named generically as short-range and long-range forces. In a more specific sense when applied to biomolecular systems, they can also be classified as interaction of electrical charges (*Coulombic* interactions, charge-dipole, dipole-dipole, etc.), van der Waals forces (vdW) and hydrophobic effects.[10, 11] This considerable number of physical interactions can be observed in proteins guiding their structure and functions.[7] As mentioned earlier, several features as the structure in these biological entities varies depending on the conditions of their neighborhood. With this in mind, resurrects the famous paradigm that says *biological functionality is determined by the structure.*[12] Here we want to focus on one of the fundamental roots of cellular life: molecular complexes, and we extended a little vision of that paradigm saying that the *"biological functionality is determined by the complexation phenomena*".[13, 14]

The protein molecular complexes are an intriguing subject and have been studied with great enthusiasm in various areas of knowledge in recent years. The reason for this is that in this molecular universe, the number of problems proposed huge challenges due to the practical applications derived from these analyses and also theoretical formulations that can be proposed to better understand the background of what is happening. Protein complexes can be formed by the physical association of different molecular species: protein-ion, protein-organic molecule, protein-nucleic acid, protein-lipid, protein-polysaccharide or protein-protein. Here we will focus on the latter. In the quaternary structure level in multimeric proteins, protein-protein interactions represent the basis of these associations. Protein complexes are found in every cellular location, including the cell organelles, the cytosol and the cell membranes. Hence, protein complexes are involved in most biological processes, such as enzymatic catalysis, signal transduction, transport of substances or structural functionality, demonstrating their importance for cell survival.[13, 15, 16, 17, 18] The complexes can be of two different types, homocomplexes (usually tend to be permanent) and heterocomplexes (which may be either permanent or non-permanent due to external factors).[13] Complex formation will be potentiated according to the *aa* sequence, the medium and the atomic arrangement they present in the space. In some cases Coulombic forces will be the basis of the association. In this type of interaction, for example, the net charges of two proteins are attracted each other by the complementarity of a positive charge and other negative.[8, 19] In a determined distance proteins undergo dipole effects that favor the attraction and subsequent aggregation of monomers. However, others mesoscopic phenomena as ion-ion correlation may also play an important role in the association, where two proteins are attracted by a organization of multivalents contra-ions around their

surfaces generating attractive forces each other due to a correlation between size and charge of counter-ions.[20, 21] Other phenomenon as the charge regulation mechanism may also give attractive interactions even for likely-charged proteins due to the proton fluctuation particularly at pH ≈ pI and low salt regimes.[5, 22, 23, 24, 25]

Understanding these physical-chemistry properties can lead to the development of optimal strategies and functionalized systems in biotechnology or in medical and pharmaceutical applications. Computational approaches have gained relevant conditions to contribute to address this problem due to their versatility and applicability, allowing the evaluation of different controllable scenarios and enabling to test the behavior of particular variables in relatively simple manner. In an effort to predict the physical nature of protein-protein interactions, theories based on statistical mechanics have shown remarkably well to describe the behavior of such systems through so-called effective models;[8, 20] where proteins can be treated as simplified objects. For the solution of these models, it has been invoked analytical[19, 26] and numerical methods [*e.g.* Monte Carlo (MC),[27, 28] molecular dynamics (MD),[29] docking approaches,[30] numerical solutions of the Poisson-Boltzmann equation (PBE),[31] etc.]. A relevant point in the computational treatment approximation is the structural representation of systems (biomolecules in our case) on which the respective calculations are made, for example, using the so-called *all-atoms*.[29] In a more simplified description the whole protein structure is reduced to a charged sphere, this model is used with frequency by the science of colloids.[5, 32] However, between these two levels of detail we find mesoscopic models known as *coarse-grained* (CG), where the atoms that constitute a specific residue are condensed to a single sphere.[33, 34, 35, 36] CG models are the central point in our discussion. To represent the solvent several models can be used.[37] However, often the water molecules can be replaced by structureless dielectric continuum, that mimics the main effect of the presence of water, *i.e.*, its screening behavior. This is appealing to reduce the cpu time and speed up the system sampling during the simulations.[8, 20] It should be noted here that the application of one model or another will be strongly dependent on the objectives to be achieved.

Bioinformatics is an extremely diverse area comprising *large scale* analysis and production of biological data using computational tools.[38] Bioinformatics has intersections with other areas of knowledge, since biological problems to be solved are not simple. To have a notion of what happens, physics and chemistry inevitably have been integrated to generate a more specific field called *structural bioinformatics*, which in turn overlaps with the approach of computational biophysics.[39] Our focus

does not escape of objectives developed in bioinformatics, which is to find sense in a biological context to the vast information generated in particular cases.[38] Protein complexes have many features that must be understood, and therefore, results obtained through of computational "experiments" on large scale, can give important information in the comprehension of these phenomena.[39]

Our main system of study is the protein S100A4 (A4) that belongs to the S100 family, a large group of calcium binding proteins.[40] A4 is a therapeutic target since this protein is involved in several metabolic cellular processes and has an important role in cancer formation and metastasis.[41] It is believed that the promotion of metastasis is correlated to the phenomenon of A4 complexation which is calcium-dependent intracellular and extracellular. It has been shown that this protein can self-associate to form oligomers of variable number (*e.g.* dimer, tetramer, hexamer, octamer, etc.).[42, 43]

In this work, our main general goal is to explain and to understand the physical foundations involved in the formation of molecular complexes, analysing the contributions responsible for this phenomenon. The emphasis is given here on short ranged (vdW) and long ranged interactions (electrostatics), particularly on how to deal with hydrophobic molecules within the continuum dielectric framework. Preliminary tests revealed the need of a better force field parametrization.[44, 45] Therefore, it was necessary to start with a critical analysis of the parametrization of the force field in order to better describe experimental data. There are evidences that the parameter routinely used can be either lacking attraction[46] or too attractive.[45] A physical-chemistry approach based on second virial coefficient B_2 (see Subsection 4.4.3) was chosen here for this propose. Due to the limited thermodynamic information available from A4, we use the parameter B_2 of a closely related protein called S100A3 (A3)[1].[48] Furthermore, following protocols established by other groups of investigation,[33] we largely worked with lysozyme (LYS), α-chymotrypsinogen (CHY), and ribonuclease A (RIB), since these proteins have a considerably quantity of available experimental information and have been studied by others.[33, 49, 50, 51, 52, 53, 54] Thus, based on these systems, we analysed and refined the calibration of the force field parameters by MC simulations. After, the homo-association of A3 proteins was employed as a test of our calibration procedure for this protein family. Finally, A4 protein relationated to metastasis process was studied.[55] Together with methodological developments, we aim to contribute with

[1] A3 is so far the only S100 protein family with available experimental values of B_2 related to complex formation in its monomeric and dimeric forms determined by light scattering assays.[47]

the understanding of physical interactions and biological issues relationated with these macromolecules.

Chapter 2

The Biological Problem

2.1 Brief overview of cancer

Many years ago science and technology have worked for the progress of humanity. They have functioned as an important way in the reduction and identification of mortality rates of different diseases, and thus, in gradual improvement of the survival of human race. Within of this huge range of diseases, the cancer tops the list as one of the most lethal and aggressive disease. Around the world, approximately 200 of each $100,000$ people die of cancer each year. While efforts in the prevention of cancer have increased and achieved significant results in recent years, the total eradication of this disease is something that has not yet been achieved and this is due to their unpredictable behavior.[56] From annual evaluations, it has been recorded that the organs where there is a greater tendency to suffer from cancer are lung, colon, breast and prostate, although any organ has the potential to suffer from this disease.[57]

The origin of the cancer is multifactorial, that is, it can occur by the influence of some external factor (*e.g.* ionizing radiation, toxic compounds, virus, etc.) or by genetic factors that disrupt the control points of the interphase of the cell cycle and lead to irregular proliferation of damaged cells. The cancer formation process is called oncogenesis or tumorigenesis. Cancer can be involved different types of organs, so the answer in each set of cells when the cancer is given, will be determined by the nature of specific metabolic pathways.[56, 58]

The cancerous cells can generate tumors that are called primary. They receive this name mainly because they remain in the organ of origin. However, in some cases, cells from the primary tumor migrate to new sites (metastasis), forming secondary tumors. Invasion of new tissue is not random, depending on the nature of cell metastasis and tissue that invade. Metastasis is facilitated if tumor cells produce

growth and angiogenesis. Generally, this complex process requires several discrete steps: (1) degradation of the basal lamina, (2) migration of endothelial cells, (3) division of these endothelial cells, and (4) formation of a new basement membrane, these processes occurs in the vicinity of a given capillary.[56, 58] Alterations in cell cycle regulation or migration of malignant cells have their origins at the molecular level where cluster of genes and proteins cooperate together to perform these mortal processes.

2.1.1 Protein of study: S100A4

In the 60s, it was isolated a subcellular fraction from bovine brain.[59] This fraction contains proteins that later were called S100, because were soluble in 100% saturated ammonium sulfate at neutral pH.[40] The S100 family is a group of proteins capable to bind calcium (Ca^{2+}),[60] constituted by 25 members.[61] In humans several proteins of this family have been identified, showing intracellular and extracellular behavior. Their functions are related to the growth and cell cycle regulation, cytoskeleton and assembly, cell differentiation, apoptosis, motility, etc.[40, 62, 63]

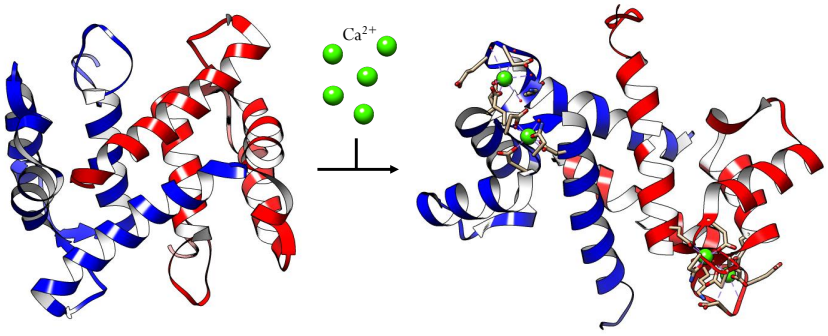


Figure 2.1: (a) Structural rearrangement of S100A4 in the presence of calcium. It is observed the A4-*apo* (PDB id 1M31) and the A4-*holo* (PDB id 3C1V) forms. Image generated by UCSF Chimera.

Within the group of S100, it is found the human protein S100A4[1] involved in the formation and spread of cancer.[41] The A4 protein has been identified in different metastatic cancers in different organs like breast,[64] pancreas,[65] prostate,[66] lung,[67] and gallbladder.[68] A4 structure is generally divided into two sites calcium (Ca^{2+}) through two structural domains EF-hand[2] called: (1) pseudo EF-hand close to NTR and (2) canonical EF-hand close to CTR.[70, 71] Due to conformational changes and

[1] Alternative names: mts1, metastasin, p9Ka, pEL98, CAPL, calvasculin, Fsp-1.[61]
[2] The EF-hand consists of two alpha helices linked by a short loop region.[69]

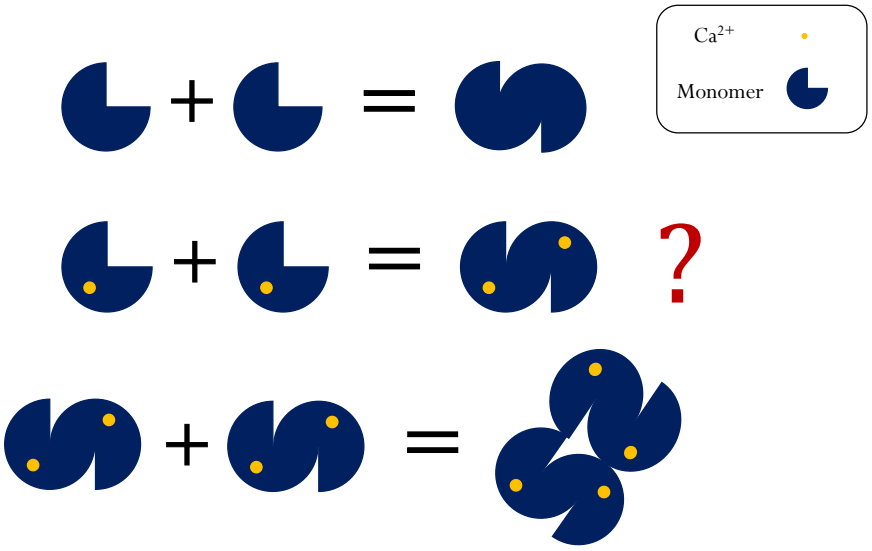


Figure 2.2: Illustration that addresses our biological problem. It is intended to understand how thermodynamic stability is given in the dimeric (*apo* and *holo*) and tetrameric (*holo*) forms.

electrostatic effects generated by the presence of Ca^{2+}, and the arrangement of hydrophobic residues on its surface in a region known as hinge (Figure 2.1),[42, 71] A4 can generate oligomers self-associate, therefore, it was found that exist a correlation between the oligomer formation and metastasis.[43] It has been established that in intracellular compartments the functional form of A4 is its dimer, which is able to interact with non-muscle myosin IIA by regulating the assembly of myosin filaments and resulting in increased cell migration.[72] The dimeric form is also involved in the conversion of plasminogen to plasmin, promoting angiogenesis.[73] However, the dimers also have the ability to self-associate by forming tetramers or even aggregates with a greater number of monomers (oligomers). The functionality of these A4 oligomers is rather extracellular. One of the hypotheses suggested that the promotion of metastasis is correlated with the increase of oligomers in plasma and synovial fluid.[74] Therefore, we propose that the understanding of the intermolecular forces that drives the complexation mechanism is the key to explain why some oligomeric forms are thermodynamically more stable than others under conditions of physiological pH and salt. We will focus on the understanding of A4 homodimerization in the absence and presence of calcium, and the complexation of these dimers to the tetrameric form in the presence of calcium too (Figure 2.2). A subsequent identification could probably serve in clinical treatments, because A4 has been proposed as a potential therapeutic target in metastasis treatment.[55]

It should be noted that other hetero-complexes can be formed between A4 and

other proteins (*e.g.* p53,[42, 62] annexin II,[42, 61] S100A1,[75] septine,[76] liprin β1,[77] among others[61, 62]). The role of A4 with each of these is functionally different and important in cellular metabolism.[61, 78]

Within the S100 family MC methods have been used in calbindin D_{9k}, where it was calculated the pK_a's of the titratable residues to determine the behavior of the structural changes in the protein and the effects of divalent ions.[79] Further, it was also evaluated by the Tandford-Kirkwood (TK) model in different salt concentrations.[27, 80] The protein A4 already has been studied computationally. A MD study was carried out in the hetero-association S100A4-myosin where the authors observed the influence of Ca^{2+} in the structure of A4, and its consequences in complex formation.[72] No theoretical study before has investigated the homo-association of A4 or any other oligomeric specie.

From the information available at the protein data bank (PDB),[81] the A4 complexes used here were solved by crystallography (at resolution of 1.50 Å) (*holo*, PDB id 3C1V) and nuclear magnetic resonance (*apo*, PDB 1M31) (Figure 2.1). The primary structure is relatively small (13 kDa) consisting of 113 residues. Following the CG phylosophy, we transformed coordinates from the atomistic structure to a mesoscopic model (see Sections 4.2 and 5.1). With this we intend to study the phenomenon of complex formation of A4. However, preliminary results indicated that a refinement of the force field used in our model was necessary, since in qualitative terms this phenomenon was not reliably reproduced.

2.2 Other studied proteins

As mentioned in Chapter 1, the lack of thermodynamical experimental data specific for the A4 protein turned necessary to study other protein systems. In fact, this was the part of the work where we needed to spend our efforts in the force field calibration process since an inappropriate set of parameters could result in an unreliable description of the proteins interactions. We used the LYS, CHY and RIB proteins for the calibration process, and the A3 protein as a test case. Surface maps showing the distribution of the titratable and hydrophobic residues on the surfaces of these four proteins are given in Figures 2.3 and 2.4, respectively. We highlight the residues that might have a stronger influence on the association process.

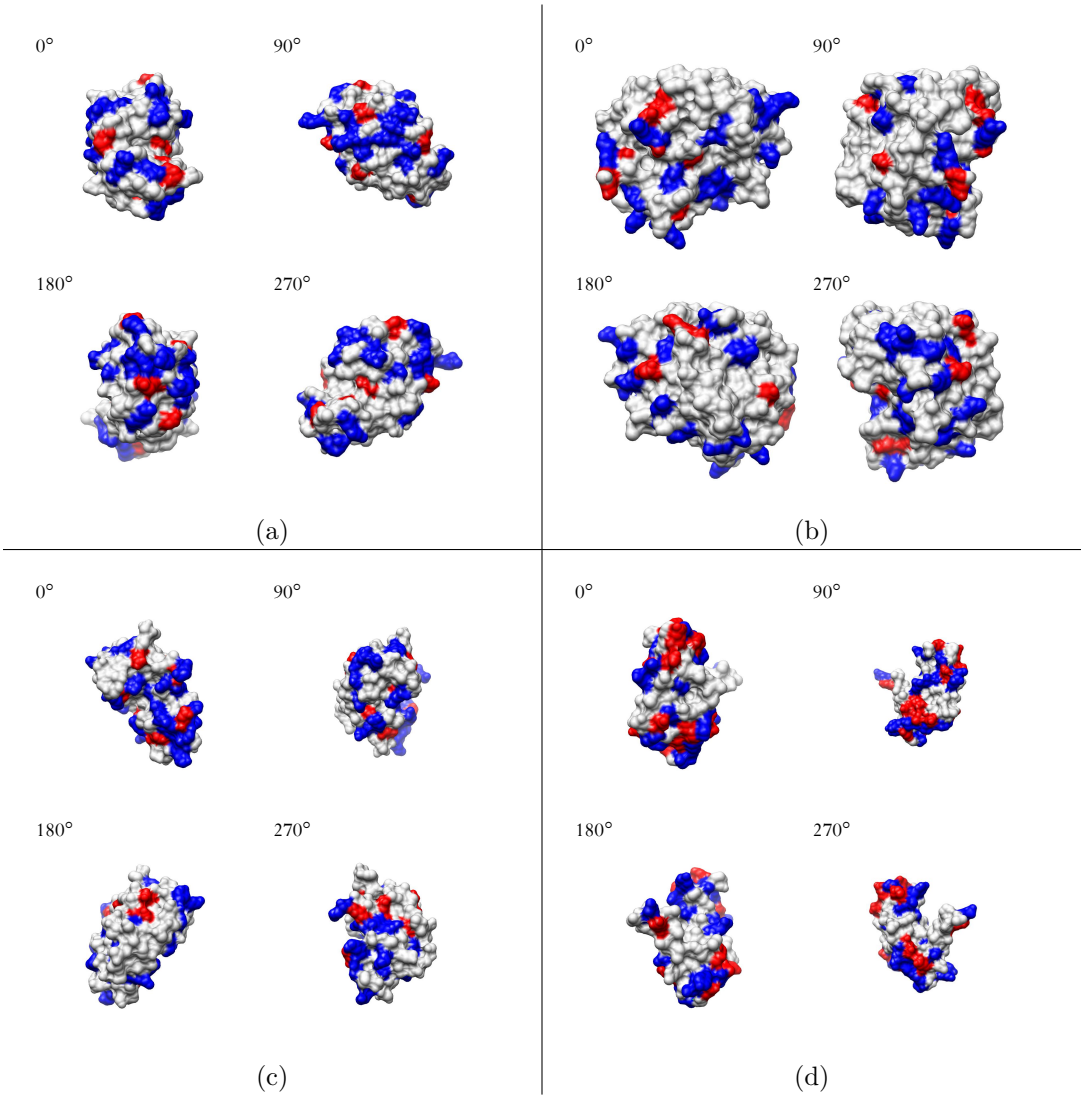


Figure 2.3: Spatial distribution of the positive (blue) and negatively (red) amino acids on the surface for (a) lysozyme (PDB id 2LZT), (b) α-chymotrypsinogen (PDB id 1CHG), (c) ribonuclease A (PDB id 1KF5), and (d) S100A3 (monomeric form) (PDB id 3NSO) viewed by rotation around the y-axis. Images generated by UCSF Chimera.

2.2.1 Protein Lysozyme

During years LYS has been used as a ideal system in studies of evolutionary biology, biochemistry and physical-chemistry. The principal function of this protein is the cleavage of glycosidic bond between two different sugar types: N-acetylmuramic acid and N-acetylglucosamine, important compounds of the bacterial peptidoglycan. This capacity in biological terms allows to a determined organism to have antibacterial defense.[15, 82] It is important to mention here that it has also been detected isopeptidase and chitinase functionality in these proteins.[83] LYS's

are present in various taxonomic groups such as animals, plants, and fungus.

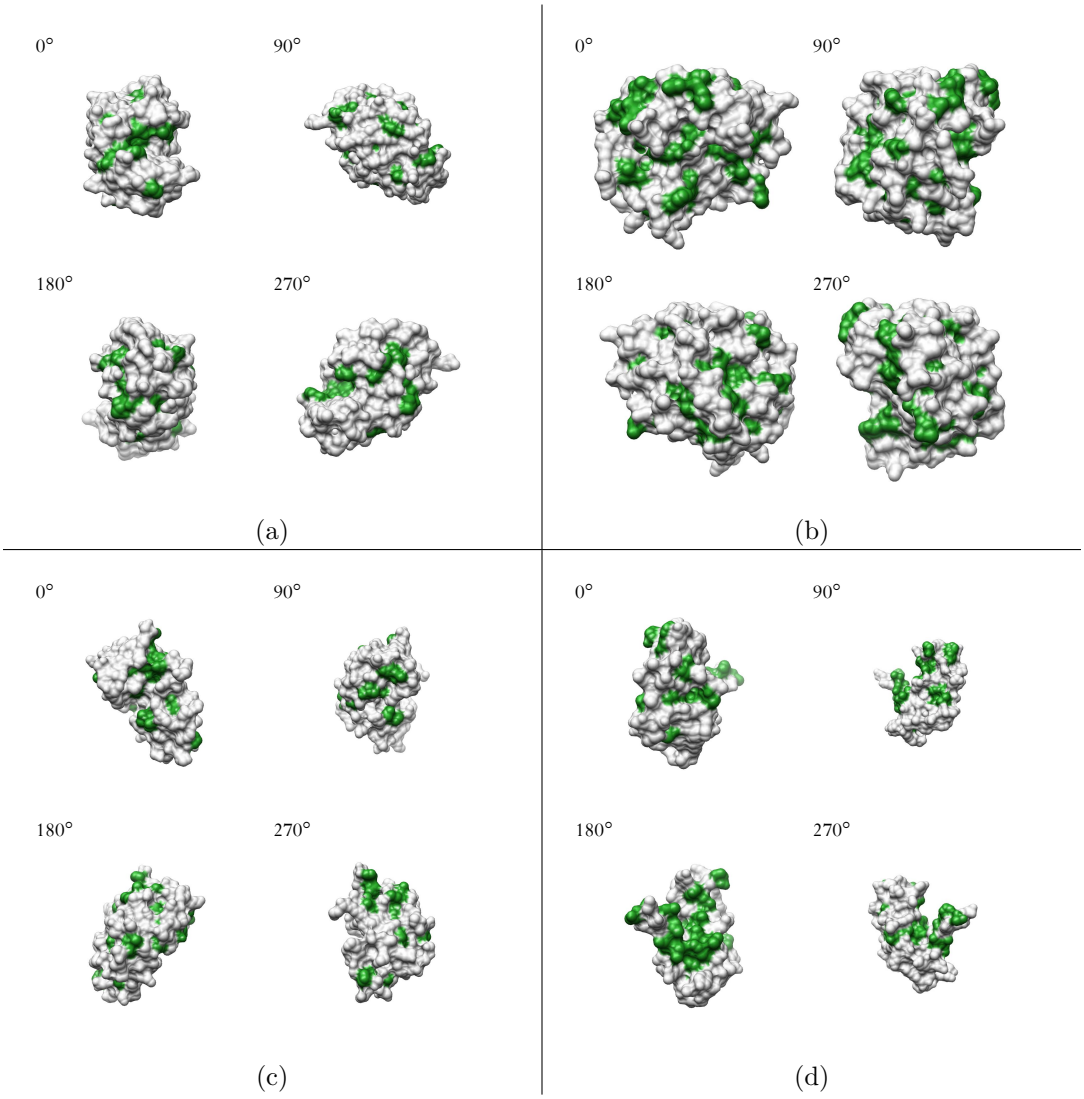


Figure 2.4: Spatial distribution of the hydrophobic (green) amino acids on the surface for (a) lysozyme (PDB id 2LZT), (b) α-chymotrypsinogen (PDB id 1CHG), (c) ribonuclease A (PDB id 1KF5), and (d) S100A3 (monomeric form) (PDB id 3NSO) viewed by rotation around the y-axis. Images generated by UCSF Chimera.

In general, these proteins have been classified in different categories, but the most estudied are c- and g- types. Normally, c-types contain between 129-130 residues, wich gives to the protein a molecular weight of $\sim$ 14 kDa. In contrast, g-types regularly presents 185 *aa* and its weight is $\sim$ 21 kDa. In the two types, Cys's form disulphide bridges allowing to maintain stable the secondary structure. A residue of Asp and other of Glu have a key role in catalitic funcionality of the enzyme. So we

clearly see that the protein in order to carry out its activity, requires specific and appropriate conditions in their environment (*e.g.* pH and low concentration of salt), to have the success in its functionality.[82, 83, 84]

In a more applied context, LYS has been a key piece in crystallization and precipitation experiments, that are important processes in industrial area. Formation of protein aggregates is strongly dependent of salt and pH concentration.[85] It has been recognized that in high concentrations of NaCl, screening favors the attraction between lysozymes. However, a strong repulsion between them can be observed in low concentrations of salt and low pH.[49, 86, 87] Electrostatic interactions are established as one of the force more relevant in this association.

The structure used in this work is the hen-egg-white lysozyme, protein determined by X-ray having a resolution of 1.97 Å (PDB id 2LZT) (Figures 2.3a and 2.4a). This protein is a LYS c-type, thereby, presents a molecular weight of 14.6 kDa and its primary structure has 129 residues.[88] Experimental and computational data under different conditions of pH and ionic strength are available in the literature,[28, 33, 49, 50, 89] showing that this is an ideal system for a force field analysis, through a thermodynamic criteria.

2.2.2 Protein α-Chymotrypsinogen

The second protein of our interest is the bovine alpha chymotrypsinogen. CHY is a pancreatic protein constituted by 245 residues. The activity of CHY occurs when it is produced the tryptic cleavage[3] in the first peptide bond formed between a residue arginine and other of isoleucine nearby to NTR. The protein remains inactivated to prevents damage to the pancreas. After the cleavage, a conformational change is produced, generating a macromolecule called chymotrypsin. In this form, the protein can be of different types such as α, γ, δ or π. Chymotrypsins function is precursor of zimogen, giving rise to other signaling cascades that allows activation of other enzymes with similar function.[90, 91]

Following ref. (Lund *et al.*[33]), we used the structure deposited in PDB was solved by X-ray to a resolution of 2.5 Å (PDB id 1CHG) (Figures 2.3b and 2.4b).[90] This structure is a globular protein with molecular weight of 25.7 kDa, consisting of 145 amino acids (coordinates are incomplete) and an isoelectric point at pH $\sim$ 9.5. In

[3] *i.e.* Catalitic reaction realized by a tripsin.

comparison with lysozyme this system have less uniformly distributed surface charge (Figures 2.3b).[51]

2.2.3 Protein Ribonuclease A

Bovine pancreatic ribonuclease A (RIB) is an endoribonuclease specialized in the catalyses and depolymerization of RNA. Generally, RIB carries out the cleavage of 3'-position of the phosphodiester bond by recognition of negatively charged phosphate groups.[92] Studies have shown the role of the pH dependence of this protein molecule regarding to its functionality, showing that its efficiency, stability, catalysis and binding capacity is strongly associated to thar physicochemical variable.[93] RIB has been an important model for the understanding of protein folding, formation of disulphide bonds, crystallography, and protein dynamics.[94] Furthermore, this protein has been a key point in cancer treatments, where it has shown toxic potential to malignant cells.[95]

RIB is a protein of a molecular weight 13.7 kDa and a sequence length of 124 amino acids. As well as on CHY, its surface generally presents a heterogeneous distribution of charged amino acids, where the ones that are in greater quantity are basic residues (Figure 2.3c).[54] For our analysis, we used the structure determined by X-ray having a resolution of 1.15 Å (PDB id 1KF5) (Figures 2.3c and 2.4c).

Finally, some important statements about the use of these calibration proteins are given in the Section 5.3.

2.2.4 Protein S100A3

Finally, the protein S100A3 also belongs to the family S100. The principal differences in comparison with A4 is that it has the ability to bind calcium and zinc ions. Furthermore, it is considered the protein with higher Cys content (10 out of 101 amino acids) between the other members of the family.[48] A3 is a protein with intracellular funcionality. So far it has not reported extracellular activity.[96] It is believed that A3 plays a important role in the cellular differentiation and several processes in follicular cells.[47, 97] In previous works it was indicated that A3 can help to counteract the damage in the hair by oxidation due to high Cys content.[98] A3 has irreversible posttranslational modifications, where Arg are citrullinated generating a alteration in its electrostatic properties. These posttranslational modifications have

been relationated with differents human diseases that include cutaneous disorders. A3 also has a close relationship with tumorigenesis in colorectal cancer, placing it as an important candidate in the metabolism of this cancer type.[99]

The complex we chose to study was determined by X-ray,[48] have a resolution of 1.45 Å (*apo*, PDB id 3NSO) (Figures 2.3d and 2.4d) and is composed of two subunits with a weight of 11.7 kDa each one. Some thermodynamics features like the second virial coefficient for the complexation of wild type and mutant of A3 were determinated by light scattering assays at salt and pH to physiological conditions.[47]

Chapter 3

Objectives

Having briefly considered some important aspects of computational approaches and the five different systems, our goals are detailed below.

3.1 General objective

1. To understand the fundamental interactions involved in the formation of molecular complexes, determining the main physicochemical properties responsible for it.

3.2 Specific objectives

1. To analyze and refine the adjustable Lennard-Jones parameter ε_{LJ} that allows a better description of the dispersive forces and hydrophobic effects based on second virial coefficients from experimental data of lysozyme, α-chymotrypsinogen, and ribonuclease A.
2. To test the refining on S100A3 proteins by comparison with experimental data.
3. To study the effect of the ionic strength, pH and calcium ions in the homo-association of S100A4 proteins.

Chapter 4

Theoretical Foundation

4.1 Intermolecular forces between biomolecules

When we observe the universe we interpret that the matter obeys certain laws and present particular properties that define its nature and behavior. We are able to distinguish various patterns in essence. Forces acting in nature and give those emergent properties to matter have been traditionally classified into four main groups: weak and strong interactions acting between neutrons, protons, electrons and other elementary particles, electromagnetic and gravitational interactions which act between atoms and molecules.[10, 11]

4.1.1 Electrostatic interactions

Net charges of proteins[1] as mentioned in Chapter 1 are expressed by the sum a set of residues *aa* that have the ability to donate or receive protons depending on the pH and other charged species present in the aqueous solution. Generally, the change in the conformation of the primary sequence of *aa* (protein folding) occurs as result of the action of them. Consequently, biology has led to the adoption of the so-called native structure, which is strongly related to biological function, and hence, to all the physiological processes of all living organisms.[7, 100, 101] In several cases, biological function is possible due to complexes of proteins. Depending on situation, interfaces protein-protein are produced by Coulomb interaction too, but is important to make a clarification, because when we consider this type of interaction, not only attractive forces appear, but also repulsive forces. Inherent to these macromolecules,

[1] Understanding that the source of the charge are provided by the protons and electrons of the atoms of wich are constituted.

the anisotropic distribution of charges of the ionizable groups in the structure can produce permanent electric dipoles, which will allow to originate attractive forces (*e.g.* between two proteins), given that exists a natural tendency to redirect these two bodies toward regions where charges are complementary to each other.[10, 102] It is worth mentioning that the dipole effect was initially described in small molecules, and we stressed that its origin is due to other physical properties in comparison with proteins (where the origin of dipoles is due to distribution of charged residues).

