The Role of Acidic Amino Acids in the Hydration and Stabilization of Halophilic Proteins

Hosein Geraili Daronkola

Univ.-Diss.
zur Erlangung des akademischen Grades "doctor rerum naturalium" (Dr. rer. nat.) in der Wissenschaftsdisziplin "Theoretische biologische Physik"

eingereicht an der Mathematisch-Naturwissenschaftlichen Fakultät Institut für Physik und Astronomie der Universität Potsdam

Potsdam, May 2021

Published online on the
Publication Server of the University of Potsdam:
https://doi.org/10.25932/publishup-51671
https://nbn-resolving.org/urn:nbn:de:kobv:517-opus4-516713

Acknowledgements

I want to thank Ana Vila Verde, my supervisor, for her guidance and support during my Ph.D. studies with her precision and analytical mind. I have learned significantly due to the independence given to me and the pragmatism shared with me under her supervision. I am very thankful to my parents, who planted the love for knowledge and academia in me because of their significant desire for wisdom and education. In the end, I would also like to thank my brother and sister, Mohammad and Zahra.

Abstract

Proteins of halophilic organisms that accumulate molar concentrations of KCl in their cytoplasm have much higher content in acidic amino acids than proteins of mesophilic organisms. It has been proposed that this excess is necessary to maintain proteins hydrated in an environment with low water activity: either via direct interactions between water and the carboxylate groups of acidic amino acids or via cooperative interactions between acidic amino acids and hydrated cations, which would stabilize the folded protein. In the course of this Ph.D. study, we investigated these possibilities using atomistic molecular dynamics simulations and classical force fields. High quality parameters describing the interaction between K^+ and carboxylate groups present in acidic amino acids are indispensable for this study. We first evaluated the quality of the default parameters for these ions within the widely used AMBER ff14SB force field for proteins and found that they perform poorly. We propose new parameters, which reproduce solution activity derivatives of potassium acetate solutions up to 2 mol/kg and the distances between potassium ions and carboxylate groups observed in x-ray structures of proteins. To understand the role of acidic amino acids in protein hydration, we investigated this aspect for 5 halophilic proteins in comparison with 5 mesophilic ones. Our results do not support the necessity of acidic amino acids to keep folded proteins hydrated. Proteins with a larger fraction of acidic amino acids indeed have higher hydration levels. However, the hydration level of each protein is identical at low ($b_{\mathrm{KCl}} = 0.15$ mol/kg) and high ($b_{\mathrm{KCl}} = 2$ mol/kg) KCl concentration. It has also been proposed that cooperative interactions between acidic amino acids with nearby hydrated cations stabilize the folded protein and slow down its solvation shell; according to this theory, the cations would be preferentially excluded from the unfolded structure. We investigate this possibility through extensive free energy calculation simulations. We find that cooperative interactions between neighboring acidic amino acids exist and are mediated by the ions in solution but are present in both folded and unfolded structures of halophilic proteins. The translational dynamics of the solvation shell is barely distinguishable between halophilic and mesophilic proteins; therefore, such a cooperative effect does not result in unusually slow solvent dynamics as has been suggested.

Table of contents

List of figures xi

List of tables xix

1 Introduction 1

2 Optimizing the force field for $K^+ \cdots$ carboxylate interactions 5

2.1 Available force fields . . . 5

2.2 Optimization approach . . . 6

2.2.1 Calculating the electrolyte activity derivative in simulations . . . 8

2.2.2 Experimental electrolyte activity derivative . . . 9

2.3 Simulation details . . . 10

2.3.1 Potassium acetate solutions . . . 10

2.3.2 Restrained ferredoxin in $c_{KCl} = 1$ mol/dm^3 . . . 11

2.4 Results and discussion . . . 12

3 Structure and dynamics of the hydration shell of mesophilic vs. halophilic proteins 19

3.1 Selecting appropriate systems for a comparative study of halophilic and mesophilic proteins . . . 20

3.2 Simulation details . . . 23

3.3 Results . . . 25

3.3.1 Structural stability . . . 25

3.3.2 Solvation layer structure . . . 25

3.3.3 Dynamics of the protein solvation shell . . . 31

3.4 Discussion . . . 33

3.5 Conclusion . . . 36

4 Are cooperative water-cation-carboxylate interactions in halophilic proteins possible? 39
4.1 Cooperative interactions in halophilic proteins . . . 39
4.2 Free energy calculations via thermodynamic integration (TI) . . . 42
4.2.1 Theoretical background . . . 42
4.2.2 Computational details of free energy calculation . . . 45
4.2.3 Identifying the best computational scheme to calculate free energies of mutation . . . 49
4.2.3.1 The effect of finite-size periodic boundary simulation box on the free energy calculation . . . 49
4.2.3.2 Simulation time and number of intermediate states . . . 52
4.2.3.3 Restraining the backbone of proteins . . . 56
4.2.3.4 Calculating the standard error of the free energies . . . 58
4.3 Replica Exchange Molecular Dynamics (REMD) . . . 58
4.3.1 Theoretical background . . . 58
4.3.2 Computational details of REMD simulation . . . 61
4.3.3 Evaluating the quality of REMD simulations . . . 63
4.4 Results . . . 65
4.4.1 Free energies of D→N and E→Q mutations with/without vicinal acidic amino acids in folded proteins . . . 65
4.4.1.1 Decomposing mutation free energy into its components . 68
4.4.2 Free energies of D→N and E→Q mutations with/without vicinal acidic amino acids in unfolded protein L. . . . 71
4.4.3 Mechanism behind synergistic effects . . . 77
4.5 Conclusion . . . 83

5 Outlook 87

References 91

Appendix A Supporting information 101
A.1 Water-protein hydrogen bonds per residue type . . . 102
A.2 Cumulative number of potassium ions as a function of distance to the protein surface . . . 102
A.3 Mean square displacement of water and of K^+ . . . 102
A.4 Ion pairing in potassium acetate and sodium acetate solutions . . . 107
A.5 Water activity in NaCl and KCl solutions . . . 107

A.6 Free energy of mutations studied in the Chapter 4 109

List of figures

1.1 Conformational fluctuations model: electrostatic repulsion (red arrows) between acidic amino acids (blue) offset the higher attraction between hydrophobic groups at high salt concentrations (not shown), to maintain flexibility (green) necessary for function. The question marks highlight that this model has not been explored. 4

2.1 Mean molal activity coefficient ($\gamma_{s,\pm}^{(b)}$) of potassium acetate as a function of the molality of KCH_3COO. The blue circles are the experimental data [83] and the red dashed line is calculated using the relevant Pitzer equation [77]. 10

2.2 (A,B,C) Molar solution activity derivative (red points) of potassium acetate solutions, as a function of the multiplicative scaling factor ($f_{R_{\mathrm{min,K^+O}}}$; see eq. 2.12) applied to the LJ $R_{\mathrm{min,K^+O,LB}}$ parameter governing the interactions between K^+ and carboxylates. The error bars are the standard error of the mean calculated from three independent production simulations. The red lines are a guide to the eye. The green line shows the experimental reference value; see also table 2.2. The shaded regions show the $\pm 7\%$ deviation from the experimental value. The two vertical dashed lines delimit the range of scaling factors acceptable for the 3 concentrations. (D) Radial distribution function of potassium and the carbon bonded to the oxygens of acetate (the same simulations as in panel A) for the indicated values of the scaling factor $f_{R_{\mathrm{min,K^+O}}}$. 13

2.3 (A) Crystal structure (pdb ID: 1DOI) of the halophilic 2Fe-2S ferredoxin from *Haloarcula Marismortui* [24]; the K^+ ions are shown in pink. The 5 acidic amino acid sites that have nearby K^+, used to parameterize $R_{\mathrm{min,K^+O}}$, are indicated. (B) Example radial distribution function of K^+ and the carboxylate oxygens of site 2, at T=298 K and c_{KCl} = 1 mol/dm^3. The distance to the first maximum is identified as $r_{\mathrm{sim.}}$. The legend shows the values of the scaling factor, $f_{R_{\mathrm{min,K^+O}}}$. (C) Unsigned relative deviation between r_{sim} and r_{cryst} for the five indicated sites of the halophilic ferredoxin. The protein site is identified by the residue number, residue name and oxygen name. The legend shows the values of the scaling factor $f_{R_{\mathrm{min,K^+O}}}$. Numerical data is shown in SI Table A.1. 15

3.1 Example simulation box, of the halophilic ferredoxin protein (pdb ID: 1DOI). Dark blue shape = New Cartoon representation of protein; Pink spheres = K^+; Green spheres: Cl^-; Transparent small blue dots: water molecules. . . . 24

3.2 Surface density of water-protein hydrogen bonds for the indicated halophilic-mesophilic proteins, identified by their pdb ID, for different KCl concentrations. 26

3.3 (A,B) Proximal number density of water molecules as a function of the distance to the surface of the indicated proteins, simulated at b_{KCl}=2 mol/kg (color) and b_{KCl}=0.15 mol/kg (dashed black lines). (C) Height of the first peak of the number density curves for b_{KCl}=2 mol/kg shown in the other panels, as a function of the surface density of acidic amino acids. Each color corresponds to a protein, identified by its pdb ID in the legend of the bottom panel. 29

3.4 (A,B) Proximal number density of K^+ as a function of the distance to the surface of the indicated proteins, in solutions with the indicated molality of KCl. (C) Height of the first peak of the number density curves for b_{KCl}=2 mol/kg shown in the other panels, as a function of the protein charge density. Each color corresponds to a protein, identified by its pdb ID in the legend of the bottom panel. 30

3.5 Diffusion coefficients of water around the indicated proteins, simulated in b_{KCl} = 2 mol/kg. First-shell: water molecules that at $t = 0$ belong to the first hydration shell of the proteins; Bulk: all water molecules in the same simulation. 32

3.6 Diffusion coefficients of potassium ions around the indicated proteins, simulated in b_{KCl} = 2 mol/kg. First-shell: potassium ions that at $t = 0$ belong to the first solvation shell of the proteins; Bulk: all potassium ions in the same simulation. 33

4.1 A snapshot of simulation box with solution showed within a radius of 10 Å around protein. Protein L.: New cartoon representation; K^+: transparent pink spheres; Cl^-: transparent blue spheres; Acidic residues: blue and red branches; Water oxygens: red dots. 41

4.2 Halophilic protein ferredoxin (pdb ID: 1DOI) in white new cartoon representation, and two acidic amino acids shown in red and blue, exemplifying the *a* and *b* positions mentioned in Equation 4.1 . Only one pair of acidic amino acids (aspartic acid residues in residue positions 81 and 83 of amino acid chain) is shown, and other residues, as well as the solution, are not shown. . 41

4.3 The ball and stick representation of acidic amino acids; blue spheres: nitrogen atom, red spheres: oxygen atoms, and cyan spheres: carbon atoms. Hydrogens are not shown. 42

4.4 Plot of $\left\langle \frac{\partial U}{\partial \lambda} \right\rangle_\lambda$ vs. λ values for the thermodynamic integration, with red filled areas indicating free energy estimates from the TI-1 and silver curve indicating interpolation via TI-3. Color intensity alternates with increasing λ index. The agreements between the alternate interpolation schemes suggest that the interpolation successfully captures the free energy change between two neighboring λ, as well as over the whole range. The subfigures correspond to the thermodynamic cycle 4.10 with (a) vdW ($\Delta G_{X^0 a Y^0}$), (b) charge ($-\Delta G_{Y^0 a Y}$), and (c) to decharge ($-\Delta G_{X a X^0}$) leg of this cycle. This figure shows the mutation corresponding to the (D81)...D83N at b_{KCl} = 2 mol/kg of 1DOI. 48

4.5 Plot of $\left\langle \frac{\partial U}{\partial \lambda} \right\rangle_\lambda$ vs. λ values for the integration. This figure shows the mutation corresponding to annihilating a potassium ion. The details of the simulation are explained in Subsection 4.2.3.1. 51

4.6 Plot of $\left\langle \frac{\partial U}{\partial \lambda} \right\rangle_\lambda$ vs. λ values for the integration. This figure shows the mutation corresponding to the (D81)...D83N of the protein 1DOI at b_{KCl} = 2 mol/kg, with 10 ns of simulation time. The details of the simulation are explained in Subsection 4.2.3.2. 53

4.7 Plot of $\left\langle \frac{\partial U}{\partial \lambda} \right\rangle_{\lambda}$ vs. λ values for the integration. This figure shows the mutation corresponding to the (D81)...D83N of the protein 1DOI at b_{KCl} = 2 mol/kg, with 30 ns of simulation time. The details of the simulation is explained in Subsection 4.2.3.2. 54

4.8 Plot of $\left\langle \frac{\partial U}{\partial \lambda} \right\rangle_{\lambda}$ vs. λ values for the integration. This figure shows the mutation corresponding to the (D81)...D83N of the protein 1DOI at b_{KCl} = 2 mol/kg, with a total of 177 λ. The details of the simulation is explained in Subsection 4.2.3.2. 55

4.9 Plot of $\left\langle \frac{\partial U}{\partial \lambda} \right\rangle_{\lambda}$ vs. λ values for the integration. This figure shows the mutation corresponding to the D34N of the protein 1DOI at b_{KCl} = 2 mol/kg. The details of the simulation are explained in Subsection 4.2.3.3. 57

4.10 A Replica exchange consisting 8 replicas. Question marks represent exchange attempts (green for accepted, and red for rejected). Figure is reproduced from ref. [14]. 59

4.11 Acceptance ratio which shows the percentage of accepted exchanges for each replica relative to all exchange attempts. 64

4.12 Temperature distribution for on of the replicas with the starting temperature of 298.98 K. 64

4.13 Change in free energy of mutation ($\Delta\Delta G \pm 1.0$; Eq. 4.20) for the indicated pairs of amino acids of the halophilic protein ferredoxin (pdb ID: 1DOI), calculated at the indicated KCl molality. The data shown in this plot is compiled in Table A.2. 66

4.14 Change in free energy of mutation ($\Delta\Delta G \pm 1.0$; Eq. 4.20) for the indicated pairs of amino acids of the halophilic protein dihydrofolate reductase (pdb ID: 2ITH), calculated at the high KCl molality. The data shown in this plot is compiled in Table A.4. 67

4.15 Change in free energy of mutation ($\Delta\Delta G \pm 1.0$; Eq. 4.20) for the indicated pairs of amino acids of the halophilic protein L. (pdb ID: 2KAC), calculated at the indicated KCl molality. The cases of D50N, E21Q, and E27Q were only simulated at high salt concentrations. The data shown in this plot is compiled in Table A.3. 68

4.16 $\Delta\Delta G_{charging} \pm 0.2, \Delta\Delta G_{decharging} \pm 1.1$ and $\Delta\Delta G_{vdW} \pm 0.04$ components (Eq. 4.22, 4.23 and 4.24 of the change in free energy of mutation ($\Delta\Delta G$; Eq. 4.20) for the indicated pairs of amino acids of the halophilic protein L., pdb ID: 2KAC, at b_{KCl} = 2 mol/kg. 70

4.17 A snapshot of REMD simulation of halophilic protein L., pdb ID:2KAC, in a denatured state, with new cartoon representation and acidic residues, blue and red branches. This is the median structure which is used for the free energy calculation. 72

4.18 This figure shows the trajectory analysis comparison between the unfolded halophilic protein L., pdb ID:2KAC, from REMD simulation in Section 4.3.2, and protein L. in the folded state, from simulation of previous chapter in Section 3.2, at 298 K and b_{KCl} = 2 mol/kg. In the x-axis, for the folded structure, each frame accounts for 100 ps for a total simulation of 1 μs, and 10 ps for each frame in the unfolded case, and a total of 100 ns. 73

4.19 Normalized histogram of the carbon of carboxylate minimum distance between two closest acidic amino acids (D=aspartic acid or E=glutamic acid), averaged over all acidic amino acids and all saved configurations, for the denatured protein L., from simulation in Section 4.3.2 and protein L. in the folded state, from simulation in Section 3.2, at 298 K, and b_{KCl} = 2 mol/kg. 74

4.20 Change in free energy of mutation ($\Delta\Delta G \pm 1.0$; Eq. 4.20) for the indicated pairs of amino acids of the denatured halophilic protein L. (pdb ID: 2KAC), calculated at the high KCl molality. The data shown in this plot is compiled in Table A.5. 76

4.21 Distance between the carbons of carboxylates of pairs of acidic amino acids as they where studied, for all the ones showing synergistic effect, soft orange color, versus when they are interfering, moderate blue. 78

4.22 Distance between the potassium ions within 10 Å of the oxygens atoms of carboxylates of pairs of acidic amino acids as they where studied, for all the ones showing synergistic effect, soft orange color, versus when they are interfering, moderate blue. 79

4.23 Distance between the potassium ions and the carbon atoms of the carboxylates of both pairs of acidic residue, potassium ions within 10 Å of these carbons are only considered, as they where studied, for all the ones showing synergistic effect, soft orange color, versus when they are interfering, moderate blue. 79

4.24 Figures (a) and (b) show the RDF of potassium ions and water molecules around the oxygens of carboxylates of vicinal acidid residues, respectively, for synergistic versus interfering cases. Figure (c) shows the distance between the potassium ions and the carbon atoms of the carboxylates of both pairs of acidic residue, potassium ions within 10 Å of these carbons are only considered. Figure (d) shows the distance between the potassium ions within 10 Å of the oxygens atoms of carboxylates of pairs of acidic amino acids. For Figures (c) and (d) cases with synergistic effect are shown with soft orange color, versus when they are interfering with moderate blue. 81

A.1 Number of water-protein hydrogen-bonds per residue, averaged over the indicated amino acid types (acidic, basic, hydrophobic, and hydrophilic (=polar, non-charged), for the indicated halophilic-mesophilic proteins, identified by their pdb IDs. 103

A.2 Cumulative number of potassium ions as a function of the distance to the heavy atoms defining the protein surface, from simulations at the indicated concentration of KCl. 104

A.3 MSD of the subpopulation of water molecules that belong to the first hydration shell of the halophilic or non-halophilic ferredoxin at $t = 0$, for the indicated salt concentration. The light blue line illustrates the diffusive limit of a Brownian particle with $D = 1/6$ Å^2/ps. 105

A.4 MSD of the subpopulation of K^+ ions that belong to the first hydration shell of the halophilic or non-halophilic ferredoxin at $t = 0$, for the indicated salt concentrations. The light blue line illustrates the diffusive limit of a Brownian particle with $D = 1/6$ Å^2/ps. 105

A.5 Diffusion coefficients of: (First-shell) water molecules that belong to the first hydration shell of the indicated proteins at $t = 0$, simulated at b_{KCl} = 0.15 mol/kg; (Bulk) all water molecules in the same simulation. 106

A.6 Diffusion coefficients of: (First-shell) potassium ions that belong to the first solvation shell of the indicated proteins at $t = 0$, simulated at b_{KCl} = 0.15 mol/kg; (Bulk) all potassium ions in the same simulation. 106

A.7 Radial distribution function between the metal ion (M^+) and the carboxylate carbon (C^-), from simulations of aqueous solutions of $NaCH_3COO$ with molality 0.5 mol/kg, or KCH_3COO at the same concentration. The parameters for the interaction between carboxylate and water and between carboxylate and Na^+ are from ref. 46; the interaction between carboxylate and K^+ is modelled using the optimized parameter shown in Table 2.1 of the main text. The remaining parameters are from GAFF and TIP3P water. 107

A.8 Experimentally determined water activity (a_w) in solutions of NaCl or KCl with the indicated molality, *b*. From ref. 82. 108

List of tables

2.1 Recommended value of $R_{\min}$ for the LJ potential between K^+ and any carboxylate oxygen, O, when using TIP3P water, the Joung and Cheatham [42] parameters for K^+ and the GAFF [110] or AMBER [65] force fields with the optimized self-interaction parameters for carboxylates from ref. 46. 16

2.2 Molal and molar solution activity derivative of aqueous solutions of KCH_3COO with the indicated molality. 17

3.1 Halophilic-mesophilic protein pairs. 21

3.2 Length (n_{aa}), number of acidic (n_{acidic}) and basic (n_{basic}) amino acids and charge of the simulated proteins. The protein charge is defined by the difference between acidic and basic amino acids, and by the charge of metal ligands if present. 22

4.1 Free energy values in units of kcal/mol, calculated from different perturbation-based, BAR , and integration-based, TI-3, methods for the cases with a total of 95 λ, and 177 λ. 53

4.2 Free energy values in units of kcal/mol, calculated from different perturbation-based, BAR , and integration-based, TI-3, methods for the cases with restrained backbone and free backbone. 56

A.1 Distance between K^+ and the indicated carboxylate oxygens of acidic amino acids in an halophilic ferredoxin (pdb ID 1DOI), from crystallography and from simulation. Related to Fig. 2.3 of the main text. 101

A.2 Free energies (kcal/mol) of mutation for halophilic protein ferredoxin. The standard error of the mean (SEM) is reported for each value. $\Delta G_{XaY} = \Delta G_{XaX^0} + \Delta G_{X^0aY^0} + \Delta G_{Y^0aY}$; see Scheme 4.10. 109

A.3 Free energy of mutation of Aspartic acid to Asparagine for halophilic protein L. 109

A.4 Free energy of mutation of Aspartic acid to Asparagine for halophilic protein Dihydrofolate reductase 110

A.5 Free energy of mutation of Aspartic acid to Asparagine for unfolded halophilic protein L. 111

Chapter 1

Introduction

The first three chapters are based on the article by Geraili and Vila Verde, "Proteins maintain hydration at high [KCl] regardless of content in acidic amino acids", accepted and pending publication at *Biophysical Journal*.

Halophilic organisms, unlike most life on Earth, have the uncanny ability to survive at molar external NaCl concentrations. Despite their relative scarcity, halophilic organisms have been found in multiple kingdoms of life: archaea [18], bacteria [105], protozoa, fungi, algae, and multicellular eukaryotes [30, 31]. Aqueous environments with NaCl mass fractions of 10 to 15% (c_{NaCl} = 2-3 mol/dm^3) are required for growth of many obligate halophiles, with optimum growth being reached above 20% (c_{NaCl} = 4 mol/dm^3) [57]. Non-obligate halophiles, in contrast, thrive at high external NaCl concentrations but can survive without them [18]. To counterbalance the high osmotic stress induced by high external NaCl concentrations, some halophiles accumulate equivalent concentrations of KCl in their cytoplasm [57]. At such high salt concentrations, the interactions that dictate the structure and structural stability of proteins differ markedly from those at the much lower salt concentrations found in most organisms. As a result, many proteins of mesophilic organisms (here termed mesophilic proteins) are poorly soluble at KCl concentrations typical of the cytoplasm of halophilic organisms [20, 32, 102]. In contrast, proteins of halophilic organisms (here termed halophilic proteins) are structurally stable and functioning at high salt concentrations, but often show lower (or no) stability and activity at KCl concentrations typical of the cytoplasm of mesophilic organisms [57]. Their resilience is related to their amino acid composition [30, 49, 57, 74, 89]. Cytoplasmic proteins of halophilic organisms are on average longer than their mesophilic counterparts; they also have a higher fraction of random coil structure at the expense of α-helix structure [112]; they are richer in small+polar and small+apolar amino acids, and poorer in large+hydrophobic amino acids; they are also poorer in cationic amino acids, which contain longer alkyl sections in their side chains than the anionic ones. The reduction in the fraction of large+hydrophobic

and cationic amino acids and the increase in small+polar and small+apolar is thought to compensate the increased attraction between non-polar groups as salt concentration increases [57].

Halophilic proteins are also much richer in acidic amino acids [30, 64] – mostly located on their surface [49] – than their mesophilic counterparts, leading to substantially negative net protein charges. Understanding the mechanisms that allow proteins to function at high salt concentrations is important, both from a fundamental perspective, and also because of the great technological potential these protein have. The role played by the excess acidic amino acids, however, is not yet well-understood. One of the most intuitive explanations is that large surface charge prevents protein aggregation at high salt concentrations. The excess surface charge in halophilic proteins would compensate for charge screening at high salt concentrations and would prevent aggregation [17, 30, 57, 96, 114]. Several results, however, suggest that it is unlikely that charge repulsion is the only – or even main – function of excess acidic amino acids. Experiments show that the stability of halophilic proteins at 0.05 mol/dm^3 NaCl and $\approx 10^{-6}$ mol/dm^3 NADH+ is as high as that reached with molar concentrations of NaCl [69]. Charge screening alone does not explain the strong dependence of the stability of the folded state of halophilic proteins, and/or their function, on salt concentrations up to several molar, because measurements [57] and calculations [17] indicate that screening of electrostatic interactions by salt is largely complete at c_{salt}= 0.5 mol/dm^3. At the high salt concentrations seen inside halophiles, electrostatic repulsion between the carboxyl groups at the surface of proteins should have only a small effect on their structural stability, as one NMR study shows [96], and likewise on their ability to aggregate.

At present, three other explanations have been proposed for the large fraction of acidic amino acids in halophilic proteins:

Ion-solvent stabilization model. Experiments have shown that some folded halophilic proteins bind large amounts of water, NaCl and KCl [12, 78]; in contrast, salt binding was not detected in mesophilic proteins in the native state [63]. Morevoer, quasielastic neutron spectroscopy (QENS) measurements of water translational dynamics in the cytoplasm of halophilic and non-halophilic bacteria suggest that halophiles have a water fraction with extremely slow dynamics that is absent from non-halophiles [40]. Based on these results, it has been proposed that the abundance of acidic amino acids enables cooperative, stabilizing interactions with cations. The acidic amino acids contribute to a net stabilization of the folded protein structure (despite intramolecular electrostatic repulsion) by forming cooperative hydrated ion networks [63, 115]. These networks keep the folded protein hydrated despite the high salt concentration and should lead to highly ordered protein hydration shells [30]. According to this view, excess acidic amino acids indeed prevents aggregation of proteins

at high salt concentration, not by charge repulsion, but because the cooperative ion-water-protein networks stabilize the protein solvation layer.

The cooperativity aspect of the stabilization is important in this model, because carboxyl groups are strongly hydrated both in the folded and in the unfolded forms of a peptide [55]; only if binding of water and ions to the folded structure is enhanced, via a cooperative effect, relative to the unfolded structure would it result in stabilization of the folded structure. Recent NMR measurements of halophilic and non-halophilic proteins also support the ion-solvent stabilization model, and have added further support to the notion that both folded and unfolded states add to haloadaptation [71]: the measurements indicate that there are weak interactions between the acidic amino acids and the cations in solution which stabilize the folded structure of halophiles; simultaneously, these interactions are absent from the unfolded state of halophiles because, the authors claim, the flexible side chains of the acidic amino acids do not have the necessary preorientation to allow for their stable interaction with cations [71]. Because the cations are excluded from the hydration layer of the protein in the unfolded state, they compete with the protein for hydration, leading to a destabilization of the unfolded state which further reinforces the stability of the folded structure.

Solvent-only stabilization model. The ion-solvent stabilization model has been challenged, however, based on multiple observations. ^{17}O magnetic relaxation experiments did not find any differences in the rotational dynamics in hydration shells of halophilic and non-halophilic versions of protein L, [79] in contrast to the predictions of the ion-solvent stabilization model. Additionally, recent experiments have shown that some extremely halophilic organisms can thrive at low external NaCl concentrations, i.e., also with cytoplasmic KCl concentrations equal to those found in non-halophiles, despite having a markedly acidic proteome [18]; this result suggests that stabilizing interactions between cations and the acidic residues are not critical. To explain these observations, it was proposed that an abundance of acidic amino acids at the surface of halophilic proteins is necessary to compete – rather than cooperate – with the ions in solution for available water, and thus ensure that the protein surface remains sufficiently hydrated – and the protein remains soluble, conformationally stable and thus functional – at high salt concentrations [18, 24]. In this model the interactions between the acidic amino acids and the cations do not play a particularly stabilizing role at high salt concentrations.

Conformational fluctuations model. Finally, an unexplored hypothesis regarding the role of acidic amino acids in protein function is their impact on protein conformational flexibility, and therefore on protein function. Protein flexibility is key for protein function [2, 22, 100]: for example, X-ray structures of proteins show only part of the pathways followed by H_2 in dehydrogenases; to fully account for the experimentally measured diffusion

rates, it is necessary to consider also transient pathways that form because of conformational fluctuations of the protein [111]. Studies on non-halophilic proteins indicate that a decrease in protein function can be correlated with increased protein stability in many cases, and that this correlation arises because of a reduction of protein flexibility [100]. The increased magnitude of the hydrophobic effect at high salt concentrations raises the possibility that conformational fluctuations in halophiles would be tendentially smaller than in non-halophiles [69]. The increased intramolecular repulsion brought by an acidic proteome might aid in the maintenance of conformational fluctuations key for protein function, as illustrated in Figure 1.1. To date, however, studies of halophilic proteins have not examined the connection between changes in protein flexibility and function.

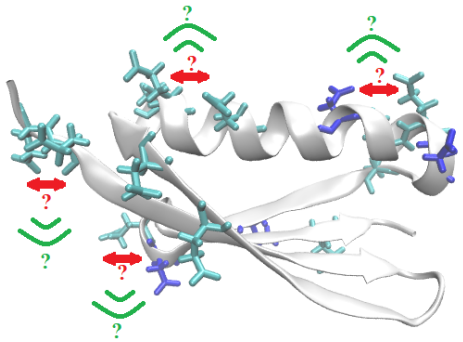


Fig. 1.1 Conformational fluctuations model: electrostatic repulsion (red arrows) between acidic amino acids (blue) offset the higher attraction between hydrophobic groups at high salt concentrations (not shown), to maintain flexibility (green) necessary for function. The question marks highlight that this model has not been explored.

Here we report a molecular dynamics study of 5 pairs of halophilic and mesophilic proteins. We comparatively examine how the structure and dynamics of protein hydration shells are affected by the concentration of KCl and the fraction of acidic amino acids in the protein. We also present parameters for the interaction between K^+ and carboxylate, critical for this study, optimized to reproduce the experimental solution activity derivative of potassium acetate solutions up to 2 mol/kg and crystallographic information of potassium ions in the vicinity of acidic amino acids. In Chapter 4, using alchemical mutation of amino acids, we examine the possibility of cooperative interactions between acidic amino acids–potassium–water to study the *Ion-solvent stabilization model*. We study the possibility and magnitude of such cooperative effects for several halophilic proteins at both high and low salt concentrations in their folded conformation and the unfolded state of halophilic protein L. as well. Throughout this study, we focus on the analysis of the two models *Ion-solvent stabilization model* and *Solvent-only stabilization model*.

Chapter 2

Optimizing the force field for $K^+ \cdots$ carboxylate interactions

2.1 Available force fields

To gain insight into the role played by acidic amino acids in halophilic proteins using simulations, we require force fields for proteins, water and ions with the correct balance of interactions between all the species, both at low and high (up to several molal) KCl concentrations. The force fields we use – the TIP3P water model [41], the AMBER ff14SB [6, 65] force field for proteins, the GAFF force field for acetate [110] and the potassium and chloride parameters of Joung and Cheatham [42] for TIP3P water – meet this requirement for most interactions, so only a few interactions were modified as described below. The AMBER ff14SB force field used with TIP3P water reproduces the hydration free energies of small-molecule analogues of the side chains of neutral amino acids to within 1 kcal/mol Root Mean Square Error of the experimental reference values [118], and the partition coefficients of the same analogues between water and several organic solvents to within 0.5 log units [117]. This combination of solvent and protein force field reproduces the secondary structure content of short peptides [65] and the backbone dynamics in the native state of globular proteins [65]. The experimental hydration free energies of K^+ and Cl^- were one of the target properties used in the parameterization done by Joung and Cheatham [42]. Their parameters reproduce the solution activity of aqueous solutions of this salt up to $b = 2$ mol/kg, [43] so they are appropriate for both dilute and concentrated solutions. Metallic ligands present in the proteins are simulated using AMBER-compatible parameters reported in the literature. The $[2Fe\text{-}2S]^{2+}$ ligand present in both ferredoxin proteins studied here is simulated based on the parameters developed by Carvalho et al. [13] for the [2Fe-2S] ferredoxin from *Mastigocladus*

laminosus, with slight modifications: i) All equilibrium angles are set to the values found in the crystal structure of our ferredoxin proteins. ii) the force constant for the Fe-S-C angle, missing in the parameter set of Carvalho et al. for the $[2Fe\text{-}2S]^{2+}$ from *Mastigocladus laminosus*, is set to the value reported by the same authors for the desulforedoxin protein from *Desulfovibrio gigas*. The desulforedoxin protein contains Fe(III) coordinated by four cysteines in the same configuration as the ferredoxins studied here. Both carbonic anhydrase proteins studied here have a 4-coordinated zinc metal center, where the zinc ion is connected to three histidine residues and one water molecule. The metal center is simulated using ZAFF (Zinc AMBER force field) [61, 75] and the Lennard Jones (LJ) parameters for the zinc ion from Li et al. [61].

Modifications to these force fields are indispensable for some of the interactions involving carboxylate groups or K^+. We modify the self-interaction LJ parameters for all carboxylate oxygens to the values proposed by Kashefolgheta et al. [46], which reproduce the hydration free energy of acetate in TIP3P water better than the original AMBER parameters; note that these parameters are used in all interactions derived using combination rules. The LJ parameters for the interaction between carboxylate and the side chain of the positively charged amino acid lysine are also modified to those in ref. 46, to reduce excessively strong salt bridges. The LJ parameters for the interaction between K^+ and carboxylate oxygens are modified as described in the following section.

Results from simulations without any modifications to the default Lennard-Jones parameters, provided for comparison, are marked as "AMBER/GAFF-only".

2.2 Optimization approach

Our tests (described in the Results section) show that the default parameters for $K^+ \cdots$ carboxylate interactions cause excessive contact between these ions. As this interaction is critical for our study, our first step was to find optimal parameters to describe it.

Intermolecular interactions between any atoms i and j have the following form in the AMBER force field:

$$V(r_{ij}) = \varepsilon_{ij}\left[\left(\frac{R_{\min,ij}}{r_{ij}}\right)^{12} - 2\left(\frac{R_{\min,ij}}{r_{ij}}\right)^{6}\right] + \frac{q_i q_j}{4\pi\varepsilon_0 r_{ij}} \tag{2.1}$$

In this expression, q_i and q_j are the charges of the atoms, r_{ij} is the distance between them, ε_0 is the dielectric permittivity in the vacuum. $R_{\min,ij}$ and ε_{ij} are parameters determining the Lennard-Jones potential, which mimics the van der Waals (vdW) interatomic interaction. Lennard-Jones parameters for the interaction between different atom types are typically

obtained from their self-interaction parameters using combination rules:

$$R_{\min,ij} = \frac{\left(R_{\min,ii} + R_{\min,jj}\right)}{2}, \quad \varepsilon_{ij} = \sqrt{\left(\varepsilon_{ii}\,\varepsilon_{jj}\right)} \tag{2.2}$$

where the indices ii and jj denote the self-interaction parameters.

In the AMBER force field, ions have their nominal charge, and the partial charges of polyatomic ions are determined using a well-defined charge fitting procedure. We opted to tune the $K^+\cdots$ carboxylate interaction by modifying the Lennard-Jones potential between K^+ and the carboxylate oxygen (O) only. The partial charges on the carboxylate atoms remain unchanged, which ensures compatibility within the AMBER force field. In principle both $R_{\min,ij}$ and ε_{ij} obtained using Eq. 2.2 could be overridden by optimized values for the $i = K^+$ $j = O$ pair. This dual optimization is, however, not feasible because of its high computational cost. We optimized only the $R_{\min,K^+O}$ parameter and left ε_{K^+O} at the value obtained with the combination rules. Our prior work confirms that substantial improvements in the description of intermolecular interactions are achieved even with this limited parameterization freedom [46]. In the Results section we show that the original AMBER parameters lead to substantial deviations between various calculated properties and the experimental reference values. Because of this substantial deviation, the $R_{\min}$ parameter, with its power of 6 and 12 in the Lennard-Jones potential, is a better choice for parameterization than ε.

The $R_{\min,K^+O}$ parameter is optimized to reproduce i) the experimental solution activity derivative of aqueous solutions of potassium acetate with molality $b_{KCH_3COO} \in \{0.5, 1, 2\}$ mol/kg, and ii) the distances between K^+ and carboxylate groups found in the crystal structure of the halophilic 2Fe-2S ferredoxin protein with pdb ID: 1DOI. This protein is one of the few for which the potassium ions present during crystallization were resolved in the crystal structure.

The second target property was added because we found that the solution activity derivative varies non-monotonically with $R_{\min,K^+O}$, to the extent that $R_{\min,K^+O}$ values leading to dramatically different solution structure yielded identical solution activity derivatives, as described below. Our final choice of $R_{\min,K^+O}$, shown in the Results section, is that which best reproduces selected distances between K^+ and carboxylate oxygens in the protein crystal structure, while yielding activity derivatives deviating not more than 7% from the reference value for all three concentrations. Our optimized LJ potential for the interactions between K^+ and carboxylate groups are thus appropriate for simulations of biomolecular systems in a wide range of KCl concentrations, and are also appropriate for simulations of systems with acetate ions modelled with the AMBER/GAFF force field.

2.2.1 Calculating the electrolyte activity derivative in simulations

The molar electrolyte activity derivative can be calculated from simulation using Kirkwood-Buff theory [51, 56], which links density fluctuations to thermodynamic properties. Below we summarize the main expressions used in this work. Examples of its application to other systems can be found in the literature [27, 29, 52, 107]. The central quantity of this theory is the Kirkwood-Buff integral, G. In this work we use expressions that enable its use with simulations of closed systems [54]. For any 2 species i and j, the integral is:

$$G_{ij}(R) = \int_0^{2R} \left[f_{ij} g_{ij}^{NVT}(r) - 1 \right] 4\pi r^2 \left(1 - 3x/2 + x^3/2\right) dr \tag{2.3}$$

where R is 1/4 of the simulation box size, $x = r/(2R)$ and $g_{ij}^{NVT}(r)$ is the radial distribution function in the canonical ensemble. The factor f_{ij} corrects for the fact that the tail of the radial distribution functions (RDFs) in closed systems does not converge to 1. Even though there is no formal relation between the RDFs in open and closed systems, the ratio of their tails is expected to be of the order of $1 \pm 1/N$ where N is the number of particles in the simulation box [36, 59, 80]. For this reason, a multiplicative correction factor is appropriate.

Because of electroneutrality, applying Kirkwood-Buff theory to an electrolyte solution requires the solution to be treated as a binary system of indistinguishable ions (called the cosolvent) and water [27, 29, 56]. The molar cosolvent activity derivative, a_c', for a solution of a 1:1 electrolyte is thus defined as

$$\begin{aligned} a_c' &= \left. \frac{\partial \ln a_c}{\partial \ln \rho_c} \right|_{p,T} \\ &= 1 + \left. \frac{\partial \ln \gamma_c}{\partial \ln \rho_c} \right|_{p,T} \end{aligned} \tag{2.4}$$

In this expression γ_c is the molar activity coefficient of the cosolvent, ρ_c is its number density, defined as $\rho_c = n_c/V$ where $n_c = n_+ + n_-$ is the total number of positive and negative ions in the solution volume V, and a_c is its molar activity ($a_c = \gamma_c \times \rho_c$). Kirkwood-Buff theory allows the calculation of the electrolyte activity derivative from simulation, as

$$a_c' = \frac{1}{1 + \rho_c (G_{cc} - G_{cw})} \tag{2.5}$$

where the subscript w refers to the water. The terms in the above expression are sums of Kirkwood-Buff integrals between the ions and water:

$$G_{cc} = \frac{1}{4}\left(2G_{+-} + G_{++} + G_{--}\right) \tag{2.6}$$

and

$$G_{cw} = G_{+w} + G_{-w} \tag{2.7}$$

2.2.2 Experimental electrolyte activity derivative

Experimental activities for electrolytes are typically reported considering the neutral salt unit – as opposed to considering a cosolvent of indistinguishable ions – and in the molal rather than the molar scale. In what follows the subscript s indicates quantities defined in terms of neutral salt units and the superscript (b) indicates quantities in the molal scale. The mean molal activity coefficient, $\gamma^{(b)}_{s,\pm}$, of potassium acetate in aqueous solution has been experimentally determined for molalities up to $b_s = 3$ mol/kg [83]. In Fig. 2.1 we show the experimental values of $\gamma^{(b)}_{s,\pm}$ as a function of the molality of KCH_3COO, together with the result of the Pitzer equation applicable to this system [77]. The Pitzer equation reproduces the experimental data excellently, so we use it to calculate the molal salt activity derivative

$$a'^{(b)}_s = \left.\frac{\partial \ln a^{(b)}_s}{\partial \ln b_s}\right|_{p,T} = \left.\frac{\partial \ln\left(\gamma^{(b)}_{s,\pm}\, b_s\right)}{\partial \ln b_s}\right|_{p,T} \tag{2.8}$$

using Mathematica.

To enable direct comparisons between experiment and simulation, the experimental molal activity derivative of the salt must first be converted to the corresponding molar quantity:

$$a'_s = \left.\frac{\partial \ln a_s}{\partial \ln \rho_s}\right|_{p,T} = 1 + \left.\frac{\partial \ln \gamma_{s,\pm}}{\partial \ln \rho_s}\right|_{p,T} \tag{2.9}$$

where ρ_s is the number density of the salt and $\gamma_{s,\pm}$ its mean molar activity coefficient. The mean molal and molar activity coefficients are related as [4]

$$\gamma_{s,\pm} = \frac{b_s \rho_w}{\rho_s}\gamma^{(b)}_{s,\pm} \tag{2.10}$$

where ρ_w is the density (mass/volume) of water and ρ_s is the number density of the salt; for a 1:1 electrolyte, $\rho_s = \rho_c/2$. Substituting eq. 2.10 into eq. 2.9 shows that the salt activity

derivatives in the two scales are related as

$$a'_s = a'^{(b)}_s \frac{\partial \ln b_s}{\partial \ln \rho_s} \tag{2.11}$$

We use eq. 2.11 with the solution density reported in ref. 11 to obtain a'_s from $a'^{(b)}_s$. The molar solution activities expressed in terms of the salt and the cosolvent are related as $a_s = 0.5a_c$ for 1:1 electrolytes; their derivatives and their activity coefficients are identical ($a'_s = a'_c$ and $\gamma_{s,\pm} = \gamma_c$); the experimental a'_s can thus be directly compared to a'_c calculated from simulation.

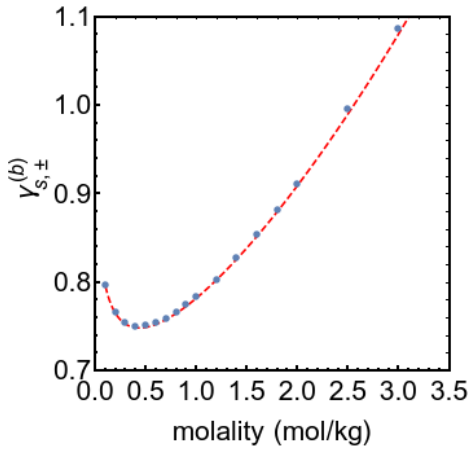


Fig. 2.1 Mean molal activity coefficient ($\gamma^{(b)}_{s,\pm}$) of potassium acetate as a function of the molality of KCH_3COO. The blue circles are the experimental data [83] and the red dashed line is calculated using the relevant Pitzer equation [77].

2.3 Simulation details

2.3.1 Potassium acetate solutions

We perform molecular dynamics (MD) simulations using the GROMACS simulation package in its 2018 version [9, 104]. Simulations are performed for three different concentrations, $b_{KCH_3COO} \in \{0.5, 1, 2\}$ mol/kg. Each simulation – minimization, equilibration and production run – is performed for a given test value of $R_{\min,K^+O}$, and multiple values are tested at each concentration. These test values are related to the default, $R_{\min,K^+O,LB}$ (obtained using the Lorentz-Berthelot combination rule; eq. 2.2), via a scaling factor:

$$R_{\min,K^+O} = f_{R_{\min,K^+O}} \cdot R_{\min,K^+O,LB} \tag{2.12}$$

For convenience, we refer to the tested values of $R_{\mathrm{min,K^+O}}$ via the scaling factor $f_{R_{\mathrm{min,K^+O}}}$.

The simulation boxes are prepared by putting 72, 144 or 288 neutral salt units, KCH_3COO – corresponding to the low, intermediate and high salt concentrations – in a cubic box with edge length $L \approx 6$ nm and solvating the systems by adding ≈ 8000 TIP3P water molecules using the editconf and solvate tools from Gromacs 2018 [9, 104]. We set a non-bonded potential cutoff at 1.2 nm for LJ interactions. Electrostatic interactions are calculated using direct summation up to the same cutoff distance; beyond it , the electrostatic interactions are calculated with the particle mesh Ewald (PME) scheme with a grid spacing of 0.12 nm, and 6^{th} order interpolation [16]. Lennard-Jones interactions are smoothly shifted to zero between 1.0 nm and 1.2 nm using the switch function available in GROMACS. Long range dispersion corrections are applied to both the energy and pressure. A leap-frog stochastic (SD) integrator [106] is used to integrate the equations of motion in all simulations, with the temperature fixed at 298 K. All bonds with H-atoms are restrained using the LINCS algorithm [33] in all simulation steps (except the minimization step with l-bfgs method), which enables integration using a 2 fs time step. The systems are equilibrated following a protocol commonly used in biological studies: i) Two initial minimization steps with the steepest-descent and l-bfgs algorithms. The latter is a quasi-Newtonian algorithm for energy minimization which converges faster than the Conjugate-Gradient algorithm. ii) A 500 ps heating equilibration simulation in the canonical ensemble (NVT) using Langevin thermostat with a coupling constant of 1.0 ps to 298 K. iii) Another 10 ns simulation is performed in the isothermal-isobaric ensemble (NpT) to equilibrate the system density at the pressure of 1 bar, using the Berendsen barostate[8] with a relaxation time of 1.0 ps. We select three distinct configurations from the NPT equilibration simulation, with a box volume similar to the average box volume obtained during that simulation. Each of these configurations is then used as the initial state of a production simulation in the NVT ensemble and lasting 50 ns, making up 150 ns for each concentration. This large simulation time is necessary to calculate activity derivatives with high precision. The trajectories of production simulations are saved every 2 ps for analysis.

2.3.2 Restrained ferredoxin in c_{KCl} = 1 mol/dm^3

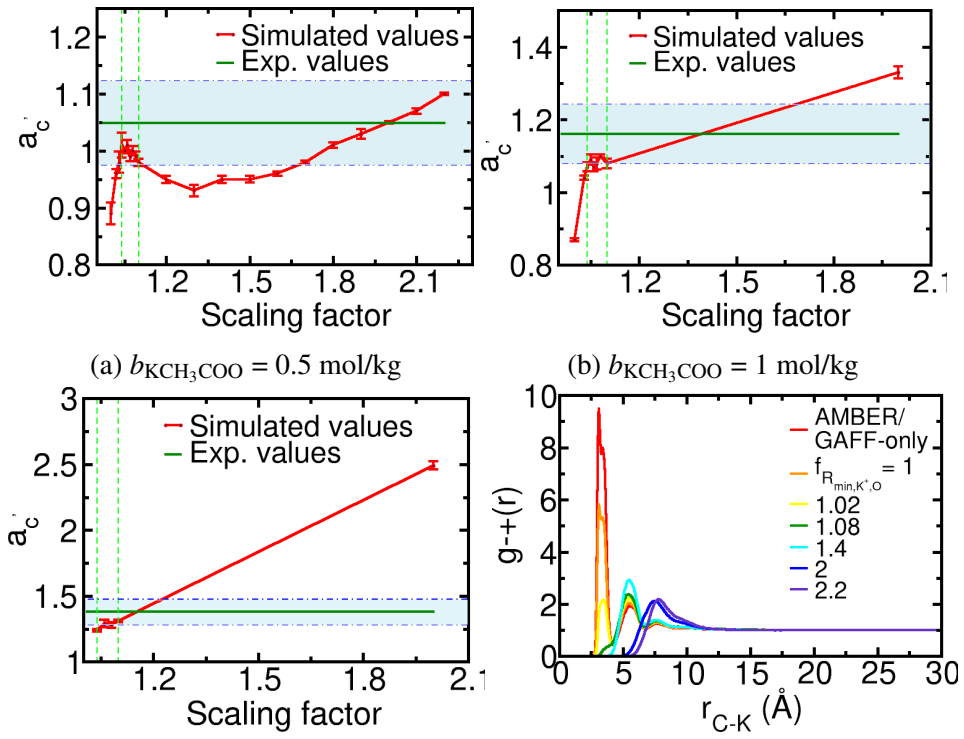
Simulations are performed with Gromacs 2018; unless otherwise noted, all simulation details are identical to those used to simulate potassium acetate solutions. Each simulation – minimization, equilibration and production run – is performed for a given test value of $R_{\mathrm{min,K^+O}}$. The simulation box is prepared by solvating the protein with a cubic box of TIP3P water, with the distance between the protein surface and the box face being at least 15 Å. The average temperature of 298 K is enforced using the Langevin thermostat with a coupling

constant of 1.0 ps; an average pressure of 1.0 bar is enforced via the Berendsen barostat with a relaxation time of 1.0 ps. During all simulation steps, the backbone atoms are restrained to their position in the crystal structure at all times, using a harmonic potential with a force constant of 1000 $kJ \cdot mol^{-1} \cdot nm^{-2}$. Applying this restraint is necessary to compare the distance of the K^+ ions to carboxylate groups between the simulated system and the crystal structure for each value of $R_{\min,K^+O}$ tested. A single NPT production simulation with duration 200 ns is performed for each $R_{\min,K^+O}$.

2.4 Results and discussion

The activity derivative of potassium acetate solutions with molality $b_{KCH_3COO} \in \{0.5, 1, 2\}$ mol/kg calculated in simulation is shown in Fig. 2.2. For each concentration, the activity derivative is calculated for different values of $R_{\min,K^+O}$, which govern the LJ interaction between K^+ and the carboxylate oxygens; these values are expressed in the figures in terms of a multiplicative scaling factor (defined in eq. 2.12). The unoptimized anion-cation parameters, corresponding to the scaling factor of 1 in Fig. 2.2, yield activity derivatives far from the reference value from experiment (green lines). This deviation demonstrates that the unoptimized parameters fail dramatically to adequately describe $K^+ \cdots CH_3COO^-$ interactions in the concentration range we tested; optimizing this interaction is indispensable. With this in mind, we first examine the solution activity derivative for $b_{KCH_3COO} = 0.5$ mol/kg (Fig. 2.2A). It varies non-monotonically with increasing scaling factor: initially it rapidly increases, then decreases slowly before again increasing. Perfect agreement with the reference experimental value at this concentration occurs for $f_{R_{\min,K^+O}} = 2.0$. Despite this agreement, $f_{R_{\min,K^+O}} = 2.0$ does *not* result in the correct solution structure at $b_{KCH_3COO} = 0.5$ mol/kg: the corresponding anion-cation radial distribution function, shown in Fig. 2.2D, would suggest that only solvent-separated (2SIP) ion pairs exist in potassium acetate solutions, and that neither contact (CIP) nor solvent-shared (SIP) occur (CIP: the two ions are in direct contact; SIP: the ions share one hydration layer; 2SIP: each ion retains its first hydration layer). The absence of CIPs disagrees with potentiometric measurements, which indicate that K^+ and CH_3COO^- associate – albeit weakly – to form neutral complexes [15, 23] best understood as pairs of ions in direct contact [26].

The anion-cation radial distribution functions in Fig. 2.2D indicate that scaling factors in the range $1.02 < f_{R_{\min,K^+O}} < 1.1$ allow the presence of CIPs in solution, i.e., yield a solution structure qualitatively in line with experiment. This range of scaling factors also yields solution activity derivatives within 7% of the target experimental values for all three concentrations (Fig. 2.2A,B,C). However, despite the fact that solution activity derivatives



(a) b_{KCH_3COO} = 0.5 mol/kg

(b) b_{KCH_3COO} = 1 mol/kg

(c) b_{KCH_3COO} = 2 mol/kg

(d) b_{KCH_3COO} = 0.5 mol/kg

Fig. 2.2 (A,B,C) Molar solution activity derivative (red points) of potassium acetate solutions, as a function of the multiplicative scaling factor ($f_{R_{min,K^+O}}$; see eq. 2.12) applied to the LJ $R_{min,K^+O,LB}$ parameter governing the interactions between K^+ and carboxylates. The error bars are the standard error of the mean calculated from three independent production simulations. The red lines are a guide to the eye. The green line shows the experimental reference value; see also table 2.2. The shaded regions show the ±7% deviation from the experimental value. The two vertical dashed lines delimit the range of scaling factors acceptable for the 3 concentrations. (D) Radial distribution function of potassium and the carbon bonded to the oxygens of acetate (the same simulations as in panel A) for the indicated values of the scaling factor $f_{R_{min,K^+O}}$.

remain constant for this range of $f_{R_{\min,K^+O}}$ values, the solution structure changes substantially: e.g., Fig. 2.2D shows that for $f_{R_{\min,K^+O}} = 1.02$, CIPs are still abundant although much less so than SIPs or 2SIPS; for $f_{R_{\min,K^+O}} = 1.1$, CIPs are residual. Clearly, randomly choosing a value of $f_{R_{\min,K^+O}}$ within the $1.02 < f_{R_{\min,K^+O}} < 1.1$ range is insufficient, and another experimental reference property is desirable to pinpoint the optimal parameter value.

The Protein Data Bank structure with pdb code 1DOI, of a halophilic ferredoxin, contains 5 co-crystalized K^+ near acidic amino acids at the protein surface. This crystal structure was obtained from both room temperature and 100 K difraction data and contains 237 water molecules, corresponding to 55% of the crystal volume. Flash freezing, typically used in crystallography to reach cryogenic temperatures, leaves the solvent in an amorphous state, but often contracts the unit cell by 2 to 7% with most of the contraction arising from the solvent [44]. These conditions mean that the protein surface and the potassium ions in this crystal structure are heavily solvated, and that the distances between K^+ and the nearby carboxylates in the crystal structure are a reasonable estimate of the distance between these ions when forming CIPs, for the solvated protein at room temperature. We used this information to select the optimal parameter within the $1.02 < f_{R_{\min,K^+O}} < 1.1$ range. To the best of our knowledge, other experimental observables that directly report on the structure of $K^+ \cdots$ carboxylate ion pairs in solvated conditions do not exist. We simulated the ferredoxin protein in c_{KCl} = 1 mol/dm^3 at $T = 298$ K. This high concentration of K^+ is within the range used in the previous parameterization step. The concentration of ions does not affect the position of the peaks of the radial distribution function, as shown in Fig. 2.2D, so other concentrations could also have been used; a high concentration is advantageous because it yields radial distribution functions with less noise, given the same simulation time. The backbone atoms were constrained to the coordinates of the 1DOI crystal structure, so that any differences in the $K^+ \cdots$carboxylate interaction between simulations with different $f_{R_{\min,K^+O}}$ values reflect the impact of the simulation parameters, and not steric changes arising from changes in protein conformation.

For each of the 5 acidic amino acids functioning as the contact site with K^+ in the crystal structure (Fig. 2.3A), we calculated the carboxylate(O)-K^+ radial distribution function.

Fig. 2.3B shows example RDFs for one of the protein sites and for different values of $f_{R_{\min,K^+O}}$. The position, r_{sim}, of the maximum of the first peak of each RDF was obtained as the mean of a Gaussian curve fitted to each peak. This fit allows a better identification of the position of the peak than taking the data point corresponding to the global maximum of each curve. The r_{sim} distances and the corresponding reference values (r_{cryst}) from the crystal are tabulated in SI table A.1 for each of the protein sites; r_{cryst} is the distance between K^+ and the carboxylate oxygen identified in Fig. 2.3C. The r_{sim} are always smaller than the

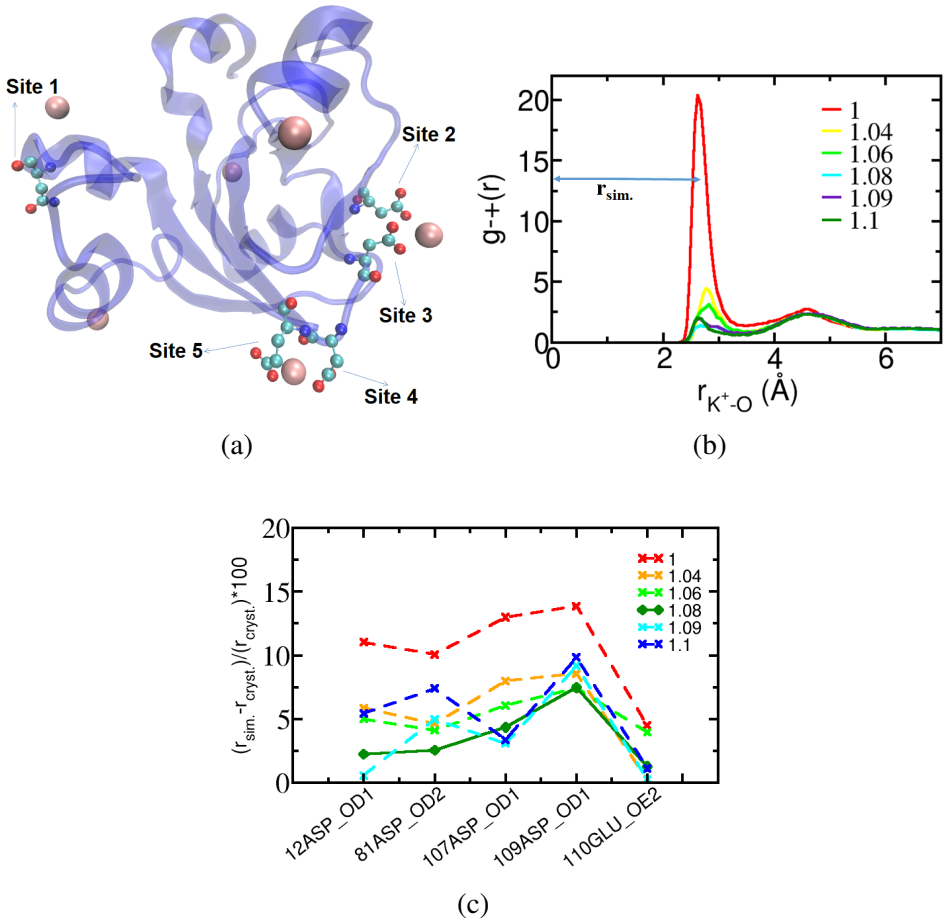


Fig. 2.3 (A) Crystal structure (pdb ID: 1DOI) of the halophilic 2Fe-2S ferredoxin from *Haloarcula Marismortui* [24]; the K^+ ions are shown in pink. The 5 acidic amino acid sites that have nearby K^+, used to parameterize $R_{\mathrm{min,K^+O}}$, are indicated. (B) Example radial distribution function of K^+ and the carboxylate oxygens of site 2, at T=298 K and c_{KCl} = 1 mol/dm^3. The distance to the first maximum is identified as $r_{\mathrm{sim.}}$. The legend shows the values of the scaling factor, $f_{R_{\mathrm{min,K^+O}}}$. (C) Unsigned relative deviation between r_{sim} and r_{cryst} for the five indicated sites of the halophilic ferredoxin. The protein site is identified by the residue number, residue name and oxygen name. The legend shows the values of the scaling factor $f_{R_{\mathrm{min,K^+O}}}$. Numerical data is shown in SI Table A.1.

reference values for all parameter values tested. The optimal value of $f_{R_{\mathrm{min,K^+O}}}$ is selected as that which leads to the smallest deviation between r_{sim}, and r_{cryst}. In Fig. 2.3C we show the unsigned relative deviation between r_{sim} and the reference value from the crystal structure, r_{cryst} for the 5 protein sites, for multiple values of $f_{R_{\mathrm{min,K^+O}}}$. A scaling factor $f_{R_{\mathrm{min,K^+O}}} = 1.08$ yields the lowest deviation over all protein sites; the corresponding value of $R^{\dagger}_{\mathrm{min,K^+O}}$ is shown in Table 2.1. Table 2.2 shows the activity derivative of potassium acetate solutions at the three concentrations investigated, for $f_{R_{\mathrm{min,K^+O}}} = 1.08$. These values deviate 4.7%, 5.3% and 7.2% from the corresponding experimental values for $b_{\mathrm{KCH_3COO}}$ = (0.5, 1, and 2) mol/kg, respectively. The position of the contact ion pair peak between the carboxylate oxygens and K^+ obtained with the optimal parameter is $r_{\mathrm{sim}} \approx 2.8$ Å, consistent with ab initio calculations using a polarizable continuum model for the solvent [3, 109], but larger than the optimal distance of 2.5 Å observed in gas phase calculations [92]. Despite the difference in optimal distance observed between those ab initio calculations, both predict that the contact ion pair between carboxylate and K^+ is less stable than that with Na^+, i.e., these cations follow the normal Hofmeister series when interacting with the carboxylate group. Those results agree with X-ray absorption measurements of acetate solutions [3], with prior molecular simulations using parameters optimized for lower concentrations [34], and with our simulations using our optimized parameters (SI Figure A.7).