4.1.2 van der Waals forces

Following the perspective from molecules, when is considered permanent electric dipoles (*e.g.* in water molecules), the complementary charged sides between them causes an attractive force at a certain distance. This phenomenon is known as Keesom interaction. Otherwise, when we examine Debye forces, if one of molecules presents a permanent dipole and other is electrically neutral, the first could to induce a momentary dipole on the other, promoting a attractive force between them. However, when it is considered London or dispersion forces, macromolecules electrically neutral at certain distance can too generate momentary dipoles by fluctuations of their electron densities. Up to this point, we can see that the net force of vdW, turns out to be the sum of these three types of interaction (Keesom, Debye, and London) and by way of conclusion: it is always an attractive force.[10, 102] Finally, if we will consider two molecules that were allowed to stay close enough, it would be expected that they completely it overlapped, but that is not the case. The "breaking point" where they are unable to superimpose is due to the electrostatic repulsion or steric forces, that prevent both bodies to occupy the same place at the same time. This is because the Pauli exclusion principle is obeyed between the atoms electrosphere.[10] The vdW force also occurs in proteins, but this effect is effectively described by functions called as interaction potential energy. The most commonly used in studies of biopolymers is the Lennard-Jones (LJ) potential.[8]

4.1.3 Charge regulation mechanism

The dynamics of a protein should not be seen only in a geometric-structural sense. The distribution of *aa* charged in these biomolecules is comparable to a constellation of residues through which protons can flow.[22] Know as the charge regulation mechanism, it has been found that in function of the pH, proton-ionizable groups

are strongly affected and this dynamic flux will result in attractive forces in pH near to pI at low salt concentrations. This phenomenon occurs at a mesoscopic level and it is enormously dependent of the *charge regulation parameter*, better known as "protein capacitance" (C), intrinsic property of any protein. The variable C depends on the number of ionizable groups, so if we assume a high quantity of these *aa*, we see that it will exist a increase of this when it is close to the pK_a of these acidic and basic residues. With the equation $\left[C \propto \frac{dZ}{dpH}\right]$, where Z is the charge of the protein at a given pH and salt solution, it is possible to experimentally obtain the parameter of regulation C.[5]

As proposed by Kirkwood and Shumaker,[22] considering two macromolecules A and B, with valences $\langle Z_A \rangle$ and $\langle Z_B \rangle$, respectively, separated by the distance r; we can calculate the free energy between them at infinite salt dilution regime[2] through the following equation,

$$\frac{A\,(r)}{k_B T} \approx \frac{l_B \langle Z_A \rangle \langle Z_B \rangle}{r} - \frac{l_B^2}{2r^2}\left(C_A C_B + C_A \langle Z_B \rangle^2 + C_B \langle Z_A \rangle^2\right), \tag{4.1}$$

where the first term corresponds to the Coulomb interaction, and the second term is the charge regulation interaction. Note that this letter term signal indicates that the mechanism of charge regulation is always attractive.[5, 25] The charge regulation parameters C_A and C_B can also be computed as $\left[\langle Z^2 \rangle - \langle Z \rangle^2\right]$. The *Boltzmann*'s constant and the temperature are, respectively, k_B (1.3807×10^{-23} J mol^{-1}K^{-1}) and T in *Kelvin*. The l_B is the *Bjerrum* length $[e^2/4\pi\epsilon_0\epsilon k_B T]$, where e is the elementar charge (1.602×10^{-19} C), ϵ_0 is the dielectric constant of the vacuum ($\epsilon_0 = 8.854 \times 10^{-12}$ C^2/N m^2) and ϵ is the dielectric constant of the aqueous solution to a specific temperature.

4.1.4 Hydrophobic effect

By observing the medium in which proteins are found, we see that exist an amount of species that interact with them and these different particles are frequently suspended in a solvent called water. In multiple cases the protein-solvent interaction comprises the affinity to make hydrogen bonds between water molecules and, for example, the surface of a particular macromolecule. However, it has been found that there is also a "inverse" effect. Some *aa* will not have preference for making these bonds (thanks to

[2] This is a limited case where electrostatic interaction do not suffer any screening from the salt particles.

its apolar nature), instead they will exert an apparent repulsion with the surrounding liquid. It has been estimated that water tends to form around a non-polar molecule, a structure that looks like a cage and in the literature has received the name of "solvation shell". This structure is formed by a rearrangement of water molecules on the apolar surface. From the perspective of nonpolar molecule, its form and structure are critical in the disposition and configuration of solvation shell. Molecules of water are "forced" to re-ordered in a new structure, wich will modify the entropy, and as consequence decreasing in several units of k_BT the free energy of the system, leading to attractive forces between hydrophobic molecules. Correlating this effect with the vdW radius (*i.e.* distance where vdW interaction begins to occur), would be expected that while the length of this gains more dimension, it will increase the hydrophobic effect, since the surface area becomes larger too. Finally, hydrophobic proteins in aqueous medium logically will create aggregates, showing that this type of effect has an important role in the formation of proteins complexes.[3, 10]

Generally, when we talk of complexation phenomena, forces mentioned above appear in a wide range of different proteins. Interestingly, it is noteworthy that despite the existence of this discrepancy, the physical mechanisms in protein-protein interactions are produced in the same way in all proteins.[103]

4.2 Modeling

When the molecular interactions in a given system are evaluated, we can necessarily make a close relationship between the properties it presents and the laws of physics by which it is dominated. As a large number of variables exists, it is possible to reduce and focus them to specific features that allow us to define how detailed should to have our system, to thereby provide a description of the total energy that describes a phenomenon of study established. For our purposes, it is considered a system mainly composed of three elements: solvent (water), solute (protein), and salt (monovalent). Systems and their interactions can be described energetically through of mathematical expressions. The equations that allows us to do it are known as *effective Hamiltonians*,[8, 104] which can be resolved favorably by computational methods.

4.2.1 Models of proteins in electrolyte solution

For the three elements, there are different approximations and levels of detail. The choice of model definitely will depend on the objectives and answers that are desired to be obtained. It is clear that every idealization highlights a particular set of physical characteristics, but adjacently should be evaluated the availability of computational resources and the desired statistical precision. Systems like biomolecules in an electrolyte solution having a large number of atoms, the simulation time can be prohibitive due to a high number of degrees of freedom. So we emphasize that there is a direct relationship between the number of details versus computational cost (Figure 4.1). Next we will brief describe each element.

4.2.1.1 Solvent

To represent the solvent it has been suggested two approaches: explicit and implicit. The first considers various molecular details of water and even is possible to have a quantum description of the same. On statistical mechanics, it has known as *Schrödinger* level (SL), which involves the evaluation of nuclei and electrons of the atoms. Nevertheless, in order to reduce the computational cost, the microscopic level or *Born-Oppenhimer* level (BOL) also can be used. In this level the input variables are the coordinates and momenta of atoms of the solvent effectively approximated (Figure 4.1).[8, 104]

In the second aproach the solvent is reduced to mathematical expressions that capture some characteristics, representing the water molecules by a relative dielectric permitivitty ϵ. In this level we note a reduction in the computational cost due to a number of atoms drastically lower. Other names for this model are *McMillan-Mayer* level (MML) or macroscopic model level (Figure 4.1).[8, 104, 105] In this work we considered the implicit approach or structureless dielectric continuum for our purposes (see Section 5.1).

4.2.1.2 Solute

The protein can be observed having account a large amount of chemical information (*i.e.* most structural details of *aa* are considered). This aproach is called all-

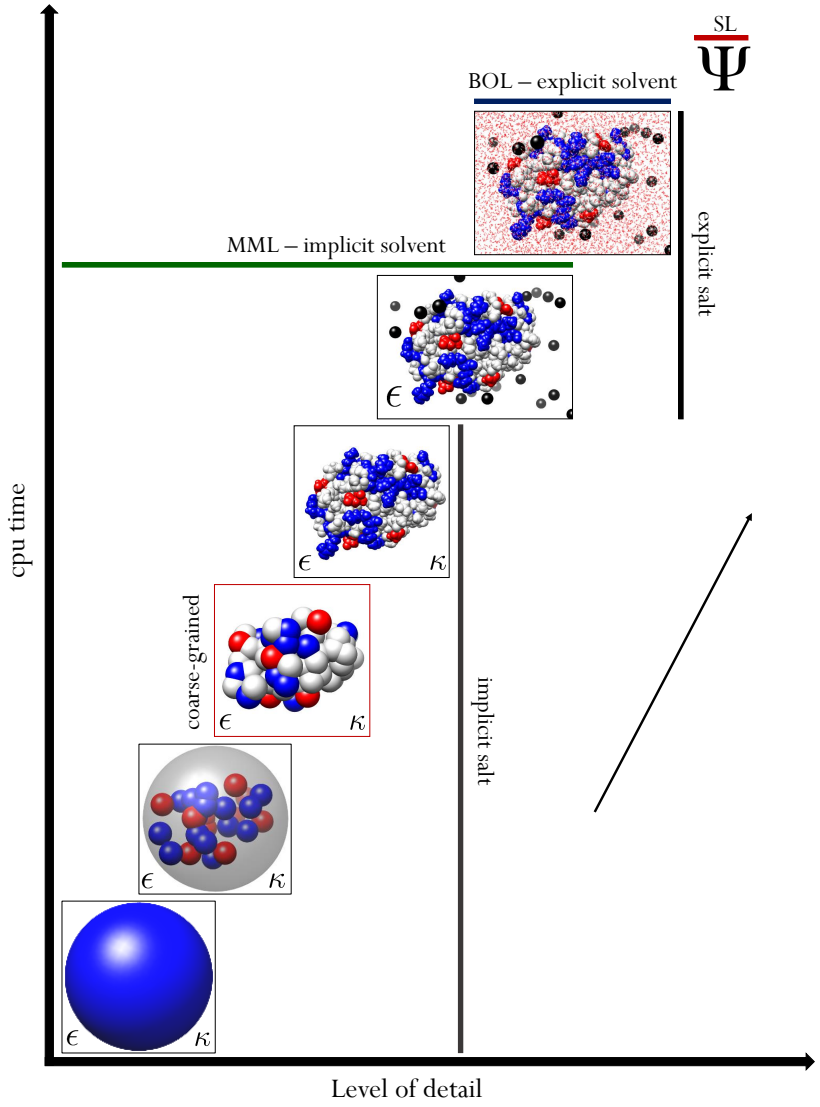


Figure 4.1: Relation between different level of detail and computational cost. The computational cost increases n^2, where n is the number of sites (*e.g.* through DLVO and TK approximations the computational cost increase 1^2 and 39^2, respectively). Positive and negatively charged amino acids are represented in blue and red, respectively, neutrals or hydrophobic amino acids are represented in white, and ions of salt are represented in black. The structure of protein used is a lysozyme (PDB id 2LZT).

atoms (Figure 4.1) and generally is accompanied of classical *forces fields*[3], that are, mathematical expresions that allow to calculate the interatomic potential energy of a system in thermal equilibrium.[29, 106] Popular examples are: Gromos,[107] Charmm,[108] OPLS,[109] Amber,[110] etc. All these popular biomolecules force fields are developed using empirical data as well, albeit frequently augmented with some

[3]Not to be confused with force field in classical physics, but rather should be understood in the context of molecular modeling (chemistry).

ab initio electronic structure simulations (see for instance, the work of Huang and MacKerell[111]).

In some cases, globular proteins in a given pH are symmetrical sufficiently, favoring to reduce the whole protein structure to a charged sphere (Figure 4.1). In a similar sense, when we use TK model, the effective radius of the protein is fixed, but is retains the constellation of ionizable residues in their relative positions. This is done in order to assess their electrostatic contributions.[27, 112] Between these two approaches it can be established any level of detail mesoscopic or coarse-grained modeling (Figure 4.1). The philosophy of CG is to simplify the entire structure in substructures reproducing the mean physico-chemical features, for example, of a functional group, an amino acid, or a region of the protein. Generally this simplification comprises the transformation of some specific part to a single sphere.

4.2.1.3 Salt

In biological systems, salt (or ionic strength) is a relevant component because its properties allow, for example, to produce action potentials in neurons by a des- and repolarization of their membranes.[56] However, other important quite feature is its capacity for screening the charged surface of a biomolecule in solution. Depending of case, screening can lead to attractive (when the system is dominated by vdW forces) or repulsive (when interactions charge-charge are more relevant) forces in proteins. In colloids science this behavior was successfully described by the electric double layer theory or DLVO (Derjaguin, Landau, Verwey and Overbeek).[102, 113] Although in principle it was applied to homogeneous surfaces, later was achieved to do approximations in biomolecules using numerical solutions of PBE by computer.[53] Like in the treatment of the solvent, ionic strength can be represented explicit or implicitly (Figure 4.1). When a salt explicit-model is used, cations and anions can interact with the biomolecule or the medium. Derived from the Debye-Hückel analysis,[114] an implicit scheme of the mobile ions can be used and given by the replacement of the inverse Debye screening length κ, which is given by formula,

$$\frac{1}{\kappa} = \sqrt{\frac{8\pi e^2}{(\epsilon_0 \epsilon k_B T) \sum_{ions} c_k z_k^2}} \tag{4.2}$$

where c_k and z_k are, respectively, the number density and charge of the mobile electrolyte specie k (counter-ions and added salt). The variable κ is interpreted as length where the influence of the charge of macroion (biomolecule) stop acting on

other charged objects. The Debye screening length of a electrolyte solution with high concentration of salt (*e.g.* 100 mM) will be shorter in comparison with regimes where salt concentration is lower (*e.g.* 5 mM).[113] In this work we used the implicit scheme (see Section 5.1). A more detailed model would be computationally prohibitive due to vast volume of necessary calculations for different system and experimental conditions.

4.2.2 Solution of the model

There are two approaches to the solution of models according to their level of detail: (a) theoretical/analytical method and (b) "computer experiment". Systems described at the quantum level (SL) are resolved via solution of the Schrödinger's equation.[115] But in systems with a large number of atoms (*e.g.* a lysozyme c-type which has between 129-130 *aa*) the solution is technically impracticable. Taking advantage of other approaches with a number lower of details (*e.g.* all-atoms or CG, with explicit or implicit salt), computational methods as MD or MC can be satisfactorily used. MD is implemented, for example, in BOL level solving numerically Newton's equations.[115, 116] However, with MC methods is possible to construct the phase space via a stochastic implementation.[117] MC methods are more convenient at model level of MML or implicit solvent.[8, 20] Therefore, due to practical reasons, in the present work we will focus in these.

4.3 Monte Carlo methods

Chemical systems are described from thermodynamics and statistical mechanics through of statistical ensembles, that are, collections of possible microstates of a thermodynamic system with a fixed set of variables (*e.g.* temperature, number of particles, volume, etc.).[114] Calculating the positions and momenta of all particles, we can define the coordinates of the so-called "phase space". With it is intended to assess the total energy of the system. The description of phase space of a chemical system can be carried out via deterministic (MD) or estocastic (MC) methods. In the MD methods, the motion of particles (*e.g.* atoms of a specific protein) is simulated according to Newton's equations.[115, 116] In contrast, MC methods are a class of stochastic computational calculations, where only the initial positions of the particles of the system are necessary. Subsequently, a perturbation

is applied (randomization), which generates a new configuration that "ignore" the intrinsic physical movement in the system (Figure 4.2). Each new configuration is dependent on the previous (Markov chain) and a criterion of choice is applied to be accepted.[8, 115, 118, 119]

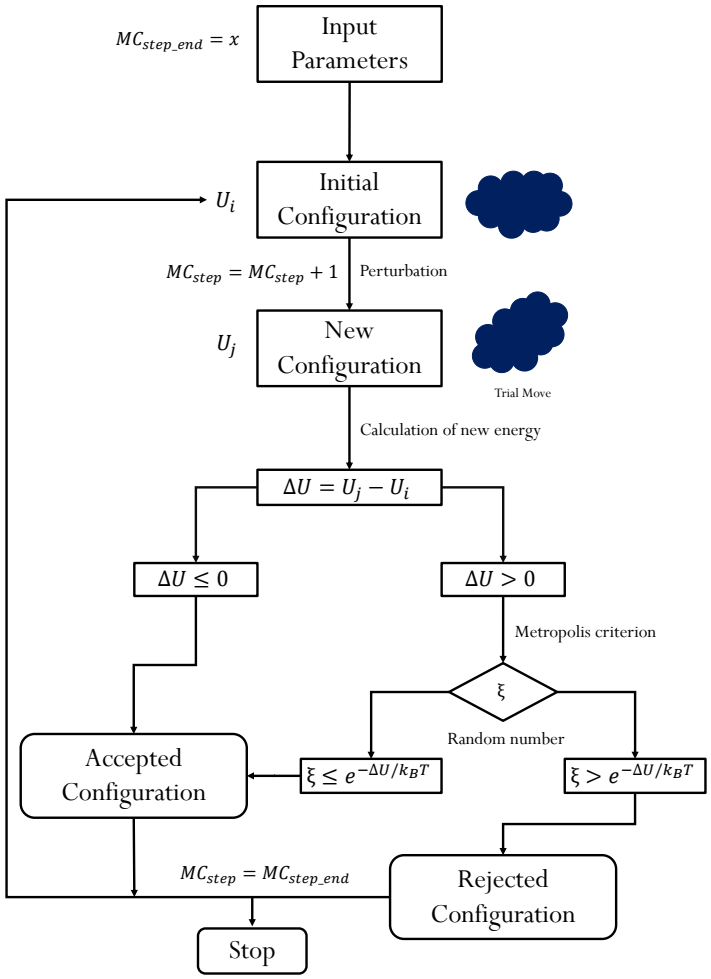


Figure 4.2: Common steps in the Monte Carlo algorithm. From input parameters the system is equilibrated. During production phases it is applied a criterion of choice to accept the new configurations. In the Metropolis criterion a random number ξ homogeneously distributed in the interval (0-1) is created.

Every time the system adopts a particular configuration and the total energy is evaluated. Under the Metropolis criterion it is understood that if the energy decreases or is equal to the initial or last configuration, the new state is accepted. However, if the energy increases, the state will be accepted with a probability proportional to $[\exp(\Delta U = k_B T)]$, where U is internal energy of the system (Figure 4.2).[8, 119, 120] MC methods have been used in complex calculations of the behavior

of biomolecules and it has proven to be an effective method in the physical and chemical description of these systems.[5, 8, 20, 44]

4.4 Principal statistical mechanics properties calculated

4.4.1 Radial distribution function

In molecular or atomistic study it is essential to understanding the structure of physical systems. The use of various experimental techniques and theoretical analysis, has been developed with the aim of elucidating one of the fundamental issues at these levels of structural organization; here we highlight the pair-correlation function or also called, radial distribution function (rdf). Several thermodynamic descriptions can be made with the rdf, allowing define the partition function, ideal in the statistical mechanics and thermodynamics studies.The rdf is a measure of the probability of finding a particle at a distance r from a reference particle, which can be obtained through MC simulations.[8, 114, 117] The rdf in proteins can be described by the equation,

$$g(r) = \frac{\rho_{ij}(r)}{f(r)}, \tag{4.3}$$

where ρ_{ij} is the number of contacts between residues i and j found in the distance r, and $f(r)$ is the number of contacts between all the residues found in the distance r. With $g(r)$ we are technically calculating the frequency of contact between two *aa* i and j according to the separation distance r.[9]

4.4.2 Potential of mean force

Densities calculated with $g(r)$ are found closely related to the potential of mean force (pmf) $w(r)$, so we can use the equation 4.4 to transform $w(r)$, between two residues in the distance r from $g(r)$.[9]

$$w(r) = -k_B T \ln g(r), \tag{4.4}$$

The pmf describes all forces of interaction between two protein molecules in electrolyte solution, therefore, it can be seen that in some sense $w(r)$, takes into

account the interaction of macromolecules and its neighborhood (*i.e.* salt ions, water molecules, etc.). The pmf accounts for short ranged and long ranged interactions. Furthermore, it is related with the work done by force $[F = -dw(r)/dr]$ to a distance r; thus, in terms more thermodynamics, $w(r)$ is *free energy* or available energy that a system can have.[9, 10]

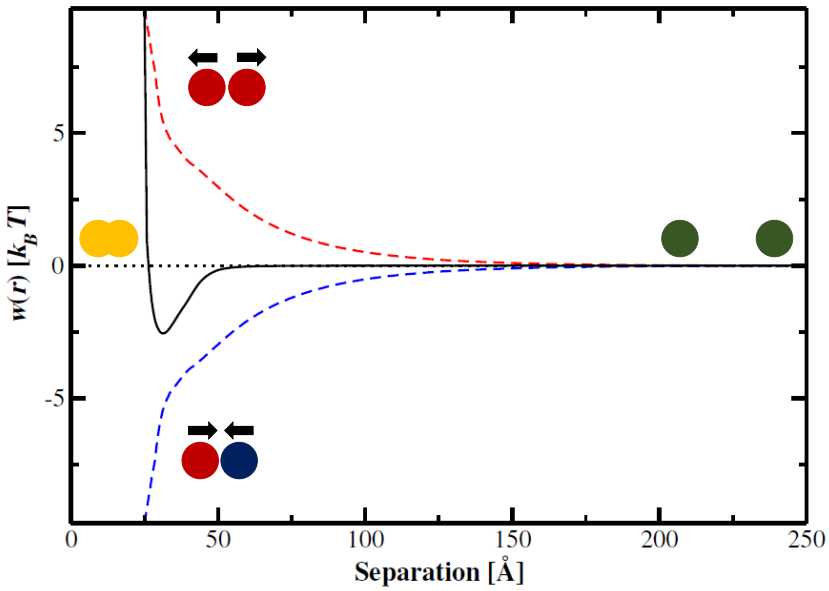


Figure 4.3: Example of typical plot of potential of mean force. In x-axis appears the separation distance in Å and y-axis the free energy of interaction in thermal energy units. Blue and red lines represent cases of attractive and repulsive interactions, respectively. Black solid line represents the sum of the two contributions: attractive and repulsive, and a hard-sphere contribution. Plot also represents the evaluation of free energy in "critical" points. For more details read the text.

Classically pmf plot obtained by MC methods is presented in Figure 4.3. Evaluating a system of two particles (*e.g.* two proteins), we can study the free energy of interaction. Based on the black line of Figure 4.3, when the distance separation is short (yellow spheres), appear steric repulsion forces and the free energy is positive (not spontaneous process). However, repulsive forces may also appear in close proximity (red line and red spheres above in Figure 4.3). Otherwise, when free energy of interactions becomes low (spontaneous process), an attractive force can act on the particles due to close proximity (blue line in Figure 4.3). In this distance a affinity interaction could be occurring, for example, between a protein with positive charge and other with negative charge at a given pH condition (blue and red spheres below in Figure 4.3). Finally, if the two particles are sufficiently distant (green spheres in Figure 4.3), the interactive force is negligible and the interaction free energy can be

regarded as zero[4] (process is in equilibrium).[9]

4.4.3 Second virial coefficient B_2

The protein-solvent system is controlled by intermolecular interactions of different nature that are implicitly described by $w(r)$, in such a way, we can (assuming a low protein concentration) express the interaction using the virial expansion for an imperfect gas. The second virial coefficient B_2 is a thermodynamic feature important in the study of complexation by allowing a numerical quantification of the interaction. This entity is obtained through various laboratory techniques, such as, light scattering [*e.g.* static light scattering (SLS), small-angle neutron scattering (SANS),[49] or small-angle x-ray scattering (SAXS)[47]], chromatography [*e.g.* self-interaction chromatography (SIC),[121] or size exclusion chromatography (SEC)[122]], and osmotic pressure [*e.g.* membrane osmometry (MO)[89]], and has proved useful theoretical analysis of the crystallization processes.[85, 123, 124, 125] Computationally and taking into account our goals, B_2 can be expressed by the next equation,

$$B_2 = B_{q+vdw} + B_{hs}, \tag{4.5}$$

where B_{q+vdw} (given in mol ml/g^2) represents the electrostatic and vdW contributions and B_{hs} (given in mol ml/g^2) represents the hard-sphere contribution.[33] The following equations show, respectively, how B_{q+vdw} and B_{hs} are computed:

$$B_{q+vdw} = -\frac{2\pi N_A}{M^2}\int_{\sigma}^{\infty}\left[\exp\left(-w\left(r\right)/kT\right)-1\right]r^2dr, \tag{4.6}$$

$$B_{hs} = \frac{N_A}{3M^2}\pi\sigma^3, \tag{4.7}$$

where M is the molecular weight (in g/mol) of proteins, σ is the sum of radii (in Å) of the two proteins and N_A is Avogadro number (6.02×10^{23} mol^{-1}). Equation 4.6 is solved through numerical integration method such as Simpson rule (see Appendix B.1), and the next (Equation 4.7), can be solved analytically. If positive, B_2 indicates that repulsive interactions dominate and if negative the net-effect is an attraction.[85]

[4]The system behaves as an ideal gas, where there is no interaction between particles and the volume is excluded.[9]

Chapter 5

Methodology

In the present work, we study the complexation phenomena of proteins through of the evaluation of configurations in a statistical thermodynamic NVT ensemble by Monte Carlo simulations.[126] Calculations were performed with a personalized version of the Faunus biomolecular simulation package.[127] The model used here is based in previous works of the group and its collaborators.[26, 33, 34, 35, 44, 46, 128] Systems used for simulation were described in Chapter 2. Structures solved by x-ray crystallography of all protein were full, except CHY, A3 (*apo*), and A4 (*holo*) which was completed by mean of homology method using the program Modeller (version 9.17).[129] Some important aspects of this process can be found in Appendix A. Principal codes and scripts used in this study are listed in Appendix B.

5.1 Model cell

There are several approaches available to study protein and protein-protein interactions in electrolyte solution (see Section 4.2). In order to perform a systematic analysis on large scale, we decided to adopt models that do not use a high computational cost per system and experimental condition, due to the need to repeat such calculations in different situations. Thus, we consider the approach of implicit solvent (or MML), where water molecules are replaced by its static dielectric constant (*i.e.* $\epsilon = 78.7$) at 298 K.[104] In the same way, counter-ions and added salt particles were represented implicitly using the screening term $[\exp(-\kappa r_{ij})]$ (see Equation 5.1), where κ is the inverse Debye length, and r_{ij} is the interparticle separation distance.

Proteins were studied here in mesoscopic scale, through of the so-called *coarse graining model*.[33, 34, 36] From a specific protein each *aa* is condensed to a charged LJ sphere of radii (R_{ai}) and valence z_{ai}. The values of the radius

Table 5.1: Radius values and molecular weight of each amino acid.

Residue[a]	Letter code	R_{ai} (Å)	Mw (g/mol)	pK_a	CG[b]	Polarity
Ala	A	3.1	66	-.-		nonpolar
Arg	R	4.0	144	12.0		basic polar
Asn	N	3.6	108	-.-		polar
Asp	D	3.6	110	4.0		acidic polar
Cys[c]	C	3.5	98	10.8		nonpolar
Gln	Q	3.8	120	-.-		polar
Glu	E	3.8	122	4.4		acidic polar
Gly	G	2.9	54	-.-		nonpolar
His	H	3.9	130	6.3		basic polar
Ile	I	3.6	102	-.-		nonpolar
Leu	L	3.6	102	-.-		nonpolar
Lys	K	3.7	116	10.4		basic polar
Met	M	3.8	122	-.-		nonpolar
Phe	F	3.9	138	-.-		nonpolar
Pro	P	3.4	90	-.-		nonpolar
Ser	S	3.3	82	-.-		polar
Thr	T	3.5	94	-.-		polar
Trp	W	4.3	176	-.-		nonpolar
Tyr	Y	4.1	154	9.6		polar
Val	V	3.4	90	-.-		nonpolar
CTR	-.-	2.0	45	2.6		-.-
NTR	-.-	2.0	16	7.6		-.-

[a]Ala, Asn, Gln, Gly, Ile, Leu, Met, Phe, Pro, Ser, Thr, Trp, and Val correspond to alanine, asparagine, isoleucine, leucine, glutamine, glycine, methione, phenylalanine, proline, serine, threonine, tryptophan, and valine respectively (other *aa* were named above in Chapter 1). [b]Positive and negatively charged, neutrals and hydrophobic amino acids are represented in blue, red, white and green respectively. [c]Not involved in SS bonds.

and molecular weight used in this study for each *aa* are presented in Table 5.1. As our interest are measures of the free energy involved in protein complexes in function of the separation distance, two macromolecules are placed in an electroneutral open cylindrical cell of radius r_{cyl} and height l_{cyl}, arranged in a line (c_{axis}) that connects by their geometric centers. We emphasize that proteins are rigid bodies without internal degrees of freedom. During the simulation they can translate in c_{axis} and rotate randomly in any direction. Each new configuration has its acceptance evaluated by the criterion of Metropolis (see Section 4.3).[120] A scheme of the model cell is given in Figure 5.1.

As will be mentioned later, the number of simulation with different parameters and experimental conditions is so high that such simplified coarse-grained model with intermediate level of details is quite appropriate in terms of efficiency and the statistical accuracy to be achieved. Other more detailed approaches such as all-atoms would be impracticable because the number of scenarios to explore is so huge and the computational cost would increase several orders of temporal magnitude (see Section 4.2 as a guide for understanding this reasoning). A particular example about the computational cost is shown in Subsection 6.4.1. On the other hand, a colloidal-like approach based on the DLVO theory would be too simple, preventing us to explore and detect all the main rich physical-chemistry contributions of each *aa*.

5.2 Coarse-Grained Force field

Considering the contribution of long ranged interactions, our model can be described by a screened Coulombic potential energy $\left[u^{el}\left(r_{ij}\right)\right]$. The $u^{el}\left(r_{ij}\right)$ between two *aa* is given by following equation,

$$u^{el}\left(r_{ij}\right) = \frac{l_B z_i z_j}{r_{ij}} \exp\left(-\kappa r_{i,j}\right), \tag{5.1}$$

where z_i and z_j is the valency of *aa* i and j.[10] Simultaneously, residues with capacity to protonate (Table 5.1[1]) on a medium with defined pH are evaluated through a fast titration scheme developed specifically for this purpose.[26]

[1]The pK_a's used here are from "isolated" amino acids obtained experimentally to 25°C, following ref. (Nozaki *et al.*[130]).