Table 2.1 Recommended value of R_{min} for the LJ potential between K^+ and any carboxylate oxygen, O, when using TIP3P water, the Joung and Cheatham [42] parameters for K^+ and the GAFF [110] or AMBER [65] force fields with the optimized self-interaction parameters for carboxylates from ref. 46.

parameter	source	value
$R^{\dagger}_{\mathrm{min,K^+O}}$ (Å)	this work	3.6355
$f^{\ddagger}_{R_{\mathrm{min,K^+O}}}$	this work	1.08
$R^{\P}_{\mathrm{min,K^+O,LB}}$ (Å)	AMBER/GAFF	3.3662

† Optimized parameter value.
‡ Optimized scaling factor, defined in eq. 2.12.
¶ From combination rules (see eqs. 2.2); provided for comparison.

Table 2.2 Molal and molar solution activity derivative of aqueous solutions of KCH_3COO with the indicated molality.

b_{KCH_3COO} (mol/kg)	$a'_s{}^{(b)\,\mathrm{exp}\dagger}$	$a'_s{}^{\,\mathrm{exp}\ddagger}$	$a'_c{}^{\,\mathrm{Opt}}$	$a'_c{}^{\,\mathrm{AMBER/GAFF}}$
0.5	1.0143	1.0488	1.00 ± 0.01	0.83 ± 0.01
1	1.1229	1.1611	1.100 ± 0.004	NC
2	1.3345	1.3798	1.28 ± 0.02	NC

† Experimental molal activity derivative, from Pitzer equations [77].

‡ Experimental molar activity derivative, from Pitzer equations and eq. 2.11.

$^{\mathrm{Opt}}$ Simulation (this work) molar activity derivative, using the optimized LJ parameters listed in Table 2.1 and the self-interaction parameters for carboxylate from ref. 46. The uncertainty is the standard error of the mean, calculated from 4 independent simulations.

$^{\mathrm{AMBER/GAFF}}$ Simulation (this work) molar activity derivative, using the AMBER/GAFF force field. "NC"= not calculated.

Chapter 3

Structure and dynamics of the hydration shell of mesophilic vs. halophilic proteins

The ion-solvent stabilization model and the solvent-only stabilization model, proposed to explain the function of excess acidic amino acids in halophilic proteins, imply that the structure and dynamics of the hydration shell of halophilic proteins differs from that of mesophilic ones: according to both models, excess acidic amino acids is indispensable in halophilic proteins so they remain hydrated at high KCl concentrations. The solvent-only stabilization model proposes that acidic amino acids enhance hydration by direct interactions with the water, and that ion-protein interactions are not relevant for hydration or for protein stability. In contrast, the ion-solvent stabilization model proposes that protein hydration is maintained, and the folded protein structure is stabilized, by cooperative interactions between the acidic amino acids, the cations in solution (K^+) and water. According to this model, these cooperative interactions are possible only in the folded conformation of halophilic protein because only specific tertiary and quaternary protein structures enable them. These cooperative interactions are responsible for the large amounts of cations bound to halophilic proteins, and should result in excessively slow dynamics of water of hydration around those proteins. In what follows we compare the structure and dynamics of the hydration shells of 5 pairs of halophilic-mesophilic proteins, and discuss the results in the context of the predictions and assumptions of both models.

3.1 Selecting appropriate systems for a comparative study of halophilic and mesophilic proteins

To gain a comprehensive view of the role played by the excess acidic amino acids of halophilic proteins on the structure and dynamics of their hydration layer, we compare 5 halophilic proteins with their mesophilic counterparts. The five protein pairs investigated here were selected considering: (i) availability of experimentally-determined structures, (ii) existence of experimental studies on their conformational changes and/or activity as a function of salt concentration, (iii) availability of parameters for simulation, (iv) diversity of size and surface charge density; (v) similarity of the amino acid sequence and structure between each halophilic protein and its mesophilic pair. In Table 3.1 we list the pdb IDs of the proteins selected for the study, and show the structural and sequence similarity between each of the 5 halophilic-mesophilic pairs. The pair sequence identity varies between 19% and 90%. Despite this broad range in sequence identity, the proteins in each pair share a common structure, visible in the narrow range of the Root Mean Square Deviation (RMSD) of the C^{α} atoms of the backbone of the amino acids between protein pairs, after structural alignment: $0.95 < \text{RMSD/Å} < 2.41$.

A comparison of 4 sets of orthologous proteins – homologous protein sequences that share the same ancestral sequence separated by a speciation event – shows that halophilic proteins have typically 17% to 20% acidic amino acids and only 8% to 10% basic amino acids, whereas their orthologous mesophilic proteins are weakly negatively charged, having 12% to 14% acidic and 10% to 12% basic amino acids. Table 3.2 shows the length of each sequence, the number of acidic and basic amino acids and the total charge of the proteins investigated here; this data shows that the sets of halophilic and mesophilic proteins are representative of their respective category. The halophilic proteins range from the very highly charged ferredoxin (-29 e) where 27% of amino acids are acidic and only 5% of amino acids are basic, to dihydrofolate reductase (-15 e) with only 19% and 9% of acidic and basic amino acids, respectively. The mesophilic proteins have an acidic amino acid content ranging from 8% (beta-lactamase) to 18% (ferredoxin) and are on average only weakly negative. These features make our choice of halophilic-mesophilic pairs representative of their respective categories. Both ferredoxin proteins are outliers: the halophilic ferredoxin is the most highly charged protein known, and even the mesophilic ferredoxin has an unusually high negative charge. The typical cytosolic environment of the halophilic ferredoxin studied here has a concentration of $c = 4$ mol/dm^3; nevertheless, this protein can remain active for NaCl concentrations as low as $c = 0.4$ mol/dm^3 [113];

In Table 3.1 we report the structural and sequence similarity of the proteins studied. The RMSD values in Table 3.1 are calculated for C^{α} atoms of the backbone of the amino acids based on which the two proteins are aligned using the program STAMP (Structural Alignment of Multiple Proteins) [86].The Q_H value is a metric for structural homology which calculates the similarity of two structures by considering their amino acid similarity and adding a term for any gap in the alignment [70]. Proteins showing a value of $Q_H > 0.6$ are considered to have good structural conservation. In Table 3.2 we show the average amino acid composition and charge for all proteins.

Table 3.1 Halophilic-mesophilic protein pairs.

Halophiles		**Mesophiles**			**Comparison**		
PDB[a]	Organism[b]	PDB[a]	Organism[b]	name[c]	RMSD (Å)[d]	Sequence identity (%)[e]	Q_H[f]
1DOI	Haloarcula marismortui	1FRD	Anabaena 7120	Ferredoxin	1.54	22	0.636
2KAC	Peptostreptococcus magnus	1HZ6	Peptostreptococcus magnus	Protein L	0.95	75	0.877
3WRT	Chromohalobacter sp. 560	1ZKJ	Enterobacter aerogenes	Beta-lactamase	2.23	44	0.680
4CNX	Bos taurus	1V9E	mammalian enzyme	Carbonic anhydrase	1.30	90	0.857
2ITH	Haloferax volcanii	2L28	Lactobacillus casei	Dihydrofolate reductase	2.41	19	0.595

[a] Protein Data Bank code

[b] Source organisms

[c] Protein name as reported in the Protein Data Bank

[d] Root Mean Square Deviation between each pair of halophilic-mesophilic proteins, calculated using VMD [38].

[e] Sequence identity between each protein pair, calculated using the MultiSeq plugin [81] in VMD, based on the algorithm described in ref. [81].

[f] A metric of structural homology.

Table 3.2 Length (n_{aa}), number of acidic (n_{acidic}) and basic (n_{basic}) amino acids and charge of the simulated proteins. The protein charge is defined by the difference between acidic and basic amino acids, and by the charge of metal ligands if present.

Halophiles					**Mesophiles**				
pdb	n_{aa}	n_{acidic}	n_{basic}	charge (e)	pdb	n_{aa}	n_{acidic}	n_{basic}	charge (e)
1DOI	128	34	7	-29	1FRD	98	18	8	-12
2KAC	64	16	1	-15	1HZ6	72	10	8	-2
3RWT	367	57	30	-27	1ZKJ	359	29	32	+3
4CNX	259	48	29	-17	1V9E	256	30	28	0
2ITH	162	30	15	-15	2L28	162	22	18	-4

3.2 Simulation details

Most simulation steps are performed using the GPU version of the pmemd engine in AMBER 2018 [14]. The only exception is the l-bfgs minimization step, which is performed on the Sander engine of the AMBER simulation package because it is not available in the pmemd engine.

The starting configurations for the simulations are prepared using the *tLeap* module of the AMBER software [14]. The protein systems are solvated with a cubic box of TIP3P [41] water with the distance between the protein surface and the box face being at least 20 Å. Each protein is simulated at two different salt concentrations: b_{KCl}= 0.15 mol/kg, corresponding to mesophilic conditions, and b_{KCl}= 2 mol/kg, corresponding to halophilic conditions. The simulation boxes are cubic, with edge length $\approx$ 100 Å. Each initial configuration is prepared by putting potassium and chloride ions in numbers corresponding to the desired concentration, and adding extra potassium or chloride for neutralization using the *addIons* tool of AMBER; the systems are then solvated with the appropriate number of TIP3P water molecules [41] using the *tLeap* tool from AMBER. The distance between the protein surface and the box face is at least 20 Å in all cases. Periodic boundary conditions are applied in the XYZ directions. The non-bonded potential cutoff distance is 12 Å for both LJ and electrostatic interactions. Beyond this cutoff distance, electrostatic interactions are calculated with the particle mesh Ewald (PME) scheme with a grid spacing of 1.0 Å, and 4^{th} order interpolation [16]. Long range dispersion corrections are applied to both the energy and pressure. All bonds with H-atoms are constrained using the SHAKE algorithm [87] during the NPT equilibration and the production simulations. We use a 2 fs time step for all simulations except for those of the carbonic anhydrase proteins, which use a 1 fs time step. The carbonic anhydrase proteins have a zinc metal center that coordinates with one water molecule; for technical reasons, the SHAKE algorithm cannot be applied to this molecule, which limits the maximum time step that can be used.

Each system is subject initially to 4 minimization cycles, each comprising 2500 minimization steps using the steepest-descent algorithm and 7500 steps using the Conjugate Gradient algorithm. The protein atoms are restrained to their positions in the pdb file, with the value of the restraint bond constants (500, 300, 100, and 50 kcal·mol^{-1}·Å^{-2}) decreasing for each cycle. This procedure removes bad contacts that may be created in the process of adding ions and water to the system. Subsequently each system is minimized for 10000 steps using the l-bfgs algorithm without any restraints. The system temperature is progressively increased from 0 to 298 K during a 500 ps simulation in the canonical ensemble (NVT), with the protein atoms restrained to their positions with a constant of 10 kcal·mol^{-1}·Å^{-2}, using the Langevin thermostat with a collision frequency of 1.0 ps^{-1}. Each system is equilibrated for 10 ns in 10

steps of 1 ns in the isothermal-isobaric ensemble (NpT), using the Berendsen barostate[8] with a relaxation time of 2.0 ps, while increasing the *skinnb* parameter[1] to 5 Å. Increasing the value of this parameter is indispensable because the GPU version of the pmemd engine is very sensitive to changes in the box size. Fig. 3.1 shows an equilibrated simulation box.

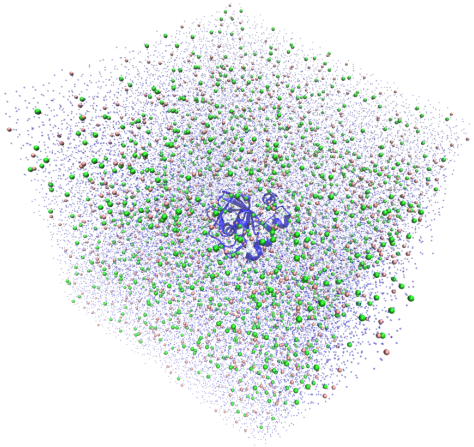


Fig. 3.1 Example simulation box, of the halophilic ferredoxin protein (pdb ID: 1DOI). Dark blue shape = New Cartoon representation of protein; Pink spheres = K^+; Green spheres: Cl^-; Transparent small blue dots: water molecules.

The starting configuration for each production simulation is selected from the last 1 ns of the NpT equilibration, under the condition that it has the box density closest to the average for that simulation. The production simulations are performed in the NVT ensemble, for 5×10^8 steps, corresponding to 0.5 μs for carbonic anhydrase and 1 μs for all other proteins. The simulations for the carbonic anhydrases were not extended to 1 μs because the proteins remain conformationally stable within the simulated time. For the production simulations, we use the Langevin thermostat with a collision frequency of 0.01 ps^{-1}. This low collision frequency is necessary to reduce the impact of the thermostat on the dynamics of the system. Prior work [5] has shown that this low collision frequency leads to self-diffusivity, rotational correlation time, and shear viscosity values within 2% of those obtained from simulations in the NVE ensemble at the same temperature. The average temperature in our simulations remains at its target value despite the low collision frequency of the thermostat. Production simulations are saved every 100 ps for analysis.

For the calculation of the Mean Square Displacement for timescales < 100 ps, we perform 3 production simulations for each protein and each KCl concentration, each lasting 1 ns and with configurations saved every 100 fs. The starting configuration for each simulation

[1] A parameter to monitor the maximum atom movement at every dynamics step, to prevent unnecessary rebuild of the pairlist when there is no possibility of missing atoms in the pairwise calculation within the cutoff.

is taken from the longer production runs, at $t \in \{100, 500, 900\}$ ns. The results for these unrestrained proteins are presented in the chapter 3, and provided here for coherency.

3.3 Results

3.3.1 Structural stability

We simulate each protein at both low ($b_{\mathrm{KCl}} = 0.15$ mol/kg) and high ($b_{\mathrm{KCl}} = 2$ mol/kg) concentration of KCl. The mesophilic proteins at high salt concentration and the halophilic ones at low salt concentration are thus under non-natural conditions. Experiments have shown that the structure of proteins is typically more stable and proteins have higher activity when in media with their natural KCl concentration [57]. Mesophilic proteins are expected to have low solubility at high KCl or NaCl concentrations; under those conditions, many mesophilic proteins have a negative osmotic second virial coefficient [20, 102], known to correlate strongly with low protein solubility [32].

Within the duration of the production simulations – 0.5 μs for the carbonic anhydrase proteins; 1 μs for the others – all proteins retain their structure irrespective of the concentration of KCl. Comparison with experimental data suggests that, under non-natural electrolyte conditions, the native fold of some of the proteins studied here indeed continues to be the stable state. This seems to be the case for the halophilic and mesophilic carbonic anhydrases [112] and the halophilic beta-lactamase [1] studied here: experimental measurements of protein activity indicate that these proteins remain active – and thus presumably folded and in soluble form – under non-natural KCl concentrations. For other proteins, however, the conformational stability seen in the simulation may reflect a metastable state with lifetime longer than the simulation time. This may be the case for the halophilic protein-L: experimental studies show that it is largely unfolded at low salt concentrations [79], even though it remains folded during the simulation.

3.3.2 Solvation layer structure

We first assess how hydrogen bonds donated by water molecules to the protein (termed water-protein hydrogen bonds) are affected by salt concentration. We focus our attention on strong hydrogen bonds [97], present if the distance between the water oxygen and the hydrogen bond acceptor (N, O) is below 3 Å, and the angle formed by the water hydroxy and the acceptor is larger than 135°. In Fig. 3.2 we show the surface density, σ_{HB}, of water-protein

hydrogen bonds, calculated as

$$\sigma_{\mathrm{HB}} = \frac{\overline{n_{\mathrm{HB}}}}{\overline{SASA}} \tag{3.1}$$

where $\overline{n_{\mathrm{HB}}}$ is the time-averaged total number of water-protein hydrogen bonds and $\overline{SASA}$ is the time-averaged solvent accessible surface area of each protein at the position of the first maximum of the hydration layer of each protein. Further details about the SASA calculation are provided below, in the text accompanying eq. 3.2.

Surface density of water-protein hydrogen bonds varies substantially with protein identity but is insensitive to KCl concentration. Fig. 3.2 shows that all halophilic proteins in this study accept significantly more hydrogen bonds from water, per unit area, than their non-halophilic counterparts. The ferredoxin proteins, in particular, have the highest surface density of water-protein hydrogen bonds than any other protein in their halophilic or mesophilic cohort. In SI Fig. A.1 we show that the number of water-protein hydrogen bonds for each amino acid type (acidic, basic, polar, apolar) is only weakly sensitive to the identity of the protein, and is highest for acidic amino acids. The simulation results are consistent with NMR experiments that demonstrate that the number of water molecules that remain unfrozen at $T << 0\ ^{\circ}\mathrm{C}$ in homopolypeptide solutions – i.e., water molecules perturbed by the homopolypeptides – is substantially higher for the acidic amino acids [55]. The variability of the surface density of water-protein hydrogen bonds with protein identity thus largely originates from the different content in acidic amino acids of each protein. This fact has

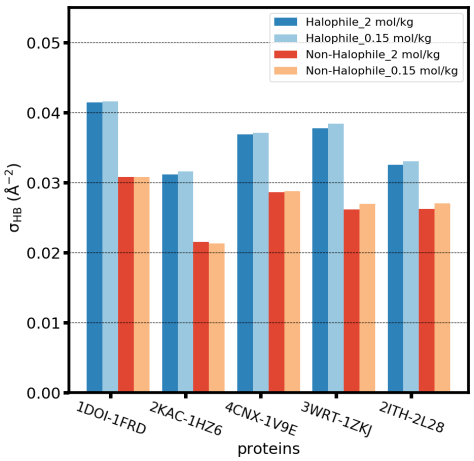


Fig. 3.2 Surface density of water-protein hydrogen bonds for the indicated halophilic-mesophilic proteins, identified by their pdb ID, for different KCl concentrations.

been used to propose that excess acidic amino acids in halophilic proteins is indispensable to compete with the electrolyte in solution for the available water, thus ensuring that the

proteins remain hydrated at high salt concentrations [18, 24, 115]. That possibility, however, is not supported by the results in Fig. 3.2 and SI Fig. A.1: the surface density of water-protein hydrogen bonds and the number of water-protein hydrogen bonds formed by acidic amino acids are almost identical at $b_{\mathrm{KCl}} = 0.15$ mol/kg and at $b_{\mathrm{KCl}} = 2$ mol/kg.

But, why does K^+ in solution not compete with the protein for available water at high KCl concentration? To answer this question, we calculate the proximal (also called the perpendicular [66]) number density, $\rho(r)$, of solvent species at distance r to the protein surface:

$$\rho_X(r) = \frac{\overline{n_X(r)}}{dr\,\overline{SASA(r)}} \tag{3.2}$$

$\overline{n_X(r)}$ is the time-averaged number of solvent species X within a shell of thickness $dr = 0.05$ Å and at distance r of the protein surface. This quantity is calculated by determining the distance between each X and the nearest non-hydrogen protein atom, as implemented in the rdf function in Gromacs 2018; the position of the oxygen atom is used in the calculation of the number density of water. $\overline{SASA(r)}$ is the time-averaged protein Solvent Accessible Surface Area, calculated using tcl scripts implemented in VMD; each point of the surface is at distance r from the nearest non-hydrogen protein atom.

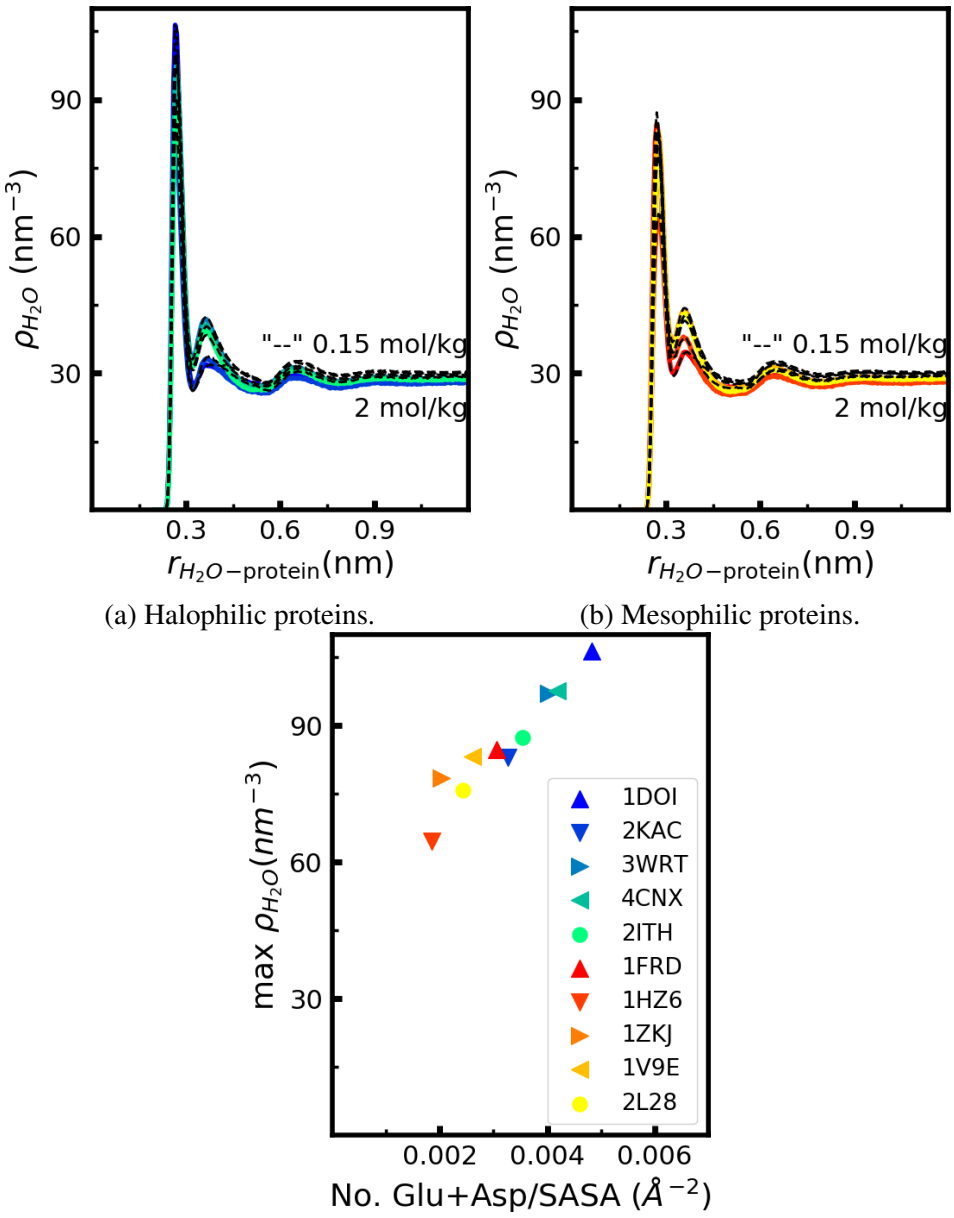
Concentration of water near proteins is highest around halophilic proteins but is insensitive to KCl concentration. In Fig. 3.3 we show the proximal water number density around the halophilic (panel A) and mesophilic (panel B) proteins. The colored curves are obtained at b_{KCl}=2 mol/kg; for $r < 0.45$ nm, overlapping each colored curve is a black, dashed one, obtained from simulations of the same protein simulated at b_{KCl}=0.15 mol/kg. The height of the peaks of the proximal water number density is sensitive to protein identity, and the peaks are on average higher around the halophilic proteins. Irrespective of the halophilic or mesophilic character of each protein, however, the concentration of water in its hydration layer is insensitive to the KCl concentration in the bulk. Moreover, for all proteins and at both KCl concentrations, the water content for $r < 0.9$ nm exceeds the concentration of water in the bulk. These results indicate that proteins do not compete with ions in solution for available water and that mesophilic proteins, despite their lower content in acidic amino acids, are able to retain their hydration layer even at high KCl concentrations.

Halophilic and mesophilic proteins perturb water structure to the same length scale. The curves in Fig. 3.3 retain the same qualitative features despite the different protein sizes and charges. This similarity does not support the hypothesis proposed by several authors [69, 78], that the hydration of halophilic proteins is qualitatively different from that of mesophilic ones and that halophilic proteins perturb water structure to particularly large

length scales [69]. Instead, it is in line with reported calculations of proximal radial distribution functions of water around different proteins, from molecular dynamics simulations and from pdb structures, which indicate that this function has a universal character [62, 66]. According to this view, differences between the proximal rdfs of different proteins reflect different protein shapes and amino acid composition [62, 66]. Fig. 3.3C shows that, for our set of proteins, there is a strong correlation between the first maximum of $\rho_{H_2O}(r)$ and the surface density of acidic amino acids, i.e., between protein hydration and its content in acidic amino acids. Our results suggest that proximal rdfs may be confidently used to aid x-ray refinements of halophilic proteins, using models such as that in ref. [66]. One way to more confidently establish whether the solvation patterns around halophilic proteins are indeed unusual would be apply those models to reconstruct solvation patterns obtained for a large set of halophilic proteins using molecular dynamics simulations. If proximal rdf models using parameters developed from a set of non-halophilic proteins perform worse than those developed using a set that includes halphilic proteins, this would hint that the solvation patterns of halophilic proteins are indeed unusual, as has been proposed [115]. That study is outside the scope of the present work.

Potassium integrates the first hydration shell of both halophilic and mesophilic proteins, in a charge-density–dependent manner. Figs. 3.4A,B show the proximal number density of K^+, $\rho_{K^+}(r)$, around each protein. The distributions clarify that K^+ accumulates in the vicinity ($r < 0.9$ nm) of each protein to concentrations between 2 and 6 times higher than in the bulk, depending on the protein identity and the bulk concentration of KCl. The position of the first peak of $\rho_{K^+}(r)$ coincides with that of $\rho_{H_2O}(r)$ at both low and high c_{KCl}. Nevertheless, K^+ cannot measurably displace water from the first hydration shell of the protein because it does not accumulate near the protein in sufficiently large amounts: the height of the peaks of $\rho_{K^+}(r)$ is always below 3 ions per cubic nanometer; in contrast, the maximum value of $\rho_{H_2O}(r)$ is $\approx$105 molecules per cubit nanometer. Rather than displacing water, K^+ integrates its own first hydration shell with that of the protein.

Figs. 3.4A,B also show that the density of K^+ in the vicinity of the protein depends strongly on protein identity and is highest for the halophilic proteins. This origin of dependence is clarified in Fig. 3.4C, which shows that the height of the first peak of $\rho_{K^+}(r)$ strongly correlates with the charge density of the protein. In other words, the total content of K^+ in the solvation shell of proteins correlates both with protein net charge and surface area. Halophilic proteins are on average larger and have a higher net negative charge than mesophilic ones, and therefore a larger amount of K^+ in their solvation layers, as quantified by the cumulative number of potassium ions in the vicinity of each protein shown in SI Fig. A.2.

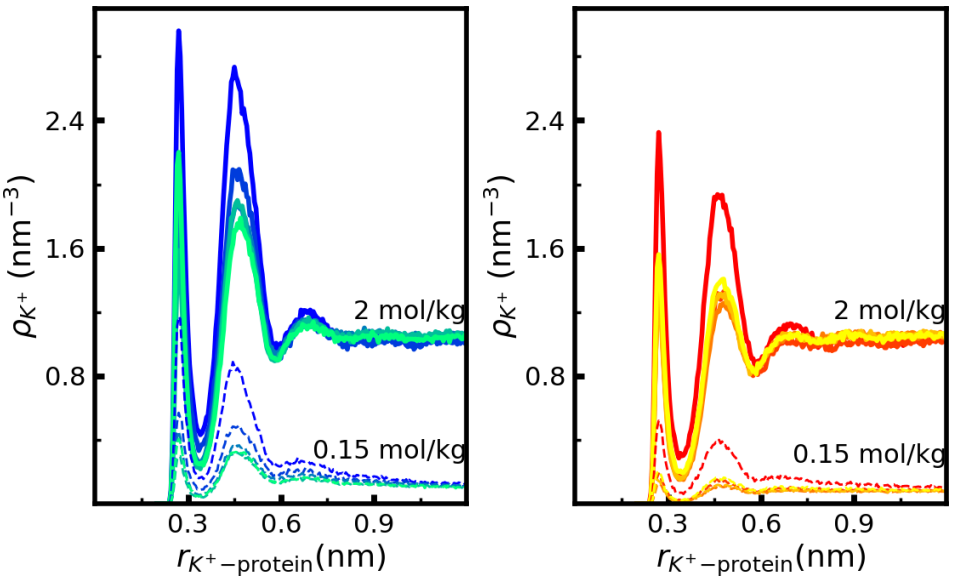


(a) Halophilic proteins.

(b) Mesophilic proteins.

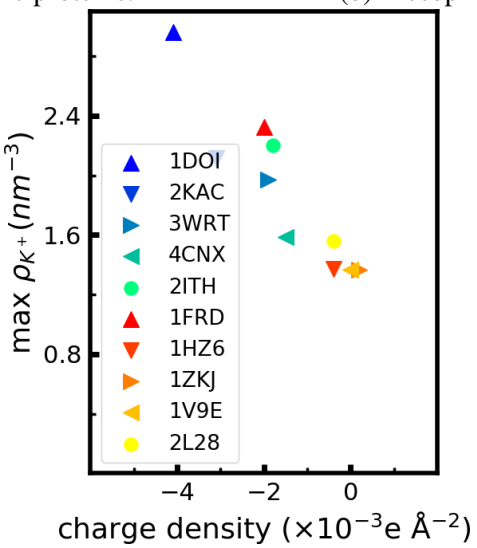
(c) Halophilic and mesophilic proteins.

Fig. 3.3 (A,B) Proximal number density of water molecules as a function of the distance to the surface of the indicated proteins, simulated at b_{KCl}=2 mol/kg (color) and b_{KCl}=0.15 mol/kg (dashed black lines). (C) Height of the first peak of the number density curves for b_{KCl}=2 mol/kg shown in the other panels, as a function of the surface density of acidic amino acids. Each color corresponds to a protein, identified by its pdb ID in the legend of the bottom panel.



(a) Halophilic proteins.

(b) Mesophilic proteins.



(c) Halophilic and mesophilic proteins.

Fig. 3.4 (A,B) Proximal number density of K^+ as a function of the distance to the surface of the indicated proteins, in solutions with the indicated molality of KCl. (C) Height of the first peak of the number density curves for b_{KCl}=2 mol/kg shown in the other panels, as a function of the protein charge density. Each color corresponds to a protein, identified by its pdb ID in the legend of the bottom panel.

3.3.3 Dynamics of the protein solvation shell

We evaluate the impact of salt concentration and protein identity on solvent translational dynamics by calculating the Mean Square Displacement, $MSD(\tau)$, of water or of K^+, according to:

$$MSD(\tau) = \left\langle (\vec{x}(t+\tau) - \vec{x}(t))^2 \right\rangle \tag{3.3}$$

$\vec{x}(t)$ is the position of each species at time t and the bracket indicates both a time and an ensemble average over the relevant species population. We calculate the MSD for water or K^+ belonging to the first solvation layer at $t = 0$ because if halophilic proteins indeed substantially slow down water dynamics, this effect should be strongest for these subpopulations. We are interested in the short-time dynamics of each species, because this dynamics should reflect the strongest impact of their initial position in the hydration shell and thus differ the most from the dynamics of the same species averaged over all its elements in the simulation box. For this reason, it is sufficient to calculate the MSD using coordinates wrapped back into the main simulation cell.

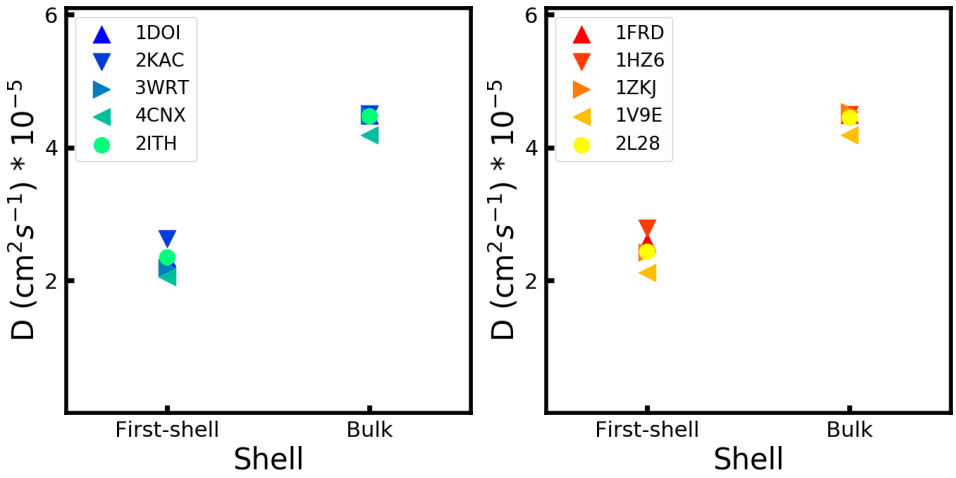
In SI Figs. A.3 and A.4 we show the MSD of water and of K^+ in the first hydration layer around the halophilic and mesophilic ferredoxin proteins; the results for the other proteins (not shown) are qualitatively similar. As expected under these conditions the curves saturate at long times, evidence of confinement to the main simulation cell. To facilitate comparisons between different proteins and simulation conditions, we calculate the diffusion coefficient, D, as

$$MSD(\tau) = 6D\tau + C \tag{3.4}$$

by fitting each $MSD(\tau)$ curve in the interval $\tau = [20, 30]$ ps. The constant C is added to account for non-diffusive movement at short times. This time interval is chosen because the dynamics of all species is already diffusive beyond 20 ps but the impact of confinement by the finite simulation box remains marginal much beyond 30 ps (see example in SI Fig. A.3). To facilitate comparisons, the diffusion coefficient averaged over all K^+ or water molecules in the simulation box (calculated using unwrapped coordinates) is estimated from fits using the same time interval.

Halophilic and mesophilic proteins have solvation layers with identical translational dynamics at both low and high KCl concentration. In Fig. 3.5 we compare the diffusion coefficients of water molecules that initially belong to the first hydration layer of each protein with the diffusion coefficient of all water molecules in each simulation, at

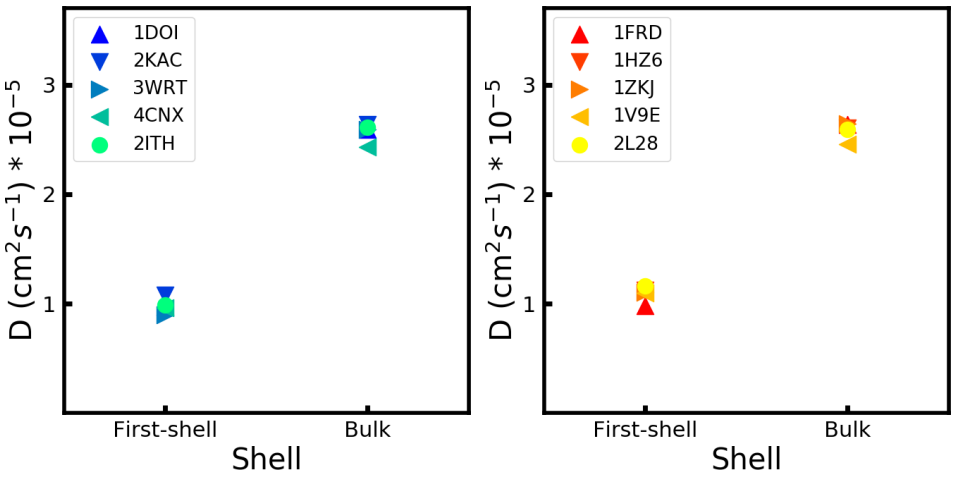
b_{KCl} = 2 mol/kg; results for b_{KCl} = 0.15 mol/kg are shown in SI Fig. A.5. The TIP3P water model is known to predict translational and rotational dynamics which are approximately twice faster than the reference experimental values [68]. Accordingly, the absolute values of D_{H_2O} reported in Fig. 3.5 are likely overestimated; only their relative magnitude is meaningful. The translational dynamics in the first hydration layer is indeed lower than that of water in the bulk, but only by a factor of $1/2$ at most. Moreover, the translational dynamics of water in the first hydration layer of the mesophilic proteins is barely faster that of the non-halophilic proteins. Our results are not consistent with the possibility that water near halophilic proteins has dramatically slower dynamics than around non-halphilic ones.



(a) Halophilic proteins. (b) Mesophilic proteins.

Fig. 3.5 Diffusion coefficients of water around the indicated proteins, simulated in b_{KCl} = 2 mol/kg. First-shell: water molecules that at $t = 0$ belong to the first hydration shell of the proteins; Bulk: all water molecules in the same simulation.

Fig. 3.6 compares the self-diffusion coefficients of potassium ions initially in the first solvation layer of the protein with those averaged over all K^+ in the simulation box, again at the highest salt concentration (see SI Fig. A.6 for the results at b_{KCl} = 0.15 mol/kg). Similarly to the trends observed for water dynamics, the translational dynamics of K^+ in the first solvation layer is lower than dynamics in the bulk, but only by a factor of $1/3$. Moreover, the diffusion coefficients of this subpopulation are very similar for the halophilic and mesophilic proteins. Even though halophilic and mesophilic proteins differ substantially in surface charge density, and these differences lead to considerable differences in the number of K^+ in the solvation layer of each protein (SI Figure A.2), they do not impact the translational dynamics of the ions near the protein.



(a) Halophilic proteins. (b) Mesophilic proteins.

Fig. 3.6 Diffusion coefficients of potassium ions around the indicated proteins, simulated in b_{KCl} = 2 mol/kg. First-shell: potassium ions that at $t = 0$ belong to the first solvation shell of the proteins; Bulk: all potassium ions in the same simulation.

3.4 Discussion

Both the solvent-only and the ion-solvent stabilization models propose that excess acidic amino acids are *necessary* in halophilic proteins to maintain their hydration at high salt concentrations. Our results do not support this scenario. Halophilic proteins do accept more hydrogen bonds from water (Fig. 3.2) and have larger concentration of water in their hydration shells (Fig. 3.3) because of their higher content in acidic amino acids (Fig. 3.3C, SI Table 3.2 and SI Fig. A.1). These results are consistent with sedimentation measurements [12, 78] which indicate that halophilic proteins bind larger amounts of water than mesophilic ones. That fact does not imply that excess acidic amino acids are necessary to maintain protein hydration at high KCl concentration, however. Our study demonstrates that all mesophilic and halophilic proteins investigated here remain equally well-hydrated at both low (b_{KCl} = 0.15 mol/kg) and high ($b_{KCl} = 2$ mol/kg) KCl concentration, as measured by the number of hydrogen bonds that they accept from water (Fig. 3.2) and the water concentration at a distance below 4.5 Å from the protein surface (Fig. 3.3). These results imply that the proteins simulated here do not compete with ions in solution for available water even at concentrations as high as $b_{KCl} = 2$ mol/kg. The force fields used here for carboxylates and K^+ (ref. 46 and the present work) were developed to reproduce experimentally-determined hydration free energies, lending confidence to our observations. The proteins selected for this study have diverse sizes, net charges and surface charge density and include both typical examples of halophilic and mesophilic proteins as well as extreme examples (the ferredoxin proteins)

of each category (SI Table 3.2). The conclusion that protein hydration is concentration-independent for $b_{KCl} \leq 2$ mol/kg and that proteins do not compete with KCl for available water thus may apply to all proteins.

Moreover, we suggest that this conclusion may hold also for aqueous NaCl solutions, because the water activity of NaCl and KCl solutions is very similar (SI Fig. A.8, based on ref. 82). The limited experimental data comparing protein hydration in NaCl vs. KCl solutions is consistent with this possibility: neutron scattering and ultracentrifugation studies show that halophilic malate dehydrogenase binds equally large amounts of water when immersed in KCl or in NaCl solutions [12]. We note, however, that Na^+ is expected to form more contact ion pairs with carboxylates than K^+ (compare the acetate..K^+ and acetate..Na^+ radial distribution functions shown in SI Fig. A.7). Differences in the stability and dynamics of halophilic proteins in NaCl and KCl solutions [99] likely reflect the different interactions between the cations in solution and the acidic amino acids in the protein. It is of interest to comparatively characterize protein solvation shells in KCl vs. NaCl solutions, using non-perturbative methods such as solvation-shell spectroscopy [94], and to understand how their differences influence the conformational stability and the functionality of enzymes. This understanding is critical to rationally develop proteins for use in biotechnological devices (e.g., to produce molecular hydrogen) that function in brackish or sea water, thus reducing pressure on our planet's fresh water resources.

The water activity decreases from 1 to 0.8 as the concentration of KCl increases to $b_{KCl} \leq 4$ mol/kg (SI Fig. A.8). Low water activity also occurs in solvent mixtures composed of water and organic liquids. At first sight, halophilic proteins offer clues to understand which sequence-structure features enable the ability to function at low water activity arising from the presence of organic solvents; in fact some halophilic proteins remain functional under these conditions [25, 45]. A closer look into the limited available data, however, suggests that protein hydration may respond very differently to the two environments, and that further studies are necessary before drawing analogies between the proteins at high salt concentration and proteins in mixtures of water with organic solvents. Our results indicate that protein hydration is insensitive to KCl concentrations up to $b_{KCl} = 2$ mol/kg, i.e., protein hydration levels are unaffected by changes in water activity between 1 and 0.9. In contrast, the fraction of water bound to proteins decreases from 60% to 40% (weight percentage of bound water in the protein+bound water system isolated by centrifugation) in solutions of water with immiscible organic solvents [60, 116], as the water activity decreases from 1 to 0.9. It is at present unclear whether the different dependence between protein hydration and water activity is real or whether it reflects a bias imposed by the different observables being compared. There is a clear need for non-perturbative experiments to comparatively

characterize the hydration shells of proteins in aqueous electrolyte media and in aqueous media containing organic solvents, to assert the extent to which halophilic proteins may be useful starting points to design enzymes and catalysts that function in low-water activity environments that occur, e.g., when producing ethanol in large quantities or when catalyzing reactions in the presence of organic solvents.

The ion-solvent stabilization hypothesis proposes that cooperative ion-solvent-protein networks arise in halophilic proteins because of their tertiary and quaternary structure [12, 78]; cooperative interactions would not exist in the unfolded or non-oligomerized state of the protein because in that state the acidic amino acids are further apart or in unfavorable orientation. These networks have been proposed to explain experimental observations (based on analysis of centrifugates) showing that halophilic proteins bind more cations than mesophilic ones [12, 78]. Our simulation results agree with experiment: halophilic proteins indeed have larger amounts of K^+ in their solvation shell than mesophilic proteins (Fig. 3.4). The fact that the maximum of the number density of K^+ around the proteins correlates linearly with the charge density of the protein (Fig. 3.4C) suggests that cation binding has a predominantly electrostatic origin. We are currently investigating whether protein-ion-solvent interactions indeed have a cooperative component.

Neutron scattering experiments have suggested that the cytoplasm of halophilic organisms has a fraction of water with translational diffusion approximately 250 times slower than that of water in the bulk; in contrast, that slow component was not observed in mesophilic organisms [40]. A fraction of water with extremely slow dynamics has been proposed to arise in the context of the ion-solvent stabilization model: the cooperative ion-water-protein networks would strongly reduce the mobility of water in the first hydration layer of the protein [69]. Our results do not support this view: the translational dynamics of water and of K^+ near halophilic and mesophilic proteins is indistinguishable, and only 2 to 3 times slower than the dynamics of the same species in the bulk (Fig. 3.5 and 3.6). Our simulations are consistent with ^{17}O magnetic relaxation measurements of halophilic and mesophilic versions of protein L (one of the protein pairs simulated here) which show that they have similar hydration dynamics [79]. This result does not imply that cooperative ion-solvent-protein interactions are necessarily absent at high KCl concentrations. Cooperative interactions may well exist, but our results indicate that they should not result on unusually slow solvent dynamics, i.e., they do not come at the expense of a high entropic price.

3.5 Conclusion

Many halophilic organisms contain molar concentrations of KCl in their cytoplasm, necessary to balance the large osmotic pressure induced by equally high external concentrations of NaCl [57]. The cytoplasmic proteins of these organisms are substantially richer in acidic amino acids than mesophilic ones [18]. As acidic amino acids can bind substantially larger amounts of water than any other natural amino acid [55], it has been proposed that halophilic proteins require acidic amino acids to remain hydrated in their low water activity environment [18, 24]. The present work does not support this possibility. Our simulations of 5 halophilic proteins and 5 mesophilic counterparts indicate that halophilic proteins indeed contain larger amounts of water in their hydration shells than mesophilic ones, and that their larger hydration level correlates with their high content in acidic amino acids. However, all proteins remained equally hydrated at low ($b_{\mathrm{KCl}} \leq 0.15$ mol/kg) and high ($b_{\mathrm{KCl}} \leq 2$ mol/kg) KCl concentration, demonstrating that a higher content in acidic amino acids is not necessary to remain hydrated at high KCl concentrations. We note, however, that to understand the connection between acidic amino acids, solvation, and protein stability, it is also necessary to investigate the unfolded state of proteins: it is the difference in protein hydration and protein-salt interactions between the folded state and the unfolded ensemble that matters [19, 91]. We will focus on this point in future studies.

Cooperative interactions between acidic amino acids, water and cations have been proposed to exist in the folded structure of halophilic proteins because their high content in acidic amino acids would enable specific, favorable ion-water-carboxylate configurations to form [12, 78, 115]. According to this view, these interactions would stabilize the folded protein configuration, they would be necessary to maintain the protein hydrated at high KCl concentration [12, 78, 115], and would manifest themselves in a fraction of water with very slow translational dynamics [40]. Our simulations show that halophilic proteins have a higher concentration of K^+ in their solvation shells than mesophilic ones, consistent with experiment [12, 78]. The concentration of K^+ in the protein solvation shell linearly correlates with the protein net-charge-to-SASA ratio. Halophilic proteins are more negative than mesophilic ones; hence the higher concentration of K^+ in their hydration shells. Despite this high concentration, the solvation shell of halophilic proteins remains as labile as that of mesophilic ones and we find no evidence of water or ions with unusually slow translational dynamics around halophilic proteins. The absence of slow dynamics, we point out, does not imply that cooperativity is impossible; it suggests, however, that if present it does not unusually slow down water dynamics. If a stabilizing interaction between acidic amino acids (mediated by K^+) indeed exists, and is more prevalent in folded rather than unfolded protein conformations, it might be one of the driving forces behind the abundance of acidic amino

acids in halophilic proteins. Conversely, if the interactions between acidic amino acids are predominantly destabilizing, they might be necessary to retain protein flexibility [69], or simply to prevent aggregation in a media that greatly enhances the hydrophobic effect [114].

In chapter 4 we investigate whether water-cation-carboxylate interactions in halophilic proteins have a cooperative nature.

Chapter 4

Are cooperative water-cation-carboxylate interactions in halophilic proteins possible?

4.1 Cooperative interactions in halophilic proteins

As discussed in more detail in Chapter 1, the *Ion-solvent stabilization model*, some experimental studies have been interpreted to indicate that, for halophilic proteins in the folded state, nearby acidic amino acids might stabilize the folded structure of the protein via a cooperative effect also involving the cations in solution [63, 71, 115]. To evaluate this possibility, we first calculate the free energy change associated with the following process:

$$\mathrm{D}_b + \mathrm{D}_a \xrightarrow{\Delta G_{(-)\mathrm{D}a\mathrm{N}}} \mathrm{D}_b + \mathrm{N}_a \tag{4.1}$$

This expression schematically describes the mutation of aspartate in position a of the amino acid sequence (denoted by D_a) into asparagine (N_a), belonging to a solvated protein, when there is at least another aspartate (D_b) in position b of the amino acid sequence which is geometrically close to position a. This process has an associated free energy, $\Delta G_{(-)\mathrm{D}a\mathrm{N}}$, where the subscript $(-)$ emphasizes that the vicinal amino acid has a negative charge. This free energy is positive (i.e., the mutation is unfavorable) because the amino acids are at the protein's surface and therefore directly contact the solvent. The dominant contribution to $\Delta G_{(-)\mathrm{D}a\mathrm{N}}$ comes from the different solvation of the amino acids and charged amino acids are more favorably solvated than uncharged ones.

We compare $\Delta G_{(-)\mathrm{D}a\mathrm{N}}$ to the free energy change associated with the mutation of the same amino acid, but where its neighbor amino acid b is an asparagine:

$$\mathrm{N}_b + \mathrm{D}_a \xrightarrow{\Delta G_{(0)\mathrm{D}a\mathrm{N}}} \mathrm{N}_b + \mathrm{N}_a \tag{4.2}$$

The subscript (0) emphasizes the zero net charge of the vicinal asparagine in position b. If amino acids a and b are geometrically distant, they will interact weakly with each other, so $\Delta G_{(-)\mathrm{D}a\mathrm{N}} \approx \Delta G_{(0)\mathrm{D}a\mathrm{N}}$. If they are nearby, conventional understanding would suggest that two charged amino acids repel each other and thus the first process should be less unfavourable than the second, i.e., $\Delta G_{(-)\mathrm{D}a\mathrm{N}} < \Delta G_{(0)\mathrm{D}a\mathrm{N}}$. If a cooperative effect with the cations in solution exists as has been proposed, however, we should observe $\Delta G_{(-)\mathrm{D}a\mathrm{N}} > \Delta G_{(0)\mathrm{D}a\mathrm{N}}$.

For quantifying such effect, we chose a limited number of halophilic proteins to study at both high and low salt concentrations to investigate the validity of the proposed model and the effect of salt on such cooperativity. Halophilic proteins 1DOI (ferredoxin), 2KAC (protein L.), and 2ITH (Dihydrofolate Reductase) only at high salt concentration, were studied. The criterium to consider two acidic residues, a and b in Equations 4.1, and 4.2, geometrically close was that the distance between the carboxylate carbons of the two acidic amino acids should be closer than ≈ 7 Å in the initial conformation. The halophilic protein L. is shown in Figure 4.1, with the acidic amino acids and the ions in solution clearly visible; we chose some of the possible pairs of vicinal amino acids for this study. Figure 4.2 shows the halophilic protein ferredoxin, with one of the studied pairs of acidic amino acids visible.

This cooperative effect, if exists, is thought to be related to electrostatic interactions of negatively charged residues with potassium ions and water molecules. To study these interactions, we mutate the negatively charged acidic residues to neutral residues of almost the same surface area and structure: aspartates are mutated to asparagine and glutamates to glutamine which their structures are shown in Figure 4.3.
The study's design is explained in the two Equations 4.1, and 4.2, and the difference between these two free energies will indicate the existence of such cooperative effect or the opposite interfering effect and its quantity for each case. Also, to comprehensively examine the *Ion-solvent stabilization model*, the quantity and existence of such cooperative effect need to be investigated in the unfolded state of halophilic proteins as well. Because studying unfolded structures is time-consuming, we only studied a subsample of the unfolded state of halophilic protein L., 2KAC. To sample this unfolded state, we used Replica Exchange, which is thoroughly explained in Section 4.3. For the main study of quantifying the free energies associated with the Equations 4.1, and 4.2 we used the method thermodynamic

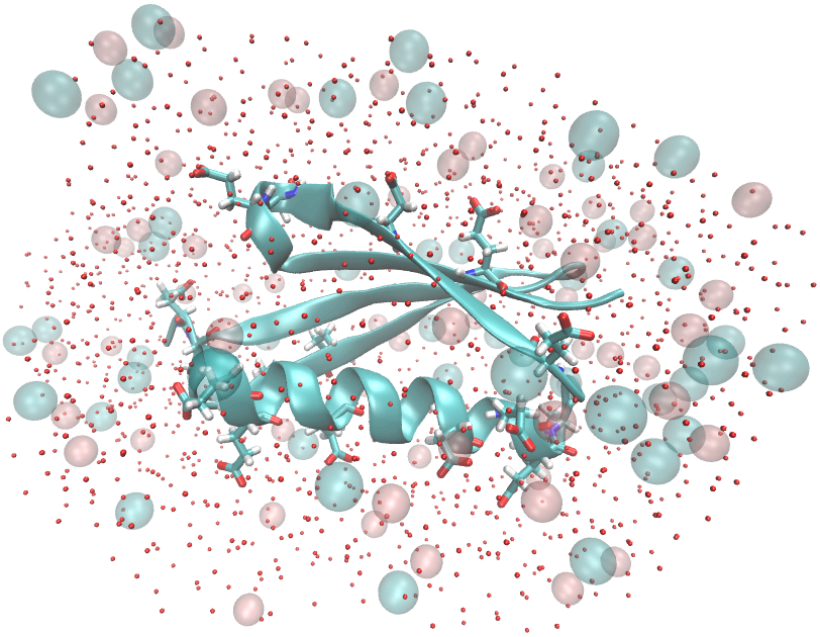


Fig. 4.1 A snapshot of simulation box with solution showed within a radius of 10 Å around protein. Protein L.: New cartoon representation; K^+: transparent pink spheres; Cl^-: transparent blue spheres; Acidic residues: blue and red branches; Water oxygens: red dots.

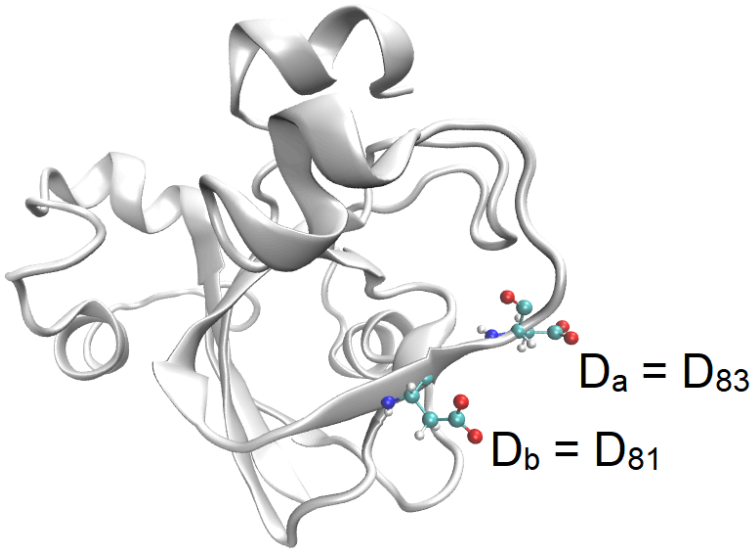
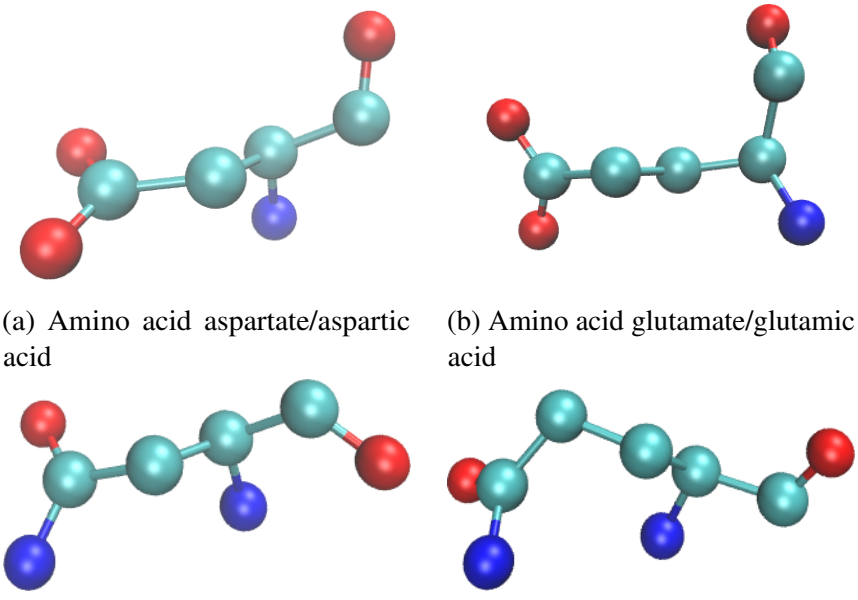


Fig. 4.2 Halophilic protein ferredoxin (pdb ID: 1DOI) in white new cartoon representation, and two acidic amino acids shown in red and blue, exemplifying the *a* and *b* positions mentioned in Equation 4.1 . Only one pair of acidic amino acids (aspartic acid residues in residue positions 81 and 83 of amino acid chain) is shown, and other residues, as well as the solution, are not shown.



(a) Amino acid aspartate/aspartic acid

(b) Amino acid glutamate/glutamic acid

(c) Amino acid asparagine

(d) Amino acid glutamine

Fig. 4.3 The ball and stick representation of acidic amino acids; blue spheres: nitrogen atom, red spheres: oxygen atoms, and cyan spheres: carbon atoms. Hydrogens are not shown.

integration available in molecular dynamics package AMBER [14], in which we mutate the acidic amino acids to their neutral counterparts thoroughly explained in the Section 4.2.

4.2 Free energy calculations via thermodynamic integration (TI)

4.2.1 Theoretical background

Free energy calculations quantify the difference between the stability of two thermodynamic end-states under study. To calculate accurate free energy difference in molecular simulations, the phase space of each simulation should have sufficient overlap with the other in a reasonable time that is only possible for only very similar end-states, which scarcely happens in biomolecular systems. To bypass such a barrier, free energy calculations are performed between a series of intermediate states between the two end-states where the neighboring states in this series have sufficient overlap. These intermediate states are constructed by modifying the functional form of potential energy relative to a mixing parameter between zero (initial end-state) and one (final end-state), but the end states themselves will use the original potential energy. Free energy calculations using these unreal intermediate states

are called alchemical transformations. After the simulation has been performed in each of these end states and the connecting intermediate states, the results can be post-processed by various methods to calculate the free energy. The two main classes of these methods are thermodynamic integration (TI) [50], and free energy perturbation (FEP) [119] methods.

For TI, the Boltzmann average of derivative of the potential energies in the series between the two end states relative to the mixing parameter in each of these states is interpolated and then integrated (see Equation 4.3). We can calculate this variable as a weighted sum shown in Equation 4.3 where W_i is the weighting factor that depends on the numerical method for interpolation and K is the number of intermediate states.

$$\Delta G = G(\lambda = 1) - G(\lambda = 0) = \int_{\lambda=0}^{\lambda=1} \left\langle \left(\frac{\partial U(\lambda)}{\partial \lambda} \right)_{\lambda_i} \right\rangle d\lambda = \sum_{i=1}^{K} W_i \left\langle \left(\frac{\partial U(\lambda)}{\partial \lambda} \right)_{\lambda_i} \right\rangle \tag{4.3}$$

Some of the widely used numerical integration schemes to calculate the free energy based on thermodynamic integration method are TI-1 and TI-3 [72] which differ in how they interpolate between data points and integrate. TI-1 uses the standard trapezoidal rule (a first-order polynomial/linear interpolation), while TI-3 fits the curve of $\langle \partial U(\lambda)/\partial \lambda \rangle_{\lambda_i}$ vs. λ_i piece-wise to a (natural) cubic spline, and then integrate it analytically using the coefficients of the fitted equation. The end states, the path chosen between them, and the resulting curve of $\langle \partial U(\lambda)/\partial \lambda \rangle_{\lambda_i}$ affect the performance of these methods.