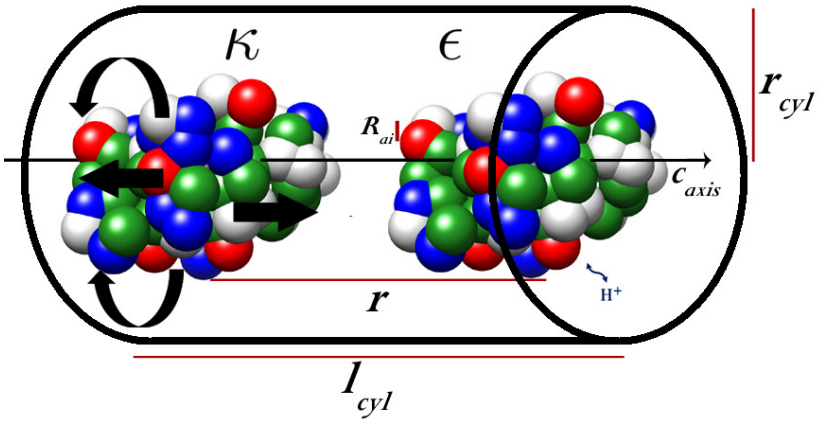


Figure 5.1: Schematic representation of model cell used in this study. Two proteins build up by a collection of charged LJ spheres of radii (R_{ai}) and valence z_{ai} mimicking aminoacids. The "box" simulation is an electroneutral open cylindrical cell of radius r_{cyl} and height l_{cyl}. The solvent is represented by its static dielectric constant ϵ. Counter-ions and added salt particles are represented by Debye's term κ. Positive and negatively charged, neutrals and hydrophobic amino acids are represented in blue, red, white and green, respectively.

Contribution of short ranged interactions are calculated with the LJ potential energy. This one capture repulsive (due to the Pauli-exclusion principle) and attractive contributions. The attraction described by this term is due to an average of all vdW interactions and an effective nonspecific contribution of the hydrophobic effect (see Section 4.1). For any two aminoacids i and j, LJ potential energy is given by,

$$u^{vdw}\left(r_{ij}\right) = 4\varepsilon_{LJ}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right], \tag{5.2}$$

where ε_{LJ} is a parameter *adjustable* and it is related with the depth of the potential well, $\sigma_{ij} = (R_{ai} + R_{aj})$ is the separation distance of amino acids i and j at contact.[10] Note that bigger aa will experience stronger attractions as expected. We assume that the density of each aa is equal to 0.9 g/ml, based on ref. (Persson *et al.*[34]). From of this value and respecting the weight molecular of each residue, the authors estimated the proper R_{ai} for each kind of molecule (Table 5.1).

Finally, the total system interaction energy for a given configuration $[U(\{r_k\})]$ is,

$$U\left(\{r_k\}\right) = \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\left(u^{el}\left(r_{ij}\right) + u^{vdw}\left(r_{ij}\right)\right), \tag{5.3}$$

where $\{r_k\}$ are amino acid positions, $N = N_A + N_B$ is the total number of residues,

N_A is the number of *aa* of protein A and N_B is the number of *aa* of protein B.

5.3 Molecular simulation

By means of Metropolis MC simulations we described several features of our model following four steps:

1. Physico-chemical characterization of each protein individually.
2. Analysis of B_2's reported in the literature by different experimental techniques.
3. Determination of dimerization process using the CG force field reported in the literature.
4. Refinement of the force field CG model from LYS, CHY, and RIB systems.
5. Test from new calibrated in proteins belonging to the S100 family.

Regarding the first step, titration curves were determined, where pH related properties were measured [the protein charge number (Z_P), the charge regulation parameter (C_P) and the dipole moment number (μ_P)], through of 10^8 cycles of MC. This allows us to have a first estimative on how the system individually behaves. Simulations were carried out at different salt concentrations: 1, 5, 20, 50, 100, 150 and 250 mM (except S100 proteins where only calculations were performed at 150 mM assuming that these proteins are in physiological environment).

For the second step of this study, we analyse the experimental works of Velev *et al.*, (SLS and SANS);[49] Rosenbaum *et al.*, (SLS);[131] Tessier *et al.*, (SIC);[121] Pjura *et al.*, (MO);[132] and Bajaj *et al.*, (SLS);[133] for LYS and CHY[2]. It were compared values of B_2 among the techniques and it was used as benchmark the work of Velev *et al.*, (SLS).[49] For the comparisons we performed calculations of ΔB_2. With this procedure, we discriminate the reliability of the experimental data for subsequent analyses. We did not consider to RIB protein in this part of the study due to little information of B_2 for this system.

For the third step, initially, we assessed the length l_{cyl} more appropriate for each system of study. For that, the displacement parameters for translation and rotation were modified with the purpose to adjust the acceptance ratio to ca. 30-50%. Further, it also was observed the mean displacement given by equation

[2]Acronyms between parentheses refer to the technique used by each author (See Section 4.4.3)

$\left[\sqrt{\langle dr_{AB}^2 \rangle}\,/\mathring{\text{A}}\right]$, where dr_{AB} is the displacement distance from the protein A respect to the protein B; in order to have values greater than a fifth of the total length of the cylinder. Once this part is defined, simulation runs to evaluate the complexation on systems LYS, CHY, and RIB were performed. We highlight that it was used in this step a ε_{LJ} equal to 0.05005 k_BT (= 0.124 kJ/mol), following ref. (Persson *et al.*[34])[3]. At each run of production. 10^8 configurations were explored, preceded by an equilibration phase with 10^6 cycles. During the production phase, rdf's were measured (see Subsection 4.4.1). They were calculated based on histograms recorded during the simulation.

The rdf's were converted into pmf by mechanical-statistical relation given by equation 4.4)[4]. From the pmf we estimated the free energy of protein-protein interaction (see Subsection 4.4.2). As our interest is on average properties depending on the distance of separation (we are not investigating the "docking" proteins, but rather the forces acting on their complexation), internal structural details can be neglected. Similar approaches have been used in other studies, showing reliable results.[33, 34, 35, 44] Additional thermodynamic properties such as the second virial coefficient B_2 were calculated (see Subsection 4.4.3). Although its partially nonspecific and semi-quantitative character, such property has proved useful and valid in several studies of molecular aggregation including protein complexation.[89, 134, 135] Since the emphasis of present work is on the association process itself where we need to sample as a function of the separation distance and not only at the binding site ("docking"), this treatment for the protein as a colloidal particle is quite appropriate.

Preliminary results allowed to determine that it was necessary a refinement of the vdW contribution in the CG force field (see Section 6.3). For the fourth step, we systematically varied ε_{LJ} (see Equation 5.2), seeking the best value that reproduces the experimental data of B_2 as given by ref. (Velev *et al.*[49]), for a set of calibration proteins (LYS and CHY). We explored different values of ε_{LJ} in a range from 0.05-0.1 k_BT (in steps of 0.0005 k_BT)[5]. However, some conditions of salt and pH required to explore other values outside this range. Concurrently, it was calculated the RMSD between the computed and experimental B_2 (given in mol ml $\times 10^4/\text{g}^2$). That allow us to have a quantified measurement of the deviation of the data (see Equation 5.4).

[3] This should correspond to a Hamaker constant of ca. 9 k_BT for amino acids pairs.[33]

[4] Codes used for this process appear in Appendices B.2 and B.3.

[5] Code used for this process appear in Appendix B.4.

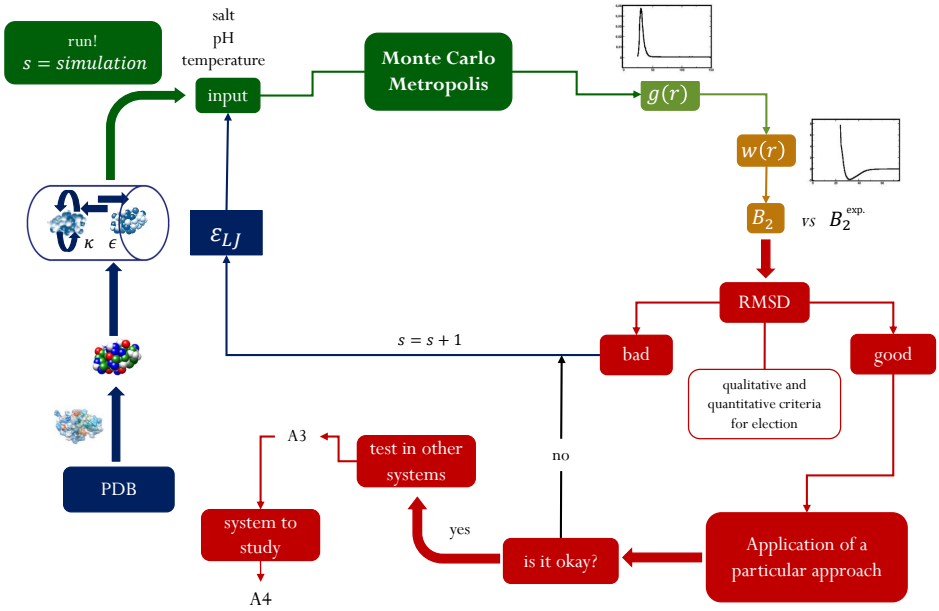


Figure 5.2: Different steps performed in the methodology proposed in this project.

$$\mathrm{RMSD}\,(a,b) = \frac{\sum \left| B_2^{\mathrm{MC}} - B_2^{\mathrm{Exp.}} \right|^2}{n}, \tag{5.4}$$

where B_2^{MC} is the value determined by MC simulations, $B_2^{\mathrm{Exp.}}$ is the data from ref. Velev *et al.*[49], and n is the total number of data used in the comparison. As an example, using the system LYS with ε_{LJ} equal to 0.08 k_BT at pH and salt concentrations of 10.5 and 100 mM, respectively, the RMSD $\approx$ 0.81 [data derived from $B_2^{\mathrm{MC}} \approx -6.41$, $B_2^{\mathrm{Exp.}} = -5.60$, and $n = 1$][6]. The selection of the best values took into account a quantitative criterion (which is intended to be RMSD values close to zero) and also a qualitative one (where it must be respected the attractive or repulsive characters of the system according to the conditions present of the electrolyte solution). Equilibration and production phases had 10^6 and 10^7 MC steps, respectively. Consequently, in order to confirm the best values in each condition of pH and salt, additional 10^8 MC cycles were performed. A flow diagram is shown in Figure 5.2, where the methodological process of calibration is represented. Following this series of logical steps, we used three strategies: (a) approach of unique value (auv), (b) approach of unique value at high salt concentration (auv$^{\mathrm{h}}$), and (c) approach of different value (adv); some important details are mentioned in Section

[6]To better understand of this computation see the code in Appendix B.5.

6.4 respect to these approaches.

Finally, regarding the fifth step, taking into account the experimental data for A3-*apo* monomers and its two types: wild (WT) and mutant (R51A),[47] we perform simulations to estimate the free energy of interaction as a function of the separation distance using $\varepsilon_{LJ} = 0.05005\ k_BT$ and $\varepsilon_{LJ} = 0.073\ k_BT$ at pH 7.5 and salt concentration of 150 mM[7]. A B_2 was then calculated and compared to the experimental B_2 from ref. (Kizawa *et al.*[47]) by using the RMSD function. Implementing the same physical-chemistry condition and the same ε_{LJ}'s used on A3, we calculate pmf's for A4 and its different forms ($A4^{m}$-*apo*, $A4^{m}$-*holo*, and $A4^{d}$-*holo*)[8].

[7]It should be mentioned that before performing this step, the acceptance calibration process was also done for these systems.

Chapter 6

Results and Discussion

6.1 Physical chemistry properties of proteins

Before we start the study of protein-protein interactions, it is necessary to understand the electrostatic behavior of each individual protein in the electrolyte solution. Different forces can act on the complexation phenomena (*e.g.* Coulombic interactions, charge-dipole, dipole-dipole, vdW forces, hydrophobic effects, etc.). Therefore, a "simplified" view we allows a better comprehension of how and why this process occur. For this part of the study, we first focus on LYS, CHY, and RIB systems (calibration proteins). Subsequently, descriptions of the main physical-chemistry properties were made for S100 proteins (*i.e.* A3 and A4).

6.1.1 Proteins used for the calibration process

We perform titrations at different pH (0-14) and salt concentrations (Figures 6.1a, 6.2a, and 6.3a). We saw that variation of Z_P goes from 16.5 to -9 for LYS, from 19.0 to -16 for CHY, and from 16.5 to -13.7 for RIB, approximately. Note that these numbers are non-dimensional. They are expressed in units of the elementar charge (*i.e.* valence). For all protein systems, we recover the typical experimental behavior with positively charged proteins at the acid regime. In all proteins we note that while the ionic strength is greater the charge increases too at extremes of pH. Mathematically, this behavior is produced because, when the Debye screening length κ becomes larger in the equation 5.1, the term tends to zero. This allows the charge of each macroion to increase recovering its ideal behavior (*i.e.* when each charged *aa* behaves as a single entity ignoring the influence of its neighborhood).[5, 26] It has been shown that this kind of model gives results that are in agreement with

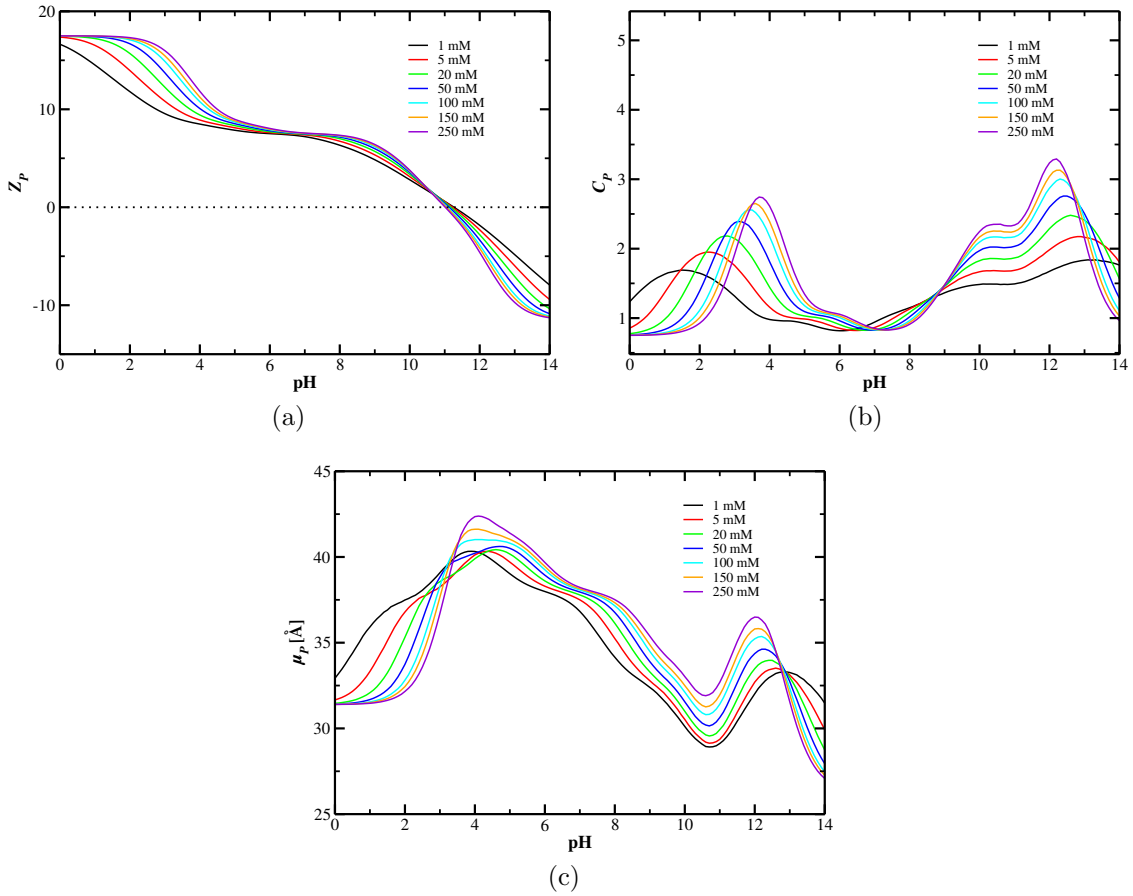


Figure 6.1: (a) Net protein charge number Z_P, (b) charge regulation parameter C_P, and (c) dipole moment number μ_P as a function of pH for lysozyme (PDB id 2LZT). Data from MC simulations at different salt concentrations.

experimental data.[136, 137]

Among the salt concentrations the average computed isoelectric points (pI) are 11.1, 9.5 and 9.5 for respectively, LYS (Figure 6.1a), CHY (Figure 6.2a), and RIB (Figure 6.3a). The pI is interesting since this indicates the functionality of the protein.[20] The experimental values are 11.0,[49], $\sim$ 9.5,[51], and 9.45,[138] for respectively, LYS, CHY, and RIB very close to our calculated pI. We conclude that the three proteins are cationic. In all MC simulations at different salt concentrations the pI changes slightly because in this model the anisotropic protein-salt interactions are neglected in the truncated spherical Kirkwood approximation.[26] However, far from the pI, the protein charges higher. Therefore, it can be deduced that repulsive forces are the most remarkable in these pH regimes for self-association proteins. For instance, if we assume two proteins LYS-LYS interacting, they have the same net charge, and probably, it will generate strong Coulombic repulsive forces.

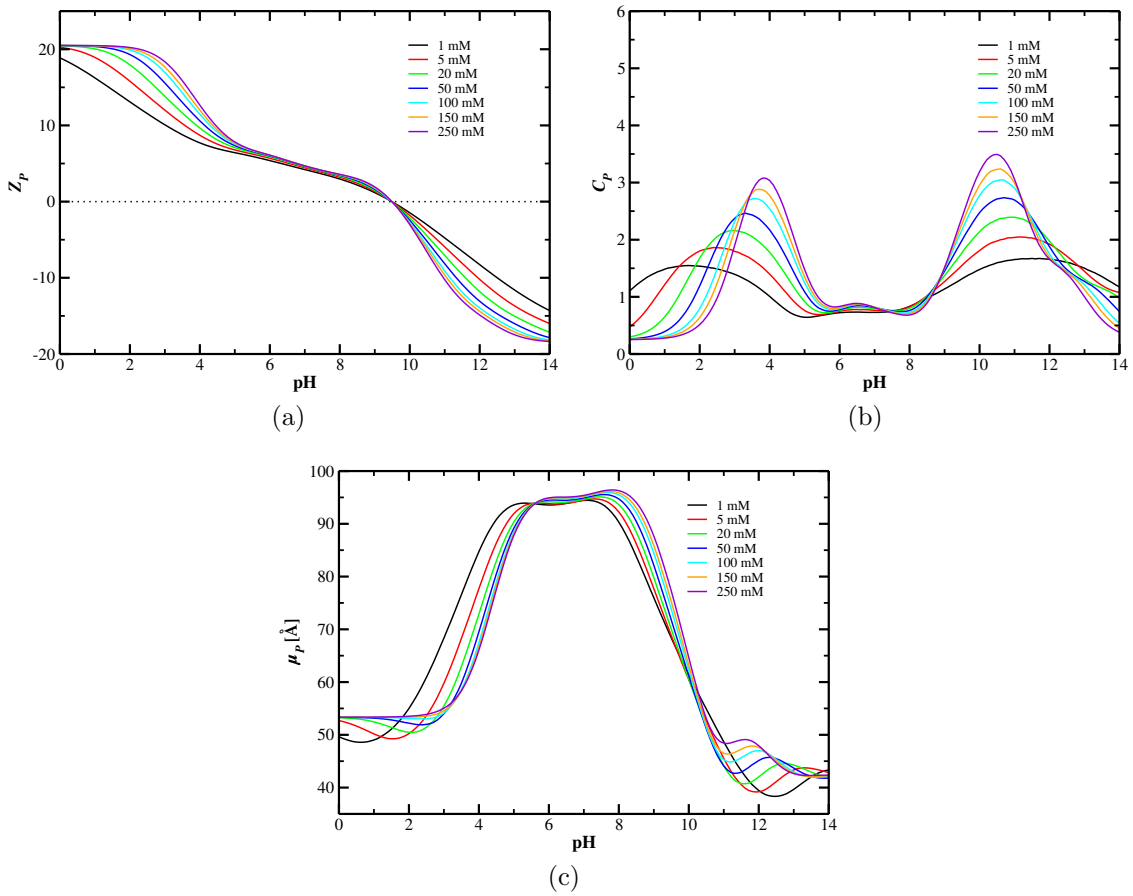


Figure 6.2: (a) Net protein charge number Z_P, (b) charge regulation parameter C_P, and (c) dipole moment number μ_P as a function of pH for α-chymotrypsinogen (PDB id 1CHG). Data from MC simulations at different salt concentrations.

Recent theoretical works have demonstrated that the charge regulation mechanism proposed by Kirkwood and Shumaker,[22] may play an important role in complexation, particularly for proteins with high values of C_P.[5, 22, 23, 24, 25] Calculated charge regulation parameter are presented in Figures 6.1b, 6.2b and 6.3b for LYS, CHY, and RIB, respectively. Note that the variable C_P does not have unity. This is because is a derivative of the valence regarding to the pH and these variables are dimensionless. Charge fluctuations are generally higher when the pH is close to pK_a's of ionizable groups (Table 5.1).

For LYS, the excess of Lys ($pK_a = 12.0$) and Arg ($pK_a = 10.4$) causes C_P to closer to their peak pK_a's. Similarly, in acidic conditions the presence of Asp ($pK_a = 4.0$) and Glu ($pK_a = 4.4$), generates another peak around pH 4.0 (Figure 6.1b). The system CHY contains 14 Lys, which implies an increase of its C_P at pH $\approx$ 10.5 (Figure 6.2b). The highest peak around of 10.5 in the RIB curve (Figure 6.3b), may be due

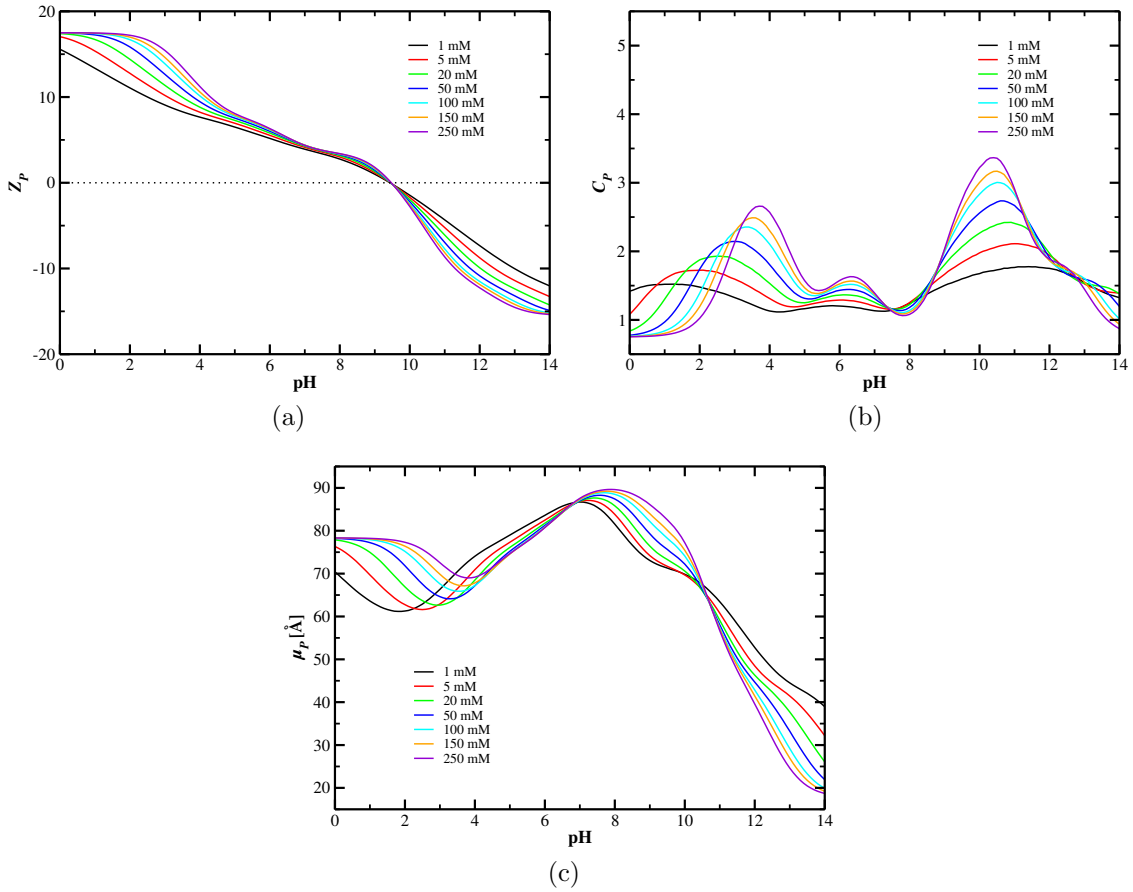


Figure 6.3: (a) Net protein charge number Z_P, (b) charge regulation parameter C_P, and (c) dipole moment number μ_P as a function of pH for ribonuclease A (PDB id 1KF5). Data from MC simulations at different salt concentrations.

to the presence of Lys and Tyr ($pK_a = 9.6$) in its structure. As C_P is the derivative of Z_P as a function of pH (see Subsection 4.1.3), significant variations on the charge directly will affect the charge regulation parameter. For instance, for LYS at the pH window from 5.2 to 7.7 at 5 mM of NaCl, Z_P does not vary much. Therefore, we see that C_P has a value around of 0.9 along that pH window, indicating a lower value in comparison with high peak ($C_P \approx 2.1$) when the variation of Z_P is higher (*i.e.* pH $\approx$ 12.7), at the same salt concentration. As previously discussed, increasing or decreasing the salt concentration has an effect on the behavior of Z_P due the electric field shielding. C_P responds in a similar manner when changing the ionic strength. On these proteins, we can expect more stable complexes in basic conditions due to charge regulation mechanism at low salt regimes.

Considering the Coulombic regime, the dipole moment number $\left[\langle\mu\rangle_0 = \sum_{i=1}^{N_P} z_i \vec{r}_i\right]$, where N_P is the number of residues of a specific protein and $\vec{r}_i$ is amino acid i position, is a important physico-chemical parameter for the molecular complexation

mechanisms.[46, 128] Dipole moment number is treated like reduced units (*i.e.* this is as function of the elementar charge), thus being represented by length units (Å). Dipole moment numbers as a function of the pH are shown in Figures 6.1c, 6.2c, 6.3c for LYS, CHY, and RIB, respectively. We can see that these three proteins do not present a very large μ_P compared to other systems such as lactoferrin (LF) ($\mu_{P\,(pI)} = 212$) or $\beta-$lactoglobulin ($\mu_{P\,(pI)} = 135$) at 20 mM salt concentration.[44] Nevertheless, among the three, CHY has a high dipole moment at a pH window from 5.0 to 9.0, proving to be a good candidate to form complexes due this physico-chemical condition. The second best candidate is RIB, since it has high values in pH's near the pI. LYS has an increase of its μ_P at pH below its pI. Likely, the μ_P on LYS is not a relevant contribution to the the formation of self aggregates. LYS is well-know to have a weaker μ_P, and the charge regulation mechanism is the most important contribution for its interaction with likely charged objects.[28] Finally, for each protein, μ_P shows a particular behavior as a function of ionic strength.

6.1.2 Proteins belonging to the S100 family

We studied the variables Z_P, C_P, and μ_P as a function of pH for two S100 proteins: A3 and A4 (Figures 6.4 and 6.5). We highlight that several explanations in respect to these calculations were given in the Subsection 6.1.1. Therefore, we only intend to emphasize in some specific cases. Due to the physiological environment in which the A3 and A4 proteins are (likely) found, calculations were performed only at a salt concentration of 150 mM.

We studied the wild (WT) and mutant (R51A) A3 monomers in its forms *apo*. In the Figure 6.4a, it is observed that the variation of the charge number along the pH scale ranges from 11.5 to -28.2, as an average between WT and R51A. The charge number is very similar among the two types. The isoelectric points are 4.5 and 4.4 for WT and R51A, respectively. Our model manages to reproduce experimental pI of WT, that is, 4.5.[139] We conclude that the general behavior of the two protein types is anionic. As expected the replacement of one basic *aa* by another neutral (*i.e.* Arg to Ala), causes the displacement of the curve toward more negative values. We will expect that this small shift to work as a greater advantage in the process of dimerization to WT at physiological pH (7.5), since in some sense the protein has its charge near to 0 in that condition (unlike to R51A) (Table 6.1). This allows in fact that vdW forces to act.

Regarding A4, we only consider its monomers (*apo* and *holo*) and dimer (*holo*) form.

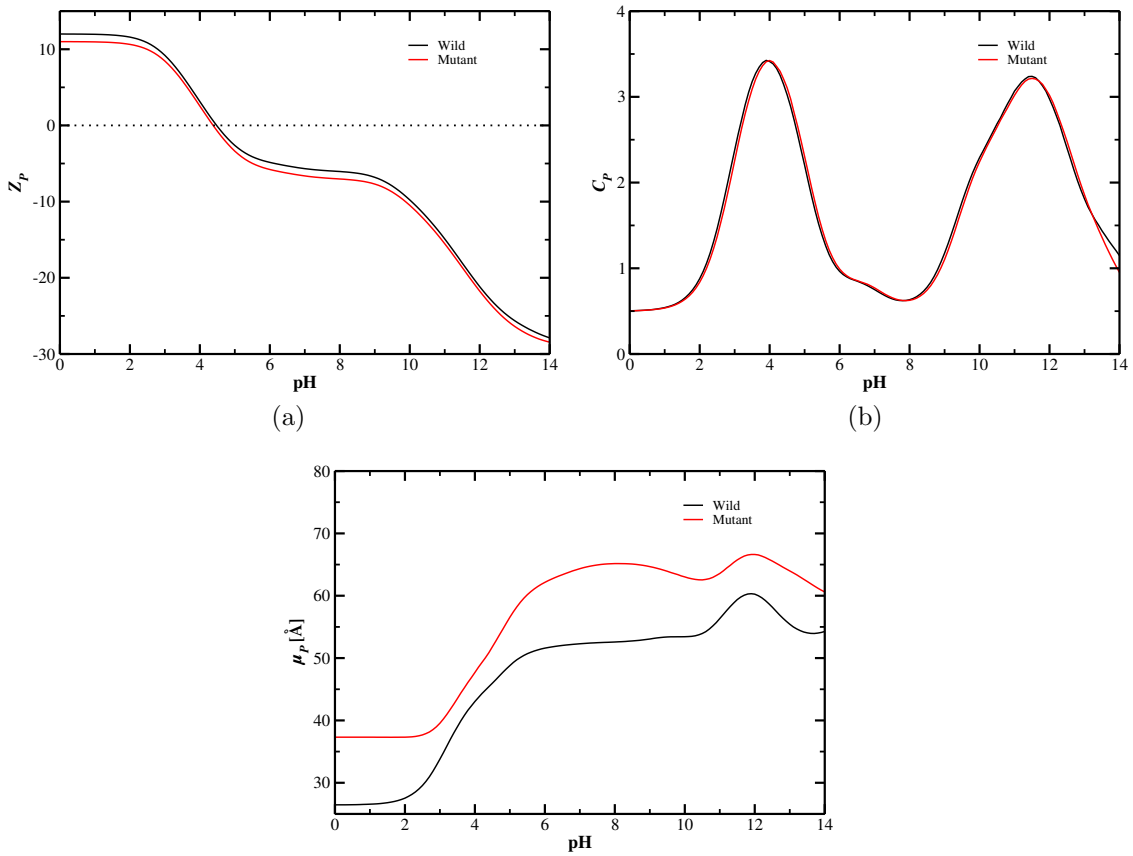


Figure 6.4: (a) Net protein charge number Z_P, (b) charge regulation parameter C_P, and (c) dipole moment number μ_P as a function of pH for S100A3 (PDB id 3NSO) wild and mutant (R51A) types. Data from MC simulations at 150 mM salt concentration.

The variation of charge number along of pH for A4$^{\mathrm{m}}$-*apo*, A4$^{\mathrm{m}}$-*holo*, and A4$^{\mathrm{d}}$-*holo* were from 16.5 to -22.3, from 20.5 to -18.3, and from 40.9 to -36.3, respectively. As shown in Figure 6.5a, the presence of Ca^{2+} in the monomeric form allows the protein to have a more basic behavior. This means that ions move the curve to more positive values for the action of their valences. Probably, the conformational change in the structure contributes to that "movement" too.