Also, perturbation-based free energy calculation includes a range of methods based on the average of the potential energies, ΔU_{ij}, between two adjacent states in the path, in its exponential form, using the Zwanzing relationship [119]:

$$\Delta G_{ij} = -\frac{1}{\beta} \ln \left\langle \exp\left(-\beta \Delta U_{ij}\right) \right\rangle_i \tag{4.4}$$

where β is $1/(k_B T)$ and k_B is the Boltzmann constant. One of the shortcomings of many perturbation-based methods is that they are direction dependent. Starting from the initial end state to the final state would give a different estimate of free energy than the opposite direction of transformation, which should not happen because free energy is a state variable. This is originated from the under-sampling of the states in the tail of the ΔU_{ij} distributions. To avoid such problems, Bennett Acceptance Ratio (BAR) [7] has been developed where both forward, ΔU_{ij}, and reverse, ΔU_{ji} (the equation is not shown), distributions are considered which eliminates the bias in the final free energy results. A development over this method (BAR) is Multistate Bennett Acceptance Ratio (MBAR) [88] where instead of considering only the data collected from the two adjacent states i and j to calculate free energy with minimum

variance in this transformation from i to j in BAR, MBAR minimizes the matrix of variances of all free energy transformations simultaneously leading to significant improvement over BAR.

The free energy estimations from the TI methods are sensitive to the smoothness of the curve $\langle \partial^2 U/\partial\lambda^2 \rangle$, and not directly a function of overlap of energy distributions of states. On the other hand, perturbation-based methods depend on the overlap of the energy distributions between adjacent states and not the smoothness of the curve. Comparing the free energy estimates from both of these post-processing families would help validate the path and the sampling time chosen.

One of the main problems in the free energy calculations using TI is that when creating or annihilating new atoms at the tail of the path, close to $\lambda = 0$ or $\lambda = 1$, respectively, there will be instabilities in the form of endpoint singularity [90]. This is because when the repulsive part of the vdW forces becomes negligible, oppositely charged species might come too close to each other.

Softcore potentials avoid this endpoint singularity by ensuring that the non-bonded energies remain finite for all the states leading to smooth free energy integration curves [10]. With this soft-core potential a linear mixing (Equation 4.5) is always used which result in $\langle \partial U(\lambda)/\partial\lambda \rangle = U_1 - U_0$:

$$U(\lambda) = (1-\lambda)U_0 + \lambda U_1 \tag{4.5}$$

where U_0 is the potential with the original Hamiltonian, initial end state, and U_1 is the potential with the perturbed Hamiltonian, final end state. The approach we have used is only non-bonded; therefore, any bond, angle, and dihedral term that involves unique atoms, appearing or disappearing atoms considered in the choice of softcore atoms, is not scaled by the coupling parameter λ and does not contribute to the calculated free energy change. For these non-bonded soft-core potentials a modified version of vdW (Equations 4.6, and 4.7) and electrostatic interactions (Equations 4.8, and 4.9) that switches off/on the disappearing/appearing, unique atoms' interactions with their common (non-soft-core atoms in the system) atom neighbors was used [14]:

$$U_{u_0,disappearing} = 4\varepsilon(1-\lambda)\left[\frac{1}{\left[\alpha\lambda + \left(\frac{r_{ij}}{\sigma}\right)^6\right]^2} - \frac{1}{\alpha\lambda + \left(\frac{r_{ij}}{\sigma}\right)^6}\right] \tag{4.6}$$

$$U_{u_1,appearing} = 4\varepsilon\lambda\left[\frac{1}{\left[\alpha(1-\lambda) + \left(\frac{r_{ij}}{\sigma}\right)^6\right]^2} - \frac{1}{\alpha(1-\lambda) + \left(\frac{r_{ij}}{\sigma}\right)^6}\right] \tag{4.7}$$

$$U_{u_0,disappearing} = (1-\lambda)\frac{q_i q_j}{4\pi\varepsilon_0\sqrt{\beta\lambda + r_{ij}^2}} \tag{4.8}$$

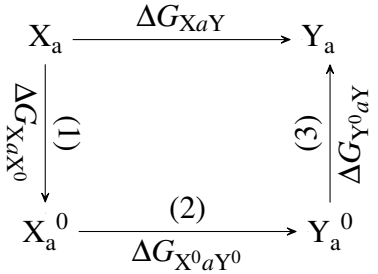
$$U_{u_1,appearing} = \lambda\frac{q_i q_j}{4\pi\varepsilon_0\sqrt{\beta\,(1-\lambda) + r_{ij}^2}} \tag{4.9}$$

This introduces new parameters α and β, which control the "softness" of the potential. Atoms that are used in the softcore potential slowly couple/decouple with their environment. For example, the "disappearing" atoms are present in U_0 but not in U_1, the interaction between these atoms themselves have vdW of 12-6 LJ and does not change with λ or contribute to $\langle\partial U(\lambda)/\partial\lambda\rangle$, but their vdW interactions with common atoms calculated from Equation 4.6 decline with increasing λ and goes to zero at λ=1. The same applies to the electrostatic potential. Transferring a molecule from solution to gas phase is a clear example of disappearing atoms of a molecule.

4.2.2 Computational details of free energy calculation

Free energy calculations were performed for halophilic proteins 1DOI, 2KAC, and 2ITH. For 2ITH, we only have simulated the system in high salt concentration of b_{KCl} = 2 mol/kg, but for the other two at both b_{KCl} = 0.15, and 2 mol/kg. Simulations used the force field parameters described in Section 2.4 that was used for the study of structure and dynamics of the solvation of these proteins (Chapter 3). The starting configurations for the free energy simulations correspond to the last saved configuration of the 1 μs production runs discussed in Section 3.2 for the respective protein at the respective concentration. During all the simulation steps explained as following, the thermodynamic integration (TI) method implemented in the AMBER 18 pmemd GPU engine was used to mutate selected acidic residue to its neutral pair. Soft-core potential was used for the whole residue as well with the AMBER default parameter values of α=0.5 and β=12.0 Å^2 which controls the softness of the soft-core potential, shown in Equations 4.6 to 4.9. Also, the SHAKE algorithm was used to constrain the length of bonds involving hydrogen atoms, except the two residues involved in the mutation, so consequently, we used a time step of 1 fs in all the simulation steps. Simulations to mutate the acidic residues, Asp (D) or Glu (E), to their neutral counterpart, Asn (N) or Gln (Q), were performed in three steps as described generically for a mutation of amino acid X into amino acid Y in scheme 4.10. In the scheme, *a* denotes the position of the

residue in the amino acid sequence. $Y_a{}^0$ and $X_a{}^0$ indicate unphysical forms of the amino acid where all atomic charges are set to zero.



(4.10)

For choosing which residues to mutate, we used the software *chimera* [76] and calculated the protein residues with a solvent accessible surface area of > 50 Å^2 to choose residues on the surface of these proteins for the study.

For the step (2) in the Scheme 4.10 which we call it from now on vdW step, 34 λ values of 0.00 (X_a^0), 0.02, 0.04, 0.06, 0.08, 0.10, 0.12, 0.14, 0.16, 0.18, 0.20, 0.22, 0.24, 0.26, 0.28, 0.30, 0.35, 0.40, 0.45, 0.50, 0.55, 0.60, 0.65, 0.70, 0.75, 0.80, 0.85, 0.88, 0.90, 0.92, 0.94, 0.96, 0.98, 1.00 (Y_a^0) was used. The value returned by this calculation is $\Delta G_{X^0aY^0}$.

For practical reasons, in the case of steps (1) and (2), the thermodynamic steps actually calculated are the reverse of those shown in Scheme 4.10. For the step (1) which we call it from now on decharge step, 37 λ values of 0.00 (X_a^0), 0.05, 0.10, 0.15, 0.20, 0.25, 0.30, 0.35, 0.40, 0.45, 0.50, 0.55, 0.60, 0.61, 0.62, 0.63, 0.64, 0.65, 0.66, 0.67, 0.68, 0.69, 0.70, 0.71, 0.72, 0.73, 0.74, 0.75, 0.76, 0.77, 0.78, 0.79, 0.80, 0.85, 0.90, 0.95, 1.00 (X_a) was used. The value returned by this calculation is $-\Delta G_{XaX^0}$.

For the step (3) which we call it from now on charge step, 24 λ values of 0.00 (Y_a), 0.02, 0.04, 0.06, 0.08, 0.10, 0.15, 0.20, 0.25, 0.30, 0.35, 0.40, 0.45, 0.50, 0.55, 0.60, 0.65, 0.70, 0.75, 0.80, 0.85, 0.90, 0.95, 1.00 (Y_a^0) was used. The value returned by this calculation is $-\Delta G_{Y^0aY}$.

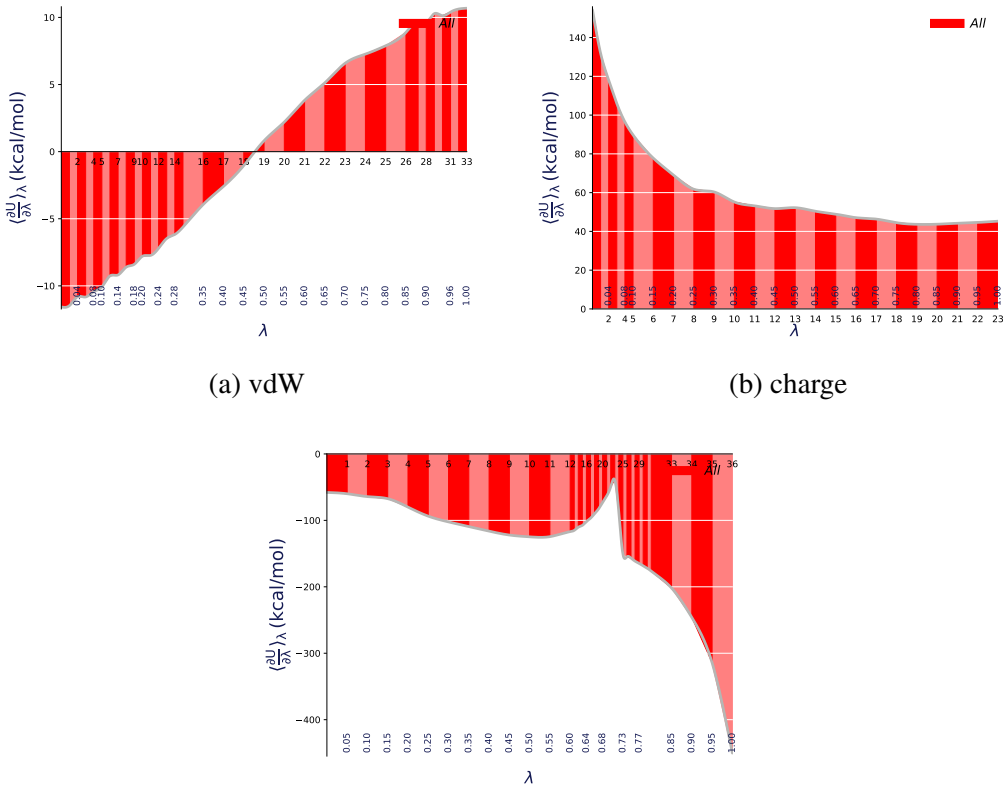
For each value of λ, the starting configuration was minimized using a steepest-decent algorithm for 10000 steps with a cutoff of 12 Å. Then in an NPT ensemble, the simulation box was heated for 1 ns with the Langevin thermostat using a coupling constant of 5.0 ps^{-1}, while the average temperature was slowly raised from 0 to 298 K, then kept at 298 K. At the same time, Berendsen barostat was used with a relaxation time of 2.0 ps to keep the average system pressure at 1.0 bar. The cutoff value used for this heating process was 10 Å. Also, in this step, a restraint was used on the protein backbone atoms of N, C_α, C, O, with 35 kcal mol^{-1} Å^2. The value of *skinbb* was increased to 5 Å. Then for the production phase of the simulation, in an NVT ensemble, we simulated the system for 10 ns with the Langevin thermostat using a coupling constant of 5.0 ps^{-1} to keep the average temperature at 298 K.

The cutoff value used for this production step was 12 Å. The same restraint on the protein backbone atoms was employed as the heating step.

The Python tool developed in ref. [53], *alchemical-analysis*, was used to calculate the free energy differences of mutation for each of the vdW, charge, and decharge steps. This module produces the results for different post-processing methods. The first 1 ns of the production simulations were ignored and considered as the equilibration time. The values reported in this thesis are from TI-3, but TI-1 integration curves alongside those from TI-3 for comparison are also shown.

In Figure 4.4, we show the result of these integrations for the mutation (D81)...D83N, mutating aspartate at position 83 to asparagine while another aspartate is present at position 81, on the surface of 1DOI protein at b_{KCl} = 2 mol/kg concentration. This figure shows the vdW (Figure 4.4a), charge (Figure 4.4b), and decharge (Figure 4.4c) curves of integration of $\langle \partial U(\lambda)/\partial\lambda \rangle_{\lambda_i}$ versus λ_i values. Usually, the integration for the charge step is smooth, and there is a sharp change around λ=0.65-0.78 in the decharge step where we used a $\Delta\lambda$ of 0.01 in this region to be able to capture the change more accurately. As for the vdW step, the integration is not always very smooth, but in general, the vdW step contributes minimal value to the total free energy change of the mutation as it can be seen in the Table A.2 in the case of 1DOI protein, and for other proteins as well.

The quantities of free energy corresponding to the Figure 4.4 calculated with TI-3 method: for the charge step is 59.50194 kcal/mol, for the decharge step is -133.78350 kcal/mol, and for the vdW step is 0.23334 kcal/mol that will result in a total free energy change of 74.5149 kcal/mol.



(a) vdW

(b) charge

(c) decharge

Fig. 4.4 Plot of $\left\langle \frac{\partial U}{\partial \lambda} \right\rangle_\lambda$ vs. λ values for the thermodynamic integration, with red filled areas indicating free energy estimates from the TI-1 and silver curve indicating interpolation via TI-3. Color intensity alternates with increasing λ index. The agreements between the alternate interpolation schemes suggest that the interpolation successfully captures the free energy change between two neighboring λ, as well as over the whole range. The subfigures correspond to the thermodynamic cycle 4.10 with (a) vdW ($\Delta G_{\mathrm{X}^0 a \mathrm{Y}^0}$), (b) charge ($-\Delta G_{\mathrm{Y}^0 a \mathrm{Y}}$), and (c) to decharge ($-\Delta G_{\mathrm{X} a \mathrm{X}^0}$) leg of this cycle. This figure shows the mutation corresponding to the (D81)…D83N at $\mathrm{b_{KCl}}$ = 2 mol/kg of 1DOI.

4.2.3 Identifying the best computational scheme to calculate free energies of mutation

Finding the appropriate procedure for an accurate free energy calculation of such a complicated system with a high salt concentration in explicit water, highly charged, and a mutation leading to charge-change in the system is a delicate and time-consuming matter. Issues such as the number of intermediate states, conformational sampling adequacy, finite-size effects in free energy calculation, post-processing free energy calculation methods, and so forth should be investigated before starting the main simulations. After finding the appropriate procedure, it should be validated by calculating the standard error corresponding to these simulations. In the subsequent computational test studies, if some simulation details are not present, the input from Section 4.2.2 is used.

The calculations of the free energy associated with an amino acid mutation are done using the 3-step protocol described schematically by the thermodynamic cycle 4.10: the charge, decharge, and vdW steps are performed separately. Mutation free energies could, in principle, have been calculated using a one-step protocol corresponding to the direct mutation of residue, ΔG_{XaY} in the thermodynamic cycle 4.10, where vdW and electrostatic interaction are both considered with softcore potential and the residue is mutated in one step. We opted not to do so because Garton et al. [28] have concluded in their study that for mutations leading to a change in charge, as it is in our case, the 3-step protocol should be used, resulting in more accurate free energy values. They have also indicated that if a change in charge with the mutation does not occur, the 3-step protocol should always be avoided.

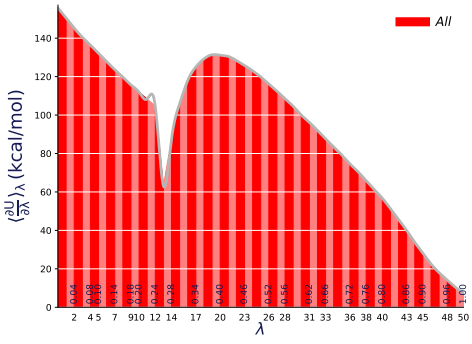
4.2.3.1 The effect of finite-size periodic boundary simulation box on the free energy calculation

Here the simulation of one potassium ion in a box of TIP3P water with different box sizes is explained. The production simulation was performed after minimization and heating, with a timestep of 2 fs for 10 ns. SHAKE algorithm was used for every bond involving one hydrogen. The average temperature was kept at 298 K using a coupling constant of 2 ps^{-1} using Langevin thermostat, and the average pressure was kept at 1 bar using Berendsen barostat with a coupling constant of 2 ps. A cutoff of 8 Å was used. In this simulation, the potassium ion was transferred to the gas phase using a one-step protocol. The integration for all the four different box sizes is shown in Figure 4.5.

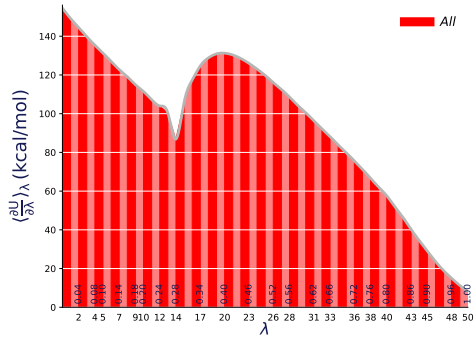
In total 51 λ values of 0.00, 0.02, 0.04, 0.06, 0.08, 0.10, 0.12, 0.14, 0.16, 0.18, 0.20, 0.22, 0.24, 0.26, 0.28, 0.30, 0.32, 0.34, 0.36, 0.38, 0.40, 0.42, 0.44, 0.46, 0.48, 0.50, 0.52, 0.54,

0.56, 0.58, 0.60, 0.62, 0.64, 0.66, 0.68, 0.70, 0.72, 0.74, 0.76, 0.78, 0.80, 0.82, 0.84, 0.86, 0.88, 0.90, 0.92, 0.94, 0.96, 0.98, 1.00 was used.

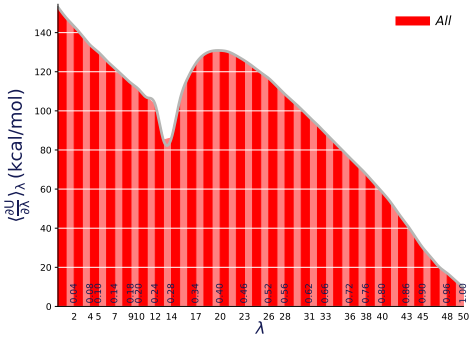
In the main simulations, with the mutation of one negatively charged amino acid, D or E, to a neutral amino acid, N or Q, the charge of the system changes after the mutation. These charge-changing free energies in explicit solvent might have a significant finite-size effect, as it has been shown to exist for similar systems [85]. The finite-size effect is the discrepancy of simulation in the periodic boundary box and ideal bulk, especially in long-range and significant electrostatic interactions treatment in simulations which causes a deviation from actual values depending mainly on the box size. This finite-size effect has its origin mainly in two issues in periodic boxes; firstly, the extra electrostatic interaction between the solute in the computational reference box, its periodic replicas, and the homogeneous background charge density; and secondly, the undersolvation of the solute in the reference box because the solvent in the periodic boxes is perturbed by the image of solute in the respective box, and is therefore unavailable for the solute in the main box [85]. Correcting schemes have been proposed analytically and numerically [39, 47, 48, 85] to correct these issues, but the implementation of these schemes, is not correctly accounted for in some implementation of lattice-sum electrostatics under periodic boundary conditions. Although in our study for every three steps of the thermodynamic cycle 4.10, and all different pairs studied at a concentration for a specific protein, the same box size is used, we have investigated this issue in the implementation of Particle Mesh Ewald (PME) [21] in AMBER [14]. We have designed this test simulation to check whether the box size affects the solvation free energy result of a charged ion. The standard error of the mean between the results in the Figure 4.5 is ± 0.18, which is negligible compared to the later reported standard error of our main calculations (see Section 4.2.3.4). This points to the fact that there is no box size-dependent effect on the result of free energies due to the finite-size effect on the PME, and correcting schemes of finite-size effect have been appropriately implemented.



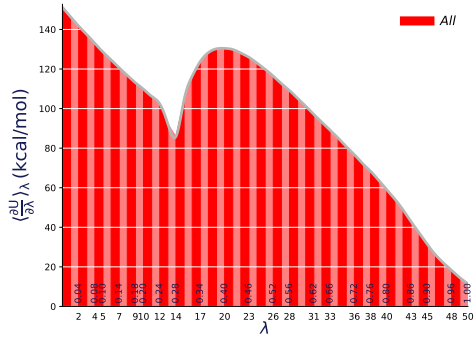
(a) Box-edge length ≈ 70 Å. Total free energy 93.78527 kcal mol^{-1}



(b) Box-edge length ≈ 60 Å. Total free energy 94.56233 kcal mol^{-1}



(c) Box-edge length ≈ 50 Å. Total free energy 94.30012 kcal mol^{-1}



(d) Box-edge length ≈ 45 Å. Total free energy 94.50033 kcal mol^{-1}

Fig. 4.5 Plot of $\left\langle \frac{\partial U}{\partial \lambda} \right\rangle_\lambda$ vs. λ values for the integration. This figure shows the mutation corresponding to annihilating a potassium ion. The details of the simulation are explained in Subsection 4.2.3.1.

4.2.3.2 Simulation time and number of intermediate states

To assess whether 10 ns simulation time for each λ was sufficient, we have designed a test to compare two simulations with 10 ns and 30 ns of production simulation time. All the simulation details are the same as Subsection 4.2.2, except that the SHAKE algorithm was not used for the heating and production runs. These simulations are performed for the same mutation (D81)...D83N of the protein 1DOI at b_{KCl} = 2 mol/kg similar to the mutation in the above-mentioned Figure 4.4.

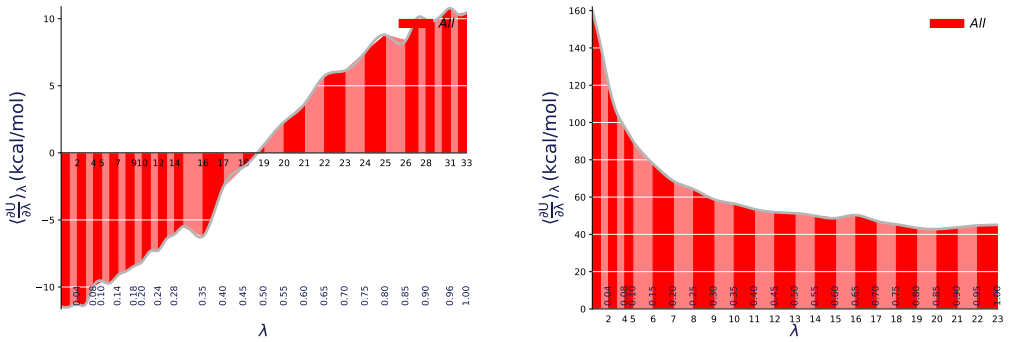
One further step for testing our results is to cross-validate the data from the integration-based method, TI-3, with the perturbation-based method, BAR. The values reported in the Table 4.1 shows the comparison of the results from Figures 4.6, and 4.7 for 10 ns and 30 ns, respectively. This comparison shows that the shape of the corresponding curve did not change by increasing simulation time to 30 ns, and the difference between the final results is less than the SEMs reported in the Section 4.2.3.4 except for the charge step where the difference is also negligible. The total free energy between the two is almost equal with a difference of 0.1 kcal/mol, which is lower than the SEM of total free energy $\pm$0.49.

Also, to assess whether the total number of 95 λ was sufficient, we performed another test simulation where the total number of λ was increased dramatically to 177. The number of λ is 59 for each vdW, decharge, and charge step reported in the following with a total of 177 λ compared to the previous simulation where for vdW 34, decharge 37, and charge 24 λ with a total of 95 λ was used. Simulation details are precisely the same as previous simulations in this subsection, with a 10 ns of total production simulation time. Figure 4.8 shows the integration curves corresponding to the different steps of this simulation, and Table 4.1 reports the values corresponding to these different steps. This comparison, 10 ns simulation for a total of 95 λ versus a total of 177 λ shows that the shape of the corresponding curve did not change, and the difference between the final results are less than at most twice of the SEMs reported in the Section 4.2.3.4 except for the vdW where the difference is also negligible, at most 0.12 kcal/mol. Considering that this increase in the number of λ is very dramatic and impossible to perform in terms of resources, such a difference in total free energy 0.65 kcal/mol slightly bigger than SEM of $\pm$0.49 is tolerable.

59 λ values of 0.00, 0.01, 0.02, 0.03, 0.04, 0.05, 0.06, 0.08, 0.12, 0.14, 0.16, 0.20, 0.24, 0.28, 0.32, 0.36, 0.40, 0.44, 0.48, 0.52, 0.56, 0.58, 0.60, 0.62, 0.64, 0.65, 0.66, 0.67, 0.68, 0.70, 0.71, 0.72, 0.73, 0.74, 0.75, 0.76, 0.77, 0.78, 0.79, 0.80, 0.81, 0.82, 0.83, 0.84, 0.85, 0.86, 0.87, 0.88, 0.89, 0.90, 0.91, 0.93, 0.94, 0.95, 0.96, 0.97, 0.98, 0.99, 1.00 for each step of vdW, decharge, charge was used for the case with a total of 177 λ.

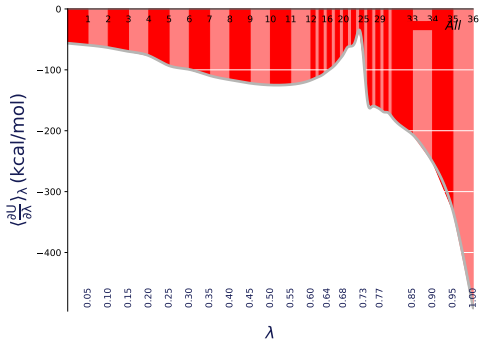
Table 4.1 Free energy values in units of kcal/mol, calculated from different perturbation-based, BAR , and integration-based, TI-3, methods for the cases with a total of 95 λ, and 177 λ.

		decharge		**charge**		**vdW**		**total**	
	Simulation time (ns)	TI-3	BAR	TI-3	BAR	TI-3	BAR	TI-3	BAR
95λ {	10 (fig. 4.6)	-135.79401	-135.49497	60.04460	59.89161	0.15400	0.19775	75.90341	75.80111
	30 (fig. 4.7)	-135.33353	-135.38722	59.66333	59.65385	0.13958	0.17432	75.80978	75.90769
177λ {	10 (fig. 4.8)	-134.75056	-134.36508	59.76601	59.73252	0.27122	0.25510	75.25577	74.88766



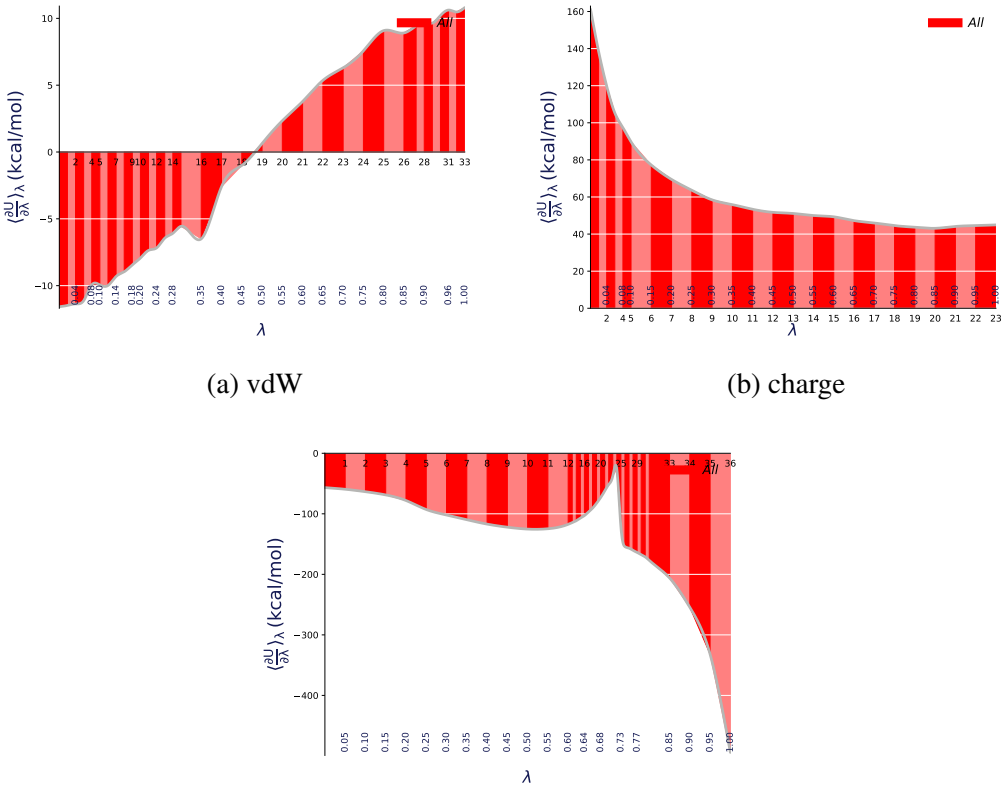
(a) vdW

(b) charge



(c) decharge

Fig. 4.6 Plot of $\left\langle \frac{\partial U}{\partial \lambda} \right\rangle_\lambda$ vs. λ values for the integration. This figure shows the mutation corresponding to the (D81)...D83N of the protein 1DOI at b_{KCl} = 2 mol/kg, with 10 ns of simulation time. The details of the simulation are explained in Subsection 4.2.3.2.