On the other hand, an increase of the charge is observed in the dimer of A4 at pH extremes, this is due to a greater amount of charged residues respect in monomeric forms that contribute significantly in this increase. The computed pI of A4$^{\mathrm{m}}$-*apo*, A4$^{\mathrm{m}}$-*holo*, and A4$^{\mathrm{d}}$-*holo* were 5.6, 9.7, and 9.8, respectively. The experimental pI following ref. (Haugen *et al.*[140]) is 5.85 for A4$^{\mathrm{m}}$-*apo*, that is relatively consistent with our results. Table 6.1 shown the values of Z_P computed for A4 at pH and salt concentrations of 7.5 and 150 mM, respectively. In all cases, the charge is

not large enough to cause considerable repulsion, particularly at the screening salt conditions of 150 mM. There is a probability for complex formation in any form in this electrolytic condition.

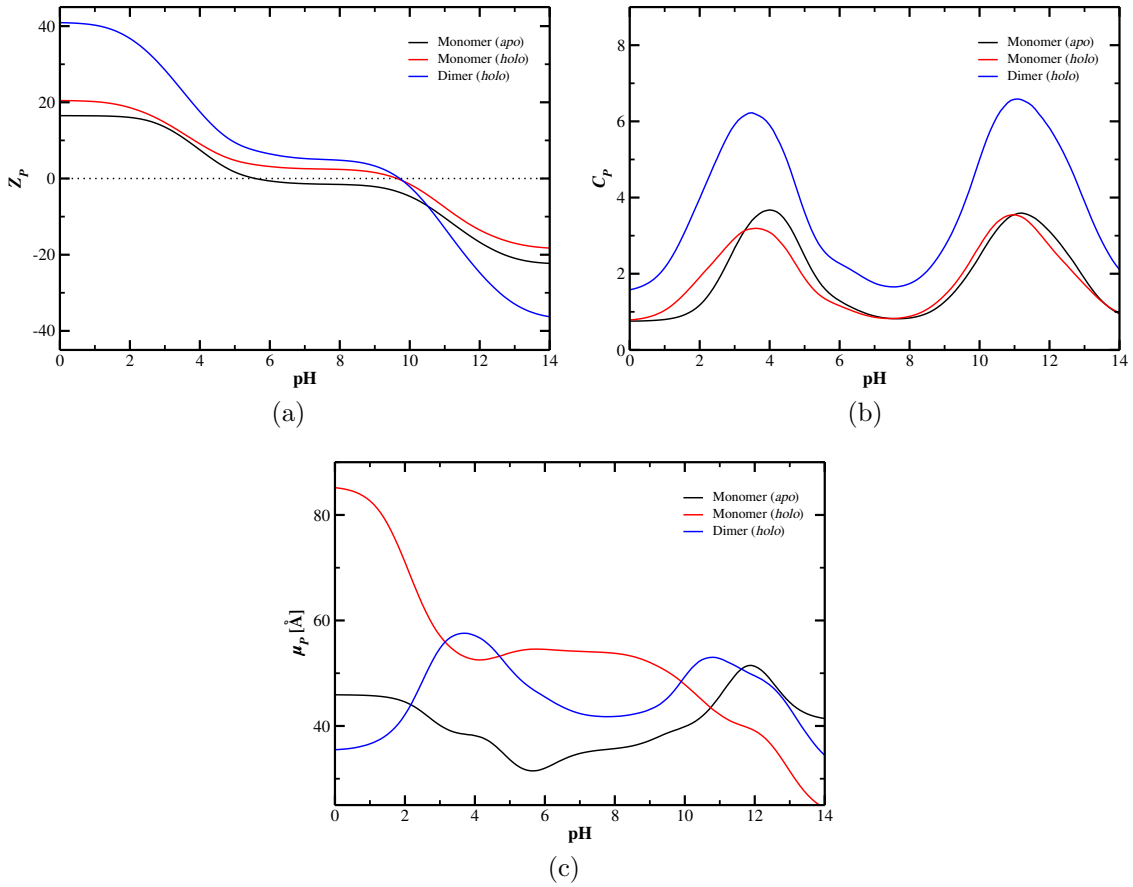


Figure 6.5: (a) Net protein charge number Z_P, (b) charge regulation parameter C_P, and (c) dipole moment number μ_P as a function of pH for A4 (monomer and dimer) *holo* form (PDB id 3C1V) and *apo* form (PDB id 1M31). Data from MC simulations at 150 mM salt concentration.

Table 6.1: Main physical chemistry properties of A3 and A4 proteins at pH and salt concentration of 7.5 and 150 mM, respectively, obtained from MC simulations.

Protein	Feature	Z_P	C_P	μ_P
A3	WT	−5.9	0.6	52.5
	R51A	−6.9	0.7	64.9
A4	[m] *apo*	−1.5	0.8	35.4
	[m] *holo*	2.5	0.8	54.0
	[d] *holo*	5.0	1.7	41.8

[a]Protein net charge. [b]Charge regulation parameter. [c]Dipole moment number.

We highlight that several explanations respect to calculations of charge regulation parameter as a function of pH were given in the Subsection 6.1.1. Therefore, we

will focus only on the values of C_P given at pH 7.5 and 150 mM of NaCl for A3 and A4[1]. For the two types of A3, the charge regulation parameter calculated are shown in the Table 6.1. The decimal variation between the two types obeys to the difference of the Z_P given each other at this point of the pH scale. Regarding A4, we saw that the presence or absence of Ca^{2+} does not affect the C_P in that electrolytic condition for the two monomers (Table 6.1). However, for the dimer A4 is seen that a greater amount of charged residues allows the parameter to increase a little in comparison to $A4^{m}$-*apo*, and $A4^{m}$-*holo*. Finally, we emphasize that this feature in these S100 proteins does probably not play an important role in complex formation at physiological conditions, due to the severe screening.

The dipole moment number as a function of pH for the two types of A3 are shown in the Figure 6.4c. For both types, the dipole becomes larger when the pH concentration is higher towards basic regimes. It reminds us of one kind of sinusoidal form in the curve. We notice a great difference of dipoles between the two types. The biggest difference between them is due to a single change of *aa* (Arg is replaced by another of Ala), which generates a more heterogeneous distribution of charged residues on the surface. The dipole moment number is increased on R51A with respect in WT; even at physiological conditions (Table 6.1). Regarding A4, each oligomer or form responds in a particular manner as a function of a specific feature. Focusing on the monomers, we noted that there is an increases of the dipole moment number for $A4^{m}$-*holo* compared to $A4^{m}$-*apo* in most pH regimes (Figure 6.5c). At low pH the saturation of protons makes to $A4^{m}$-*holo* positively charged producing that increase its μ_P by anisotropy of positive residues and additional Ca^{2+} ions. The increase of dipole moment is neglected for A4 at extremely acidic or basic regimes, because under these conditions the protein functionality is probably equal to zero in this system. Finally, at physiological conditions, the $A4^{d}$-*holo* has a considerable reduction of its μ_P compared to $A4^{m}$-*holo*. The dipole seems an important feature in the dimerization processes of A3 and A4.

6.2 Comparison of published B_2

In this part of this study, we consider the experimental works of Velev *et al.*, (SLS and SANS);[49] Rosenbaum *et al.*, (SLS);[131] Tessier *et al.*, (SIC);[121] Pjura *et al.*, (MO);[132] and Bajaj *et al.*, (SLS);[133] for LYS and CHY[2]. Using these results as

[1]Some important aspects of the description of the Figures 6.4b and 6.5b can be found above.
[2]Acronyms between parentheses refer to the technique used by each author (See Section 4.4.3).

our referential, we analysed the inherent differences that could be found in the experimental works. A plot of second virial coefficient B_2 as a function of pH is presented in Figure 6.6. It is seen that for all cases there are dissimilarities between each one of results obtained from different studies. However, different research groups have shown that B_2 is a quite sensitive description of the complexation phenomena. Thus, relatively discrepant results between techniques and methods used to obtain B_2 data were reported in the past by others.[87, 121, 133, 141, 142] For instance, analyses of literature made by Ahamed and collaborators,[141] showed that B_2 measures may have an error depending on the technique used: via MO, SLS, SEC or SIC, the estimated errors are, respectively, ± 1.0, ± 2.0, ± 3.0, and $\pm 1.0 \times 10^{-4}$mol ml/g^2. It should not be forgotten that B_2 values determined and reported by laboratory experiments represent the mean value of all interactions that may be occurring between pairs of proteins.[85] This could be a source of error and bias, especially, when working with low salt concentrations, the most critical regime.[49, 121]

Table 6.2: Comparison between experimental values of second virial coefficient B_2 at different pH and salt (5 mM) concentrations for lysozyme and α-chymotrypsinogen.

Protein	pH	$B_2^{\text{Exp.}}$[a,b]		ΔB_2[c]	Reference
LYS	4.5	33.3	31.0	−2.3	Velev *et al.*, (SANS)
	6.0	28.5	21.0	−7.5	Velev *et al.*, (SANS)
	9.0	15.6	17.0	1.4	Velev *et al.*, (SANS)
CHY	3.0	7.8	4.46	−3.34	Velev *et al.*, (SANS)

[a]Second virial coefficients given in mol ml $\times 10^{-4}$/g^2. [b]Values of left and right are B_2's from ref. Velev *et al.*, (SLS); and reported by other studies, respectively. [c]Data are calculated as the difference between B_2's reported by other studies and B_2's from ref. Velev *et al.*, (SLS).

In the Tables 6.2 and 6.3, we summarized the B_2 values for LYS and CHY from the above mentioned references. It is observed that most values in the work from ref. Velev *et al.*, (SLS) have an attractive trend compared to the other studies reported here. We could say that these differences are more marked when comparing the B_2 reported by Velev *et al.*, (SLS); and Tessier *et al.*, (SIC) (except for CHY at pH 3.0, and salt 100 mM). However, for LYS at 5 mM, the data from Velev *et al.*, using SANS, show a more attractive behavior. A generalization of the B_2 quantitative behavior for any physical-chemistry condition is problematic. It could be said that the main root of this problem is that there is no study where B_2's are determined systematically using a set of different techniques and experimental conditions. Furthermore, it is evident that in essence each technique proposes a different methodology and the discrepancy is also correlated to that.[141] For instance,

Table 6.3: Comparison between experimental values of second virial coefficient B_2 at different pH and salt (100 mM) concentrations for lysozyme and α-chymotrypsinogen.

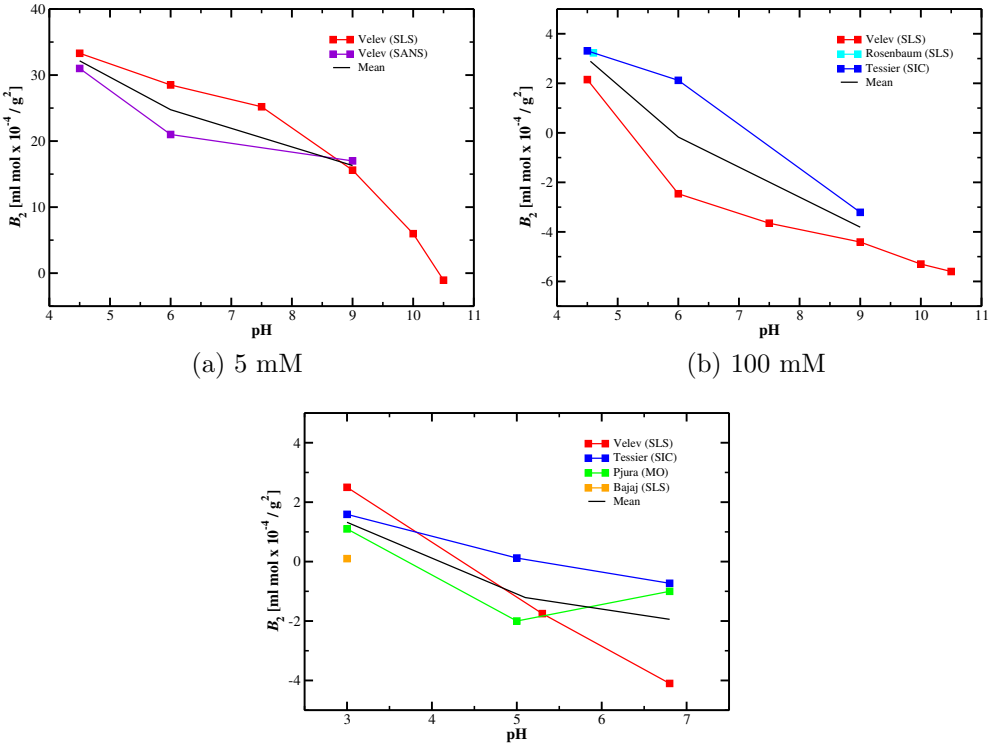
Protein	pH	$B_2^{\text{Exp.}}$ [a,b]		ΔB_2 [c]	Reference
LYS	4.5	2.15	3.23[e]	1.08	Rosenbaum *et al.*, (SLS)
			3.31	1.16	Tessier *et al.*, (SIC)
	6.0	−2.46	2.12	4.58	Tessier *et al.*, (SIC)
	9.0	−4.41	−3.21	1.2	Tessier *et al.*, (SIC)
CHY	3.0	2.50	1.59	−0.91	Tessier *et al.*, (SIC)
			1.10	−1.4	Pjura *et al.*, (MO)
			0.10	−2.4	Bajaj *et al.*, (SLS)
	5.3	−1.75[d]	0.12[f]	1.87	Tessier *et al.*, (SIC)
			−2.00[f]	−0.25	Pjura *et al.*, (MO)
	6.8	−4.10	−0.73	3.37	Tessier *et al.*, (SIC)
			−1.00	3.1	Pjura *et al.*, (MO)

[a]Second virial coefficients given in mol ml $\times 10^{-4}$/g^2. [b]Values of left and right are B_2's from ref. Velev *et al.*, (SLS); and reported by other studies, respectively. [c]Data are calculated as the difference between B_2's reported by other studies and B_2's from ref. Velev *et al.*, (SLS). [d]Common interpolated value determined by ref. Velev *et al.* [e]pH 4.6. [f]pH 5.0.

lower amounts of protein concentrations are used when implementing SIC, unlike the experiments involving light scattering (*e.g.* SLS or SANS).[121] In other cases, through SIC it has been proved that under the same electrolytic condition, B_2 values for the pair LYS-LYS are relatively different depending on the chromatographic peak used.[87] Another possible reason for these variations is the purity of the samples, which may play an important role in the interference and interpretation of B_2 measured by experimental procedures.[121]

Despite problems seen in the literature, in some sense, it has been established that in many cases the qualitative description is fulfilled for several types of protein molecules.[87] By comparison with the experimental data proposed here, and focusing on ref. Velev *et al.*, (SLS);[49] we deduced from our analysis that the same qualitative description is given in the majority of pH and salt concentrations. The dependence of pH for the two systems is a feature that is relatively replicated in all the experimental works. An average behavior between the common points can be observed as black lines in Figure 6.6. Again, we want to highlight the semi-quantitative character of B_2 in favor of the natural tendency observed *per se* for the two systems.

The following analyses were performed using the work of Velev *et al.*, (SLS).[49] In spite of the differences found, this work obtained interesting data in the sense that for B_2's estimated there was a significant correlation with the crystallization capacity of LYS and CHY. With that data, there was a relevant evidence supporting the



(a) 5 mM (b) 100 mM

(c) 100 mM

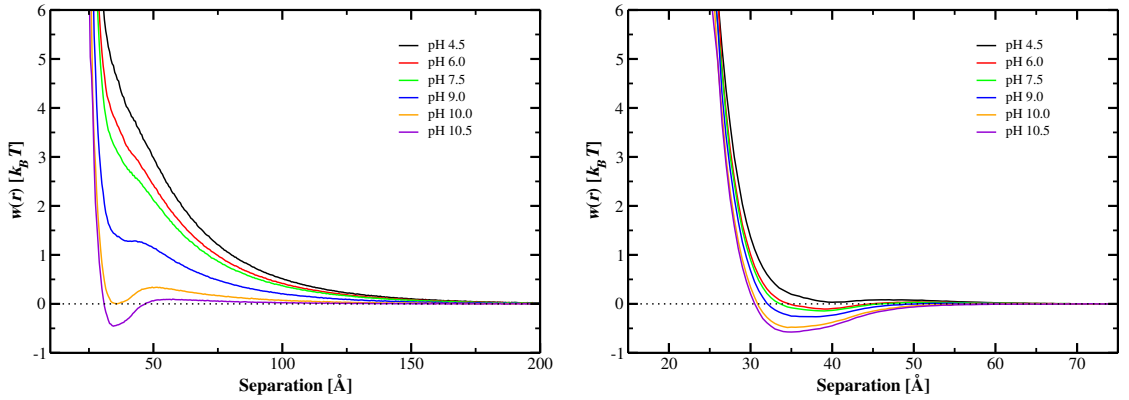
Figure 6.6: Second virial coefficient B_2 as a function of pH for (a) & (b) lysozyme, and (c) α-chymotrypsinogen at different salt concentrations determined by different authors.

George's hypothesis of the so-called "crystallization slot".[85] Evidently, it is known that the crystallization capacity of a protein is linked to many other factors (*e.g.* the chemical nature of the salt).[141] Based on that experimental work we see that there is a good direction in the sense that it achieves a description of what happens in the self-interaction of LYS and CHY at electrolyte solution. For the remaining system (*i.e.* RIB), we used the work of Tessier *et al.*, (SIC).[54] We do not carry out an exhaustive analysis of experimental B_2 values for this protein molecule since data reported are scarce.

6.3 Initial estimation of protein-protein interactions

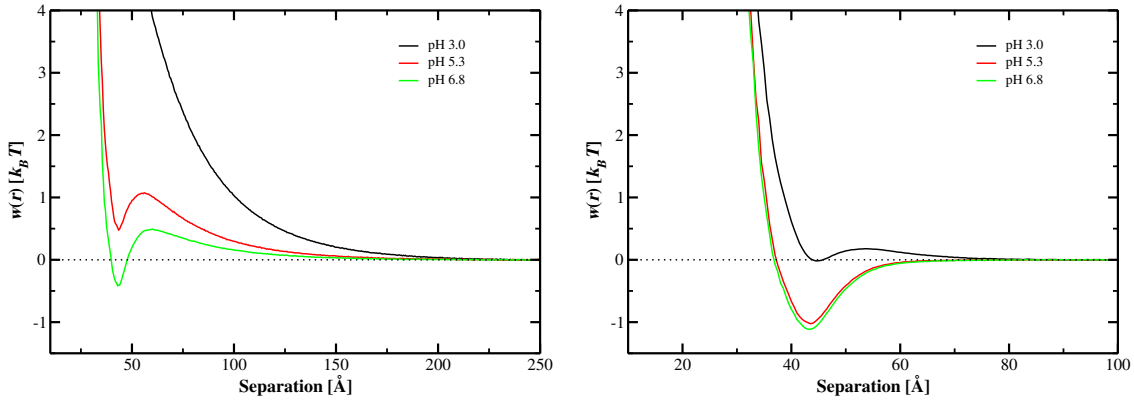
The dimerization process was investigated by means of free energy derivatives. As mentioned above, considering the experimental works of Velev *et al.*, (SLS);[49] and Tessier *et al.*, (SIC);[54] we calculated by MC simulations the $w(r)$ between pairs

of (a) LYS, (b) CHY, and (c) RIB. We used the same pH values for LYS (4.5, 6.0, 7.5, 9.0, 10.0, and 10.5), CHY (3.0, 5.3, and 6.8), and RIB (3.0, 4.0, 5.0, 6.5, and 8.0) as given by the experimental works. Further, we studied the same systems at salt concentrations of 5 and 100 mM for LYS and CHY;[49] and 50 and 100 mM for RIB.[54] Figures 6.7, 6.8, and 6.9 shown the $w(r)$ for LYS, CHY, and RIB respectively, determined with ε_{LJ} equal to 0.05005 k_BT. This is the ε_{LJ} assumed to be ideal in the literature.[34]



(a) 5 mM (b) 100 mM

Figure 6.7: Potential of mean force as a function of the center-center separation distance for two lysozymes at different pH solutions and salt concentrations. Data for MC simulations with ε_{LJ} equal to 0.05005 k_BT.

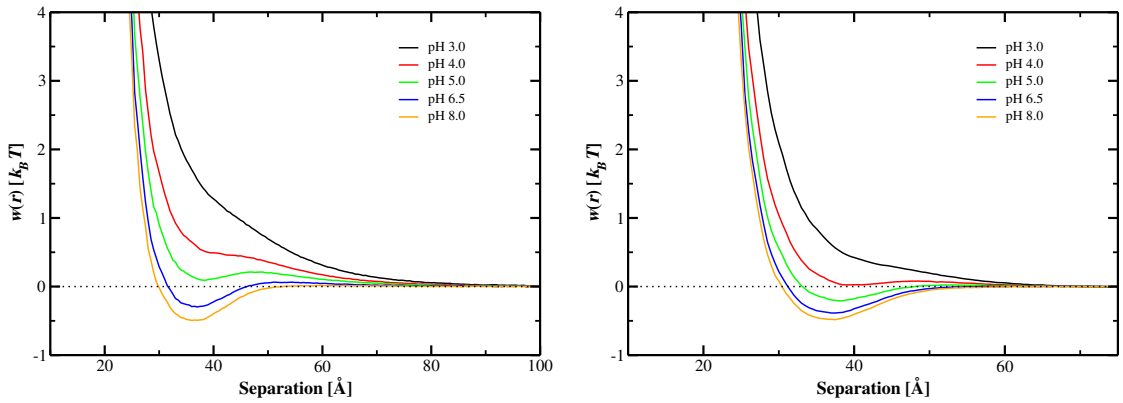


(a) 5 mM (b) 100 mM

Figure 6.8: Potential of mean force as a function of the center-center separation distance for two α-chymotrypsinogens at different pH solutions and salt concentrations. Data for MC simulations with ε_{LJ} equal to 0.05005 k_BT.

At 5 mM of salt concentration, we see that systems are very repulsive between pH's 4.5 and 10.0 $[w(r) > 0]$ for the pair LYS-LYS. However, an unstable association is

possible when pH is equal to 10.5 ($-0.4\ k_BT$ at a separation distance of 34.5 Å) (Figure 6.7a). A similar pattern is observed for the pair CHY-CHY at 5 mM of NaCl (Figure 6.8a). Positive values for the free energies of interactions are observed at pH's 3.0 and 5.3 at separation distances closer to σ (= 29.5 Å); but at pH 6.8, a minimum occurs at 43 Å ($-0.4\ k_BT$). We see that the self-association of RIB has a repulsive behavior at 50 mM of NaCl when the pH is further away from the pI (Figure 6.9a). Some pH solutions, for example, 6.5 and 8.0, have peaks of attraction at a distance less than 35 Å, but these values are not sufficient to generate a stable dimer ($< -1\ k_BT$).



(a) 50 mM (b) 100 mM

Figure 6.9: Potential of mean force as a function of the center-center separation distance for two ribonucleases A at different pH solutions and salt concentrations. Data for MC simulations with ε_{LJ} equal to 0.05005 k_BT.

At 100 mM of NaCl, the $w(r)$ have minimal peaks less than 0 for the three systems at most pH solutions: 6.0, 7.5, 9.0, 10.0, and 10.5 for LYS (Figure 6.7b); 5.3, and 6.8 for CHY (Figure 6.8b); and 5.0, 6.5, and 8.0 for RIB (Figure 6.9b). Nevertheless, in those particular cases the attraction given by the interaction is not enough to generate a stable self-association ($< -1\ k_BT$). For instance, for the pair LYS-LYS, when pH = 10.5, the free energy of interaction is equal to $-0.57\ k_BT$ at $r = 34$ Å.

As exposed in Section 4.1, when pH $\approx$ pI the charge regulation mechanism tends result in attractive mechanism.[5, 23, 25] Adjacently, dispersion forces are acting too, since when Z_P tends to zero this kind of interaction becomes dominant.[10] Therefore, we see that LYS, CHY, and RIB are governed mainly by vdW forces and charges regulation mechanism at this condition. As discussed in Subsection 6.1.1, CHY and RIB in comparison with LYS has one additional electrostatic attractive force given by the dipolar interaction (*e.g.* the averaged μ_P for LYS and CHY are, 34.8 and

Table 6.4: Comparison between values of calculated and experimental second virial coefficient B_2 at different pH and salt concentrations for lysozyme, α-chymotrypsinogen, and ribonuclease A using ε_{LJ} reported by Persson *et al.*

Protein	Salt[a]	pH	$Z_P \times Z_P$[b]	B_2^{MC}[c,d]	RMSD	$B_2^{\mathrm{Exp.}}$[c,e]
LYS (PDB id 2LZT)	5	4.5	72.3	88.5 $\pm$ 0.4	55.22	33.3
		6.0	57.8	78.3 $\pm$ 0.5	49.85	28.5
		7.5	50.4	71.0 $\pm$ 0.3	45.88	25.2
		9.0	28.1	46.1 $\pm$ 0.4	30.52	15.6
		10.0	9.6	19.7 $\pm$ 0.2	13.77	5.98
		10.5	3.2	6.7 $\pm$ 0.5	7.80	-1.06
	100	4.5	88.4	5.419 $\pm$ 0.006	3.27	2.15
		6.0	62.4	4.58 $\pm$ 0.05	7.05	-2.46
		7.5	54.8	4.19 $\pm$ 0.08	7.84	-3.65
		9.0	38.4	3.4 $\pm$ 0.1	7.86	-4.41
		10.0	13.7	2.0 $\pm$ 0.2	7.38	-5.30
		10.5	4.0	1.62 $\pm$ 0.07	7.30	-5.60
CHY (PDB id 1CHG) $\mathrm{pH_{crys}} = 6.3$[f]	5	3.0	144.0	43.5 $\pm$ 0.1	35.75	7.8
		5.3	42.3	18.88 $\pm$ 0.06	20.64	-1.75[g]
		6.8	22.1	11.01 $\pm$ 0.06	19.66	-8.65
	100	3.0	289.0	3.35 $\pm$ 0.07	0.85	2.5
		5.3	49.0	0.50 $\pm$ 0.04	2.25	-1.75[g]
		6.8	25.0	0.182 $\pm$ 0.002	4.28	-4.10
RIB (PDB id 1KF5) $\mathrm{pH_{crys}} = 7.1$[f]	50	3.0	169	15.0 $\pm$ 0.8	10.45	4.61
		4.0	90.3	10.4 $\pm$ 0.2	7.02	3.43
		5.0	56.3	7.5 $\pm$ 0.4	6.25	1.27
		6.5	26.0	3.9 $\pm$ 0.1	4.63	-0.64
		8.0	10.89	1.73 $\pm$ 0.06	2.44	-0.7
	100	3.0	193.2	7.28 $\pm$ 0.09	3.60	3.68
		4.0	102.0	4.9 $\pm$ 0.2	2.82	2.13
		5.0	60.8	3.0 $\pm$ 0.1	1.89	1.18
		6.5	27.0	1.9011 $\pm$ 0.0009	1.39	0.51
		8.0	11.6	1.14 $\pm$ 0.09	0.87	0.28

[a]Salt concentration given in mM. [b]Protein net charge product. [c]Second virial coefficients given in mol ml $\times 10^{-4}$/g^2. [d]B_2 estimated by MC simulations using ε_{LJ} equal to 0.05005 k_BT (mean $\pm$ SD, $n = 3$). [e]Experimental data from ref. Velev *et al.,* [SLS (LYS and CHY)]; and Tessier *et al.,* [SIC (RIB)]. [f]Value of pH used in the x-ray experiment. [g]Common interpolated value determined by ref. Velev *et al.*

80.0, respectively). A summary of the main physical-chemistry features of the three proteins is shown in the Table 6.6[3]. At high ionic strengths, the electrostatic salt screening reduces the protein-protein repulsion for three systems at extremes of pH allowing attractive forces to dominate.[102] Finally, studies where atomic force fields

[3]In Table 6.6, the estimation of percentage (%) of the solvent-accessible surface area of side chains of hydrophobic *aa*, was calculated using VADAR program.[143]

are used together explicit solvent models and solved by MD to estimate $w(r)$ for LYS-LYS,[50] showed a similar pattern.

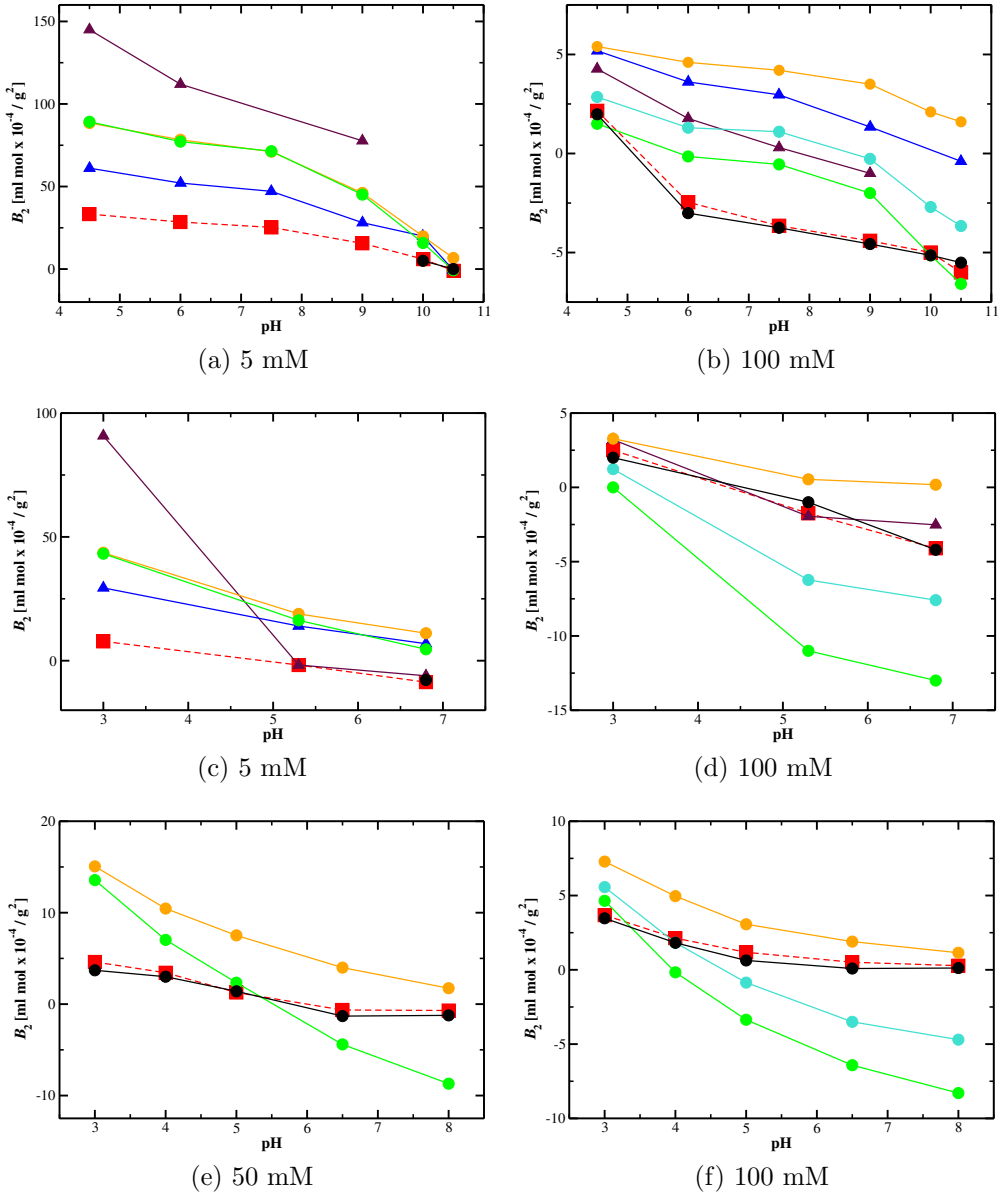


Figure 6.10: Second virial coefficient B_2 as a function of pH for (a) & (b) lysozyme, (c) & (d) α-chymotrypsinogen, and (e) & (f) ribonuclease A at different salt concentrations. Square, triangle, and circle lines, represent the experimental data, other computational models, and our simulations, respectively. Red lines represent the experimental data from Velev *et al.*, [SLS (LYS and CHY)]; and Tessier *et al.*, [SIC (RIB)]. Violet, and blue lines are data from Velev *et al.*, using DLVO theory, and data from Lund *et al.*, using other CG model, respectively. Orange, green, cyan, and black lines represent data estimated by MC simulations using ε_{LJ}'s equal to 0.05005 k_BT (from Persson *et al.*), 0.08 k_BT (from auv), 0.073 k_BT (from auv$^{\text{h}}$), ideal ε_{LJ} (from adv), respectively.

Values of second virial coefficient B_2 by MC simulations using ε_{LJ} equal to 0.05005 k_BT are shown in the Table 6.4, and Figure 6.10; for LYS, CHY, and RIB. The

comparison between experimental and calculated B_2's indicates that the simulated outcomes predict stronger repulsion in general. Even at higher salt concentrations, the calculated B_2 are always higher than the experimental values. This suggests that a better description of the vdW contribution in the coarse-grained force field is necessary for the three protein molecules. We also can see that pH values are inversely proportional to RMSD's calculated (especially at 5 mM for LYS and CHY systems, and at 50 mM, and 100 mM for RIB). The error is lower when pH tends to pI. In this regime of pH, the physical description of the model is facilitated, because the protein conformation (assumed to be rigid in the model) is closer to the one from the input structure.

Also, the balance between electrostatic and vdW interactions becomes more reasonable when the protein net charge is smaller and less unfolded. The folding process at extreme pH's would expose more hydrophobic *aa* that would require stronger attractive vdW parameters to reproduce the experimentally observed stronger attraction.[144, 145] At 100 mM salt concentration the electrostatic screening allows vdW forces to be stronger. We observe that the estimated B_2 values are relatively closer to the experimental data (Figures 6.10b, 6.10d, and 6.10f). From a qualitative perspective, we have evaluated the influence on the interaction using an ε_{LJ} equal to 0.05005 k_BT in the all three systems. In the Table 6.6, this information is summarized.

Calculation of the protein-protein interactions using analytic models based on spheres of radius and electrostatic properties similar to LYS and CHY were also done by Velev and co-workers (Figure 6.10).[49] This approximation had a behavior similar to reported in our study. On the other hand, comparing our results with the study of Lund *et al.*,[33] where were also applied other type of CG model; our data behave similarly to that study (Figure 6.10).

6.4 Refining the coarse-grained force field

6.4.1 Approach of unique value: attempting a generic description

Our first estimates have been aimed to reproduce the experimental data from a single value of ε_{LJ} (approach of unique value or auv). The intention was that this value

had to "mimic" the experimental B_2 from ref. Velev *et al.*, (SLS);[49]and Tessier *et al.*, (SIC);[54] for all pH and salt concentrations. By means of a *large scale* analysis, we explored several ε_{LJ} values in a range from 0.05-0.1 k_BT (in steps of 0.0005 k_BT). This means that the total number of ε_{LJ}'s explored for the three systems and different physical-chemistry conditions was 2828. Figure 6.11 shows the RMSD obtained from this approach in function of different ε_{LJ} values.

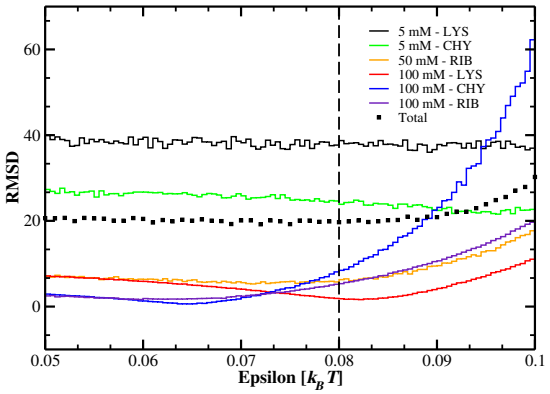
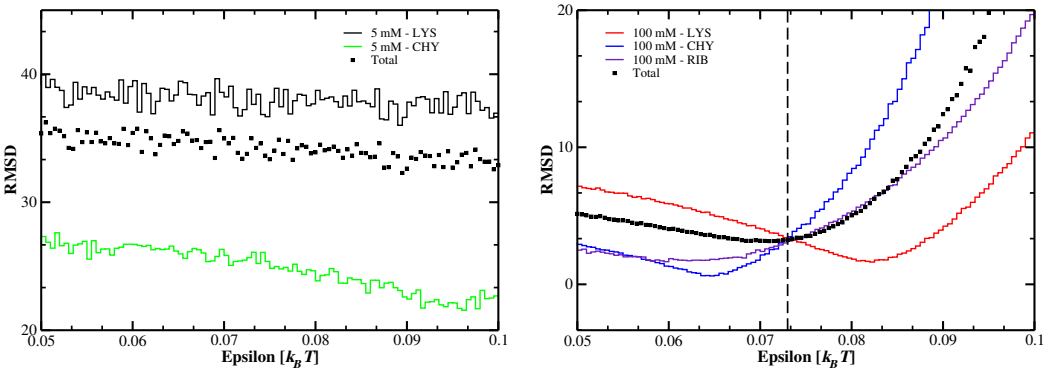


Figure 6.11: RMSD values as a function of ε_{LJ} for lysozyme, α-chymotrypsinogen, and ribonuclease A at different salt concentrations. Square lines represent the *total* RMSD calculated for all cases. Dashed black lines represent the best value: 0.08 k_BT.

The hydrophobic residues fewer on LYS in respect to CHY (see respectively, Figures 2.4a and 2.4b, and Table 6.6), reflects the intrinsic correlation of hydrophobic effect and the vdW radius of these *aa* on the surface of these systems. We found that RMSD values are lower on CHY in respect to LYS at ionic strength of 5 mM. However, based on Figure 6.12a it is shown that effective replication of experimental B_2 is difficult to achieve for the two systems. This is verified by observation of RMSD's of LYS and CHY oscillating around of 38 and 24, respectively. When salt concentration is equal to 100 mM, the dipolar interaction and hydrophobic effect on CHY and RIB allow to have RMSD's lower in comparison with estimations given for LYS (until $\sim$ 0.065 k_BT in the Figure 6.12b). Above that value of ε_{LJ} (*i.e.* $\sim$ 0.065 k_BT), data from MC simulations are strongly deviated from experimental B_2's on CHY and RIB. The curve with an "exponential trend" corresponds to an excessive increase of attraction, due to the increase of vdW forces contribution in the force field by salt screening, and besides, by the presence of "many" hydrophobic *aa* and high dipoles on these systems. The same effect occurs in LYS but fewer. With this, we conclude that the systematic variation of B_2 is also a function of the ionic strength, and therefore of ε_{LJ} parameter. For example, regard to RIB the best effective values of ε_{LJ} at 50 mM and 100 mM of NaCl for all pH solutions are

$\sim 0.072\ k_BT$ and $\sim 0.06\ k_BT$, respectively (Figure 6.11). It is observed that while the salt concentration is increased, ε_{LJ}'s become lower.

By means of this approach (auv), we observe separately the best ε_{LJ} for each protein and its respective salt and pH concentrations: for LYS, CHY, and RIB its ε_{LJ}'s were $0.083\ k_BT$ (RMSD = 26.49), $0.071\ k_BT$ (RMSD = 17.79), and $0.068\ k_BT$ (RMSD = 4.13), respectively. We see with this the underlying difference in the behavior of each system in electrolyte solution. Nevertheless, despite of limitations encountered (*e.g.* at 5 mM of NaCl), and taking into account the quantitative and qualitative criteria proposed by our methodology, it was established that the "best value" that represents the experimental data for all cases is $0.08\ k_BT$ (RMSD[4] =19.62).



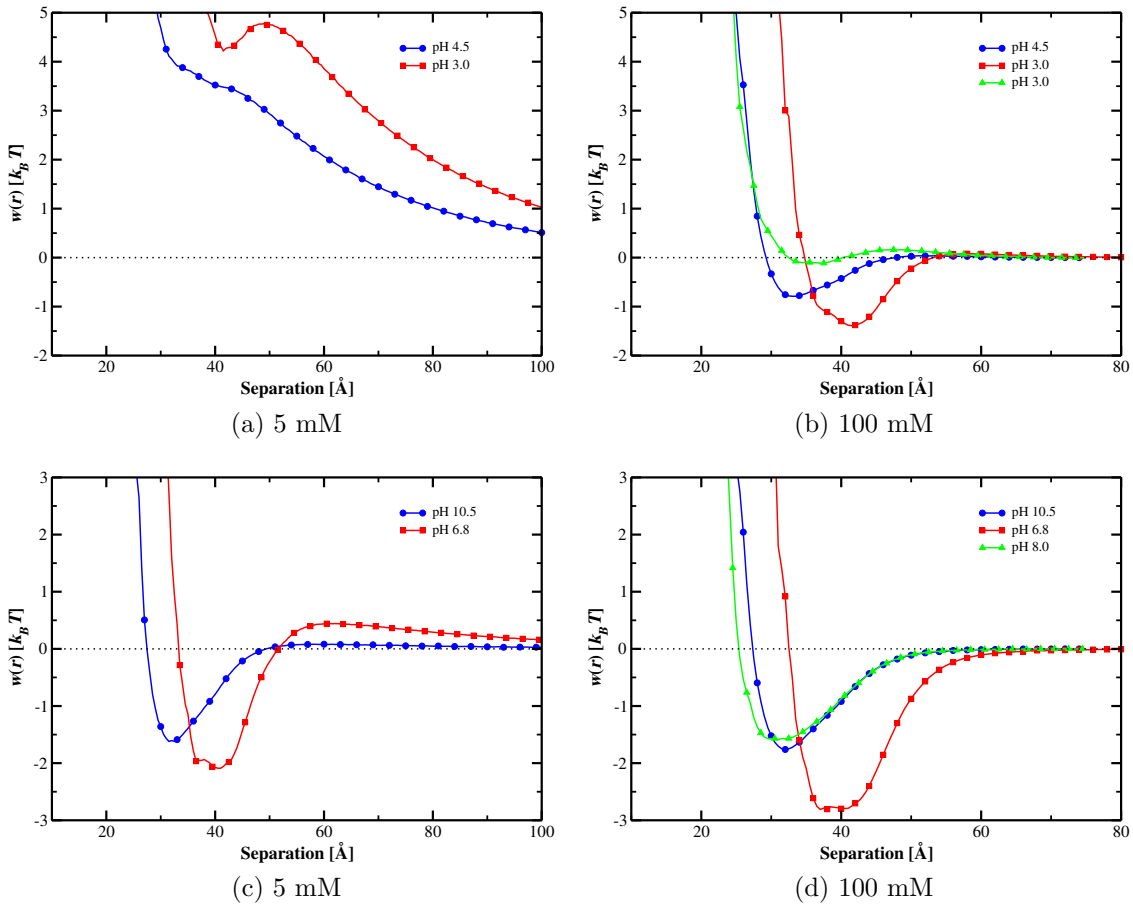
(a) 5 mM (b) 100 mM

Figure 6.12: RMSD values as a function of ε_{LJ} for lysozyme, α-chymotrypsinogen, and ribonuclease A at different salt concentrations. Square lines represent the *total* RMSD calculated for each one particular case. Dashed black lines in (b) represent the best value at 100 mM for the three proteins: $0.073\ k_BT$.

A more detailed view of $w(r)$ using a ε_{LJ} equal to $0.08\ k_BT$ is shown in Figure 6.13 for the three protein molecules (only the cases with extreme pH regimes are shown). In general, with increasing the ε_{LJ} parameter, as expected, it was increased the attractive forces which decrease the minima of the free energies of interaction by several units of k_BT in all experimental conditions for the all three systems. Observing the curves at 5 mM of NaCl, repulsive interactions are observed in acid pH for LYS (at pH = 4.5), and CHY (at pH = 3.0) in distance close to σ (Figure 6.13a). However, a higher decrease of $w(r)$ occurs at pH near the pI. LYS reaches a global minimum around $-1.6\ k_BT$ ($r = 32$ Å), and CHY around $-2.1\ k_BT$ ($r = 40$

[4]Data derived from B_2's obtained by MC simulations and experimental data, using the three systems and all experimental results avaliable.

Å) (Figure 6.13c). Compared to MC calculations using ε_{LJ} equal to 0.05005 k_BT, this new pmf in the "basic conditions" evaluated here, seems more reasonable for LYS than for CHY. An overestimate of the attraction is given for CHY. Computational calculations using CHY have shown that the change of repulsive interactions (*e.g.* given at pH 3.0) to strongly attractive interactions (*e.g.* given at pH 6.8) at low salt concentration is apparently due to a high geometric complementarity on the pair CHY-CHY.[52]



(a) 5 mM (b) 100 mM

(c) 5 mM (d) 100 mM

Figure 6.13: Potential of mean force as a function of the center-center separation distance for two lysozymes (blue lines), two α-chymotrypsinogens (red lines), and two ribonucleases A (green lines) at different salt concentrations. Acid and basic regimes of pH are represented in plots above and bellow, respectively. Data for MC simulations with ε_{LJ} equal to 0.08 k_BT.

It is important to mention that LYS seems to be a tricky system as shown by data reported in the literature. By employing different experimental techniques (SAXS and SANS) and using pH physiological regimens and low salt concentration, it was found that LYS tends to form aggregates due to the action of short-range

interactions in the dimerization process.[146] However, a later study in which the same techniques and experimental conditions were used showed that repulsive electrostatic interactions dominate, preventing the formation of aggregates.[147]

At 100 mM of NaCl, there are negative free energies of interaction close to σ at all pH solutions for the three systems. At low pH very stable complexes are given for LYS and CHY (they have a energy $< -1\ k_BT$), but an unstable complex is given to RIB (Figure 6.13b). On the other hand, we can observe that when the pH is near of pI, the minima are very negatives: LYS has a minimum around $-1.8\ k_BT$ ($r = 32$ Å), CHY around $-2.8\ k_BT$ ($r = 37$ Å), and RIB around $-1.5\ k_BT$ ($r = 29$ Å) (Figure 6.13d). Summarizing, when this new ε_{LJ} value is used the attraction is considerably lower at 5 mM for the three proteins, and higher at 100 mM specifically, for CHY and RIB (Table 6.6).

In the Figure 6.10, it can be observed B_2's calculated using the auv approach for the three protein pairs. At 5 mM of NaCl for LYS and CHY, the pattern of these curves is very similar, compared with curves when $\varepsilon_{LJ} = 0.05005\ k_BT$. At this salt concentration (Table 6.5), the estimated error calculated via RMSD analysis is higher when the solution pH is more acid. As previously mentioned, unfolding process occurs at acid regimes. The rigid approximation model used here becomes cruder as pH departs from pI. In order to demonstrate more accurately the pH effect on the protein conformation, we could use other approaches more sophisticated that invokes an all-atom description of the system. Nevertheless, constant pH methods in this direction (based on atomistic models) still suffer from slow convergence.[148] Moreover, the study of several experimental conditions of electrolyte solutions, as done in this work, makes prohibitive this class of methods due to the quite high cpu time required. An estimation of the cpu costs is easily obtained analysing the present calculations. For example, using our CG model for the pair LYS-LYS, a single simulation with one equilibration (10^6 MC steps) and two production phases (10^7 MC steps) take about 8 hours of cpu time (Intel core *i*5-3330 cpu, and 3.00 GHz). So, if we multiply this value by the number of ε_{LJ} explored at any physical-chemistry condition (*i.e.* 101 times[5]), the final computational cost will be 808 hours of cpu time (around of 33.7 days!). Through the all-atoms approach this value of time would be even higher and impractical. For the LYS-LYS system at 100 mM of NaCl, the reproduction of the experimental data considerably improved with the new ε_{LJ}. Meanwhile for the pair CHY-CHY and RIB-RIB, the attraction is overestimated (Tables 6.5 and 6.6, and Figure 6.10). For instance, at pH equal to 6.8 for CHY,

[5]In a range from 0.05-0.1 k_BT in steps of 0.0005 k_BT.

the $B_2^{0.05005 k_B T}$ and $B_2^{0.08 k_B T}$ are 0.18 and -13.6×10^{-4}mol ml/g^2, respectively. The attraction is 13 times larger. Considering RIB at 50 mM of NaCl, when pH is equal to 6.5 and 8.0 the attraction between RIB pairs is excessive. At pH equal to 3.0, and 4.0 do not have sufficient attraction. The remaining pH solution (*i.e.* 5.0), seems to be near to the experimental B_2.

Table 6.5: Comparison between values of calculated and experimental second virial coefficient B_2 at different pH and salt concentrations for lysozyme, α-chymotrypsinogen, and ribonuclease A using ε_{LJ} estimated via auv and auv[h].

Protein	Salt[a]	pH	0.08 k_BT		0.073 k_BT		$B_2^{\text{Exp.}}$[b,e]
			B_2^{MC}[b,c]	RMSD	B_2^{MC}[b,d]	RMSD	
LYS	5	4.5	89.2 ± 0.3	55.94	-.-	-.-	33.3
		6.0	77.2 ± 0.3	48.77	-.-	-.-	28.5
		7.5	71.4 ± 0.7	46.29	-.-	-.-	25.2
		9.0	45.10 ± 0.05	29.50	-.-	-.-	15.6
		10.0	15.8 ± 0.2	9.82	-.-	-.-	5.98
		10.5	−0.7 ± 2.0	0.31	-.-	-.-	−1.06
	100	4.5	1.5 ± 0.1	0.56	2.854 ± 0.007	0.70	2.15
		6.0	−0.15 ± 0.05	2.30	1.3 ± 0.1	3.85	−2.46
		7.5	−0.55 ± 0.09	3.09	1.10 ± 0.05	4.75	−3.65
		9.0	−2.2893 ± 0.0005	2.12	−0.27 ± 0.01	4.13	−4.41
		10.0	−5.14 ± 0.01	0.15	−2.7 ± 0.1	2.53	−5.30
		10.5	−6.58 ± 0.03	0.98	−3.660 ± 0.004	1.93	−5.60
CHY	5	3.0	43.23 ± 0.02	35.43	-.-	-.-	7.8
		5.3	16.9 ± 0.3	18.65	-.-	-.-	−1.75[f]
		6.8	4.6 ± 0.1	13.31	-.-	-.-	−8.65
	100	3.0	−0.14 ± 0.08	2.64	1.23 ± 0.05	1.27	2.5
		5.3	−11.71 ± 0.01	9.96	−6.23 ± 0.04	4.48	−1.75[f]
		6.8	−13.6 ± 0.1	9.57	−7.59 ± 0.01	3.50	−4.10
RIB	50	3.0	13.566 ± 0.001	8.95	-.-	-.-	4.61
		4.0	7.0 ± 0.1	3.59	-.-	-.-	3.43
		5.0	2.34 ± 0.04	1.07	-.-	-.-	1.27
		6.5	−4.4 ± 0.1	3.76	-.-	-.-	−0.64
		8.0	−8.7 ± 0.2	8.00	-.-	-.-	−0.7
	100	3.0	4.64 ± 0.04	0.96	5.57 ± 0.1	1.89	3.68
		4.0	−0.1 ± 0.2	2.82	1.824 ± 0.005	0.30	2.13
		5.0	−3.3 ± 0.1	4.54	−0.86 ± 0.08	2.04	1.18
		6.5	−6.4 ± 0.1	6.93	−3.5 ± 0.1	4.03	0.51
		8.0	−8.29 ± 0.05	8.57	−4.7 ± 0.07	4.98	0.28

[a]Salt concentration given in mM. [b]Second virial coefficients given in mol ml $\times 10^{-4}$/g^2. [c]B_2 estimated by MC simulations using ε_{LJ} equal to 0.08 k_BT (mean ± SD, $n = 3$) (auv). [d]B_2 estimated by MC simulations using ε_{LJ} equal to 0.073 k_BT (mean ± SD, $n = 3$) (auv[h]). [e]Experimental data from ref. Velev *et al.,* [SLS (LYS and CHY)]; and Tessier *et al.,* [SIC (RIB)]. [f]Common interpolated value determined by ref. Velev *et al.*

Table 6.6: Qualitative evaluation of ε_{LJ} by comparison of the calculated and experimental B_2's for lysozyme, α-chymotrypsinogen, ribonuclease A, spidroin, and lactoferrin. Several features relationated with electrostatics and vdW forces are shown too.

Protein	Salt[a]	pH	Electrostatics				vdW	Interaction[d,e]		
			pI − pH[b]	Z_P	C_P	μ_P	Hb[c]	ref.	auv	auv[h]
LYS	5	4.5	6.6	8.5	1.0	40.2	19.6	↓	↓	-.-
		6.0	5.1	7.6	0.9	38.3		↓	↓	-.-
		7.5	3.6	7.2	1.0	36.6		↓	↓	-.-
		9.0	2.1	5.3	1.4	32.3		↓	↓	-.-
		10.0	1.1	3.1	1.6	30.5		↓	↓	-.-
		10.5	0.6	1.9	1.6	29.3		↓	↓	-.-
	100	4.5	6.6	9.4	1.6	41.0	19.6	↓	↑	↓
		6.0	5.1	7.9	1.0	39.3		↓	↓	↓
		7.5	3.6	7.4	0.8	37.7		↓	↓	↓
		9.0	2.1	6.2	1.5	34.5		↓	↓	↓
		10.0	1.1	3.7	2.0	32.1		↓	↓	↓
		10.5	0.6	2.0	2.2	31.0		↓	↑	↓
CHY	5	3.0	6.5	12.0	1.8	59.8	25.5	↓	↓	-.-
		5.3	4.2	6.5	0.7	93.5		↓	↓	-.-
		6.8	2.7	4.7	0.8	94.2		↓	↓	-.-
	100	3.0	6.5	17.0	2.3	54.1	25.5	↓	↑	↑
		5.3	4.2	7.0	0.9	91.9		↓	↑	↑
		6.8	2.7	5.0	0.8	94.9		↓	↑	↑
RIB	50	3.0	6.5	13.0	2.1	64.6	16.8	↓	↓	-.-
		4.0	5.5	9.5	1.8	67.2		↓	↓	-.-
		5.0	4.5	7.5	1.3	75.2		↓	↓	-.-
		6.5	3.0	5.1	1.4	84.3		↓	↑	-.-
		8.0	1.5	3.3	1.1	87.4		↓	↑	-.-
	100	3.0	6.5	13.9	2.3	67.6	16.8	↓	↓	↓
		4.0	5.5	10.1	2.1	67.2		↓	↑	↑
		5.0	4.5	7.8	1.4	74.7		↓	↑	↑
		6.5	3.0	5.2	1.5	84.2		↓	↑	↑
		8.0	1.5	3.4	1.1	88.5		↓	↑	↑
NTD[f]	5	6.0	−1.5	−3.5	0.7	119.8	29.8	↓	-.-	-.-
LF[g]	5	7.0	2.6	16.6	1.3	317.9	21.3	↓[h]	-.-	-.-
	300	7.0	2.6	19.7	1.4	351.5	21.3	↑[h]	-.-	-.-

[a]Salt concentration given in mM. [b]Distance in pH units between the pH and pI in each system (pI = 11.1 for LYS, pI = 9.5 for CHY, pI = 9.5 for RIB, pI = 4.5 for NTD, and pI = 9.6 for LF). [c]Percentage (%) of the solvent-accessible surface area of side chains of hydrophobic *aa.* [d]Evaluation of the interaction from B_2's estimated by MC simulations using ε_{LJ}'s equal to 0.05005 (from Persson *et al.*), 0.08 (from auv), and 0.073 (from auv[h]) k_BT . [e]The symbols ↑ and ↓ refer to cases where ε_{LJ} was relatively overestimated (*i.e.* interaction has "excessive" attraction) or underestimated (*i.e.* interaction has "excessive" repulsion), respectively. [f]Data extracted from ref. Da Silva *et al.*, the evaluation of the interaction is derived of pmf. [g]Electrostatic properties calculated by us. [h]Data extracted from ref. Li *et al.*

Finally, we confirm that the use of an ε_{LJ} equal to 0.08 k_BT does not converge in an optimal solution for all cases. We think that the overestimation in the attraction is due to a decompensation between values that technically remain unmovable at low salt (Figure 6.12a), versus values that are strongly affected at high salt (Figure 6.12b). This undoubtedly represents a source of error and consequent deviation in several B_2's. Therefore, to have a relatively reasonable generic description, we work with B_2's given at high salt concentrations (approach of unique value at high salt concentration or auv[h]), since at low salt is unreliable (see also the discussion about experimental measurements at low salt regime in Section 6.2).

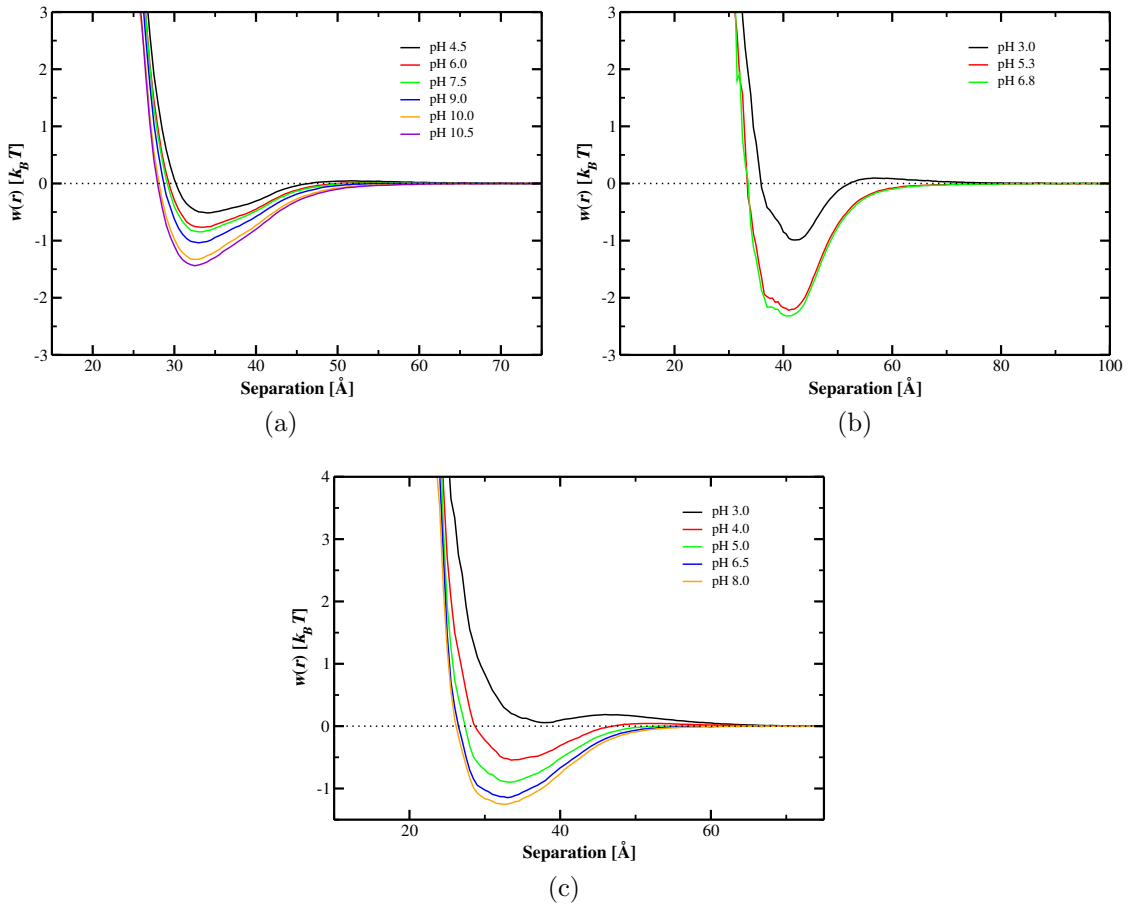


Figure 6.14: Potential of mean force as a function of the center-center separation distance for (a) two lysozymes, (b) two α-chymotrypsinogens, and (c) two ribonucleases A at different pH solutions and salt concentrations of 100 mM. Data for MC simulations with ε_{LJ} equal to 0.073 k_BT.

Limiting the three systems to high salt concentrations and applying the methodology provided in this section, we found that the best ε_{LJ} value is 0.073 k_BT (RMSD = 3.23[6]) (Figure 6.12b). Interpretation of these results using a concentration of 100

[6]Data derived from B_2's obtained by MC simulations and experimental data, using the three systems at all pH solutions (for each one), and salt concentration of 100 mM.

mM has already been given. The Figure 6.14 shown pmf's between pairs of (a) LYS, (b) CHY, and (c) RIB obtained via MC simulations. In general, for the three proteins, negative values of free energy interaction are observed (with the exception of RIB at pH 3.0). For LYS at pH's equal to 4.5, 6.0, and 7.5; and for RIB at pH's equal to 4.0 and 5.0; unstable complexes are observed at a separation distance of $\sim$ 33 Å for the two systems. Regarding the remaining pH solutions for these two systems, attractions are observed at $\sim$ 32 Å. Too attractive complexes CHY-CHY are formed with $\varepsilon_{LJ} = 0.073\ k_BT$. Its minima reach values of approximately $-1.0\ k_BT$ (at pH 3.0 at a separation distance of $r = 42$ Å) and $-2.0\ k_BT$ (at pH's 5.3 and 6.8, at separation distance $r \approx 40$ Å). In general, observing the calculation of B_2 (Table 6.5, and Figure 6.10), we see that values computed by MC simulations exhibit a particular behavior to each protein and are translated to more positive numbers, compared to the B_2's calculated with $\varepsilon_{LJ} = 0.08\ k_BT$.

It should be emphasized that through auv[h] the value of $0.073\ k_BT$ represents the point of inflection between the three proteins (Figure 6.12b). Hence the B_2's obtained are less attractive for LYS, and more attractive for CHY and RIB (Table 6.6). That ε_{LJ} represents a "mean" value between the three systems. Thus, we highlight that this is a probable ε_{LJ} between LYS, CHY, and RIB, since in some sense, despite not having a strong precise quantitative description, the qualitative descriptions for each is reasonable (this is expected in models with CG approach).

Studies using CG model for LF have shown similar problems as found by us.[45] Their results using a ε_{LJ} to $0.05005\ k_BT$ turned out to be in a very attractive interaction between two LF (Table 6.6). MC simulations of two N-terminal domains (NTD) of the spidroin protein using the same ε_{LJ} showed an underestimation of the system's attractive capacity (Table 6.6).[46] On the other hand, other studies where is incorporated the solvent-accessible surface area in the CG force field for the same protein, converge to an overestimate of the attraction.[128] This supports the idea that a CG force field must be refined or approximated for each system because each one presents specific intrinsic properties as shown in the Table 6.6.