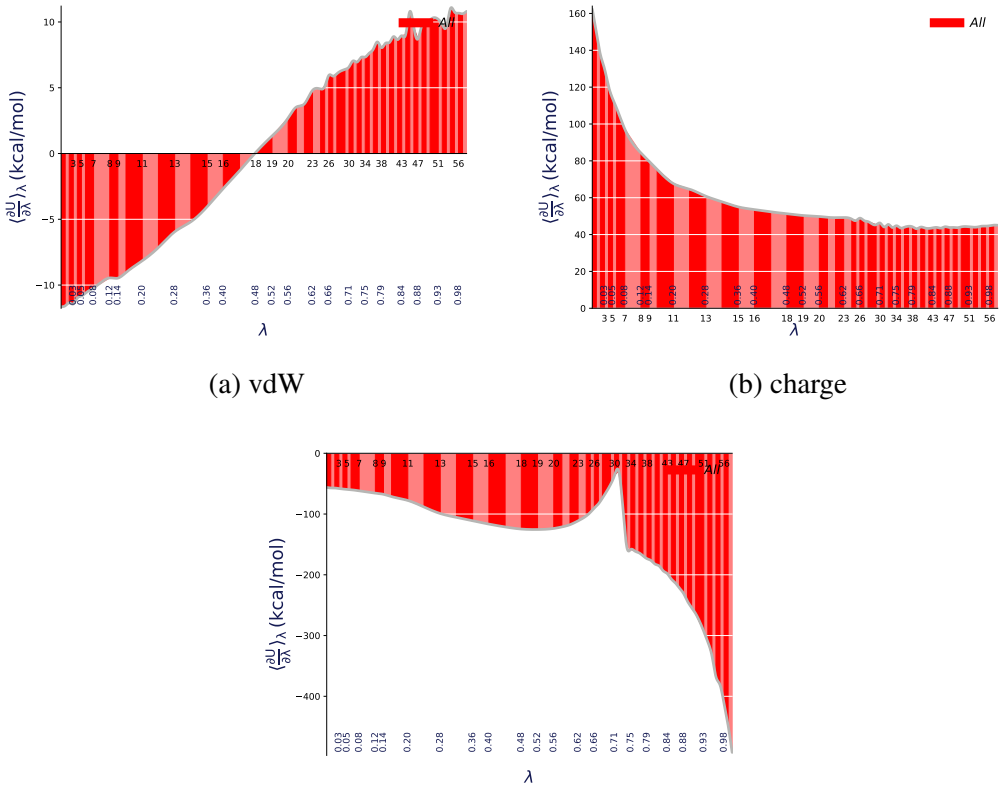


(a) vdW

(b) charge

(c) decharge

Fig. 4.7 Plot of $\left\langle \frac{\partial U}{\partial \lambda} \right\rangle_\lambda$ vs. λ values for the integration. This figure shows the mutation corresponding to the (D81)...D83N of the protein 1DOI at b_{KCl} = 2 mol/kg, with 30 ns of simulation time. The details of the simulation is explained in Subsection 4.2.3.2.



(a) vdW

(b) charge

(c) decharge

Fig. 4.8 Plot of $\left\langle \frac{\partial U}{\partial \lambda} \right\rangle_\lambda$ vs. λ values for the integration. This figure shows the mutation corresponding to the (D81)…D83N of the protein 1DOI at b_{KCl} = 2 mol/kg, with a total of 177 λ. The details of the simulation is explained in Subsection 4.2.3.2.

4.2.3.3 Restraining the backbone of proteins

Here we compare two simulations for the protein 1DOI, one with the backbone restrained and the other without restraint. In the main simulation, Section 4.2.2 we used restraint on the backbone atoms to avoid insufficient conformational sampling.

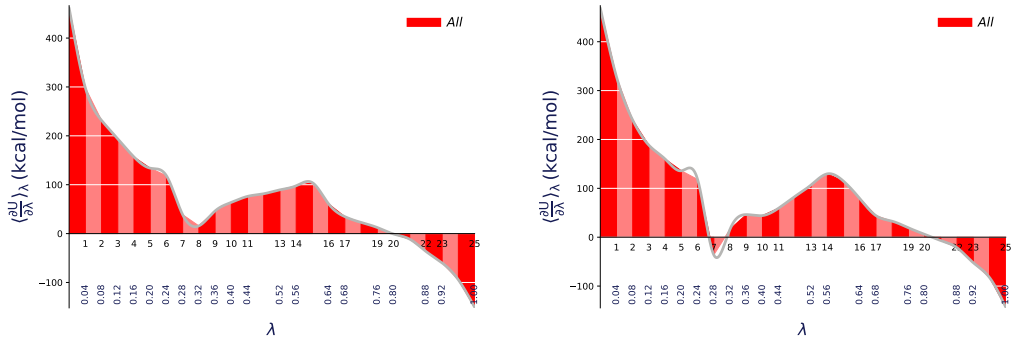
For simulation, first, we heated the system 300 ps with a timestep of 1 fs without using SHAKE on the system, using a Langevin thermostat with a coupling constant 1 ps^{-1}, we slowly heated the system from 0 to 298 K and then kept the average temperature at 298 K. A cutoff of 12 Å was used in this step. In this step, a restraint of 10 kcal mol^{-1} $Å^2$ on all heavy atoms of protein was used. Then 30 ns of production simulation using a timestep of 2 fs because the SHAKE algorithm was used for all the bonds connected to the hydrogen atoms, even the mutating residues, was performed. We used the same cutoff as the heating step. The Langevin thermostat was used to keep the average temperature at 298 K using a coupling constant of 2 ps^{-1}, and Berendsen barostat to keep the average pressure at 1 bar using a coupling constant of 2 ps. For one simulation, a restraint of 50 kcal mol^{-1} $Å^2$ on the backbone atoms of N, C_α, C, and O was used, and for the other, the backbone was free.

In total 26 λ values of 0.00 (X), 0.04, 0.08, 0.12, 0.16, 0.20, 0.24, 0.28, 0.32, 0.36, 0.40, 0.44, 0.48, 0.52, 0.56, 0.60, 0.64, 0.68, 0.72, 0.76, 0.80, 0.84, 0.88, 0.92, 0.96, 1.00 (Y) corresponding to direct mutation in the thermodynamic cycle 4.10 was used.

Comparing the results, firstly, the shape of the derivation of mixed potential versus λ is similar and somewhat smoother, with less sharp change, in the case where the backbone is restrained with a better overlap between TI-1 and TI-3 in the Figure 4.9a, compared to when it is free in the Figure 4.9b. Another point is that when the backbone is restrained, the difference between different methods of free energy calculation, TI-3 and BAR, reported in the Table 4.2 is smaller, and the large difference in the case of free backbone points to the insufficiency of simulation time because TI-3 results are different between the two cases.

Table 4.2 Free energy values in units of kcal/mol, calculated from different perturbation-based, BAR , and integration-based, TI-3, methods for the cases with restrained backbone and free backbone.

Backbone	TI-3	BAR
Restrained (fig. 4.9a)	72.75423	73.61249
Free (fig. 4.9b)	76.03225	73.29420



(a) Backbone restrained (b) Backbone free

Fig. 4.9 Plot of $\left\langle \frac{\partial U}{\partial \lambda} \right\rangle_{\lambda}$ vs. λ values for the integration. This figure shows the mutation corresponding to the D34N of the protein 1DOI at b_{KCl} = 2 mol/kg. The details of the simulation are explained in Subsection 4.2.3.3.

4.2.3.4 Calculating the standard error of the free energies

For calculating the Standard Error of the Mean (SEM) in our free energy calculation, we have repeated the simulation for the mutation of (E26)...D29N from the Table A.2 at b_{KCl} = 2 mol/kg of protein 1DOI. We have performed the same simulation with simulation detail explained in the Section 4.2.2 for 5 times, and then used the following final results to calculate the SEM for each step of the thermodynamic cycle 4.10:

$-\Delta G_{Y^0{}_{aY}}$ = charge = [56.26149, 56.54705, 56.56241, 56.45222, 56.99566]
$-\Delta G_{X_{aX^0}}$ = decharge = [-132.62702, -131.24322, -133.03272, -133.62545, -134.52607]
$\Delta G_{X^0{}_{aY^0}}$ = vdW = [-0.64023, -0.61521, -0.73233, -0.61889, -0.60338]
$\Delta G_{X_{aY}}$ = total = [75.7253, 74.08096, 75.73798, 76.55434, 76.92703]

Then we calculated the SEM from the following equations:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N-1}} \tag{4.11}$$

$$\text{SEM} = \frac{\sigma}{\sqrt{N}} \tag{4.12}$$

where N is the number of data points, x_i is the data points, $\bar{x}$ is the mean of all data, and σ is the standard deviation. Standard deviation, Equation 4.11, is an estimate of the variability of the data. The standard error of the mean, Equation 4.12, is a measure of the precision of the sample mean.

$\text{SEM}_{\text{charge}} = \pm\ 0.12$
$\text{SEM}_{\text{decharge}} = \pm\ 0.55$
$\text{SEM}_{\text{vdW}} = \pm\ 0.02$
$\text{SEM}_{\text{total}} = \pm\ 0.49$

4.3 Replica Exchange Molecular Dynamics (REMD)

4.3.1 Theoretical background

The number of local minimum-energy that a particular conformation might get trapped in for an ensemble of unfolded states is enormous compared to a folded configuration, making it difficult to sample unfolded states at low temperatures accurately. To overcome the energy barriers of these many minimas and sample larger phase space, some methods have been developed to perform a simulation where there could be a random walk in energy

space by using a non-Boltzmann probability distribution. The Replica Exchange Molecular Dynamics (REMD) [37, 93, 95, 101] is one of these methods which, by exchanging multiple non-interacting system copies at different temperatures, attempts to solve the multiple-minima problem. REMD is one of the many expanded ensemble methods, trying to perform random walks in energy space, in which the sampling of the system is much larger than a typical molecular dynamics conformational sampling of an ensemble. In expanded ensemble techniques, not only a continuous conformational sampling of the system is performed, but also a discrete thermodynamic state space is explored. Therefore, the phase space in these methods consists of the particle's position and conjugate momentum, and thermodynamic state variables. Figure 4.10 shows how REMD works, where boxes are system copies or replicas at different temperatures with no interaction between them, and in the course of the simulation, there are specific points in time where these replicas exchange with each other. The arrows in this figure show how these exchanges happen between neighboring replicas in the space of the thermodynamic state variable.

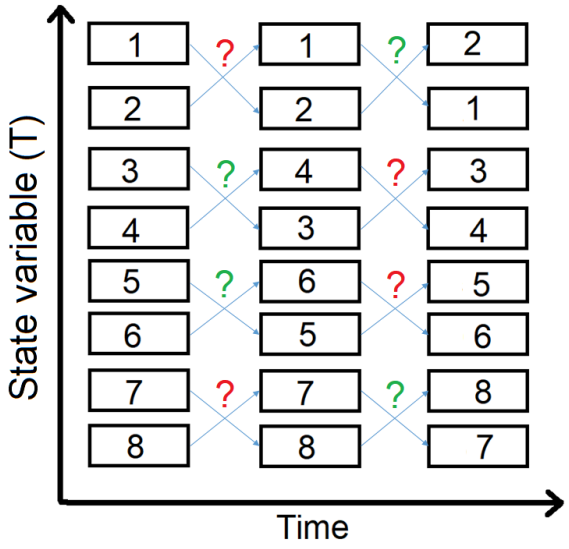


Fig. 4.10 A Replica exchange consisting 8 replicas. Question marks represent exchange attempts (green for accepted, and red for rejected). Figure is reproduced from ref. [14].

In the following, the formulation of these exchanges is explained. Boltzman weight factor of a canonical state at temperature T is:

$$W_B(x;T) = e^{-\beta H(q,p)}, \tag{4.13}$$

where $H(q,p)$ shows the Hamiltonian of the state x with q as position vector, and p as momentum vector of the particles of the system, and β is equal $1/(k_B T)$ (k_B is the Boltzmann constant) [93].

The average kinetic energy at temperature T is:

$$\langle K(p)\rangle_T = \left\langle \sum_{k=1}^{N} \frac{p_k^2}{2m_k} \right\rangle_T = \frac{3}{2} N k_B T. \tag{4.14}$$

where m is the mass of the particles, and N is the number of the particles in the system [93].

A REMD simulation has copies of the canonical ensemble at M different temperatures that are called replicas. The simulation performed in each of these replicas does not interact with another replica, so the forces on the particles in one replica do not affect particles of other replicas. During the REMD simulation, these replicas try to exchange their temperature with each other. Consider the following system:

$$x_m^{[i]} \equiv \left(q^{[i]}, p^{[i]}\right)_m. \tag{4.15}$$

Where the state x has the temperature m and is replica number i. In this state the N atoms have coordinates of $q^{[i]}$ and momentum of $p^{[i]}$. An exchange of this replica with another replica can be written as:

$$\begin{cases} x_m^{[i]} \equiv \left(q^{[i]}, p^{[i]}\right)_m \rightarrow x_n^{[i]'} \equiv \left(q^{[i]}, p^{[i]'}\right)_n, \\ x_n^{[j]} \equiv \left(q^{[j]}, p^{[j]}\right)_n \rightarrow x_m^{[j]'} \equiv \left(q^{[j]}, p^{[j]'}\right)_m, \end{cases} \tag{4.16}$$

where $p^{[i]'}$ and $p^{[j]'}$ are defined in Equation 4.17 [93]. In this process, a pair of temperatures T_m and T_n corresponding to replicas i and j have been exchanged. In such a process, both the potential energy function $E(q)$ and the momentum p are taken into account for allowing such exchange to take place [93]. The following assignment after a successful exchange is performed to scale the velocities of all atoms in both replicas with the condition in Equation 4.14:

$$\begin{cases} p^{[i]'} \equiv \sqrt{\frac{T_n}{T_m}} p^{[i]}, \\ p^{[j]'} \equiv \sqrt{\frac{T_m}{T_n}} p^{[j]}, \end{cases} \tag{4.17}$$

The criteria for the exchange to be accepted is based on satisfying detailed balance, Equation 4.18, to converge an equilibrium of the REMD simulation. To calculate the probability of the exchange between the two states $w\left(X \rightarrow X'\right)$:

$$W_X \; w_{X \to X'} = W_{X'} \; w_{X' \to X} \tag{4.18}$$

$$w\left(X \to X^{'}\right) \equiv w\left(x_m^{[i]} \middle| x_n^{[j]}\right) = \begin{cases} 1, & for\ \Delta \leq 0, \\ exp\left[-\left[\beta_n - \beta_m\right]\left(E\left(q^{[i]}\right) - E\left(q^{[j]}\right)\right)\right], & for\ \Delta > 0 \end{cases} \tag{4.19}$$

Where the Equation 4.19 is the Metropolis criterion to satisfy the detailed balance, and the Δ is the negative of the argument of the *exp* in the second line of the Equation 4.19 [93].

A REMD simulation is then performed by first simulating each replica in the canonical ensemble with the specific temperature assigned to it, *simultaneously* and *independently*, for a specified number of steps. Secondly, in temperature space, each pair of neighboring replicas attempt an exchange with the probability calculated in the Equation 4.19. These two steps are alternately carried out. In REMD implementation, only the neighboring temperatures are attempted to exchange because the transition probability between them decreases exponentially with the difference in temperature shown in Equation 4.19. For example, if replicas 5 and 6 attempt to exchange at one time, then the next time replicas 6 and 7, as well as 4 and 5, will attempt to exchange. The choice of the temperature space is essential to the convergence of REMD simulations [93].

4.3.2 Computational details of REMD simulation

We performed REMD simulation to sample an unfolded state ensemble of halophilic protein L. (pdb ID: 2KAC) at the high salt concentration of b_{KCl} = 2 mol/kg to investigate the cooperative effect in the unfolded state and compare it with the results from the folded state.

The simulation was performed using the AMBER simulation package 2018[14] version of the pmemd engine on GPU. The only exception was the l-bfgs minimization step, which was performed on the Sander engine of the AMBER simulation package because it is not available on the pmemd engine. All the simulation boxes are cubic with periodic boundary conditions applied in the XYZ directions. We set a non-bonded potential cutoff at 12 Å distance for vdW and electrostatic interactions, and beyond this cutoff distance, the electrostatic interactions are calculated with the particle mesh Ewald (PME) scheme with a grid spacing of 1.0 Å, and 4^{th} order of interpolation [16]. Long-range dispersion corrections were applied to both the energy and pressure. The crystal structure of the halophilic protein of 2KAC at a high concentration of 2 mol/kg of KCl in a box of TIP3P water with the edge length of $\approx$ 110 Å was used. All bonds with H-atoms were constrained using the SHAKE algorithm [87] in the NPT and REMD simulation. Four initial minimizations with 2500 steps of the algorithm of

steepest-decent and 7500 steps of Conjugate Gradient while using four descending restraints on protein atoms of 500, 300, 100, and 50 kcal mol^{-1} $Å^{-2}$ were performed to remove the bad contacts from the experimental structure and in the process of adding ions and water to the system. Another step of minimization using the l-bfgs algorithm without any constraints or restraints for 10000 steps was performed.

Then a 50 ns heating simulation in a canonical ensemble (NVT) using Langevin thermostat with a collision frequency of 1.0 ps^{-1} to slowly increase the temperature of the system from 0 to 800 K was performed to help denature the protein. Then in another heating simulation in a canonical ensemble (NVT) for 2.5 ns, the temperature was decreased again to reach 298 K using Langevin thermostat with the same collision frequency. No SHAKE algorithm was used for these two heating simulations, so consequently, a timestep of 1 fs was used. Another 10 ns of equilibration simulation, in 10 steps of 1 ns simulations, was performed for equilibrating density in the isothermal-isobaric ensemble (NPT) using Berendsen barostat[8], and the Langevin thermostat, to keep the average temperature at 298 K, and pressure at 1 bar.

The final part of the simulation, REMD simulation, was performed in NVT ensemble, and as the input coordination we used a frame from the the last 1 ns NPT equilibration simulation in which the density of the box was closest to the average. Also for the REMD simulation we used Langevin thermostat with a collision frequency of 1 ps^{-1}. The number of exchange attempts that was performed for each replica was 20000, while using 2500 as the number of MD steps performed between each exchange attempt. Consequently the total number of steps for each replica performed is equal to $2500 \times 20000 = 5 \times 10^7$ which considering the timestep of 2 fs results in 100 ns of simulation time for each replica. A total of 221 replica with different temperatures was used as 298.15, 298.98, 299.82, 300.65, 301.49, 302.33, 303.18, 304.02, 304.87, 305.71, 306.56, 307.42, 308.27, 309.12, 309.98, 310.84, 311.70, 312.57, 313.43, 314.30, 315.17, 316.04, 316.91, 317.79, 318.67, 319.55, 320.43, 321.31, 322.20, 323.09, 323.97, 324.87, 325.76, 326.66, 327.55, 328.45, 329.35, 330.26, 331.16, 332.07, 332.98, 333.89, 334.81, 335.73, 336.64, 337.57, 338.49, 339.41, 340.34, 341.27, 342.20, 343.13, 344.07, 344.97, 345.91, 346.86, 347.80, 348.75, 349.69, 350.64, 351.60, 352.55, 353.51, 354.47, 355.43, 356.39, 357.36, 358.33, 359.30, 360.27, 361.24, 362.22, 363.20, 364.18, 365.17, 366.15, 367.14, 368.13, 369.12, 370.12, 371.11, 372.11, 373.12, 374.12, 375.12, 376.13, 378.16, 37.16, 380.18, 381.20, 382.22, 383.25, 384.28, 385.30, 386.34, 387.37, 388.41, 389.44, 390.48, 391.53, 392.57, 393.62, 394.68, 395.73, 396.79, 397.84, 398.90, 399.97, 401.03, 402.10, 403.18, 404.25, 405.33, 406.40, 407.48, 408.57, 409.65, 410.74, 411.83, 412.92, 414.02, 415.12, 416.22, 417.32, 418.42, 419.53, 420.64, 421.75, 422.87, 423.99, 425.11, 426.23, 427.36, 428.48, 429.62, 430.75, 431.89, 433.03, 434.17, 435.31, 436.46, 437.61, 438.76, 439.91, 441.07, 442.2,

443.39, 444.56, 445.73, 446.90, 448.07, 449.25, 450.43, 451.61, 452.79, 453.98, 455.17, 456.36, 457.55, 458.75, 459.95, 461.15, 462.36, 463.57, 464.78, 466.00, 467.21, 468.44, 469.66, 470.88, 472.11, 473.35, 474.57, 475.81, 477.05, 478.29, 479.54, 480.79, 482.04, 483.29, 484.55, 485.81, 487.07, 488.34, 489.61, 490.88, 492.15, 493.43, 494.71, 495.99, 497.28, 498.57, 499.86, 501.16, 502.46, 503.76, 505.06, 506.37, 507.68, 508.99, 510.31, 511.63, 512.95, 514.28, 515.61, 516.94, 518.27, 519.61, 520.95, 522.29, 523.64, 524.99, 526.35, 527.70, 529.06, 530.43, 531.79, 533.16, 534.53, 535.91, 537.00. This means a total of 22.1 μs of simulation. This temperature distribution was taken from the website http://folding.bmc.uu.se/remd-temperature-generator/ [73].

At such high temperatures, (> 500) unwanted rotations around the peptide bond might occur, leading to non-physical chiralities. To prevent this, we used chirality restraints of 50 kcal mol^{-1} Å^{-2} on the backbone ω dihedrals, consists of C_α, C, N, C_α, to retain it in *trans* configuration to keep it planar, and do not allow its rotation to *cis* configuration.

In the end, we extracted a trajectory from only those frames of all replica trajectories that corresponded to 298 K and performed the subsequent analysis on this trajectory.

4.3.3 Evaluating the quality of REMD simulations

To examine whether the REMD simulation performed accurately with the temperature distribution, the number of replicas, and the simulation details used in our study, we examine the acceptance ratios of exchanges of adjacent pairs. Figure 4.11 shows the acceptance ratio of the exchanges between replicas, considering up and down exchanges, with more than 40% acceptance for all replicas. The values are relatively uniform (all around 40-56% of acceptance probability), pointing to the fact that exchanges between neighboring replicas were likely in all parts of the temperature distribution. Also, the acceptance ratios are large enough (> 40%), leading to a sufficient number of exchanges during the simulation.

Figure 4.12 shows the temperature distribution for one of the replicas with an initial temperature of 298.98 K, where we observe that this replica has visited higher temperatures at the end of the REMD simulation, but it is far from a completely random walk that obviously needs a much longer simulation time. Considering these analyses, we have sampled a subset of the unfolded structure ensemble, which considering our purpose, would suffice. We want to study one frame from the resulting trajectory to study the possibility of a cooperative effect in a denatured conformation of the halophilic protein and use the trajectory for simple structural analysis. These simulations can be continued for a longer timescale to study the structure and dynamics of the hydration shell in the properly sampled unfolded state to compare with the folded structure in the future.

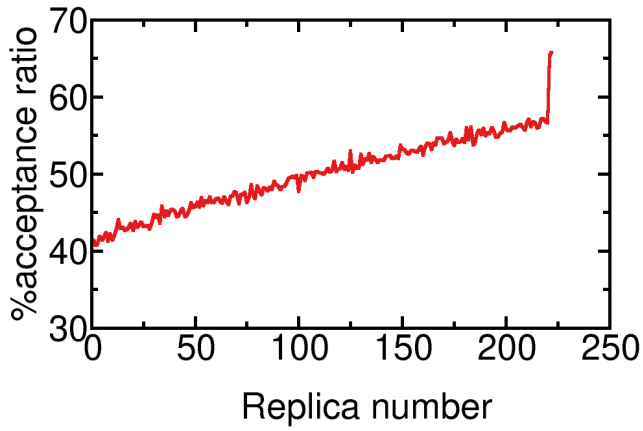


Fig. 4.11 Acceptance ratio which shows the percentage of accepted exchanges for each replica relative to all exchange attempts.

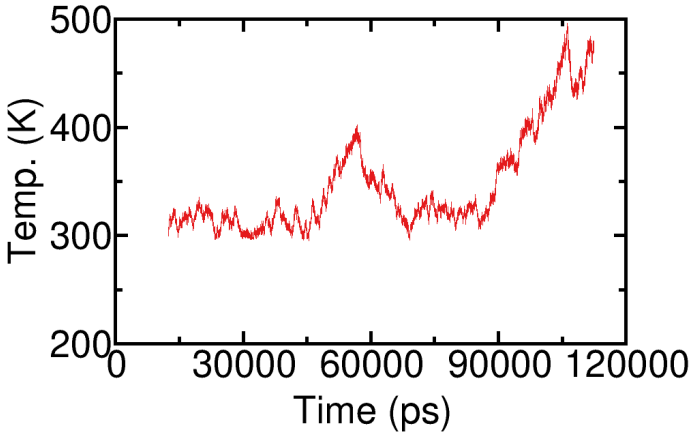


Fig. 4.12 Temperature distribution for on of the replicas with the starting temperature of 298.98 K.

4.4 Results

4.4.1 Free energies of D→N and E→Q mutations with/without vicinal acidic amino acids in folded proteins

As described in Section 4.1, we evaluate the extent of cooperativity between two vicinal acidic residues by comparing the free energy of mutation of one acidic residue to a neutral one in the presence or absence of its acidic neighbor. The two mutation processes are described by Eq. 4.1 and 4.2 and have associated free energy differences $\Delta G_{(-)\mathrm{X}a\mathrm{Y}}$ and $\Delta G_{(0)\mathrm{X}a\mathrm{Y}}$, where the symbol in parenthesis emphasizes the charge of the vicinal amino acid and X and Y will indicate amino acids using one-letter amino acid notation. Amino acids are mutated to the natural amino acid that most resembles the original: aspartate (D) to asparagine (N) and glutamate (E) to glutamine (Q). Likewise, when calculating $\Delta G_{(0)\mathrm{X}a\mathrm{Y}}$, the (now absent) acidic neighbor has been replaced by its closest neutral amino acid.

We compare the two mutation processes via the difference

$$\Delta\Delta\mathrm{G} = \Delta G_{(0)\mathrm{X}a\mathrm{Y}} - \Delta G_{(-)\mathrm{X}a\mathrm{Y}} \tag{4.20}$$

If ΔΔG is negative, the free energy change $\Delta G_{(-)\mathrm{X}a\mathrm{Y}}$ (Eq. 4.1) is more positive than $\Delta G_{(0)\mathrm{X}a\mathrm{Y}}$ (Eq. 4.2), which means it is harder to mutate an aspartate to an asparagine when it has another acidic residue close to it compared to when it does not. In other words, it means that having two vicinal negatively charged amino acids is more energetically favorable than having a neutral residue close to a negatively charged one. This is how we define and quantify synergistic, or stabilizing, or cooperative effects between acidic amino acids. A positive ΔΔG indicates that nearby negative charges repel each other (as expected); we call this an interfering interaction.

The figures that follow show ΔΔG for multiple amino acid pairs. The amino acid pairs and the mutations performed are indicated using the condensed notation: e.g., (D,N36)...E41Q indicates that the glutamate in position 41 of the amino acid sequence is mutated to glutamine; the vicinal amino acid is in position 36 and is either an aspartate (when calculating $\Delta G_{(-)\mathrm{E}a\mathrm{Q}}$) or an asparagine (when calculating $\Delta G_{(0)\mathrm{E}a\mathrm{Q}}$).

We measured ΔΔG for selected pairs of acidic residues close to each other on the surface of some of the halophilic proteins we have already studied in previous chapters: halophilic protein ferredoxin, protein L., and protein dihydrofolate reductase. Figure 4.13 shows the results for the protein ferredoxin at high (b_{KCl} = 2 mol/kg), and low (b_{KCl} =0.15 mol/kg) KCl molality. As one can see, in all the cases at low KCl concentration, the ΔΔG is positive but varies substantially (between 0.5 kcal/mol and 2 kcal/mol) with the pair of amino acids

being investigated. At high KCl concentration, most values of ΔΔG remain positive, although they are clearly lower than those at low KCl concentration, reflecting the expected higher electrostatic shielding brought by the more concentrated electrolyte solution. Most amino acid pairs thus show interfering, repulsive interactions between the acidic residues because the ΔΔG is positive, and this is according to expectations: two negative residues close to each other should repel each other and lead to the destabilization of the structure as we can see here. This destabilization effect of having two negative residues is higher at low KCl concentration because there is less electrostatic shielding. The only exceptions to this trend are the (D,N109)...E110Q amino acid pair, which shows synergistic interaction, and the (D,N83)...D81N pair, which shows barely repulsive interactions. These exceptions confirm that strong destabilizing effects induced by electrostatic repulsion between neighboring acidic amino acids are not always the case at high salt concentrations. Thus, synergistic interactions between acidic amino acids in proteins are indeed possible for some protein sites.

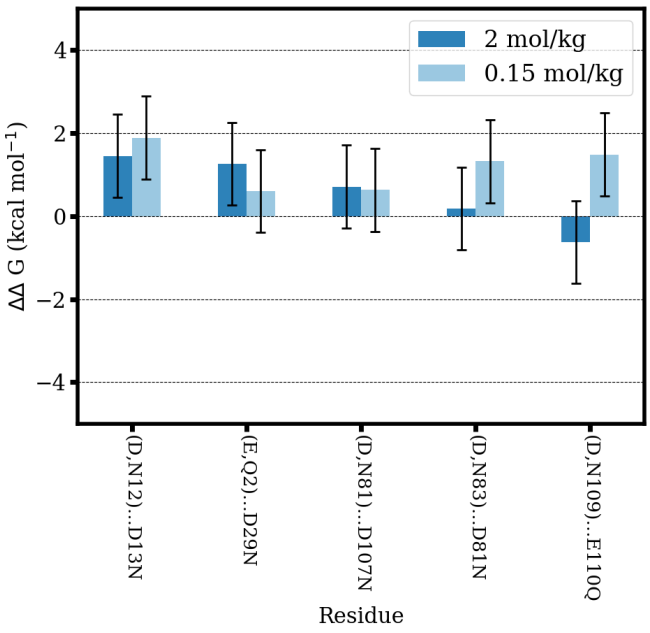


Fig. 4.13 Change in free energy of mutation (ΔΔG±1.0; Eq. 4.20) for the indicated pairs of amino acids of the halophilic protein ferredoxin (pdb ID: 1DOI), calculated at the indicated KCl molality. The data shown in this plot is compiled in Table A.2.

The presence of synergistic effects between vicinal acidic amino acids at high salt concentration was also observed for a second halophilic protein, the dihydrofolate reductase with pdb ID:2ITH; these results are shown in Figure 4.14. This protein has a substantially lower net charge (-15e) than the halophilic ferredoxin (-29e). One pair of acidic residues

clearly shows a strong synergistic effect, two show strong interfering effects, and two other pairs show only weak ($\leq$ 1 kcal/mol) interfering effects.