In our calculations, there are a remarkable difference between LYS and the other two proteins. Studies suggest that the electrostatic behavior of LYS is in some sense "atypical". However, similar features have also been found in other small proteins, such as α-crystallins, ATCase, and BMV.[149] On the other hand, CHY and RIB are proteins that share several features among which the most relevant is the heterogeneous distribution of their charges on the surface (Figures 2.3b and 2.3c, respectively).[51, 54] This distribution allows the dipolar moment to be relatively

high on these proteins molecules (Table 6.6). Velev and collaborators,[49] propose an alternative classification between groups of proteins. The first group includes proteins with features where the electrostatic repulsion is dominant below its pI (*e.g.* LYS). The second group includes proteins, where electrostatic repulsion controls at low pH and attractive forces are dominant close to pI (*e.g.* CHY). By means of this classification, our ε_{LJ} value via auv^{h} can be useful, as long as the studied system is similar or is within the range of those two probable groups. It has also been argued by means of calculation of first principles that appropriate Hamaker constants can be 3.1 and 23.4 k_BT.[150] From the point of view of our calculated ε_{LJ} (*i.e.* 0.073 k_BT equivalent to a Hamaker constant of ca. 13 k_BT), this could be within those two values.

As mentioned in Section 6.3, analytical calculations of LYS and CHY using the DLVO approximation were performed by Velev and co-authors.[49] For these calculations, different Hamaker's constants were tested for each protein (13.8 k_BT and 10.1 k_BT for LYS and CHY, respectively). We see that Hamaker's constant determined for us [13 k_BT (when $\varepsilon_{LJ} = 0.073$ k_BT)] is not far from the reported value in the literature. However, for each specific physical-chemistry condition different Hamaker's constants were not tested in their work. We saw that there are thermodynamic differences underlying for LYS, CHY, and RIB (Table 6.6), and this allows them to have a specific and different response according to the pre-established electrolyte conditions. For this reason, it was necessary to search an ε_{LJ} at each specific condition for our CG model, in order to not leave empty this discussion.

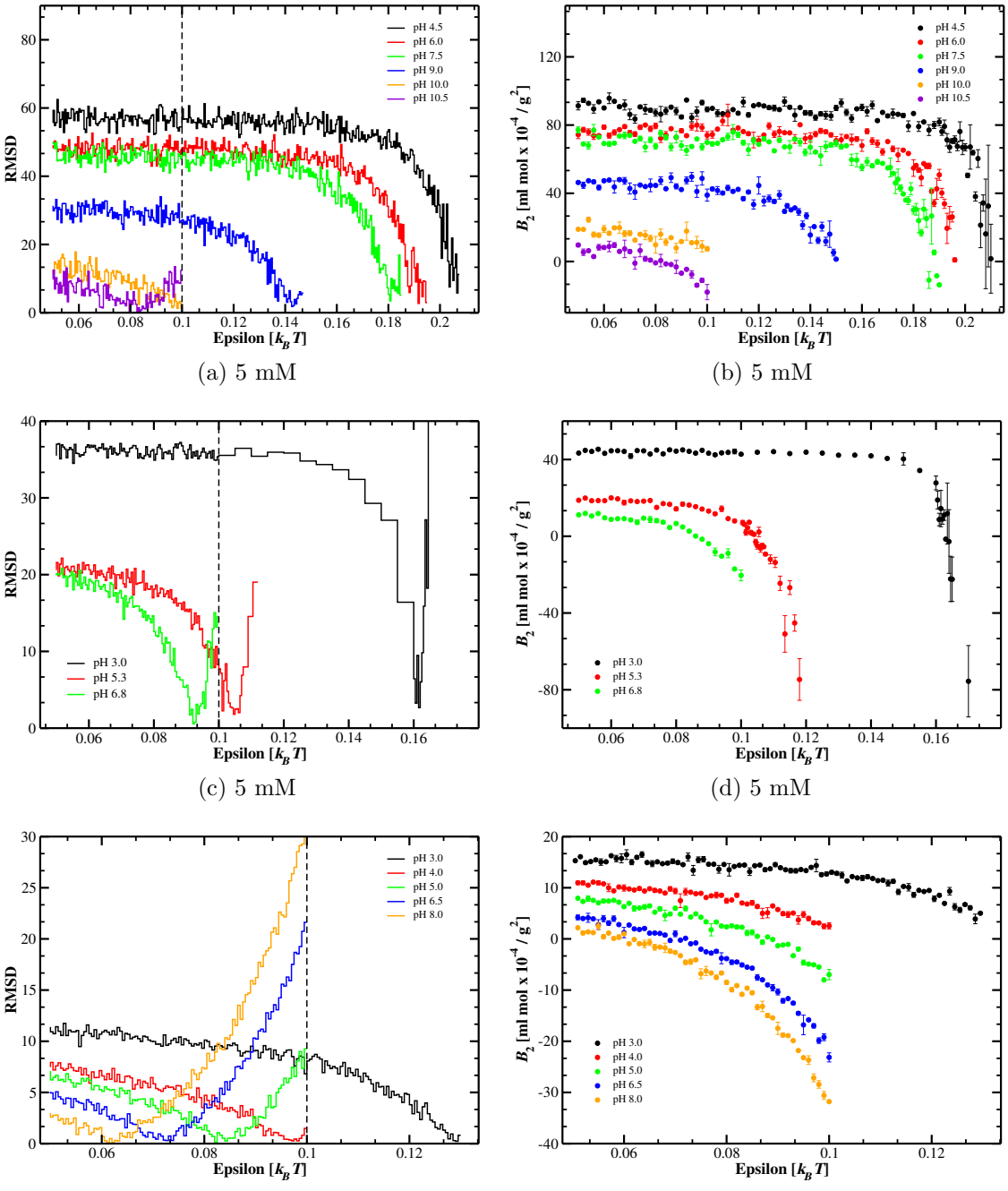
6.4.2 Approach of different values: about pH dependence and effects of the ε_{LJ} parameter

We explored the possibility to adopt different values of ε_{LJ}'s in order to reproduce the experimental data for each physical-chemistry condition (approach of different values or adv). The idea was to investigate ε_{LJ} as a function of pI$-$pH could be used as an adjustable parameter to incorporate all phenomena not explicitly described in the effective Hamiltonian. We know that it is more difficult for rigid model to describe B_2 far from pI. As anticipated there was greater difficulty to calibrate the force field at acidic regimes and salt concentration of 5 mM for LYS and CHY systems. We explored ε_{LJ} values outside the pre-set range (proposed in auv) at pH's equal to 4.5, 6.0, 7.5, and 9.0 on LYS (Figure 6.15a), and pH's equal to 3.0 and 5.3 on CHY (Figure 6.15c). By means of standard deviation measures between

B_2's, we find that being close to the experimental value, at the same time we lose resolution in the calculation of B_2 at those pH values (*i.e.* statistically, the data are widely dispersed). The error bar diagrams show huge errors at low pH and 5 mM concentration for LYS and CHY (Figures 6.15b, and 6.15d). We assume that the source of this error is based on a poor statistic in rdf's in those electrolyte conditions. By "obliging" the system to a more attractive representation, insufficient sampling is carried out via MC, and therefore, poor statistics prevail.

We must clarify that a higher number of MC steps per production phase ($\geq 10^8$ MC steps) could be performed, and a consequent decrease of the error will be visible. However, as discussed above with respect to the number of conditions that we are exploring that number of MC steps would be impractical (even on systems as small as LYS). As these data seem unreliable, they have been discarded for the next analysis. At 5 mM of NaCl, we only conserved ε_{LJ}'s estimated at pH 10.0 and 10.5 for LYS, and at pH 6.8 for CHY. By contrast, a rapid convergence between ε_{LJ} values were found at 50 mM for RIB (Figure 6.15e), and at 100 mM for all systems (Figures 6.16a, 6.16c, and 6.16e). In addition, we see that the statistic in these salt concentrations is very good for all three systems (Figures 6.15f, 6.16b, 6.16d, and 6.16f). Large deviations start at $\varepsilon_{LJ} \approx 0.095$ for CHY at pH 5.3 and 6.8, but an ideal ε_{LJ} value in that order is not of concern since smaller values are needed in that protein molecule. Through this approach it is achieved reliably reproduce the B_2 experimental data (Table 6.7, and Figure 6.10). RMSD values are evidently lower compared to previous approaches carried out. Regarding RIB at 50 mM the average of ideals ε_{LJ}'s is equal to 0.089 k_BT. The average values of ideals ε_{LJ}'s for LYS, CHY, and RIB are 0.083, 0.065, and 0.068 k_BT at 100 mM.

Figure 6.17 shows the potential of mean force obtained by means of adv approach for LYS, CHY, and RIB at 100 mM of NaCl. It could also be said that this is the ideal representation of a $w(r)$ for this type of system using a CG model approximation. We observe stable complexes for LYS with free energies of interaction in a range from -1.5 (at pH 6.0) to -1.6 (at pH 10.5) k_BT at a separation distance of ~ 32 Å (Figure 6.17a). At pH 4.5, an unstable complex with a free energy equal to -0.7 k_BT ($r = 33$ Å) occurs. Regarding CHY, formation of stable complexes are observed with minima of -1.6 [at pH 5.3 ($r = 42$ Å)], and -1.9 [at pH 6.8 ($r = 41$ Å)] k_BT. Complex unstable at pH 3.0 is observed (Figure 6.17b). Finally, all RIB-RIB pairs are unstable and are present in a range of free energies interaction from -0.4 (at pH 3.0) to -0.7 (at pH 6.5) k_BT at a separation distance of ~ 34 Å (Figure 6.17c). Since these pmf's are based on B_2's reported in the literature, we see a higher repulsivity



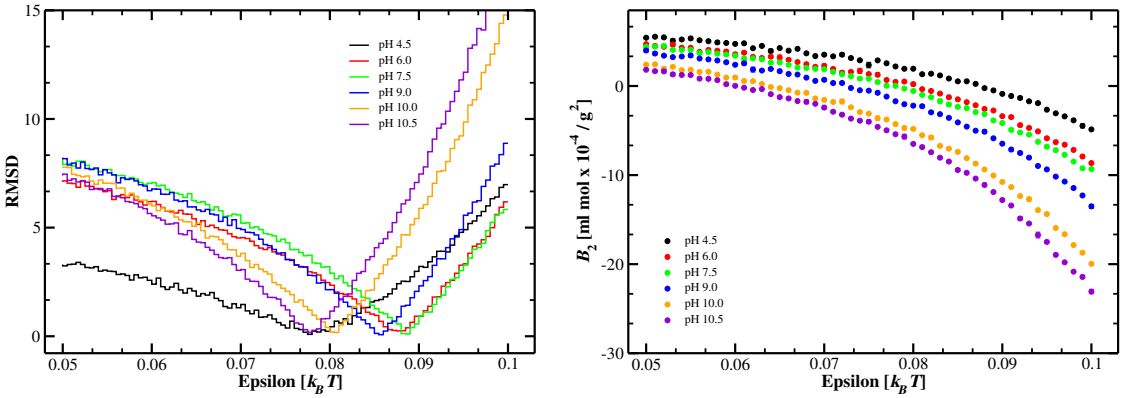
(a) 5 mM (b) 5 mM

(c) 5 mM (d) 5 mM

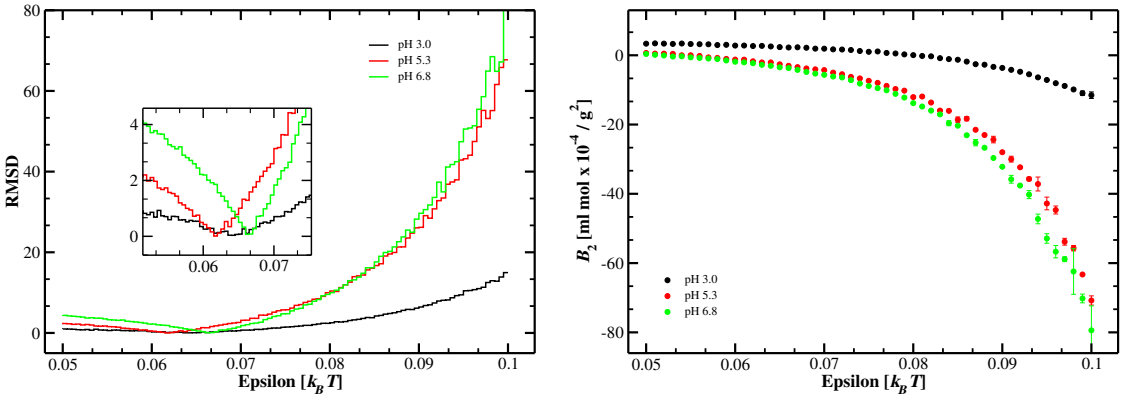
(e) 50 mM (f) 50 mM

Figure 6.15: Plots on the left: different RMSD values as a function of ε_{LJ}. Plots on the right: Second virial coefficient B_2 as a function of ε_{LJ} with error bars (not all data were included for a better readability). MC simulations via adv using (a) & (b) lysozyme, (c) & (d) α-chymotrypsinogen, and (e) & (f) ribonuclease A at different salt concentrations.

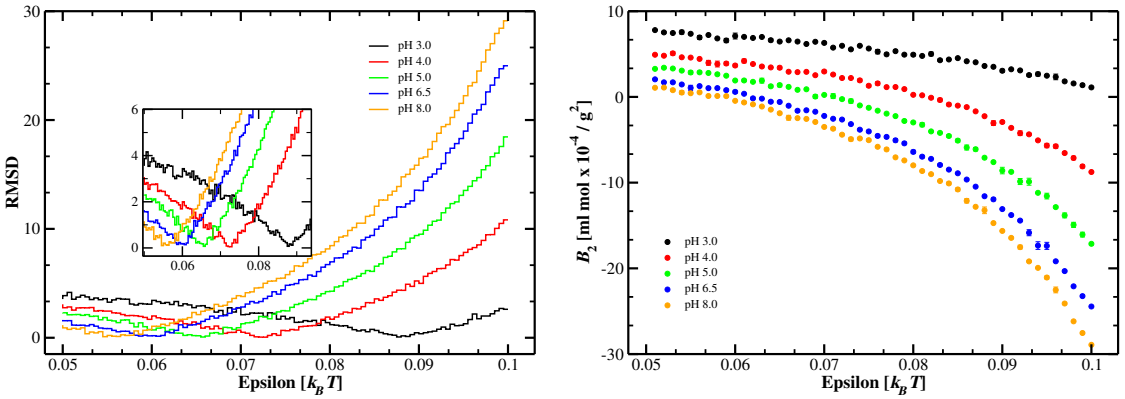
at pH equal to 8.0 in comparison to $w(r)$ given at pH's 5.0 and 6.5, despite being closer to the pI. In the Figure 6.18, an example of the pmf's for LYS at pH 10.5 and 100 mM of NaCl is shown. In all cases, it is clearly increased the attraction between



(a) 100 mM (b) 100 mM



(c) 100 mM (d) 100 mM



(e) 100 mM (f) 100 mM

Figure 6.16: Plots on the left: different RMSD values as a function of ε_{LJ}. Plots on the right: Second virial coefficient B_2 as a function of ε_{LJ} with error bars (not all data were included for a better readability). MC simulations via adv using (a) & (b) lysozyme, (c) & (d) α-chymotrypsinogen, and (e) & (f) ribonuclease A at 100 mM of salt concentration.

LYS's via auv, auv$^{\mathrm{h}}$, and adv in comparison with the results using $\varepsilon_{LJ} = 0.05005$ k_BT.[34] More stable complexes are also observed by the different approaches applied here.

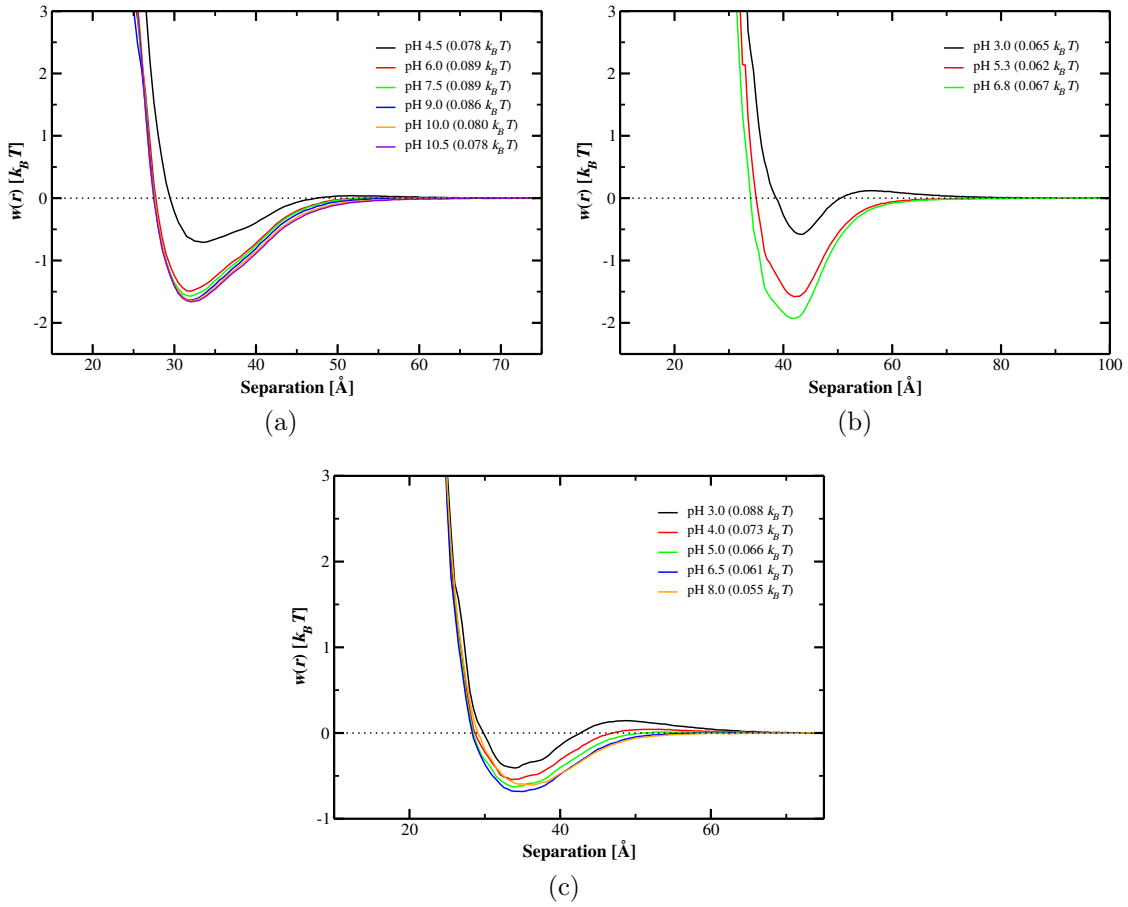


Figure 6.17: Potential of mean force as a function of the center-center separation distance for (a) two lysozymes, (b) two α-chymotrypsinogens, and (c) two ribonucleases A at different pH solutions and salt concentration of 100 mM. Data for MC simulations with ideal ε_{LJ}.

Notable differences among ideals ε_{LJ}'s (via adv) are present from one protein to another. Due to an absence of points at 5 mM, and since there are only one estimates of B_2's values for RIB at 50 mM, we have decided to focus on the estimates given at 100 mM for the three protein molecules. Deeper analysis using as a reference point the ideal ε_{LJ} of each pH condition for the three protein as a function of pI $-$ pH at 100 mM, is shown in the Figure 6.19. Initially, curves drawn are very different, showing a differential behavior among systems as already tested by others analyses. Nevertheless, an approximation is applied based on the mean value between the curves (Figure 6.19). Through this mean value a regression was calculated assuming linearity. The equation for this calculation is given as following,

$$y = 0.0015x + 0.066, \tag{6.1}$$

where x represent a determined value in units of pH at a given pI$-$pH. With this we

Table 6.7: Comparison between values of calculated and experimental second virial coefficient B_2 at different pH and salt concentrations for lysozyme, α-chymotrypsinogen, and ribonuclease A using ε_{LJ} estimated via adv.

Protein	Salt[a]	pH	$\varepsilon_{LJ}^{\text{ideal}}$ [b]	B_2^{MC} [c,d]	RMSD	$B_2^{\text{Exp.}}$ [c,e]
LYS	5	10.0	0.098	5.9 ± 0.6	0.01	5.98
		10.5	0.083	−0.2 ± 0.7	0.80	−1.06
	100	4.5	0.078	1.983 ± 0.004	0.16	2.15
		6.0	0.089	−3.01 ± 0.03	0.55	−2.46
		7.5	0.089	−3.75 ± 0.01	0.10	−3.65
		9.0	0.086	−4.57 ± 0.02	0.16	−4.41
		10.0	0.080	−5.14 ± 0.01	0.15	−5.30
		10.5	0.078	−5.51 ± 0.05	0.08	−5.60
CHY	5	6.8	0.093	−7.8 ± 0.3	0.79	−8.65
	100	3.0	0.065	2.38 ± 0.02	0.11	2.5
		5.3	0.062	−1.75 ± 0.06	0.00	−1.75[f]
		6.8	0.067	−4.20 ± 0.06	0.10	−4.10
RIB	50	3.0	0.130	3.74 ± 0.07	0.86	4.61
		4.0	0.098	3.0 ± 0.3	0.40	3.43
		5.0	0.083	1.47 ± 0.03	0.20	1.27
		6.5	0.073	−1.31 ± 0.03	0.67	−0.64
		8.0	0.063	−1.2 ± 0.2	0.53	−0.7
	100	3.0	0.088	3.47 ± 0.02	0.18	3.68
		4.0	0.073	1.824 ± 0.005	0.30	2.13
		5.0	0.066	0.6 ± 0.1	0.44	1.18
		6.5	0.061	0.0 ± 0.1	0.53	0.51
		8.0	0.055	0.1 ± 0.1	0.06	0.28

[a]Salt concentration given in mM. [b]New calibrated value of ε_{LJ} given in k_BT (adv). [c]Second virial coefficients given in mol ml $\times 10^{-4}/\text{g}^2$. [d]$B_2$ estimated by MC simulations using ideal ε_{LJ} (mean ± SD, $n = 3$) (adv). [e]Experimental data from ref. Velev *et al.,* [SLS (LYS and CHY)]; and Tessier *et al.,* [SIC (RIB)]. [f]Common interpolated value determined by ref. Velev *et al.*

can have a notion of the best ideal ε_{LJ} (y) that probably represents the experimental values in high salt concentrations for these two groups of proteins.

In Section 6.2, we performed an analysis of the discrepancies of B_2 found between experimental data. Based on this literature, we apply the same procedures given in this part of the study, and find the ideal ε_{LJ}'s for each experimental work using only LYS and CHY proteins (Figure 6.20). We see that even taking into account the B_2 reported by the literature, higher ε_{LJ} values are required on LYS in respect to CHY. Each protein exhibits a particular behavior when is far from the pI. The pH points common to each experimental work were used and an mean value between them was calculated. Again, assuming linearity we calculate a regression for LYS and CHY, which are represented by equations 6.2 and 6.3, respectively.

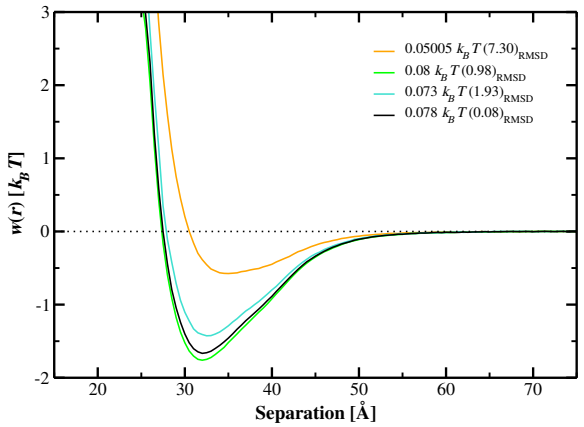


Figure 6.18: Potential of mean force as a function of the center-center separation distance between two lysozymes at pH 10.5 and 100 mM of NaCl. Black, cyan, green, and orange lines represent data for MC simulations using ε_{LJ}'s from adv, auv^{h}, auv, and Persson *et al.*, respectively. Values in parentheses are calculated RMSD's between simulated and experimental B_2's.

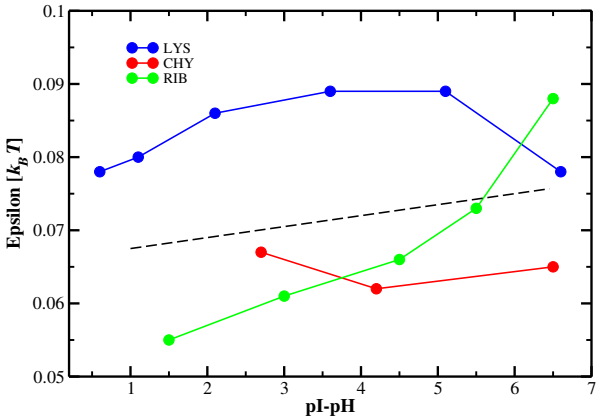
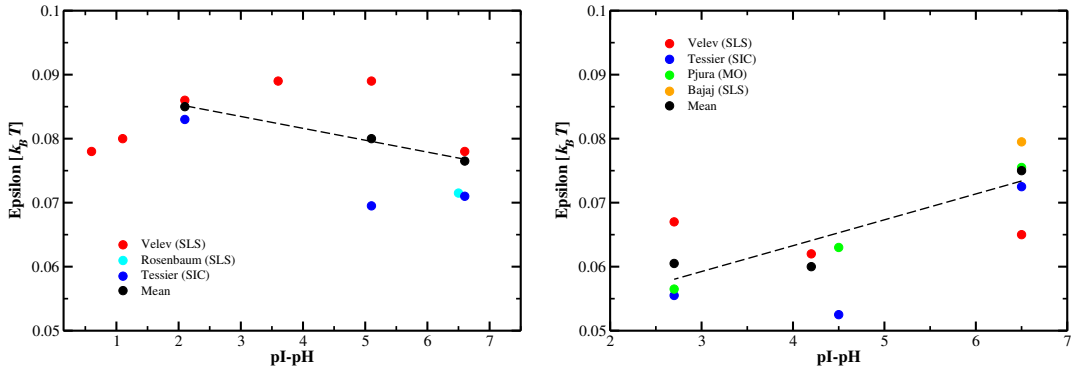


Figure 6.19: Ideals ε_{LJ}'s as a function of pI − pH for lysozyme, α-chymotrypsinogen, and ribonuclease A at 100 mM. Mean value between the curves is represented as black circles. The dashed line correspond to a linear regression from mean values.

$$y = -0.0019x + 0.089, \tag{6.2}$$

$$y = 0.0040x + 0.047, \tag{6.3}$$

Through all these analyses we propose that a single ε_{LJ} is possible, but will always be limited to one condtition. This is clear in CG models! On the other hand, one can find ideal parameter specific for a given protein. However, this would be of small practical use. Instead, we suggest a consensus strategy using the physico-chemical sense. The results obtained here for the calibration of a force field in CG models



(a) (b)

Figure 6.20: Ideals ε_{LJ}'s as a function of pI-pH from different experimental work for (a) lysozyme, and (b) α-chymotrypsinogen at 100 mM. Mean value between common points is represented as black circles. The dashed lines correspond to a linear regression from mean values.

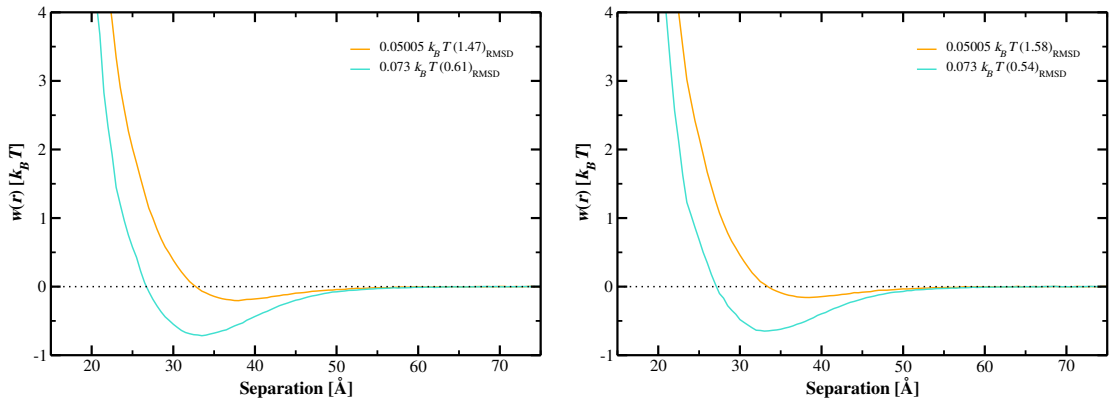
contributes to form the basis for this model and future studies. This is the first time that such systematic search for various ε_{LJ} was carried out for this model.

6.5 Molecular complexation for S100 proteins

Following the methodological order proposed in this project, we performed pmf's by MC simulations between two A3-*apo* monomers on its wild (WT) and mutant (R51A) types. As proposed by the ref. (Kizawa *et al.*[47]) the pH and salt concentrations used were, respectively, 7.5 and 150 mM (physiological concentrations). As in the previous protein systems simulations with $\varepsilon_{LJ} = 0.05005\ k_BT$ were performed (Figure 6.21). We note for WT and R51A that at a separation distance of 37 Å and 38 Å, the free energies of interaction are equal to -0.18 and $-0.17\ k_BT$, respectively. The process of dimerization with this ε_{LJ} is dominated by electrostatic repulsion. From this result and using the experimental values of $B_2^{\mathrm{WT}} = -1.7$ and $B_2^{\mathrm{R51A}} = -0.8\ \times 10^{-4}$mol ml/g^2,[47] we calculate B_2's for the two types from the $w(r)$. For WT and R51A were obtained B_2's equal to 1.882 $\pm$0.009 (RMSD = 1.46) and 0.76 $\pm$0.02 (RMSD = 1.58) $\times 10^{-4}$mol ml/g^2. With this we reiterate the need for a better fit for our CG force field.

Since with a greater electrostatic salt screening the repulsion is lower considering self-interaction, vdW forces begin to dominate even far from the pI.[102] It has been shown for certain systems that B_2 values at high salt have a tendency to converge in different pH solutions.[49, 87, 45] Using that as a principle, we apply the value

of ε_{LJ} refined by us (*i.e.* 0.073 k_BT) on A3. As we see in the Figure 6.21, the increase of this variable in the LJ's force field allowed to increase the attraction between monomers of each type. Through of $w(r)$ are observed minima of -0.7 ($r = 33$ Å) and -0.6 ($r = 33$ Å) k_BT for WT and R51A, respectively. It is verified with the calculation of B_2 a greater affinity with the experimental values: $B_2^{\mathrm{WT}} = -1.06 \pm 0.01 \times 10^{-4}$mol ml/g^2 (RMSD = 0.61), and $B_2^{\mathrm{R51A}} = -0.26 \pm 0.01 \times 10^{-4}$mol ml/g^2 (RMSD = 0.54). From a qualitative point of view, B_2's seem reasonable too. Since the charge number is closer to the 0 on WT compared to R51A at pH and salt concentrations of 7.5 and 150 mM, respectively (Table 6.1), that small change favors the dimerization process on WT type. Although the mutant type has a more heterogeneous charge distribution favoring a larger dipole, it is known that at high salt this force is diminished reducing the probability of complex formation (see experimental data for CHY in the Table 6.4, as an example).[49, 52, 46] We believe that on A3-*apo* the vdW forces are the most important contribution.



(a) (b)

Figure 6.21: Potential of mean force as a function of the center-center separation distance between two S100A3 at pH 7.5 and 150 mM of NaCl. (a) Wild and (b) mutant types are shown. Cyan and orange lines represent data for MC simulations using ε_{LJ}'s equal to 0.073 k_BT (from auv$^{\mathrm{h}}$), and 0.05005 k_BT (from Persson *et al.*), respectively. Values in parentheses are calculated RMSD's between simulated and experimental B_2's.

MC simulations using $\varepsilon_{LJ} = 0.05005$ k_BT and $\varepsilon_{LJ} = 0.073$ k_BT at pH 7.5 and salt concentration of 150 mM, were also performed on A4 and its different forms (A4$^{\mathrm{m}}$-*apo*, A4$^{\mathrm{m}}$-*holo*, and A4$^{\mathrm{d}}$-*holo*). The Figure 6.22a shows the $w(r)$ for all A4 forms studied here with a $\varepsilon_{LJ} = 0.05005$ k_BT. In summary, we see that systems are governed by repulsive forces. This is contradictory, since the formation of complexes on A4 have already been reported by different experimental techniques.[42, 43, 55, 71] Increasing the attraction in the LJ potential (*i.e.* 0.073 k_BT), a more realistic and

probable description of what happens in each of the forms is observable. Minima with different free energies of interaction were detected for A4^{m}-*apo*, A4^{m}-*holo*, and A4^{d}-*holo*: -0.73 ($r = 35$ Å), -0.96 ($r = 32$ Å), and -1.46 ($r = 39$ Å) k_BT, respectively (Figure 6.22b). The use of a $\varepsilon_{LJ} = 0.073\ k_BT$ seems reasonable for A4. Taking advantage of the fact A4^{m}-*holo* presents similar features in comparison with proteins used for the refinement of the force field, we estimate the distance to pI$^{A4^{m}-holo}$ (9.7) when the pH is equal to 7.5 (*i.e.* 2.2). By replacement in equation 6.1 we would get an ε_{LJ} ideal about of 0.07 k_BT. This value is very close to the estimated via auv^{h}, so we believe that these estimates of the free energy of interaction are reliable and allow the proper characterization of the system.

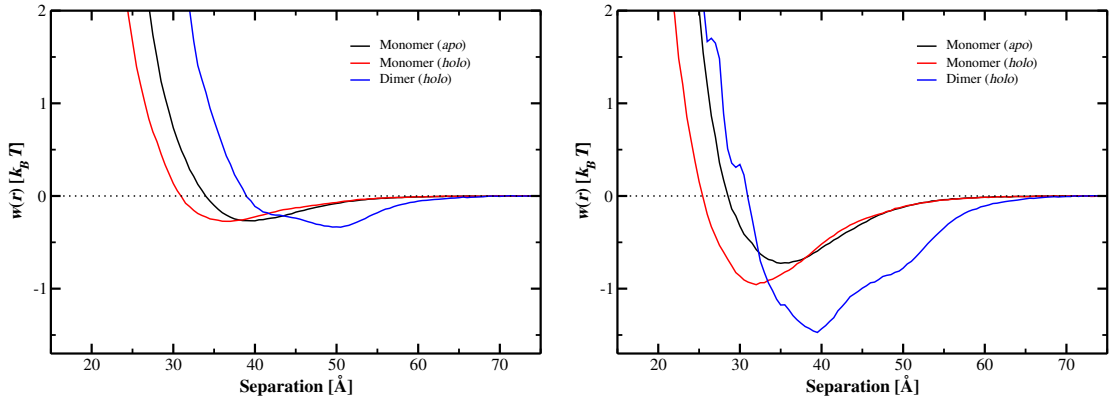


Figure 6.22: Potential of mean force as a function of the center-center separation distance between two S100A4 in its different forms at pH 7.5 and 150 mM of NaCl. MC simulations using ε_{LJ}'s equal to (a) 0.05005 k_BT (from Persson *et al.*), and (b) 0.073 k_BT (from auv^{h}).

A4 monomers form unstable complexes in the absence of Ca^{2+}, since their physico-chemical features do not favor greater stability as seen in the Table 6.1. However, more stable complexes are formed between monomers in the presence of Ca^{2+}. Highlighting the hypothesis of "abrupt" rearrangement of the structural surface of the A4 protein in presence of calcium.[70, 78] The new folding allows hydrophobic residues to be more exposed on the surface and, therefore, vdW forces act more easily. It is interesting from the point of view of the mesoscopic model, that in spite of having a reduction in the degrees of freedom, this feature is distinguishable. Probably, for A4^{m}-*holo* under physiological conditions the second most important force is the dipole.

Finally, very attractive forces are observed on A4 dimers in the presence of calcium. We assume that this may be due to a prevalence of hydrophobic *aa* on the surface

and to stable configurations. This can be confirmed seeing in the Figure 6.22b the roughness in the pmf at a distance relatively equal to σ (= 24.5). We see that in terms of thermodynamic stability, the dimeric form is the most stable among the different forms that A4 can adopt. However, as shown by ref. (Abdali *et al.*[43]), it may be that the probability of forming dimers in the presence of calcium is more frequent compared to the formation of tetramers, since it is likely that a contribution of dipolar interaction in the monomeric form was higher (Table 6.1). In addition, it is likely that the influence of the hydrophobic effect was higher in $A4^{m}$-*holo* compared to $A4^{d}$-*holo*, since its percentages of the solvent-accessible surface area of hydrophobic *aa* are 36% and 27%, respectively[7]. These contributions should significantly affect to the dimerization process of the monomeric forms with Ca^{2+} resulting in a greater chance for their complexation.

[7]Percentages was calculated using VADAR program.[143]

Chapter 7

Conclusions

- The behavior of LYS, CHY, RIB, A3, and A4 proteins in electrolyte solutions was evaluated. We observed that each protein responds different and specifically to the pH and salt conditions. Generally, LYS, CHY, RIB, A4^{m}-*holo*, and A4^{d}-*holo* are cationic, and the two types of A3 (*i.e.* WT and R51A) and A4^{m}-*apo* are anionic, classification given in function of pI. Regarding the proteins used for the force field evaluation, we highlight that the dipolar moment in CHY and RIB may play an important role in the complexation phenomena. vdW forces have a significant influence on the dimerization process, and there is a correlation between this type of interaction and the number of hydrophobic residues exposed on the surface of proteins as expected.

- An analysis of the published data of B_2 for LYS and CHY in electrolyte solution allowed us to detect inherent discrepancies between a series of experimental works involving different techniques (it should not be ignored that these remarkable differences were also reported in other studies). The origin of these discrepancies between B_2 values is multifactorial. However, the semi-quantitative character of B_2 is relatively preserved between each one of the authors.

- We confirm that a unique description of force field parameters for CG models at all salt and pH conditions is unlikely possible. In particular, larger problems were found at low salt concentrations (as observed in the laboratory experiments) and away from the pI. Because repulsive electrostatic interactions dominate the self-interaction, reliable values are not obtained. On the other hand, by restricting of physical-chemistry conditions at high salt concentrations a more reliable description is given for the force field parameters. This is due to a combination of effects including the salt screening that allows vdW forces to dominate. Since differences between

experimental data prevail, we explored for each particular case an ideal ε_{LJ} as a function of pI-pH, and found a pattern that could be used as an adjustable parameter to incorporate all phenomena not explicitly described in the effective Hamiltonian. This semi-quantitative approach provides a guided for future studies with other proteins.

- MC simulations using S100 proteins involved in cancer were performed using the new ε_{LJ} determined in the present refinement process. With this new value, a more probable representation of the dimerization process of these S100 proteins was found. With this results we highlight that a sum of interactions (electrostatics and vdW) dominate in these proteins. Furthermore, presence of calcium optimizes the processes of dimerization and tetramerization. Tetramer formation of A4 in presence of calcium is thermodynamically more stable than the other two oligomeric forms studied here. This could serve as a basis in future pharmacological analyses that aim to find clinical targets which may prevent complex formation in these proteins.

Bibliography

[1] Chang, R. *Physical chemistry for the chemical and biological sciences*. University Science Books, 2000.

[2] Branden, C. I. and Tooze, J. *Introduction to protein structure*. Garland Science, 1999.

[3] Gregory, R. *Protein-solvent interactions*. CRC Press, 1995.

[4] Prabhu, N. and Sharp, K. Protein-solvent interactions. *Chemical Reviews*, 106(5):1616–1623, 2006.

[5] Jönsson, B., Lund, M., and Da Silva, F. L. B. Electrostatics in macromolecular solutions. In *Conference on Food Colloids 2006*, volume 302, pages 129–154. Royal Society of Chemistry, 2007.

[6] Harris, T. K. and Turner, G. J. Structural basis of perturbed pK_a values of catalytic groups in enzyme active sites. *IUBMB Life*, 53(2):85–98, 2002.

[7] Perutz, M. Electrostatic effects in proteins. *Science*, 201(4362):1187–1191, 1978.

[8] Da Silva, F. L. B. *Statistical Mechanical Studies of Aqueous Solutions and Biomolecular Systems*. PhD thesis, University of Lund, 2000.

[9] McQuarrie, D. A. and Simon, J. D. *Physical chemistry: a molecular approach*, volume 1. Sterling Publishing Company, 1997.

[10] Israelachvili, J. N. *Intermolecular and surface forces: revised third edition*. Academic Press, 2011.

[11] Stone, A. *The theory of intermolecular forces*. Oxford University Press, 2013.

[12] Aziz, M. F., Caetano-Anollés, K., and Caetano-Anollés, G. The early history and emergence of molecular functions and modular scale-free network behavior. *Scientific Reports*, 6, 2016.

[13] Jones, S. and Thornton, J. M. Principles of protein-protein interactions. *Proceedings of the National Academy of Sciences*, 93(1):13–20, 1996.

[14] Reva, B., Antipin, Y., and Sander, C. Determinants of protein function revealed by combinatorial entropy optimization. *Genome Biology*, 8(11):R232, 2007.

[15] Cox, M. M. and Nelson, D. L. *Lehninger principles of biochemistry.* New York, Worth, 2000.

[16] Jones, S. and Thornton, J. M. Protein-protein interactions: a review of protein dimer structures. *Progress in Biophysics and Molecular Biology*, 63(1):31–65, 1995.

[17] Keskin, O., Gursoy, A., Ma, B., and Nussinov, R. Principles of protein-protein interactions: what are the preferred ways for proteins to interact? *Chemical Reviews*, 108(4):1225–1244, 2008.

[18] Milroy, L. G., Grossmann, T. N., Hennig, S., Brunsveld, L., and Ottmann, C. Modulators of protein–protein interactions. *Chemical Reviews*, 114(9):4695–4748, 2014.

[19] Tavares, F., Bratko, D., Striolo, A., Blanch, H., and Prausnitz, J. Phase behavior of aqueous solutions containing dipolar proteins from second-order perturbation theory. *The Journal of Chemical Physics*, 120(20):9859–9869, 2004.

[20] Lund, M. *Electrostatic Interactions in and Between Biomolecules.* PhD thesis, University of Lund, 2006.

[21] Lund, M. and Jönsson, B. Driving forces behind ion-ion correlations. *Journal of Chemical Physics*, 125(23):6101, 2006.

[22] Kirkwood, J. G. and Shumaker, J. B. Forces between protein molecules in solution arising from fluctuations in proton charge and configuration. *Proceedings of the National Academy of Sciences of the United States of America*, 38(10):863–871, 1952.

[23] Lund, M. and Jönsson, B. On the charge regulation of proteins. *Biochemistry*, 44(15):5722–5727, 2005.

[24] Lund, M. and Jönsson, B. Charge regulation in biomolecular solution. *Quarterly Reviews of Biophysics*, 46(03):265–281, 2013.

[25] Da Silva, F. L. B. and Jönsson, B. Polyelectrolyte–protein complexation driven by charge regulation. *Soft Matter*, 5(15):2862–2868, 2009.

[26] Teixeira, A. A. R., Lund, M., and Da Silva, F. L. B. Fast proton titration scheme for multiscale modeling of protein solutions. *Journal of Chemical Theory and Computation*, 6(10):3259–3266, 2010.

[27] Da Silva, F. L. B., Jönsson, B., and Penfold, R. A critical investigation of the Tanford-Kirkwood shceme by means of Monte Carlo simulations. *Protein Science*, 10(7):1415–1425, 2001.

[28] Da Silva, F. L. B., Lund, M., Jönsson, B., and Åkesson, T. On the complexation of proteins and polyelectrolytes. *The Journal of Physical Chemistry B*, 110(9):4459–4464, 2006.

[29] Karplus, M. and McCammon, J. A. Molecular dynamics simulations of biomolecules. *Nature Structural & Molecular Biology*, 9(9):646–652, 2002.

[30] Yuriev, E., Agostino, M., and Ramsland, P. A. Challenges and advances in computational docking: 2009 in review. *Journal of Molecular Recognition*, 24(2):149–164, 2011.

[31] Bashford, D. and Gerwert, K. Electrostatic calculations of the pK_a values of ionizable groups in Bacteriorhodopsin. *Journal of Molecular Biology*, 224(2):473–486, 1992.

[32] De Vries, R. and Stuart, M. C. Theory and simulations of macroion complexation. *Current Opinion in Colloid & Interface Science*, 11(5):295–301, 2006.

[33] Lund, M. and Jönsson, B. A mesoscopic model for protein-protein interactions in solution. *Biophysical Journal*, 85(5):2940–2947, 2003.

[34] Persson, B. A., Lund, M., Forsman, J., Chatterton, D. E., and Åkesson, T. Molecular evidence of stereo-specific Lactoferrin dimers in solution. *Biophysical Chemistry*, 151(3):187–189, 2010.

[35] Kurut, A., Persson, B. A., Åkesson, T., Forsman, J., and Lund, M. Anisotropic interactions in protein mixtures: self assembly and phase behavior in aqueous solution. *The Journal of Physical Chemistry Letters*, 3(6):731–734, 2012.

[36] Voth, G. A. *Coarse-graining of condensed phase and biomolecular systems*. CRC Press, 2008.

[37] Sedlmeier, F., Horinek, D., and Netz, R. R. Spatial correlations of density and structural fluctuations in liquid water: A comparative simulation study. *Journal of the American Chemical Society*, 133(5):1391–1398, 2011.

[38] Luscombe, N. M., Greenbaum, D., and Gerstein, M. What is bioinformatics? A proposed definition and overview of the field. *Methods of Information in Medicine*, 40(4):346–358, 2001.

[39] Samish, I., Bourne, P. E., and Najmanovich, R. J. Achievements and challenges in structural bioinformatics and computational biophysics. *Bioinformatics*, 31(1):146–150, 2015.

[40] Zimmer, D. B., Cornwall, E. H., Landar, A., and Song, W. The S100 protein family: history, function, and expression. *Brain Research Bulletin*, 37(4):417–429, 1995.

[41] Semov, A., Moreno, M. J., Onichtchenko, A., Abulrob, A., Ball, M., Ekiel, I., Pietrzynski, G., Stanimirovic, D., and Alakhov, V. Metastasis-associated protein S100A4 induces angiogenesis through interaction with Annexin II and accelerated plasmin formation. *Journal of Biological Chemistry*, 280(21):20833–20841, 2005.

[42] Garrett, S. C., Varney, K. M., Weber, D. J., and Bresnick, A. R. S100A4, a mediator of metastasis. *Journal of Biological Chemistry*, 281(2):677–680, 2006.

[43] Abdali, S., De Laere, B., Poulsen, M., Grigorian, M., Lukanidin, E., and Klingelhöfer, J. Toward methodology for detection of cancer-promoting S100A4 protein conformations in subnanomolar concentrations using Raman and SERS. *The Journal of Physical Chemistry C*, 114(16):7274–7279, 2010.

[44] Delboni, L. A. and Da Silva, F. L. B. On the complexation of whey proteins. *Food Hydrocolloids*, 55:89–99, 2016.

[45] Li, W., Persson, B. A., Morin, M., Behrens, M. A., Lund, M., and Zackrisson Oskolkova, M. Charge-induced patchy attractions between proteins. *The Journal of Physical Chemistry B*, 119(2):503–508, 2015.

[46] Da Silva, F. L. B., Pasquali, S., Derreumaux, P., and Dias, L. G. Electrostatics analysis of the mutational and pH effects of the N-terminal domain self-association of the major ampullate Spidroin. *Soft Matter*, 12(25):5600–5612, 2016.

[47] Kizawa, K., Jinbo, Y., Inoue, T., Takahara, H., Unno, M., Heizmann, C. W., and Izumi, Y. Human S100A3 tetramerization propagates Ca^{2+}/ Zn^{2+} binding states. *Biochimica et Biophysica Acta (BBA)-Molecular Cell Research*, 1833(7):1712–1719, 2013.

[48] Unno, M., Kawasaki, T., Takahara, H., Heizmann, C. W., and Kizawa, K. Refined crystal structures of human Ca^{2+}/ Zn^{2+} binding S100A3 protein characterized by two disulfide bridges. *Journal of Molecular Biology*, 408:477–490, 2011.

[49] Velev, O., Kaler, E., and Lenhoff, A. Protein interactions in solution characterized by light and neutron scattering: comparison of Lysozyme and Chymotrypsinogen. *Biophysical Journal*, 75(6):2682–2697, 1998.

[50] Pellicane, G., Smith, G., and Sarkisov, L. Molecular dynamics characterization of protein crystal contacts in aqueous solutions. *Physical Review Letters*, 101(24):248102, 2008.

[51] Coen, C., Blanch, H., and Prausnitz, J. Salting out of aqueous proteins: phase equilibria and intermolecular potentials. *AIChE Journal*, 41(4):996–1004, 1995.

[52] Neal, B., Asthagiri, D., and Lenhoff, A. Molecular origins of osmotic second virial coefficients of proteins. *Biophysical Journal*, 75(5):2469–2477, 1998.

[53] Bashford, D. and Karplus, M. pK_a's of ionizable groups in proteins: atomic detail from a continuum electrostatic model. *Biochemistry*, 29(44):10219–10225, 1990.

[54] Tessier, P. M., Johnson, H. R., Pazhianur, R., Berger, B. W., Prentice, J. L., Bahnson, B. J., Sandler, S. I., and Lenhoff, A. M. Predictive crystallization of Ribonuclease A via rapid screening of osmotic second virial coefficients. *Proteins: Structure, Function, and Bioinformatics*, 50(2):303–311, 2003.

[55] Klingelhöfer, J., Grum-Schwensen, B., Beck, M. K., Knudsen, R. S. P., Grigorian, M., Lukanidin, E., and Ambartsumian, N. Anti-S100A4 antibody suppresses metastasis formation by blocking stroma cell invasion. *Neoplasia*, 14(12):1260–1268, 2012.

[56] Lodish, H., Berk, A., Zipursky, S. L., Matsudaira, P., Baltimore, D., and Darnell, J. *Molecular cell biology*, volume 4. WH Freeman New York, 2000.

[57] Byers, T. Two decades of declining cancer mortality: progress with disparity. *Annual Review of Public Health*, 31:121–132, 2010.

[58] Weinberg, R. *The biology of cancer*. Garland science, 2013.

[59] Moore, B. W. A soluble protein characteristic of the nervous system. *Biochemical and Biophysical Research Communications*, 19(6):739–744, 1965.

[60] Schäfer, B. W. and Heizmann, C. W. The S100 family of EF-hand calcium-binding proteins: functions and pathology. *Trends in Biochemical Sciences*, 21(4):134–140, 1996.

[61] Tarabykina, S., L Griffiths, T., Tulchinsky, E., Mellon, J., Bronstein, I., and Kriajevska, M. Metastasis-associated protein S100A4: spotlight on its role in cell migration. *Current Cancer Drug Targets*, 7(3):217–228, 2007.

[62] Chen, H., Xu, C., Jin, Q., and Liu, Z. S100 protein family in human cancer. *American Journal of Cancer Research*, 4(2):89–115, 2014.

[63] Donato, R. S100: a multigenic family of calcium-modulated proteins of the EF-hand type with intracellular and extracellular functional roles. *The International Journal of Biochemistry & Cell Biology*, 33(7):637–668, 2001.

[64] Pedersen, K., Nesland, J., Fodstad, Ø., and Maelandsmo, G. Expression of S100A4, E-cadherin, α- and β-catenin in breast cancer biopsies. *British Journal of Cancer*, 87(11):1281–1286, 2002.

[65] Rosty, C., Ueki, T., Argani, P., Jansen, M., Yeo, C. J., Cameron, J. L., Hruban, R. H., and Goggins, M. Overexpression of S100A4 in pancreatic ductal adenocarcinomas is associated with poor differentiation and dna hypomethylation. *The American Journal of Pathology*, 160(1):45–50, 2002.

[66] Gupta, S., Hussain, T., MacLennan, G. T., Fu, P., Patel, J., and Mukhtar, H. Differential expression of S100A2 and S100A4 during progression of human prostate adenocarcinoma. *Journal of Clinical Oncology*, 21(1):106–112, 2003.

[67] Kimura, K., Endo, Y., Yonemura, Y., Heizmann, C., Schafer, B., Watanabe, Y., and Sasaki, T. Clinical significance of S100A4 and E-cadherin-related adhesion molecules in non-small cell lung cancer. *International Journal of Oncology*, 16(6):1125–1156, 2000.

[68] Nakamura, T., Ajiki, T., Murao, S., Kamigaki, T., Maeda, S., Ku, Y., and Kuroda, Y. Prognostic significance of S100A4 expression in gallbladder cancer. *International Journal of Oncology*, 20(5):937–941, 2002.

[69] Kretsinger, R. H. and Nockolds, C. E. Carp muscle calcium-binding protein II. Structure determination and general description. *Journal of Biological Chemistry*, 248(9):3313–3326, 1973.

[70] Pathuri, P., Vogeley, L., and Luecke, H. Crystal structure of metastasis-associated protein S100A4 in the active calcium-bound form. *Journal of Molecular Biology*, 383(1):62–77, 2008.

[71] Gingras, A. R., Basran, J., Prescott, A., Kriajevska, M., Bagshaw, C. R., and Barsukov, I. L. Crystal structure of the Ca^{2+}-form and Ca^{2+}-binding kinetics of metastasis-associated protein S100A4. *FEBS letters*, 582(12):1651–1656, 2008.

[72] Duelli, A., Kiss, B., Lundholm, I., Bodor, A., Petoukhov, M. V., Svergun, D. I., Nyitray, L., and Katona, G. The C-terminal random coil region tunes the Ca^{2+}-binding affinity of S100A4 through conformational activation. *PloS One*, 9(5):e97654, 2014.

[73] Boye, K. and Mælandsmo, G. M. S100A4 and metastasis: a small actor playing many roles. *The American Journal of Pathology*, 176(2):528–535, 2010.

[74] Klingelhöfer, J., Šenolt, L., Baslund, B., Nielsen, G. H., Skibshøj, I., Pavelka, K., Neidhart, M., Gay, S., Ambartsumian, N., and Hansen, B. S. Up-regulation of metastasis-promoting S100A4 (mts-1) in rheumatoid arthritis: Putative involvement in the pathogenesis of rheumatoid arthritis. *Arthritis & Rheumatism*, 56(3):779–789, 2007.

[75] Wang, G., Rudland, P. S., White, M. R., and Barraclough, R. Interaction in vivo and in vitro of the metastasis-inducing S100 protein, S100A4 (p9ka) with S100A1. *Journal of Biological Chemistry*, 275(15):11141–11146, 2000.

[76] Koshelev, Y. A., Kiselev, S., and Georgiev, G. Interaction of the S100A4 (mts1) protein with septins Sept2, Sept6, and Sept7 in vitro. *Doklady Biochemistry and Biophysics*, 391(1):195–197, 2003.

[77] Kriajevska, M., Fischer-Larsen, M., Moertz, E., Vorm, O., Tulchinsky, E., Grigorian, M., Ambartsumian, N., and Lukanidin, E. Liprin β1, a member of the family of LAR transmembrane tyrosine phosphatase-interacting proteins, is a new target for the metastasis-associated protein S100A4 (mts1). *Journal of Biological Chemistry*, 277(7):5229–5235, 2002.

[78] Santamaria-Kisiel, L., Rintala-Dempsey, A. C., and Shaw, G. S. Calcium-dependent and -independent interactions of the S100 protein family. *Biochemical Journal*, 396(2):201–214, 2006.

[79] Juffer, A. H. and Vogel, H. J. pK_a calculations of Calbindin D_{9k}: Effects of Ca^{2+} binding, protein dielectric constant, and ionic strength. *Proteins: Structure, Function, and Bioinformatics*, 41(4):554–567, 2000.

[80] Kesvatera, T., Jönsson, B., Thulin, E., and Linse, S. Binding of Ca^{2+} to Calbindin D_{9k}: structural stability and function at high salt concentration. *Biochemistry*, 33(47):14170–14176, 1994.

[81] Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T., Weissig, H., Shindyalov, I. N., and Bourne, P. E. The protein data bank. *Nucleic Acids Research*, 28(1):235–242, 2000.

[82] Jollès, P. and Jollès, J. What's new in Lysozyme research? *Molecular and Cellular Biochemistry*, 63(2):165–189, 1984.

[83] Callewaert, L. and Michiels, C. W. Lysozymes in the animal kingdom. *Journal of Biosciences*, 35(1):127–160, 2010.

[84] Chang, K. Y. and Carr, C. W. Studies on the structure and function of Lysozyme: I. The effect of pH and cation concentration on Lysozyme activity. *Biochimica et Biophysica Acta (BBA)-Protein Structure*, 229(2):496–503, 1971.

[85] George, A. and Wilson, W. W. Predicting protein crystallization from a dilute solution property. *Acta Crystallographica Section D: Biological Crystallography*, 50(4):361–365, 1994.

[86] Ruppert, S., Sandler, S., and Lenhoff, A. Correlation between the osmotic second virial coefficient and the solubility of proteins. *Biotechnology Progress*, 17(1):182–187, 2001.

[87] Quigley, A. and Williams, D. The second virial coefficient as a predictor of protein aggregation propensity: a self-interaction chromatography study. *European Journal of Pharmaceutics and Biopharmaceutics*, 96:282–290, 2015.

[88] Ramanadham, M., Sieker, L., and Jensen, L. Refinement of triclinic Lysozyme: II. the method of stereochemically restrained least squares. *Acta Crystallographica Section B: Structural Science*, 46(1):63–69, 1990.

[89] Moon, Y. U., Curtis, R. A., Anderson, C. O., Blanch, H. W., and Prausnitz, J. M. Protein–protein interactions in aqueous ammonium sulfate solutions. Lysozyme and Bovine Serum Albumin (BSA). *Journal of Solution Chemistry*, 29(8):699–718, 2000.

[90] Freer, S., Kraut, J., Robertus, J., and Wright, H. T. Chymotrypsinogen: 2.5 Å crystal structure, comparison with α-chymotrypsin, and implications for zymogen activation. *Biochemistry*, 9(9):1997–2009, 1970.

[91] Neurath, H. and Walsh, K. A. Role of proteolytic enzymes in biological regulation (a review). *Proceedings of the National Academy of Sciences*, 73(11):3825–3832, 1976.

[92] Nogués, M. V., Vilanova, M., and Cuchillo, C. M. Bovine pancreatic Ribonuclease A as a model of an enzyme with multiple substrate binding sites. *Biochimica et Biophysica Acta (BBA)-Protein Structure and Molecular Enzymology*, 1253(1):16–24, 1995.

[93] Berisio, R., Sica, F., Lamzin, V., Wilson, K., Zagari, A., and Mazzarella, L. Atomic resolution structures of Ribonuclease A at six pH values. *Acta Crystallographica Section D: Biological Crystallography*, 58(3):441–450, 2002.

[94] Marshall, G. R., Feng, J. A., and Kuster, D. J. Back to the future: Ribonuclease A. *Peptide Science*, 90(3):259–277, 2008.

[95] Rutkoski, T. J., Kurten, E. L., Mitchell, J. C., and Raines, R. T. Disruption of shape-complementarity markers to create cytotoxic variants of Ribonuclease A. *Journal of Molecular Biology*, 354(1):41–54, 2005.

[96] Donato, R., Cannon, B., Sorci, G., Riuzzi, F., Hsu, K., Weber, D., and Geczy, C. Functions of S100 proteins. *Current Molecular Medicine*, 13(1):24, 2013.

[97] Takizawa, T., Takizawa, T., Arai, S., Kizawa, K., Uchiwa, H., Sasaki, I., and Inoue, T. Ultrastructural localization of S100A3, a cysteine-rich, calcium binding protein, in human scalp hair shafts revealed by rapid-freezing immunocytochemistry. *Journal of Histochemistry & Cytochemistry*, 47(4):525–532, 1999.

[98] Kizawa, K., Inoue, T., Yamaguchi, M., Kleinert, P., Troxler, H., Heizmann, C. W., and Iwamoto, Y. Dissimilar effect of perming and bleaching treatments on cuticles: advanced hair damage model based on elution and oxidation of S100A3 protein. *International Journal of Cosmetic Science*, 27(6):355–355, 2005.

[99] Liu, B., Sun, W.-Y., Zhi, C.-Y., Lu, T.-C., Gao, H.-M., Zhou, J.-H., Yan, W.-Q., and Gao, H.-C. Role of S100A3 in human colorectal cancer and the anticancer effect of cantharidinate. *Experimental and Therapeutic Medicine*, 6(6):1499–1503, 2013.

[100] Selkoe, D. J. Folding proteins in fatal ways. *Nature*, 426(6968):900–904, 2003.

[101] Dobson, C. M. Protein folding and misfolding. *Nature*, 426(6968):884–890, 2003.

[102] Russel, W. B., Saville, D. A., and Schowalter, W. R. *Colloidal dispersions*. Cambridge University Press, 1992.

[103] Zacharias, M. *Protein-protein complexes: Analysis, modeling and drug design*. World Scientific, 2010.

[104] Friedman, H. L. Electrolyte solutions at equilibrium. *Annual Review of Physical Chemistry*, 32(1):179–204, 1981.

[105] McMillan Jr, W. G. and Mayer, J. E. The statistical thermodynamics of multicomponent systems. *The Journal of Chemical Physics*, 13(7):276–305, 1945.

[106] Ponder, J. W. and Case, D. A. Force fields for protein simulations. *Advances in Protein Chemistry*, 66:27–86, 2003.

[107] Oostenbrink, C., Villa, A., Mark, A. E., and Van Gunsteren, W. F. A biomolecular force field based on the free enthalpy of hydration and solvation: the GROMOS force-field parameter sets 53a5 and 53a6. *Journal of Computational Chemistry*, 25(13):1656–1676, 2004.

[108] MacKerell Jr, A. D., Bashford, D., Bellott, M., Dunbrack Jr, R. L., Evanseck, J. D., Field, M. J., Fischer, S., Gao, J., Guo, H., and Ha, S. All-atom empirical potential for molecular modeling and dynamics studies of proteins. *The Journal of Physical Chemistry B*, 102(18):3586–3616, 1998.

[109] Jorgensen, W. L., Maxwell, D. S., and Tirado-Rives, J. Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. *Journal of the American Chemical Society*, 118(45):11225–11236, 1996.

[110] Cornell, W. D., Cieplak, P., Bayly, C. I., Gould, I. R., Merz, K. M., Ferguson, D. M., Spellmeyer, D. C., Fox, T., Caldwell, J. W., and Kollman, P. A. A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *Journal of the American Chemical Society*, 117(19):5179–5197, 1995.

[111] Huang, J. and MacKerell, A. D. CHARMM36 all-atom additive protein force field: Validation based on comparison to NMR data. *Journal of Computational Chemistry*, 34(25):2135–2145, 2013.

[112] Tanford, C. and Kirkwood, J. G. Theory of protein titration curves. I. General equations for impenetrable spheres. *Journal of the American Chemical Society*, 79(20):5333–5339, 1957.

[113] Duncan, J. S. *Introduction to colloid and surface chemistry.* Butterworth Heinemann: New York, 1992.

[114] McQuarrie, D. A. *Statistical Mechanics.* Harper Collins, New York, 1976.