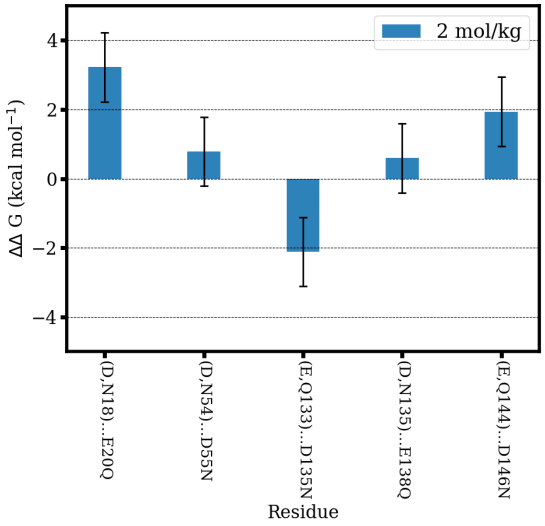


Fig. 4.14 Change in free energy of mutation ($\Delta\Delta G \pm 1.0$; Eq. 4.20) for the indicated pairs of amino acids of the halophilic protein dihydrofolate reductase (pdb ID: 2ITH), calculated at the high KCl molality. The data shown in this plot is compiled in Table A.4.

Figure 4.15 shows the $\Delta\Delta G$ results for protein L. This protein also has a net charge of -15e but is substantially smaller (only 64 amino acids) than either the halophilic dihydrofolate reductase (162 amino acids) or the halophilic ferredoxin (128 amino acids). A total of 8 amino acid pairs were investigated at high KCl concentration, but only 5 pairs were studied at low concentration. Given the high computational cost of these calculations, we increased the number of data points at the high salt concentration, for which synergistic effects appear to occur more frequently. Like ferredoxin at low KCl concentration, 3 out of the 5 pairs investigated in protein L. at low KCl concentration have a destabilizing influence and are interfering. Surprisingly, we see synergistic effects and weak interfering effects (i.e., quasi non-interfering effects) between vicinal negative amino acids in most cases at high salt concentration.

The results in Figures 4.13, 4.14 and 4.15 demonstrate that a synergistic effect between neighboring acidic amino acids are possible and may even be more frequent than the expected interfering interaction, at least for some proteins and depending on the electrolyte concentration. First of all, a synergistic effect arises from interactions between the acidic amino acids and the electrolyte in solution because it happens much less frequently at low KCl concentration. Secondly, the abundance of weak interfering effects (i.e., quasi non-interfering

effects) is also surprising: two negative residues (e.g., Asp…Asp) would be expected to destabilize each other, rather than to interact similarly to a charged-non-charged pair (e.g., Asp…Asn).

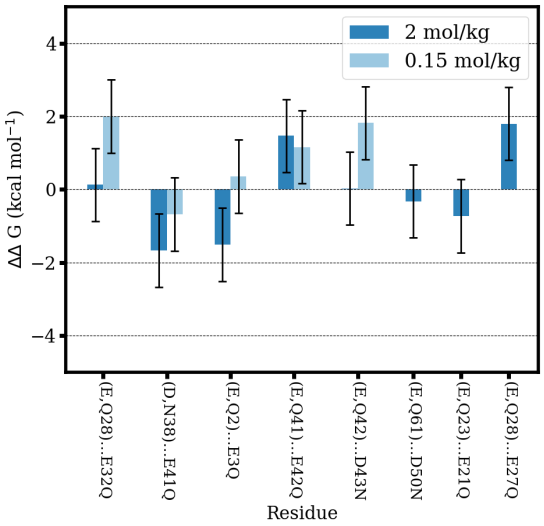


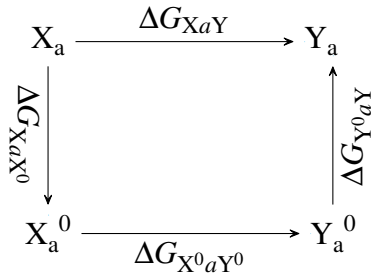
Fig. 4.15 Change in free energy of mutation (ΔΔG±1.0; Eq. 4.20) for the indicated pairs of amino acids of the halophilic protein L. (pdb ID: 2KAC), calculated at the indicated KCl molality. The cases of D50N, E21Q, and E27Q were only simulated at high salt concentrations. The data shown in this plot is compiled in Table A.3.

4.4.1.1 Decomposing mutation free energy into its components

To understand the origin of the synergistic effects, we will show the decomposition of the change in mutation free energy, $\Delta\Delta G$, for the various amino acid pairs investigated for protein L. at high KCl concentration. The free energy associated with a generic mutation XaY is calculated in three steps

$$\Delta G_{\mathrm{X}a\mathrm{Y}} = \Delta G_{\mathrm{X}a\mathrm{X}^0} + \Delta G_{\mathrm{X}^0a\mathrm{Y}^0} + \Delta G_{\mathrm{Y}^0a\mathrm{Y}} \tag{4.21}$$

following the thermodynamic cycle shown in Eq. 4.10. That scheme is repeated here for ease of vieweing:



The term ΔG_{XaX^0} is the free energy change associated with decharging amino acid X, i.e., setting all atomic charges of this residue to zero; this term includes only the contribution of the residue charges to the interactions with its environment (the rest of the protein and the solvent). The term $\Delta G_{X^0aY^0}$ corresponds to mutating the decharged residue X^0 into the decharged residue Y^0; this term includes only the contributions of the changes in the Lennard-Jones potentials of the residue to the interactions with its environment and is often termed the van der Waals (vdW) contribution. The final term ΔG_{Y^0aY} is the free energy change associated with reinstating the atomic charges of amino acid Y; similarly to the first term, this term includes only the contribution of the residue charges to the interactions with its environment.

The change in mutation free energy, $\Delta\Delta G$ (defined in Eq. 4.20) associated with the presence of a neutral or charged vicinal amino acid, is decomposed into its decharging, vdW and charging components, as:

$$\Delta\Delta G_{\text{decharging}} = \Delta G_{(0)XaX^0} - \Delta G_{(-)XaX^0} \tag{4.22}$$

$$\Delta\Delta G_{\text{vdW}} = \Delta G_{(0)X^0aY^0} - \Delta G_{(-)X^0aY^0} \tag{4.23}$$

$$\Delta\Delta G_{\text{charging}} = \Delta G_{(0)Y^0aY} - \Delta G_{(-)Y^0aY} \tag{4.24}$$

The subscripts (0) and $(-)$ indicate the net charge of the vicinal amino acid, as indicated previously. These components are shown in Figure 4.16 for the halophilic protein L. at b_{KCl} = 2 mol/kg. The $\Delta\Delta G_{\text{vdW}}$ component takes both positive and negative values, but its absolute value is always below ≈ 0.5 kcal/mol. The vdW component is thus essentially independent of the identity of the vicinal amino acid, for changes between glutamate/glutamine and aspartate/asparagine. The sign of $\Delta\Delta G$ is thus determined by the balance between the charging and the decharging steps. The charging step, which introduces the atomic charges in the final (neutral) amino acid, predominantly has a positive contribution to $\Delta\Delta G$. In contrast, the decharging step, which removes the atomic charges in the initial (negative)

amino acid, negatively contributes to $\Delta\Delta G$. Figure 4.16 thus clarifies that a negative ΔΔG –

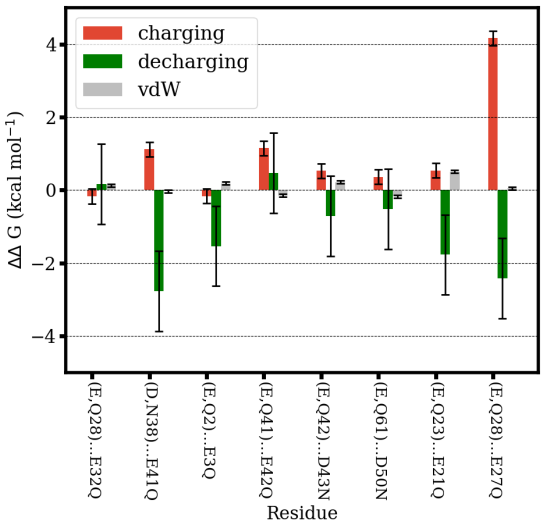


Fig. 4.16 $\Delta\Delta G_{\text{charging}} \pm 0.2, \Delta\Delta G_{\text{decharging}} \pm 1.1$ and $\Delta\Delta G_{\text{vdW}} \pm 0.04$ components (Eq. 4.22, 4.23 and 4.24 of the change in free energy of mutation ($\Delta\Delta G$; Eq. 4.20) for the indicated pairs of amino acids of the halophilic protein L., pdb ID: 2KAC, at b_{KCl} = 2 mol/kg.

the synergistic interaction between two acidic amino acids – is associated with the decharging step. A negative value of $\Delta\Delta G_{\text{decharging}}$ means the decharging step of a $(-)\text{X}a\text{Y}$ mutation, corresponding to Equation 4.25 where the vicinal amino acid is aspartate (or glutamate) which has a negative charge, has more positive energy (is more unfavorable) than the decharging step of a $(0)\text{X}a\text{Y}$ mutation, corresponding to Equation 4.26 where the vicinal amino acid is asparagine (or glutamine) which is neutral. A negative value of $\Delta\Delta G_{\text{decharging}}$ can arise either from i) unexpected *stabilizing* electrostatic interactions between the two vicinal acidic amino acids, "reactants" in Equation 4.25 or ii) *destabilizing* electrostatic interactions between vicinal acidic and neutral amino acids, , "reactants" in Equation 4.26, strong enough to compensate the expected electrostatic repulsion between two negative amino acids.

$$\text{D}_b + \text{D}_a \xrightarrow{\Delta G_{(-)\text{D}a\text{D}^0}} \text{D}_b + \text{D}_a^0 \tag{4.25}$$

$$\text{N}_b + \text{D}_a \xrightarrow{\Delta G_{(0)\text{D}a\text{D}^0}} \text{N}_b + \text{D}_a^0 \tag{4.26}$$

In either case, the interactions are clearly mediated by the solvent (water and ions in solution), as indicated by the strong dependence of $\Delta\Delta G$ with KCl concentration. We propose that the first possibility is the one at play here, based on indirect insight obtained from the charging step. This step indicates it is more favorable to introduce atomic charges in the final

amino acid (which is net neutral) when its neighbor has a negative charge, corresponding to Equation 4.27, than when its neighbor is neutral, corresponding to Equation 4.28. Considering that i) the electrostatic interaction between two neutral amino acids is well-described as a dipole-dipole interaction; ii) this interaction decays as $1/r^3$ where r is the distance between the dipoles; iii) the amino acids in question are at least 5 Å apart; iv) we are considering interactions at high KCl concentrations, we can assume that the electrostatic energy between 2 neutral amino acids, corresponding to the "products" in the Equation 4.28, is approximately zero; consequently, $-\Delta\Delta G_{\text{charging}}$ is a reasonable estimate of the electrostatic interaction between a neutral (N or Q) and an acidic (D or E) amino acid; this value is negative, indicating that the possibility ii) in the discussion of $\Delta\Delta G_{\text{decharging}}$ can be safely rejected, and as mentioned the possibility of unexpected *stabilizing* electrostatic interactions between the two vicinal acidic amino acids, can be accepted. These results indicate that synergistic effects between neighboring acidic amino acids, as measured by $\Delta\Delta G$, reflect this surprising electrostatically favorable interactions between vicinal acidic amino acids.

$$\mathrm{D}_b + \mathrm{N}_a^0 \xrightarrow{\Delta G_{(-)\mathrm{N}^0 a\mathrm{N}}} \mathrm{D}_b + \mathrm{N}_a \tag{4.27}$$

$$\mathrm{N}_b + \mathrm{N}_a^0 \xrightarrow{\Delta G_{(0)\mathrm{N}^0 a\mathrm{N}}} \mathrm{N}_b + \mathrm{N}_a \tag{4.28}$$

The *Ion-solvent stabilization hypothesis* claims a stabilizing effect like that which was observed in the simulations is the reason that the surface of halophilic proteins has more acidic residues, as they seem to have a synergistic effect with the high concentration of potassium ions in solution, which we show leads to the protein stabilization. I note, however, that the same model proposes this effect would lead to slower water content in the first shell of the protein, which we show in Chapter 3 this is not correct. The proponents of the ion-solvent stabilization hypothesis claim that a synergistic effect is exclusive to the folded state and is absent from the unfolded state, leading to the higher stabilization of the folded structure. In the next section, we investigate the presence of synergistic effects in the unfolded protein structures.

4.4.2 Free energies of D→N and E→Q mutations with/without vicinal acidic amino acids in unfolded protein L.

To calculate the free energy of mutation for the unfolded protein L., we used the extracted trajectory from REMD simulation, explained in the Section 4.3.2, corresponding to the temperature 298 K. This trajectory is combined of 10^5 frames with 1 ps distance between these frames making up a total of 100 ns. Theseus [103] was then used to find the frame with

the structure most similar to the average structure in this trajectory. This can be considered to be the most "typical" structure in the ensemble, shown in Figure 4.17.

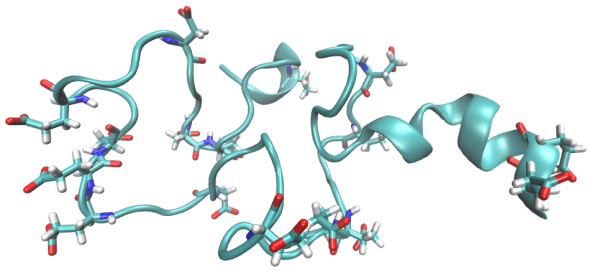
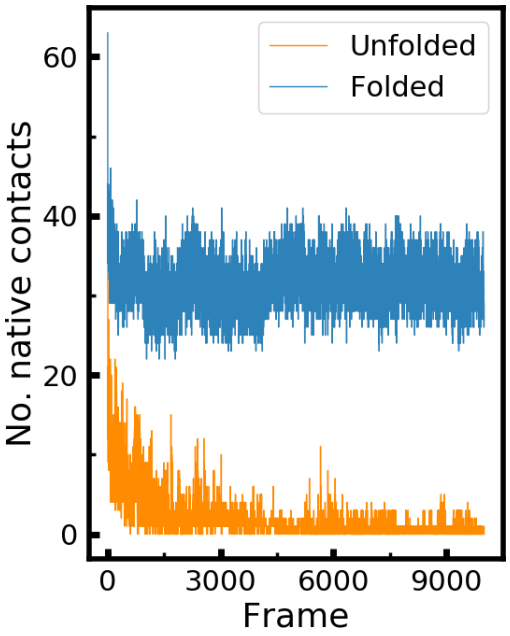


Fig. 4.17 A snapshot of REMD simulation of halophilic protein L., pdb ID:2KAC, in a denatured state, with new cartoon representation and acidic residues, blue and red branches. This is the median structure which is used for the free energy calculation.

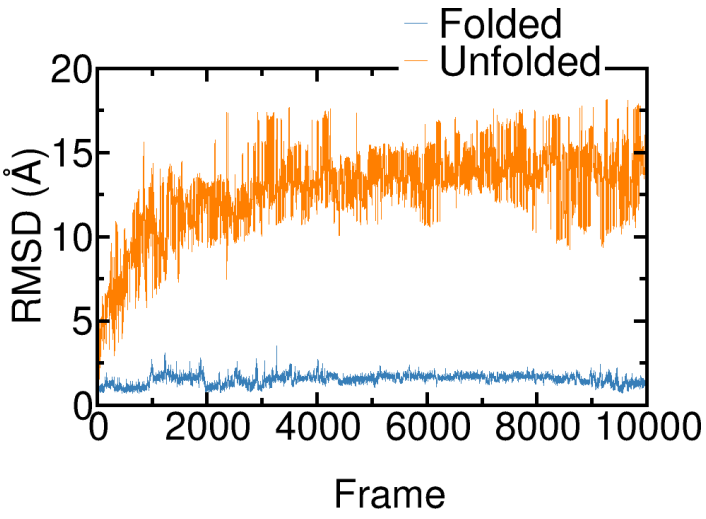
Figure 4.18a shows the number of native contacts present in the simulation of halophilic protein L., pdb ID: 2KAC, compared between the unfolded and the folded structures. This figure shows that the folded structure, from simulation in Section 3.2, has a steady and much higher number of native contacts in each frame compared to the unfolded structure, from simulation in Section 4.3.2, usually having a very low number of native contacts. Figure 4.18a and the RMSD from the Figure 4.18b show with their large frequent fluctuations in the unfolded case that indeed Replica-exchange simulation sampled many different conformations of the unfolded ensemble throughout the trajectory.

So far, we can say that a synergistic effect between acidic residues, likely involving the potassium ions in solution, is often present in folded proteins. We also need to show that such a synergistic effect does not exist in the unfolded form of these proteins to conclude that this effect stabilizes the folded structure of halophilic proteins and might be a reason for their high content in acidic residues. One of the ways to show the absence of such cooperative effect would be to show that the distribution of minimum distances between acidic residues in the unfolded state of the halophilic protein L. does not allow for such cooperative effect relative to the folded structure because the acidic amino acids are on average so distant that their electrostatic interactions are expected to be well-described by the screened electrostatics model. Figure 4.19 shows the distribution of the minimum distance between acidic residues for the folded and unfolded halophilic protein L. at T=298.15 K.

The minimum distance between acidic residues is shifted to higher values as expected, but the difference between the folded state and the unfolded ensemble is not substantial, and the distribution of the distance for the unfolded protein is almost 1 Å wider, compared to the folded protein. This small difference would suggest that synergistic effects are in principle possible also in the unfolded ensemble. However, after performing these simulations, we



(a) The number of native contacts, considered when two heavy atoms between residues spaced 4 or more residues apart in the sequence, comes into contact closer than 3.5 Å in each frame.



(b) RMSD of the backbone atoms N, C_α, C, and O.

Fig. 4.18 This figure shows the trajectory analysis comparison between the unfolded halophilic protein L., pdb ID:2KAC, from REMD simulation in Section 4.3.2, and protein L. in the folded state, from simulation of previous chapter in Section 3.2, at 298 K and b_{KCl} = 2 mol/kg. In the x-axis, for the folded structure, each frame accounts for 100 ps for a total simulation of 1 μs, and 10 ps for each frame in the unfolded case, and a total of 100 ns.

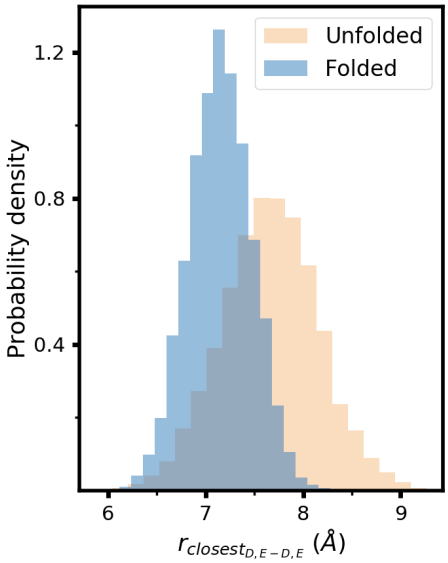


Fig. 4.19 Normalized histogram of the carbon of carboxylate minimum distance between two closest acidic amino acids (D=aspartic acid or E=glutamic acid), averaged over all acidic amino acids and all saved configurations, for the denatured protein L., from simulation in Section 4.3.2 and protein L. in the folded state, from simulation in Section 3.2, at 298 K, and b_{KCl} = 2 mol/kg.

became aware that the force field we used for this study incorrectly predicts an excessively collapsed unfolded protein ensemble [84]. The distance distribution shown in Figure 4.19 is thus not representative of the unfolded state ensemble of protein L., but at best only of the small subset of this ensemble that is a collapsed, globular state. We could not repeat the REMD simulations using a more appropriate force field because before doing so, we would have to optimize the parameters for the interaction between potassium and carboxylate for the water model used in that force field [84], and that procedure is very time consuming. The level of the collapse of the denatured ensemble of protein L., and its impact on the likelihood of the occurrence of configurations enabling synergistic effects in unfolded configurations, should be investigated in the future.

The *Ion-solvent stabilization model* claims that synergistic interactions between acidic amino acids are only possible in particular configurations of those amino acids; according to this model, these configurations occur much more frequently in the folded structure rather than in a collapsed but unfolded state. We test this possibility by calculating mutation free energies for multiple pairs of acidic amino acids of an unfolded structure of halophilic protein L. and comparing the frequency of occurrence and magnitude of synergistic effects relative to those observed for the folded protein. The free energy calculation for this median or typical structure, a denatured halophilic protein L., 2KAC, is precisely the same as what we used previously for the folded proteins, explained in Section 4.2.2. During the free energy calculation, the protein's backbone was restrained, similarly to what was done for the analogous calculations for the folded protein L.

Figure 4.20 shows the $\Delta\Delta G$ result for the denatured structure of the halophilic protein L. at 2 mol/kg. Surprisingly, almost all of the pairs of amino acids studied show synergistic effects, and the value of $\Delta\Delta$G is more negative in some cases than the synergistic cases in the folded structure. Our results do not support the claim that synergistic effects between pairs of amino acids are enabled by specific amino acid configurations that predominate in the folded protein structure; on the contrary, it seems that pairs of acidic residues assume relative distances and conformations appropriate for the acidic residues-potassium-water synergistic interaction just as easily in denatured conformations of halophilic protein L. as in the folded protein.

Every free energy mutation value calculated in this study is presented in the Tables A.2, A.3, A.4, and A.5 in the Section A.6 of SI.

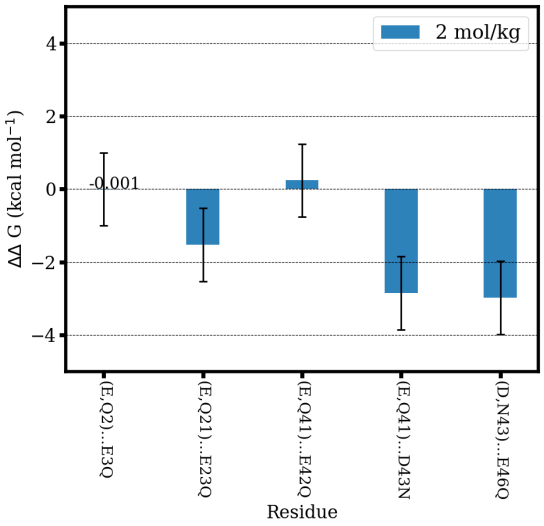


Fig. 4.20 Change in free energy of mutation ($\Delta\Delta G \pm 1.0$; Eq. 4.20) for the indicated pairs of amino acids of the denatured halophilic protein L. (pdb ID: 2KAC), calculated at the high KCl molality. The data shown in this plot is compiled in Table A.5.

4.4.3 Mechanism behind synergistic effects

So far, we have established that synergistic interactions between the vicinal acidic residues-cation-water on the surface of both folded and unfolded halophilic proteins exist. A significant result was the concentration dependence of such interactions, which points out that salt ions in solution are involved in these synergistic interactions. It has been claimed [71] that the unfolded state lacks such stabilizing synergistic interactions because acidic residues in the unfolded structure do not have the necessary preorientaion for stable interaction with cations. The lack of such synergistic interactions would per se destabilize the unfolded structure; furthermore, it would cause the preferential ion-exclusion from the surface of the unfolded structure leading to the competition between the unfolded structure and salt ions for hydration as well, which would further destabilize the unfolded structure. These two effects, by destabilizing the unfolded structure, would help the stabilization of the folded structure. However, we have observed even a more substantial synergistic effect in a denatured configuration of the halophilic protein L. This observation contradicts the claim of the lack of necessary preorientation for cation-acidic residue interaction in the unfolded structure. Our results show that the possibility of such synergistic interactions in an unfolded conformation exists, but the simulation was performed with a structure whose backbone was restrained; therefore, it cannot be directly compared with the experimental result on the issue of preferential ion-exclusion [71]. For such comparison, we need to have a trajectory of the unfolded ensemble without any restraint to quantify the solvation of such structure and the possibility of observing synergistic interactions. Obtaining this simulation result needs an appropriate force field, significant computational resource, and a change in the setup of free energy calculations. This can be the subject of a future study.

Another critical question that needs to be further investigated is the mechanism behind such a synergistic effect. More specifically, how can two neighboring acidic residues, in the presence of hydrated salts, interact in a stabilizing manner, despite having a similar charge? To gain some insight into this issue, we simulated the folded halophilic protein L., 2KAC, at a high KCl concentration of 2 mol/kg, with the same parameters as for the mutation free energy simulations, Section 4.2.2, to obtain a long, continuous, trajectory with frequently saved configurations for detailed analysis. Simulations were started from the last frame of the production simulation described in Section 3.2. Then this starting configuration was minimized and heated with the exact simulation details as for the main free energy simulations, Section 4.2.2, except that a TI simulation was not performed. Then for the production phase of the simulation, in an NVT ensemble, we simulated the system for 400 ns, again with the exact same simulation parameters as in the free energy simulations. The final trajectory contains 4×10^4 frames with 10 ps distance combining 400 ns of trajectory.

We analyzed the trajectory from the above simulation to investigate whether characteristic configurations of acidic amino acids and potassium ions differ between pairs of amino acids showing synergistic vs. interfering effects. The synergistic pairs are (E,Q23)...E21Q, (E,Q61)...D50N, (E,Q2)...E3Q, (D,N38)...E41Q; the interfering ones are: (E,Q28)...E27Q, (E,Q42)...D43N, (E,Q41)...E42Q, (E,Q28)...E32Q. The categorization between interfering and synergistic follows the $\Delta\Delta G$ values in Table A.3 from SI for protein L., 2KAC, at high KCl concentration of 2 mol/kg.

Figures 4.21, 4.22, and 4.23 compare the histogram, which is normalized, of distances between certain atoms in the proximity of different interfering and synergistic above-mentioned pairs of acidic amino acids. The corresponding $\Delta\Delta G$ values for every pair are mentioned above the figures as well for comparison. For simplicity, only the numbers corresponding to the position of neighboring residues, the pairs mentioned above, are used in the figures' legend to represent the pair of acidic amino acids studied. A correlation between the calculated structural observables with the values of $\Delta\Delta G$ cannot be discerned from these figures. Widely different distributions of the distance between the carboxylate groups within a pair, the distance between carboxylate groups and potassium, and the distance between 2 potassium ions in the hydration shell of a pair of acidic amino acids, are associated with synergistic effects of the same magnitude.

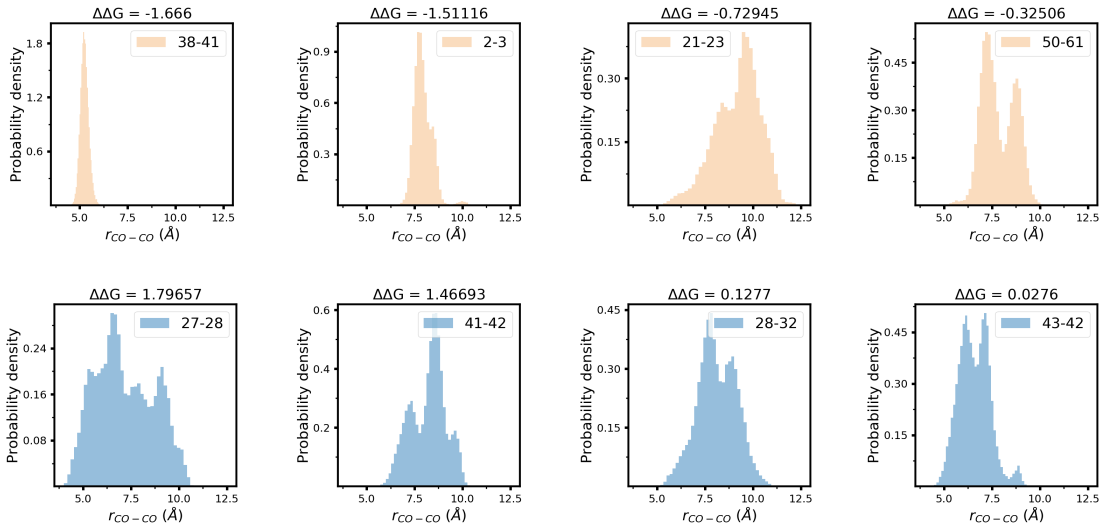


Fig. 4.21 Distance between the carbons of carboxylates of pairs of acidic amino acids as they where studied, for all the ones showing synergistic effect, soft orange color, versus when they are interfering, moderate blue.

Figure 4.24a, and 4.24b show the RDF calculated from the trajectory of the simulation as mentioned above, where it shows the distribution of potassium ions and water molecules

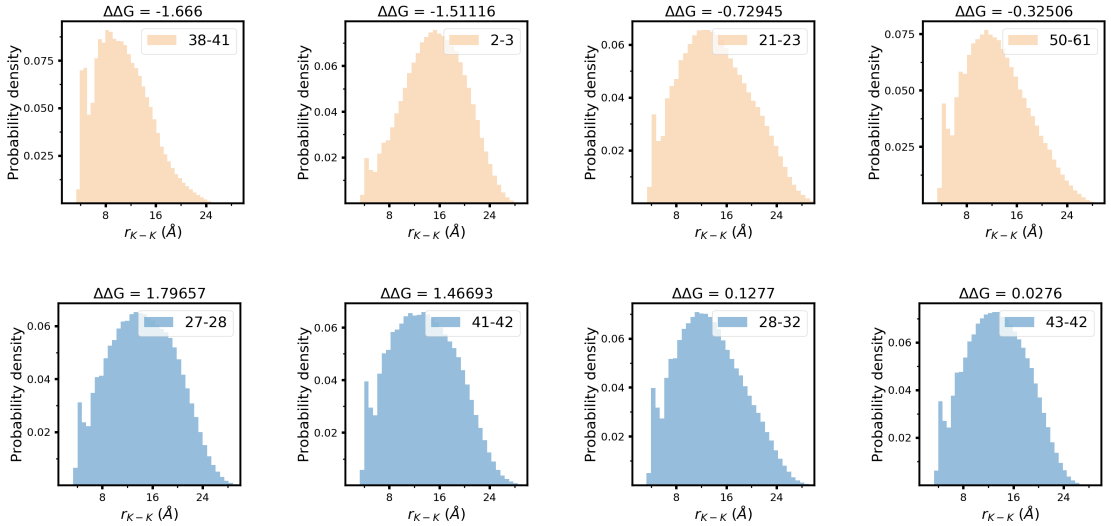


Fig. 4.22 Distance between the potassium ions within 10 Å of the oxygens atoms of carboxylates of pairs of acidic amino acids as they where studied, for all the ones showing synergistic effect, soft orange color, versus when they are interfering, moderate blue.