[115] Cramer, C. J. *Essentials of computational chemistry: theories and models.* John Wiley & Sons, 2013.

[116] Der Spoel, D., Lindahl, E., and Hess, B. Gromacs user manual version 4.6.7, 2014.

[117] Allen, M. P. and Tildesley, D. J. *Computer simulation of liquids.* Oxford University Press, 1989.

[118] Satoh, A. *Introduction to practice of molecular simulation: molecular dynamics, Monte Carlo, Brownian dynamics, Lattice Boltzmann and dissipative particle dynamics*. Elsevier, 2010.

[119] De Carvalho, S. J. *Estudo dos aspectos eletrostáticos da interação entre polieletrólitos macroíons*. PhD thesis, Universidade Estadual Paulista, 2008.

[120] Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6):1087–1092, 1953.

[121] Tessier, P. M., Lenhoff, A. M., and Sandler, S. I. Rapid measurement of protein osmotic second virial coefficients by self-interaction chromatography. *Biophysical Journal*, 82(3):1620–1631, 2002.

[122] Bloustine, J., Berejnov, V., and Fraden, S. Measurements of protein-protein interactions by size exclusion chromatography. *Biophysical Journal*, 85(4):2619–2623, 2003.

[123] Neal, B., Asthagiri, D., Velev, O., Lenhoff, A., and Kaler, E. Why is the osmotic second virial coefficient related to protein crystallization? *Journal of Crystal Growth*, 196(2):377–387, 1999.

[124] Tessier, P. M. and Lenhoff, A. M. Measurements of protein self-association as a guide to crystallization. *Current Opinion in Biotechnology*, 14(5):512–516, 2003.

[125] Dumetz, A. C., Snellinger-O'Brien, A. M., Kaler, E. W., and Lenhoff, A. M. Patterns of protein–protein interactions in salt solutions and implications for protein crystallization. *Protein Science*, 16(9):1867–1877, 2007.

[126] Kalos, M. H. and Whitlock, P. A. *Monte Carlo methods*. John Wiley & Sons, 2008.

[127] Lund, M., Trulsson, M., and Persson, B. Faunus an object oriented framework for molecular simulation. *Source Code for Biology and Medicine*, 3(1):1, 2008.

[128] Kurut, A., Dicko, C., and Lund, M. Dimerization of terminal domains in spiders silk proteins is controlled by electrostatic anisotropy and modulated by hydrophobic patches. *ACS Biomaterials Science & Engineering*, 1(6):363–371, 2015.

[129] Webb, B. and Sali, A. Comparative protein structure modeling using modeller. *Current Protocols in Bioinformatics*, pages 5–6, 2014.

[130] Nozaki, Y. and Tanford, C. Examination of titratation behavior. *Methods Enzymol*, 11:715–734, 1967.

[131] Rosenbaum, D. and Zukoski, C. Protein interactions and crystallization. *Journal of Crystal Growth*, 169(4):752–758, 1996.

[132] Pjura, P., Lenhoff, A., Leonard, S., and Gittis, A. Protein crystallization by design: Chymotrypsinogen without precipitants. *Journal of Molecular Biology*, 300(2):235–239, 2000.

[133] Bajaj, H., Sharma, V. K., and Kalonia, D. S. Determination of second virial coefficient of proteins using a dual-detector cell for simultaneous measurement of scattered light intensity and concentration in SEC-HPLC. *Biophysical Journal*, 87(6):4048–4055, 2004.

[134] Boström, M., Tavares, F. W., Bratko, D., and Ninham, B. Specific ion effects in solutions of globular proteins: comparison between analytical models and simulation. *The Journal of Physical Chemistry B*, 109(51):24489–24494, 2005.

[135] Kastelic, M., Kalyuzhnyi, Y. V., Hribar-Lee, B., Dill, K. A., and Vlachy, V. Protein aggregation in salt solutions. *Proceedings of the National Academy of Sciences*, 112(21):6766–6770, 2015.

[136] Kuehner, D. E., Engmann, J., Fergg, F., Wernick, M., Blanch, H. W., and Prausnitz, J. M. Lysozyme net charge and ion binding in concentrated aqueous electrolyte solutions. *The Journal of Physical Chemistry B*, 103(8):1368–1374, 1999.

[137] Medda, L., Barse, B., Cugia, F., Boström, M., Parsons, D. F., Ninham, B. W., Monduzzi, M., and Salis, A. Hofmeister challenges: ion binding and charge of the BSA protein as explicit examples. *Langmuir*, 28(47):16355–16363, 2012.

[138] Tanford, C. and Hauenstein, J. D. Hydrogen ion equilibria of Ribonuclease. *Journal of the American Chemical Society*, 78(20):5287–5291, 1956.

[139] Kizawa, K., Troxler, H., Kleinert, P., Inoue, T., Toyoda, M., Morohashi, M., and Heizmann, C. W. Characterization of the cysteine-rich calcium-binding S100A3 protein from human hair cuticles. *Biochemical and Biophysical Research Communications*, 299(5):857–862, 2002.

[140] Haugen, M. H., Flatmark, K., Mikalsen, S.-O., and Malandsmo, G. M. The metastasis-associated protein S100A4 exists in several charged variants suggesting the presence of posttranslational modifications. *BMC Cancer*, 8(1):172, 2008.

[141] Ahamed, T., Ottens, M., Dedem, G. W., and Wielen, L. A. Design of self-interaction chromatography as an analytical tool for predicting protein phase behavior. *Journal of Chromatography A*, 1089(1):111–124, 2005.

[142] Wilson, W. W. and DeLucas, L. J. Applications of the second virial coefficient: protein crystallization and solubility. *Acta Crystallographica Section F: Structural Biology Communications*, 70(5):543–554, 2014.

[143] Willard, L., Ranjan, A., Zhang, H., Monzavi, H., Boyko, R. F., Sykes, B. D., and Wishart, D. S. VADAR: a web server for quantitative evaluation of protein structure quality. *Nucleic Acids Research*, 31(13):3316–3319, 2003.

[144] Booth, D. R., Sunde, M., Bellotti, V., and Robinson, C. V. Instability, unfolding and aggregation of human Lysozyme variants underlying amyloid fibrillogenesis. *Nature*, 385(6619):787, 1997.

[145] Chi, E. Y., Krishnan, S., Randolph, T. W., and Carpenter, J. F. Physical stability of proteins in aqueous solution: mechanism and driving forces in nonnative protein aggregation. *Pharmaceutical Research*, 20(9):1325–1336, 2003.

[146] Stradner, A., Sedgwick, H., Cardinaux, F., Poon, W. C., Egelhaaf, S. U., and Schurtenberger, P. Equilibrium cluster formation in concentrated protein solutions and colloids. *Nature*, 432(7016):492–495, 2004.

[147] Shukla, A., Mylonas, E., Di Cola, E., Finet, S., Timmins, P., Narayanan, T., and Svergun, D. I. Absence of equilibrium cluster phase in concentrated Lysozyme solutions. *Proceedings of the National Academy of Sciences*, 105(13):5075–5080, 2008.

[148] Chen, W., Huang, Y., and Shen, J. Conformational activation of a transmembrane proton channel from constant pH molecular dynamics. *The Journal of Physical Chemistry Letters*, 7(19):3961–3966, 2016.

[149] Finet, S., Skouri-Panet, F., Casselyn, M., Bonnete, F., and Tardieu, A. The hofmeister effect as seen by *saxs* in protein solutions. *Current Opinion in Colloid & Interface Science*, 9(1):112–116, 2004.

[150] Roth, C. M., Neal, B. L., and Lenhoff, A. M. van der *waals* interactions involving proteins. *Biophysical Journal*, 70(2):977–987, 1996.

[151] Biasini, M., Bienert, S., Waterhouse, A., Arnold, K., Studer, G., Schmidt, T., Kiefer, F., Cassarino, T. G., Bertoni, M., and Bordoli, L. *swiss − model*: modelling protein tertiary and quaternary structure using evolutionary information. *Nucleic Acids Research*, page gku340, 2014.

[152] Kelley, L. A., Mezulis, S., Yates, C. M., Wass, M. N., and Sternberg, M. J. The Phyre^2 web portal for protein modeling, prediction and analysis. *Nature Protocols*, 10(6):845–858, 2015.

Appendix A

Homology modeling

In the method of homology modeling, it is necessary to have two principal components: (1) the target (sequence to model) and (2) the template (protein resolved by x-ray crystallography or nuclear magnetic resonance which has a defined three-dimensional structure on which the model is built). Several programs can be used to perform modeling by this method (*e.g.* Modeller,[129] Swiss-Model,[151] Phyre^2,[152] etc.). We chosen the program Modeller (version 9.17), which can be downloaded for free from its website [https://salilab.org/modeller/]. Modeller is good option for homology modelling, considering that query squence and template have a identity around the 30%. For practical reasons, Modeller gives flexibility in the template selection, loop refinement, and model quality assessment.

Due to the fact that we already knew the template necessary for the modeling of CHY, A3 (*apo*), and A4 (*holo*) (*i.e.* the structures registered PDB id 1CHG, 3NSO, and 3C1V, respectively), were used the same crystallographic structures as a template despite not having regions with missing residues that correspond to small sections of sequence. We were not very specific in the selection of the template in this part of the procedure, since in our CG model the *aa* are reduced to spheres and differences of rotameros do not represent problem. The following steps were carried out to complete the process as described below:

Table A.1: Estimated models through of Modeller.

Protein	[a]DOPE score	[b]GA341 score	[c]RMSD$_{BB}$
CHY	−23704.90820	1.0000	1.14
A3	−11182.33594	1.0000	0.99[d]
A4	−23501.74023	1.0000	1.20[d]

[a] DOPE score is a function that evaluates energy, lowest score corresponds to best models predicted. [b] GA341 scores always range from 0.0 (worst) to 1.0 (native-like). [c]The RMSD$_{BB}$ was calculated comparing the backbones of both model and template. [d]Data from the chain A.

- Target-template alignment: using the script *align2d.py*[1], files were generated (*.ali* and *.pap*). These have alignments between target and templates that the program uses to build the model.
- Model building and evaluation: using the script *model-single.py*, ten different models were estimated for each protein. We selected the best model for having the best DOPE score and GA341 score (Table A.1). The RMSD_{BB} was calculated comparing the backbones of both model and template.

[1]Scripts are also available on its website [https://salilab.org/modeller/tutorial/basic.html].

Appendix B

Scripts and codes for analysis of data

Codes were written in Fortran 77. Scripts were written in Bash (Linux).

B.1 Code to calculate the Second Virial Coefficient B_2

```
*   ////////////////////////////////////////////////////////////////////////////   *
*   COPYRIGHT 2015                                                                 *
*   BY: SERGIO ALEJANDRO POVEDA CUEVAS                                             *
*   SUPERVISOR: FERNANDO LUIS BARROSO DA SILVA                                     *
*   LABORATORY OF COMPUTATIONAL BIOPHYSICAL CHEMISTRY                              *
*    UNIVERSITY OF SAO PAULO - RIBEIRAO PRETO, SP, BRAZIL                          *
*                                                                                  *
*   NO PART OF THIS CODE MAY BE COPIED OR REDISTRIBUTED WITHOUT                    *
*   THE WRITTEN PERMISSION OF THE COPYRIGHT OWNER                                  *
*   THE COPYRIGHT OWNER DOES NOT TAKE ANY RESPONSIBILITY FOR ANY                   *
*   ERRORS IN THE CODE OR DOCUMENTATION.                                           *
*   ////////////////////////////////////////////////////////////////////////////   *
program virial
implicit none
integer, parameter :: num = 10000
integer i, j, var, num1 real wr(num), rad(num)
real sig, x1, x2, y1, y2, f1, f2
real term1, term2, term3, term4, term5, sim
real smt
```

```
real (KIND=4) :: b2el

real (KIND=4) :: b2er

real(KIND=4) :: b2

double precision, parameter :: pi = 3.141592654d0

double precision, parameter :: avg = 6.022E23

double precision, parameter :: mw = 11713.42 ! for LYS in g/mol

open(2,file='pfm.dat')

open(3,file='integral')

var = 0

80 do while (var < num)

var = var + 1 read (2,*,end=99)

rad(var), wr(var)

go to 80

end do

99 continue

var = 0

num1 = num - 1 x1 = 0 x2 = 0

do while (var < num1)

do i = 1, 2

j = j + 1

var = var + 1

if (var .EQ. 1) then

sig = rad(var)

!write(*,*) var, rad(var), sig

end if

if (i .EQ. 1) then

x1 = rad(var)

y1 = wr(var)

else

x2 = rad(var)

y2 = wr(var)

end if
```

```
!write(*,*) x1, x2

if (x2 .NE. 0) then

if (x1 .NE. x2) then

f1 = (exp(-y1) - 1) * x1**2

f2 = (exp(-y2) - 1) * x2**2

term1 = (x2 - x1) / 6

term2 = ((f2 + f1) / 2 ) * 4

sim = term1 * (f1 + term2 + f2)

smt = smt + sim

!write(3,*) sim, smt

end if

end if

end do

var = var - 1

end do

write(*,*) 'Integral: ', smt

term3 = (-avg*pi*2)/(mw*mw)

b2el = (smt * 1E-24) * term3

write(*,'(" B2 (ele + VdW): ",E12.4 )') b2el

term4 = avg/(3*mw*mw)

term5 = pi * (sig+sig)**3 * 1E-24

b2er = term4 * term5

write(*,'(" B2 (er): ",E12.4 )') b2er

b2 = (b2el + b2er) * 1E4

write(*,*) 'B2:', b2

write(3, '(F13.7 )') b2

close (2)

close (3)

end program virial
```

B.2 Code to calculate pmf

```
*   ////////////////////////////////////////////////////////////////////////////   *
*   COPYRIGHT 2015                                                                 *
*   BY: SERGIO ALEJANDRO POVEDA CUEVAS                                             *
*   SUPERVISOR: FERNANDO LUIS BARROSO DA SILVA                                     *
*   LABORATORY OF COMPUTATIONAL BIOPHYSICAL CHEMISTRY                              *
*    UNIVERSITY OF SAO PAULO - RIBEIRAO PRETO, SP, BRAZIL                          *
*                                                                                  *
*   NO PART OF THIS CODE MAY BE COPIED OR REDISTRIBUTED WITHOUT                    *
*   THE WRITTEN PERMISSION OF THE COPYRIGHT OWNER                                  *
*   THE COPYRIGHT OWNER DOES NOT TAKE ANY RESPONSIBILITY FOR ANY                   *
*   ERRORS IN THE CODE OR DOCUMENTATION.                                           *
*   ////////////////////////////////////////////////////////////////////////////   *
program pfm

implicit none

integer, parameter :: num = 10000

integer var, num1, num2

real wr(num), rad(num)

real y, x1, x2, y1, y2, med

open(2,file='rdf_p2p.dat') open(3,file='output_1.dat')

var = 0

80 do while (var < num)

var = var + 1 read (2,*,end=99)

rad(var), wr(var)

go to 80

end do

99 continue

var = 0

do while (var < num)

var = var + 1

if (rad(var) .NE. 0) then

y = -log(wr(var) + 0.0000000001)
```

```
write(3,*) rad(var), y

end if

end do

close (2)

close (3)

end program pfm
```

B.3 Script to transform rdf to pmf

```
*   ////////////////////////////////////////////////////////////////////////////   *
*   COPYRIGHT 2015                                                                 *
*   BY: SERGIO ALEJANDRO POVEDA CUEVAS                                             *
*   SUPERVISOR: FERNANDO LUIS BARROSO DA SILVA                                     *
*   LABORATORY OF COMPUTATIONAL BIOPHYSICAL CHEMISTRY                              *
*    UNIVERSITY OF SAO PAULO - RIBEIRAO PRETO, SP, BRAZIL                          *
*                                                                                  *
*   NO PART OF THIS CODE MAY BE COPIED OR REDISTRIBUTED WITHOUT                    *
*   THE WRITTEN PERMISSION OF THE COPYRIGHT OWNER                                  *
*   THE COPYRIGHT OWNER DOES NOT TAKE ANY RESPONSIBILITY FOR ANY                   *
*   ERRORS IN THE CODE OR DOCUMENTATION.                                           *
*   ////////////////////////////////////////////////////////////////////////////   *
cp ../rdf_p2p.dat ./

wc -l rdf_p2p.dat > hola1

read i j < hola1

j=$[ ${i} - 1 ]

cp rdf_p2p.dat rdf_p2p_1.dat

head -n ${i} rdf_p2p_1.dat | tail -n ${j} > rdf_p2p.dat

chmod +x pfm med

./pfm

./med

grep ^“ ” media > med.dat

read n < med.dat

sed ‘s/y = -log(wr(var) + 0.0000000001)/y = -log(wr(var) + 0.0000000001) - “”${n}””/’
```

```
pfm.f > pfm_1.f

gfortran pfm_1.f -o pfm_1

chmod +x pfm_1 ; ./pfm_1

mv output_1.dat pfm.dat

cp pfm.dat ../

rm pfm_1 media med.dat pfm_1.f rdf_p2p.dat rdf_p2p_1.dat pfm.dat hola1
```

B.4 Script to generate files to search ε_{LJ}

```
*   ////////////////////////////////////////////////////////////////////////////   *
*   COPYRIGHT 2015                                                                 *
*   BY: SERGIO ALEJANDRO POVEDA CUEVAS                                             *
*   SUPERVISOR: FERNANDO LUIS BARROSO DA SILVA                                     *
*   LABORATORY OF COMPUTATIONAL BIOPHYSICAL CHEMISTRY                              *
*    UNIVERSITY OF SAO PAULO - RIBEIRAO PRETO, SP, BRAZIL                          *
*                                                                                  *
*   NO PART OF THIS CODE MAY BE COPIED OR REDISTRIBUTED WITHOUT                    *
*   THE WRITTEN PERMISSION OF THE COPYRIGHT OWNER                                  *
*   THE COPYRIGHT OWNER DOES NOT TAKE ANY RESPONSIBILITY FOR ANY                   *
*   ERRORS IN THE CODE OR DOCUMENTATION.                                           *
*   ////////////////////////////////////////////////////////////////////////////   *
for ((pHnum=1; pHnum <= 6; pHnum+=1)) do

if [ ${pHnum} -eq 1 ]; then pH=4.5 fi

if [ ${pHnum} -eq 2 ]; then pH=6.0 fi

if [ ${pHnum} -eq 3 ]; then pH=7.5 fi

if [ ${pHnum} -eq 4 ]; then pH=9.0 fi

if [ ${pHnum} -eq 5 ]; then pH=10.0 fi

if [ ${pHnum} -eq 6 ]; then pH=10.5 fi

rm -fr $pH ; mkdir $pH ; cd $pH

for ((i=5000; i <= 10000; i += 50))

do

epsi=0$(echo "scale=5; ${i}/100000" | bc); epsihp=0$(echo "scale=0; ${epsi} * 2.47773" | bc)

cp -r ../base $epsi ; cd $epsi
```

```
cp cyl_prod1.run cyl_1.run ; sed 's/for pH in 4.5/for pH in '${pH}'/' cyl_1.run |
sed 's/epsi_hydrophob=0.05005/epsi_hydrophob='${epsi}'/' |
sed 's/epsi_hydrophil=0.124/epsi_hydrophil='${epsihp}'/' > cyl_prod1.run ; rm cyl_1.run
cp cyl_prod2.run cyl_1.run
sed 's/for pH in 4.5/for pH in '${pH}'/' cyl_1.run |
sed 's/epsi_hydrophob=0.05005/epsi_hydrophob='${epsi}'/' |
sed 's/epsi_hydrophil=0.124/epsi_hydrophil='${epsihp}'/' > cyl_prod2.run ; rm cyl_1.run
cd ../
directory=$[ $directory+1 ]
done
echo
cd ../
done
echo 'Done!'
echo "Total number of directories for simulation: $directory" > total_dir
exit
```

B.5 Script to calculate RMSD

```
*   ////////////////////////////////////////////////////////////////////////////   *
*   COPYRIGHT 2015                                                                 *
*   BY: SERGIO ALEJANDRO POVEDA CUEVAS                                             *
*   SUPERVISOR: FERNANDO LUIS BARROSO DA SILVA                                     *
*   LABORATORY OF COMPUTATIONAL BIOPHYSICAL CHEMISTRY                              *
*    UNIVERSITY OF SAO PAULO - RIBEIRAO PRETO, SP, BRAZIL                          *
*                                                                                  *
*   NO PART OF THIS CODE MAY BE COPIED OR REDISTRIBUTED WITHOUT                    *
*   THE WRITTEN PERMISSION OF THE COPYRIGHT OWNER                                  *
*   THE COPYRIGHT OWNER DOES NOT TAKE ANY RESPONSIBILITY FOR ANY                   *
*   ERRORS IN THE CODE OR DOCUMENTATION.                                           *
*   ////////////////////////////////////////////////////////////////////////////   *
salt=0.005
rm rmsd.dat
```

```
echo '---------- ---------- ---------- ----------' > rmsd.dat

echo 'Eps_hfb (kT) Eps_hfl (kJ/mol) RMSD average RMSD' >> rmsd.dat

echo '---------- ---------- ---------- ----------' >> rmsd.dat

i=0

for ((i=5000; i <= 10000; i += 50))

do

epsi=0$(echo "scale=5; ${i}/100000" | bc) ; epsihp=0$(echo "scale=0; ${epsi} * 2.47773" | bc)

x=0 ; y=0 ; n=0 ; b2pos=0 ; b2pos1=0

for ((pHnum=1; pHnum <= 6; pHnum+=1)) do

if [ ${pHnum} -eq 1 ]; then pH=4.5 b2_exp=33.3 fi

if [ ${pHnum} -eq 2 ]; then pH=6.0 b2_exp=28.5 fi

if [ ${pHnum} -eq 3 ]; then pH=7.5 b2_exp=25.2 fi

if [ ${pHnum} -eq 4 ]; then pH=9.0 b2_exp=15.6 fi

if [ ${pHnum} -eq 5 ]; then pH=10.0 b2_exp=5.98 fi

if [ ${pHnum} -eq 6 ]; then pH=10.5 b2_exp=-1.06 fi

echo "Epsilon" 'KT=' $epsi 'kJ/mol=' $epsihp

echo "pH $pH"

grep -A 0 "Production 1" zcu_simul.pH$pH.salt$salt.epsi$epsihp.dat | tail -n1 | cut -f 6 > hola

read j < hola

x=$(echo "scale=10; $x+(${b2_exp}+(-1*${j}))^2" | bc)

n=$(echo "scale=10; (${b2_exp}*${j}*10)" | bc)

echo $x

echo $n

n1=${n%%.*}

echo $n1

if [ ${n1} -gt 0 ]; then ; b2pos=$[ $b2pos+1 ] fi

echo '++/--' $b2pos

grep -A 0 "Production 2" zcu_simul.pH$pH.salt$salt.epsi$epsihp.dat | tail -n1 | cut -f 6 > hola

read k < hola

y=$(echo "scale=10; $y+(${b2_exp}+(-1*${k}))^2" | bc)

m=$(echo "scale=10; (${b2_exp}*${k}*10)" | bc)

echo $y
```

```
echo $m

m1=${m%%.*}

echo $m1

if [ ${m1} -gt 0 ]; then ; b2pos1=$[ $b2pos1+1 ] fi

echo '++/--' $b2pos1

done

rmsd_x=$(echo "scale=10; sqrt(${x}/6)" | bc)

rmsd_y=$(echo "scale=10; sqrt(${y}/6)" | bc)

rmsd_med=$(echo "scale=6; (${rmsd_x}+${rmsd_y})/2" | bc)

echo "$rmsd_med" >> rmsd_score1

echo " $epsi $epsihp $rmsd_x | $rmsd_y $rmsd_med" >> rmsd.dat

echo " ++/--: $b2pos | ++/--: $b2pos1" >> rmsd.dat

echo " |" >> rmsd.dat

done

echo '----------' >> rmsd.dat

echo '---------- ---------- ---------- ----------' > rmsd1.dat

echo 'Eps_hfb (kT) Eps_hfl (kJ/mol) RMSD average RMSD ' >> rmsd1.dat

echo '---------- ---------- ---------- ----------' >> rmsd1.dat

wc -l rmsd.dat > fila

read f g < fila

sort -n rmsd_score1 | head -n1 > rmsd_score2

read u v < rmsd_score2

u=$(echo "$u * 1000000" | bc)

u=${u%%.*}

for ((e=4; e < ${f}; e += 3))

do

more rmsd.dat | head -n${e} | tail -n1 | cut -f2 > hola2

read r1 r2 < hola2

r1=$(echo "$r1 * 1000000" | bc)

r1=${r1%%.*}

if [ ${u} -eq ${r1} ]; then

more rmsd.dat | head -n${e} | tail -n1 > best1
```

```
echo " <<<<" >> best2

paste best1 best2 >> rmsd1.dat

e1=$[ $e + 1]

more rmsd.dat | head -n${e1} | tail -n1 >> rmsd1.dat

e1=$[ $e1 + 1]

more rmsd.dat | head -n${e1} | tail -n1 >> rmsd1.dat

else

more rmsd.dat | head -n${e} | tail -n1 >> rmsd1.dat

e1=$[ $e + 1]

more rmsd.dat | head -n${e1} | tail -n1 >> rmsd1.dat

e1=$[ $e1 + 1]

more rmsd.dat | head -n${e1} | tail -n1 >> rmsd1.dat

fi

done

echo '----------------------------------------------------------------------------------------------' >> rmsd1.dat

mv rmsd1.dat rmsd.dat

rm hola hola2 rmsd_score1 rmsd_score2 fila best1 best2
```

B.6 Script to calculate the standard deviation between three production

```
*  ////////////////////////////////////////////////////////////////////////////  *
*  COPYRIGHT 2016                                                                *
*  BY: SERGIO ALEJANDRO POVEDA CUEVAS                                            *
*  SUPERVISOR: FERNANDO LUIS BARROSO DA SILVA                                    *
*  LABORATORY OF COMPUTATIONAL BIOPHYSICAL CHEMISTRY                             *
*   UNIVERSITY OF SAO PAULO - RIBEIRAO PRETO, SP, BRAZIL                         *
*                                                                                *
*  NO PART OF THIS CODE MAY BE COPIED OR REDISTRIBUTED WITHOUT                   *
*  THE WRITTEN PERMISSION OF THE COPYRIGHT OWNER                                 *
*  THE COPYRIGHT OWNER DOES NOT TAKE ANY RESPONSIBILITY FOR ANY                  *
*  ERRORS IN THE CODE OR DOCUMENTATION.                                          *
*  ////////////////////////////////////////////////////////////////////////////  *
```

```
salt=0.005

x=0 ; x1=0 ; x11=0 ; y=0 ; y1=0 ; y11=0 ; z=0 ; z1=0 ; z11=0

for ((pHnum=1; pHnum <= 6; pHnum+=1)) do

if [ ${pHnum} -eq 1 ]; then pH=4.5 fi

if [ ${pHnum} -eq 2 ]; then pH=6.0 fi

if [ ${pHnum} -eq 3 ]; then pH=7.5 fi

if [ ${pHnum} -eq 4 ]; then pH=9.0 fi

if [ ${pHnum} -eq 5 ]; then pH=10.0 fi

if [ ${pHnum} -eq 6 ]; then pH=10.5 fi

epsi_function(){

epsi=0$(echo "scale=5; ${e}/100000" | bc)

epsihp=0$(echo "scale=0; ${epsi} * 2.47773" | bc)

}

for ((varep=1; varep <= 3; varep+=1)) do

if [ ${varep} -eq 1 ]; then e=5005 epsi_function fi

if [ ${varep} -eq 2 ]; then e=7800 epsi_function fi

if [ ${varep} -eq 3 ]; then e=8000 epsi_function fi

file_zcu=zcu_simul.pH${pH}.salt${salt}.epsi${epsi}.dat

grep -A10 "simul.pH" $file_zcu | tail -n1 | cut -f6 > hola1 # Production1

read x < hola1

grep -A13 "simul.pH" $file_zcu | tail -n1 | cut -f6 > hola2 # Production2

read y < hola2

grep -A16 "simul.pH" $file_zcu | tail -n1 | cut -f6 > hola3 # Production3

read z < hola3

average=$(echo "scale=10; (${x}+${y}+${z})/3" | bc)

x1=$(echo "scale=10; (sqrt((${x}+(-1*${average}))^2))" | bc)

echo "(±$x1)" > x_file

y1=$(echo "scale=10; (sqrt((${y}+(-1*${average}))^2))" | bc)

echo "(±$y1)" > y_file

z1=$(echo "scale=10; (sqrt((${z}+(-1*${average}))^2))" | bc)

echo "(±$z1)" > z_file

st_des=$(echo "scale=10; sqrt(((${x}+(-1*${average}))^2 + (${y}+(-1*${average}))^2 + (${z}+(-1*${average}))^2)/2)"
| bc)
```

```
wc -l $file_zcu > line

read n m < line

head -n8 $file_zcu | tail -n8 | cut -f1-6 > zcu_1.dat

for ((i=9; i <= ${n}; i+=3)) do

head -n${i} $file_zcu | tail -n1 | cut -f6 > adios1 # Production1

read a < adios1

a=$(echo "$a * 1000000" | bc)

a1=${a%%.*}

# Equilibration

i1=0

if [ ${i} -eq 9 ]; then

i1=$[ $i + 2]

head -n${i1} $file_zcu | tail -n3 | cut -f1-6 >> zcu_1.dat fi

# Production1 - X

x11=$(echo "$x * 1000000" | bc) x11=${x11%%.*}

i1=0

if [ ${a1} -eq ${x11} ]; then

head -n${i} $file_zcu | tail -n1 | cut -f1-6 >> zcu_1.dat

i1=$[ $i + 1]

head -n${i1} $file_zcu | tail -n1 | cut -f1-4 >> zcu_2.dat

paste zcu_2.dat x_file >> zcu_1.dat

i1=$[ $i1 + 1]

head -n${i1} $file_zcu | tail -n1 | cut -f1-6 >> zcu_1.dat

fi

# Production2 - Y

y11=$(echo "$y * 1000000" | bc)

y11=${y11%%.*}

i1=0

if [ ${a1} -eq ${y11} ]; then

head -n${i} $file_zcu | tail -n1 | cut -f1-6 >> zcu_1.dat

i1=$[ $i + 1]

head -n${i1} $file_zcu | tail -n1 | cut -f1-4 >> zcu_3.dat
```

```
paste zcu_3.dat y_file >> zcu_1.dat

i1=$[ $i1 + 1]

head -n${i1} $file_zcu | tail -n1 | cut -f1-6 >> zcu_1.dat

fi

# Production3 - Z

z11=$(echo "$z * 1000000" | bc)

z11=${z11%%.*}

i1=0

if [ ${a1} -eq ${z11} ]; then

head -n${i} $file_zcu | tail -n1 | cut -f1-6 >> zcu_1.dat

i1=$[ $i + 1]

head -n${i1} $file_zcu | tail -n1 | cut -f1-4 >> zcu_4.dat

paste zcu_4.dat z_file >> zcu_1.dat

i1=$[ $i1 + 1]

head -n${i1} $file_zcu | tail -n1 | cut -f1-6 >> zcu_1.dat

fi

done

more $file_zcu | cut -f7 > time

for ((j=1; j <= 7; j+=1)) do

paste zcu_1.dat zcu_5.dat > zcu

echo > space

echo "SD:" $st_des > st_des

echo "MD:" $average > avr

paste space avr st_des >> zcu

mv zcu $file_zcu

mv zcu ${varep}.${pH}

rm line hola* adios* x_file y_file z_file zcu_1.dat zcu_2.dat zcu_3.dat zcu_4.dat zcu_5.dat time space st_des avr

done

done
```