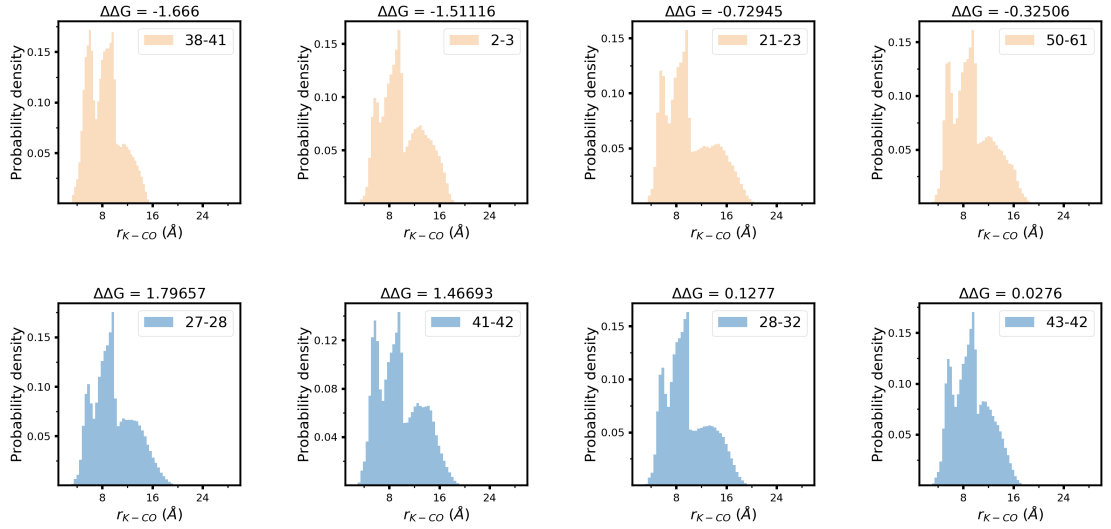
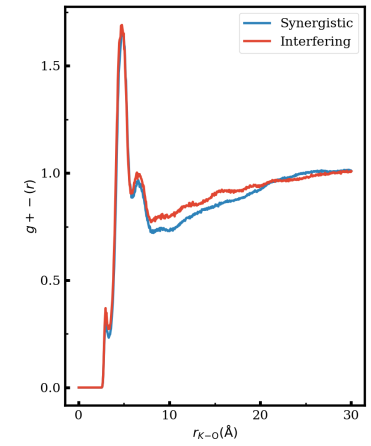


Fig. 4.23 Distance between the potassium ions and the carbon atoms of the carboxylates of both pairs of acidic residue, potassium ions within 10 Å of these carbons are only considered, as they where studied, for all the ones showing synergistic effect, soft orange color, versus when they are interfering, moderate blue.

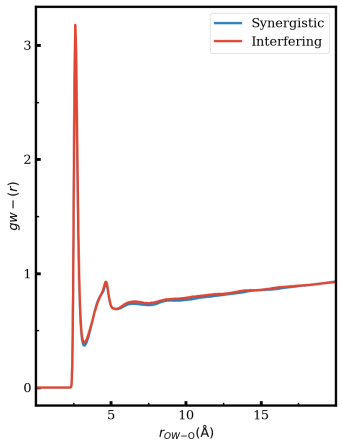
around the oxygens of the carboxylates of interfering versus synergistic residues. Figures 4.24c, and 4.24d show the histogram from raw data of distances between potassium ions-potassium ions and carboxylates-potassium ions in the proximity of different interfering and synergistic pairs mentioned above, of acidic amino acids.
One interesting observation from these figures is that, although very small, the peak of the carboxylate-potassium RDF (Figure 4.24a) is slightly higher for the interfering cases. This difference is not present in the case of water molecules. This increase in the concentration of potassium ions around the acidic residues involved in the interfering interactions happens in the solvent-separated shell of oxygens of carboxylates (> 6 Å) and continues to long distances (almost 20 Å), and then is compensated in longer distances.

Figures 4.24c, and 4.24d show the same data as in Figures 4.22 and 4.23, in aggregated form and is not normalized form. Comparison with the RDF in Figure 4.24a indeed suggests that interfering and synergistic pairs differ in their long-range interactions with potassium ions which affects long-range potassium-potassium distances. The small differences seen in Figures 4.21 to 4.24 between the interfering vs. synergistic pairs do not give direct insight into the mechanism behind synergistic effects. They allow us only to conclude that synergistic effects are not associated with particular types of ion pairs (solvent shared vs. solvent separated) or particular distances between carboxylates. The origin of the synergistic effect needs to be further investigated, which is not possible with the limited data and time that is currently available.

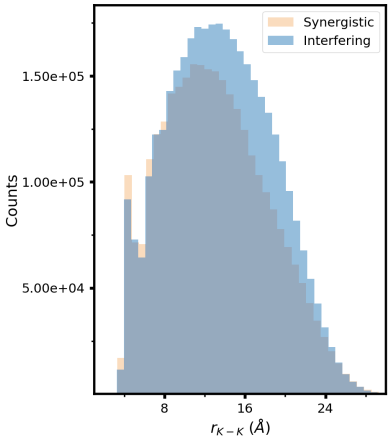
To find the mechanism, we need to formulate the possible components of synergistic interactions. We can compare the case in our study with the more studied subject of cooperativity in polyvalent biological interactions. Multiple simultaneous interactions are called polyvalent, which have different properties from their separate monovalent interactions that are qualitatively different. The average free energy of interaction between a ligand moiety and a receptor moiety in a polyvalent interaction ($\Delta G_{avg}^{poly} = \Delta G_N^{poly}/N$, N number of monovalent) can be greater than, equal to, or less than the free energy in the analogous monovalent interaction (ΔG^{mono}), and we call these classes of polyvalent interactions positively cooperative (synergistic), noncooperative (additive), or negatively cooperative (interfering), respectively. We use these definitions as given in Mammen et al. [67] to formulate the synergistic interactions using the clear definitions given. The parameter (ΔG_N^{poly}) is made up of enthalpic (ΔH_N^{poly}) and entropic (ΔS_N^{poly}) components [67]. When binding a second ligand to a receptor is more favorable (less enthalpy) than the first one, like binding four oxygen molecules to the tetrameric hemoglobin where each next oxygen binds more favorably than the previous one; this is called enthalpically enhanced bindings. The opposite of this case is enthalpically diminished binding, which happens when two polyvalent have



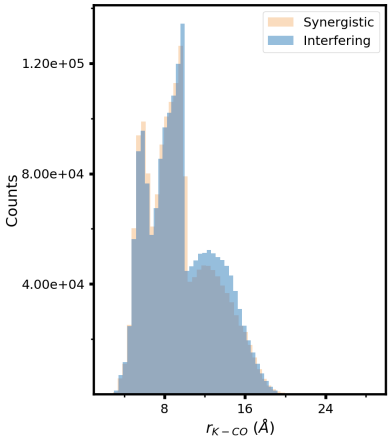
(a) potassium ions.



(b) water molecules.



(c) $K^+ \ldots K^+$ distance



(d) $K^+ \ldots CO$ distance

Fig. 4.24 Figures (a) and (b) show the RDF of potassium ions and water molecules around the oxygens of carboxylates of vicinal acidid residues, respectively, for synergistic versus interfering cases. Figure (c) shows the distance between the potassium ions and the carbon atoms of the carboxylates of both pairs of acidic residue, potassium ions within 10 Å of these carbons are only considered. Figure (d) shows the distance between the potassium ions within 10 Å of the oxygens atoms of carboxylates of pairs of acidic amino acids. For Figures (c) and (d) cases with synergistic effect are shown with soft orange color, versus when they are interfering with moderate blue.

energetically unfavorable conformation for interaction. For entropy, it is more complicated due to various components of entropy. Firstly, if we consider a bivalent ligand and receptor connected having a rigid linkage, the second of the bivalent interactions has no rotational and translational cost because the first interaction already paid the cost. However, if we consider a flexible linking group, an unfavorable entropy upon interaction due to the loss of this flexibility is compensatory with the total rotational and translational entropic cost of the bivalent interaction. Interestingly, the enthalpic and entropic parts have compensating effects on the polyvalent interaction's cooperativity: for example, in a linked ligand and receptor, the increase in the conformational flexibility increases the entropic cost upon complexation, which at the same time increases the favorable molecular conformation for binding. For the cooperative effect seen here involving neighboring acidic amino acids, a possible enthalpic enhancement could originate from stronger hydrogen bonds involving water for particular configurations of the cations and the carboxylate groups. These configurations are necessarily transient because of the dynamics of the protein, but would be more frequent for some pairs of amino acids because of their particular local environment at the surface of the protein. The effect might also have an entropic contribution, e.g., arising from the lower entropic cost of a second cation-carboxylate interaction if the cation is already interacting with another carboxylate group.

Finding the exact mechanism by which such synergistic interaction between vicinal acidic amino acids happens need more analysis which, because of time limitations, we did not perform to the extent necessary. For example, one helpful analysis could be calculating a Potential of Mean Force (PMF) between two acidic amino acids versus a PMF between one acidic amino acid and one neutral residue corresponding to the present study's design with two separate comparisons in $b_{KCl} = 0.15$ mol/kg, and one in $b_{KCl} = 2$ mol/kg to check if one observes the same difference due to the salt concentration. These PMFs can be analyzed to understand the underlying mechanism of cooperativity by comparing the changes in free energies between the two amino acids versus distance. Another interesting analysis could be calculating the local enthalpic and entropic contributions to the solvation free energies around acidic residue pairs of synergistic versus interfering cases, using, e.g., spatially resolved solvation energy for enthalpic contribution and position-based approximation [58] for the entropic contribution. Ref. [35] reviews the methods that have been developed so far to calculate such local thermodynamic properties. However, further methodological developments might be necessary as currently applying existing methods to a solution containing salt ions remains problematic.

4.5 Conclusion

It has been shown that the effect of salt concentration on halophilic proteins is mainly on the interaction between these enzymes and the solution, ion and water, rather than significant conformational changes [78]. Therefore the research on the mechanism of halophilic protein stability and function was mainly focused on the interaction of halophile-specific composition of amino acids with the medium and its effect on the stability and function. The most striking amino acid composition exclusive to halophilic enzymes is the abundance of acidic amino acids on the surface of these proteins.

In this chapter, we have investigated a well-established possibility, the *Ion-solvent stabilization model* [115]. This model claims that halophilic enzymes have evolved a specific folded structure that can co-ordinate hydrated salt ions in order to avoid competing with the high concentration of salt, natural to the halophilic cytoplasm, for hydration. Carboxylates (the termini of the side-chain of acidic amino acids), on the surface of halophilic proteins, create strong hydrogen bonds [108] with a network of water and salt ions. It is thought that this network of water+cation+carboxylate is more structured than an analogous network of water and ions would be in the bulk solvent, because the anchoring of the acidic residues in the protein may limit the number of available configurations and slow down the hydrogen bond dynamics [115]. Such interaction is claimed [115] to increase the hydration of halophilic proteins at such high salt concentration, preventing their aggregation and increasing their stabilization.

The more subtle aspect of this specific interaction is that it is proposed to be predominant in the quaternary configuration of halophilic proteins, and therefore it should only stabilize the folded structure and not the unfolded structure. NMR studies of halophilic and non-halophilic proteins also support this model and have added further support to the notion that both folded and unfolded states add to haloadaptation [71]: the measurements indicate that there are weak interactions between the acidic amino acids and the cations in solution, which stabilize (not for all carboxylates) the folded structure of halophiles. These interactions might be partially enabled by the fairly short side-chain of the acidic amino acids, which would reduce the entropic cost of their interaction with cations. Simultaneously, these interactions are absent from the unfolded state of halophiles because the authors claim, the flexible side chains of the acidic amino acids do not have the necessary preorientation to allow for their stable interaction with cations [71]. Because the cations are excluded from the hydration layer of the protein in the unfolded state, they compete with the protein for hydration, leading to a destabilization of the unfolded state, which further reinforces the stability of the folded structure. Based on the claim of the lack of necessary preorientation for cation-carboxylate interaction, and consequently the absence of synergistic interactions in the unfolded state, we

point to the following two differences between the folded and unfolded structures to test this claim. i) There are interactions between the spatially close acidic amino acids, exclusive to the folded conformation, and a network of ion and water, absent in an unfolded conformation due to the loss of such spatial vicinity. ii) Multitude of conformations with a short lifetime and flexible backbone is specific to the unfolded ensemble compared to the robust quaternary configuration of the folded protein with a comparatively rigid backbone capable of *retaining* necessary preorientation for such synergistic interactions if exists.

We have designed a study to test these possibilities, where we have chosen several pairs of acidic amino acids close to each other on the surface of halophilic proteins such as ferredoxin, protein L., and dihydrofolate reductase. Then we have mutated one of these acidic residues to a structurally similar neutral amino acid in the presence of the neighboring acidic residue and in the absence of this neighboring acidic amino acid substituted with a neutral residue. If there is a cooperation between the neighboring acidic residues that help stabilize the structure of a halophilic protein through its interaction with a network of salt and water, it can be observed in the difference of these two alternative mutation free energies.

Surprisingly, many of the pairs studied show cooperative interaction meaning that the presence of two acidic amino acids close to each other stabilized the protein compared to having an acidic residue and a neutral residue. This is surprising because the cooperative effect should be larger than the destabilizing repulsive interaction between two acidic residues to result in a stable structure, although the electrostatic repulsion in high concentration is relatively shielded. The number and magnitude of such cooperative interactions decrease in the low salt concentration since there is less shielding effect in this environment. Also, the dependence of occurrence and quantity of such synergistic interactions to the salt concentration, points to the engagement of salt ions in the solution with the acidic amino acids for these interactions. One important point is that observing some local environments on the surface of halophilic proteins show cooperative effect and some others do not, which previously has been observed as well [71], should be considered with the fact that there is a compensatory relation between the electrostatic repulsion and cooperative interaction. In other words, even for the pair of acidic amino acids for which we do not observe such cooperative effect or the quantity of cooperative interaction is low, it does not mean that there is no cooperation at all; it might be that this cooperative interaction is weaker than the electrostatic repulsion.

On the other hand, for an unfolded structure of the halophilic protein L., despite the claim [71] that there is no cooperative interaction in the unfolded conformation due to the lack of necessary preorientation for the stable cation-carboxylate interaction, we have observed cooperative interactions. Surprisingly, the magnitude and the probability of such cooperative interactions are higher in the unfolded structure than in the folded structure of

protein L. This observation rejects the possibility mentioned above, number i). Concerning the possibility number ii) on the *retaining* of such preorientation, because our simulation was performed with a structure whose backbone was restrained, we cannot directly use our data to suggest anything on this matter. This issue of *retaining* the necessary preorientation which might be related to the observation of preferential ion-exclusion from the surface of the unfolded structure suggested in the ref. [71] should be further investigated. But considering our observation of the existence of the necessary preorientation of acidic amino acids for cooperative interaction in the unfolded conformation, we can point to the possibility of the influence of these cooperative interactions on the halophilic protein folding pathway and its kinetics.

More investigation is needed to understand the underlying mechanism for such a cooperative interaction between neighboring acidic amino acids and a network of hydrated cations. Also, an extensive study of the unfolded ensemble's hydration, as well as the study of non-restrained unfolded protein's cooperativity, is required to understand the exact role of the synergistic interactions in the unfolded state, and therefore the net contribution of these cooperative interactions to the folded halophilic protein's stability. Finally, we conclude that many acidic residues on the surface of the halophilic proteins contribute to the stabilization of the folded halophilic proteins through synergistic interactions with hydrated cations and may also help stabilize some of the unfolded conformations of these proteins on their folding pathway from the unfolded state to the folded state.

Chapter 5

Outlook

Studies of halophilic proteins and other extremophiles are experimentally challenging because procedures have typically been developed for non-extremophile conditions and cannot be seamlessly used under extreme conditions. The same is true for the computational studies of halophiles because most of the force field parameters for molecular dynamics simulations are developed for the typical environment, such as low salt concentration, and cannot be assumed to describe extreme environments adequately. Such limitations caused many contradictory and partial investigations in the study of halophilic proteins where usually a minimal number of proteins – usually one – is studied, and the results and their conclusion are generalized to the whole proteome. In this study, we have tried to thoroughly understand what is the physical mechanism driving the abundance of acidic amino acids in halophilic proteins, although we have finished only a part of it. However, the foundation of the study has produced reliable parameters for all-atom, fixed-charge force fields, and relatively comprehensive analysis of currently well-established models. The force field parameters governing the dominant interactions in the environment specific to the halophilic proteins have been optimized using experimental data. Then for the main study, we have used a wide variety of halophilic proteins and their counterparts to reliably and comprehensively test the proposed models.

First, we studied the structure and dynamics of the solvation shell of the folded halophilic proteins. The first model that we have studied proposed that halophilic proteins have a more significant number of acidic amino acids at their surface to maintain protein hydration in the low water activity environment that is the halophilic cytoplasm. We found that halophilic proteins contain more significant amounts of water in their hydration shell than their mesophilic counterparts, which correlates with their higher number in acidic amino acids. However, the hydration of the halophilic protein does not change in high and low KCl concentration, pointing to the fact that higher content of acidic amino acids, although attracting more water, is not necessary for the competition with salt ions at high KCl

concentration. The potassium cations substantially integrate the solvation shell of the protein, often in the form of solvent shared and the solvent separated ion pairs with the carboxylate groups of the acidic amino acids. Our results so far thus do not support the proposed model. However, a definitive conclusion can only be reached after characterizing the structure of the solvation shell of the unfolded state of halophilic proteins and comparing it with that of the folded protein structure. Future analysis of the same type that has been conducted in Chapter 3 of this work for the unfolded state will shed light on the mechanism behind protein halotolerance.

Subsequently, we investigated another widely accepted model, that there is a cooperative stabilizing interaction between acidic residues-cations-water exclusive to the folded structure in halophilic proteins. We have shown that if such a cooperative interaction exists, it does not slow down the water and ion dynamics in the hydration layer of halophilic proteins compared to the non-halophilic counterparts, as has been proposed [98]. We investigated the existence of those cooperative interactions in both folded and unfolded states of halophilic proteins in high and low salt concentrations, using free energy calculations. We observed cooperative interactions between neighboring acidic amino acids in both states. We demonstrated that the cooperative effect between neighboring amino acids is connected to the salt solution because, for low KCl concentration, the number and magnitude of these cooperative interactions decrease. Given that cooperative interaction between acidic amino acids also exist in the unfolded protein conformations, the hypothesis that these interactions are enabled by particular amino acid configurations that are more often found in the folded conformation is not supported. Moreover, cooperative effects are not clearly associated with characteristic distances between neighboring carboxylate groups or between carboxylate groups and potassium ions; on the contrary, cooperative interactions were observed for very different distributions of these distances. Time limitations prevented us from gaining further insight into the mechanism of the cooperative interactions between acidic amino acids, but we outlined possible simulation studies that can be done to this end.

Our results are insufficient to indicate whether cooperative interactions between acidic amino acids have a net stabilizing, destabilizing or neutral effect on the protein structure. Drawing this conclusion would require knowledge of the frequency and magnitude of these interactions in the unfolded state ensemble. Quantifying cooperative effects in the unfolded state ensemble with this level of detail is particularly challenging because of our limited understanding of this ensemble and the number and time consuming nature of the simulations required for that study. Irrespective of this issue, we point out that the presence of cooperative interactions between acidic amino acids in unfolded protein conformations may impact the protein folding pathway and its kinetics.

Beyond these two models proposed to explain the large number of acidic amino acids in halophilic proteins, others have not been studied in this thesis. Firstly, could the repulsive interactions between many negatively charged acidic residues on halophilic proteins' surface be necessary to retain protein's flexibility necessary for protein function [69]? Secondly, in media with high KCl concentration, the hydrophobic effect is enhanced. One possibility mentioned in the literature is that a high content in acidic amino acids is necessary to enhance the electrostatic repulsive interaction between halophilic proteins to prevent their aggregation [114]. Tadeo et al. [96] have observed using point mutation studies that a reduction in the size of the residues on the surface of halophilic proteins correlates with their ability to remain folded at a higher salt concentration (their halophilicity), whereas changing the charge of these residues and the net charge of the protein does not correlate with halophilicity. They propose that the highly hydrophilic acidic amino acids are necessary to maintain protein solubility by decreasing its solvent accessible surface area, leading to a decrease in protein's hydrophobic effect in a highly hydrophobic environment and by increasing the protein net charge. Both these hypotheses draw on our intuitive understanding of electrostatic interactions and the hydrophobic effect in media with high ionic strength, but neither has been investigated for a diverse set of proteins. Further simulation work is necessary to elucidate the possibly multiple roles of acidic amino acids in halophilic proteins.

References

[1] Arai, S., Yonezawa, Y., Okazaki, N., Matsumoto, F., Shibazaki, C., Shimizu, R., Yamada, M., Adachi, M., Tamada, T., Kawamoto, M., Tokunaga, H., Ishibashi, M., Blaber, M., Tokunaga, M., and Kuroki, R. (2015). Structure of a highly acidic β-lactamase from the moderate halophile Chromohalobacter sp. 560 and the discovery of a Cs(+)-selective binding site. *Acta Crystallographica Section D: Biological Crystallography*, 71:541–554.

[2] Arroyo-Manez, P., Bikiel, D. E., Boechi, L., Capece, L., Lelia, S. D., Estrin, D. A., Marti, M. A., Moreno, D. M., Nadra, A. D., and A. A. Petruk (2011). Protein dynamics and ligand migration interplay as studied by computer simulation. *Biochimica et Biophysica Acta-proteins and Proteomics*, 1814(8):1054–1064.

[3] Aziz, E. F., Ottosson, N., Eisebitt, S., Eberhardt, W., Jagoda-Cwiklik, B., Vácha, R., Jungwirth, P., and Winter, B. (2008). Cation-specific interactions with carboxylate in amino acid and acetate aqueous solutions: X-ray absorption and ab initia calculations. *Journal of Physical Chemistry B*, 112(40):12567–12570.

[4] Bart, H.-J. (2001). *Reactive Extraction*, chapter Appendix A, B. Springer-Verlag.

[5] Basconi, J. E. and Shirts, M. R. (2013). Effects of temperature control algorithms on transport properties and kinetics in molecular dynamics simulations. *Journal of Chemical Theory and Computation*, 9(7):2887–2899.

[6] Bayly, C. I., Merz, K. M., Ferguson, D. M., Cornell, W. D., Fox, T., Caldwell, J. W., Kollman, P. A., Cieplak, P., Gould, I. R., and Spellmeyer, D. C. (1995). 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.

[7] Bennett, C. H. (1976). Efficient estimation of free energy differences from Monte Carlo data. *Journal of Computational Physics*, 22(2):245–268.

[8] Berendsen, H. J. C., Postma, J. P. M., van Gunsteren, W. F., Di Nola, A., and Haak, J. R. (1984). MD with coupling to an external bath. *J Chem Phys*, 81(8):3684–3690.

[9] Berendsen, H. J. C., van der Spoel, D., and van Drunen, R. (1995). GROMACS: A message-passing parallel molecular dynamics implementation. *Computer Physics Communications*, 91(1-3):43–56.

[10] Beutler, T. C., Mark, A. E., van Schaik, R. C., Gerber, P. R., and van Gunsteren, W. F. (1994). Avoiding singularities and numerical instabilities in free energy calculations based on molecular simulations. *Chemical Physics Letters*, 222(6):529–539.

[11] Bury, C. R. and Parry, G. A. (1935). The densities of aqueous solutions of potassium acetate and laurate. *J. Chem. Soc.*, 0:626–628.

[12] Calmettes, P., Eisenberg, H., and Zaccai, G. (1987). Structure of halophilic malate dehydrogenase in multimolar KCl solutions from neutron scattering and ultracentrifugation. *Biophysical Chemistry*, 26(2-3):279–290.

[13] Carvalho, A. T., Teixeira, A. F., and Ramos, M. J. (2013). Parameters for molecular dynamics simulations of iron-sulfur proteins. *Journal of Computational Chemistry*, 34(18):1540–1548.

[14] Case, D., Ben-Shalom, I., Brozell, S., Cerutti, D., Cheatham III, T., Cruzeiro, V., Darden, T., Duke, R., Ghoreishi, D., Gilson, M., Gohlke, H., Goetz, A., Greene, D., Harris, R., Homeyer, N., Huang, Y., Izadi, S., Kovalenko, A., Kurtzman, T., ..., and Kollman, P. A. (2018). *AMBER 2018*. University of California, San Francisco.

[15] Daniele, P. G., De Robertis, A., De Stefano, C., Sammartano, S., and Rigano, C. (1985). On the possibility of determining the thermodynamic parameters for the formation of weak complexes using a simple model for the dependence on ionic strength of activity coefficients: Na^+, K^+, and Ca^{2+} complexes of low molecular weight ligands in aqueous solution. *J. Chem. Soc., Dalton Trans.*, (11):2353–2361.

[16] Darden, T., York, D., and Pedersen, L. (1993). An N-log(N) method for Ewald sums in large systems. *J. Chem. Phys.*, 98(12):10089–10092.

[17] Date, M. S. and Dominy, B. N. (2013). Modeling the influence of salt on the hydrophobic effect and protein fold stability. *Communications In Computational Physics*, 13(1):90–106.

[18] Deole, R., Challacombe, J., Raiford, D. W., and Hoff, W. D. (2013). An extremely halophilic proteobacterium combines a highly acidic proteome with a low cytoplasmic potassium content. *Journal of Biological Chemistry*, 288(1):581–588.

[19] Diddens, D., Lesch, V., Heuer, A., and Smiatek, J. (2017). Aqueous ionic liquids and their influence on peptide conformations: Denaturation and dehydration mechanisms. *Phys. Chem. Chem. Phys.*, 19(31):20430–20440.

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

[21] Essmann, U., Perera, L., Berkowitz, M. L., Darden, T., Lee, H., and Pedersen, L. G. (1995). A smooth particle mesh Ewald method. *J. Chem. Phys.*, 103(19):8577–8593.

[22] Fields, P. A. (2001). Review: Protein function at thermal extremes: Balancing stability and flexibility. *Comp. Biochem. Physiol. A Mol. Integr. Physiol.*, 129(2-3):417–431.

[23] Fournier, P., Gout, R., and Oelkers, E. H. (2005). A Raman spectrographic and potentiometric study of aqueous lithium and potassium acetate complexation at temperatures from 20 to 200 °c. *Journal of Solution Chemistry*, 34(8):881–898.

[24] Frolow, F., Harel, M., Sussman, J. L., Mevarech, M., and Shoham, M. (1996). Insights into protein adaptation to a saturated salt environment from the crystal structure of a halophilic 2Fe-2S ferredoxin. *Nature Structural Biology*, 3(5):452–458.

[25] Fukushima, T., Mizuki, T., Echigo, A., Inoue, A., and Usami, R. (2005). Organic solvent tolerance of halophilic alpha-amylase from a Haloarchaeon, Haloarcula sp. strain S-1. *Extremophiles*, 9(1):85–89.

[26] Fuoss, R. M. (1958). Ionic Association. III. The Equilibrium between Ion Pairs and Free Ions. *Journal of the American Chemical Society*, 80(19):5059–5061.

[27] Fyta, M. and Netz, R. R. (2012). Ionic force field optimization based on single-ion and ion-pair solvation properties: Going beyond standard mixing rules. *Journal of Chemical Physics*, 136(12):124103–124111.

[28] Garton, M., Corbi-Verge, C., Hu, Y., Nim, S., Tarasova, N., Sherborne, B., and Kim, P. M. (2019). Rapid and accurate structure-based therapeutic peptide design using GPU accelerated thermodynamic integration. *Proteins: Structure, Function and Bioinformatics*, 87(3):236–244.

[29] Gee, M. B., Cox, N. R., Jiao, Y., Bentenitis, N., Weerasinghe, S., and Smith, P. E. (2011). A kirkwood-buff derived force field for aqueous alkali halides. *J. Chem. Theory Comput.*, 7(5):1369–1380.

[30] Graziano, G. and Merlino, A. (2014). Molecular bases of protein halotolerance. *Biochimica et Biophysica Acta-proteins and Proteomics*, 1844(4):850–858.

[31] Gunde-Cimerman, N., Zalar, P., De Hoog, S., and Plemenitaš, A. (2000). Hypersaline waters in salterns - Natural ecological niches for halophilic black yeasts. *FEMS Microbiology Ecology*, 32(3):235–240.

[32] Guo, B., Kao, S., McDonald, H., Asanov, A., Combs, L., and William Wilson, W. (1999). Correlation of second virial coefficients and solubilities useful in protein crystal growth. *J. Cryst. Growth*, 196(2):424 – 433.

[33] Hess, B., Bekker, H., Johannes, H. J. C. B., and Fraaije, G. E. M. (1997). LINCS: A linear constraint solver for molecular simulations. *J. Comput. Chem.*, 18(12):1463–1472.

[34] Hess, B. and van der Vegt, N. F. A. (2009). Cation specific binding with protein surface charges. *Proceedings of the National Academy of Sciences*, 106(32):13296–13300.

[35] Heyden, M. (2019). Disassembling solvation free energies into local contributions—Toward a microscopic understanding of solvation processes. *WIREs Comput. Mol. Sci.*, 9(2):e1390.

[36] Hill, T. L. (1987). *Statistical Mechanics: principles and selected applications*. Dover.

[37] Hukushima, K. and Nemoto, K. (1996). Exchange Monte Carlo Method and Application to Spin Glass Simulations. *Journal of the Physical Society of Japan*, 65(6):1604–1608.

[38] Humphrey, W., Dalke, A., and Schulten, K. (1996). VMD: Visual molecular dynamics. *Journal of Molecular Graphics*, 14(1):33–38.

[39] Hünenberger, P. H. and McCammon, J. A. (1999). Ewald artifacts in computer simulations of ionic solvation and ion-ion interaction: A continuum electrostatics study. *Journal of Chemical Physics*, 110(4):1856–1872.

[40] Jasnin, M., Stadler, A., Tehei, M., and Zaccai, G. (2010). Specific cellular water dynamics observed in vivo by neutron scattering and NMR. *Phys. Chem. Chem. Phys.*, 12(35):10154–10160.

[41] Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W., and Klein, M. L. (1983). Comparison of simple potential functions for simulating liquid water. *J. chem. Phys.*, 79(2):926.

[42] Joung, I. S. and Cheatham, T. E. (2008). Determination of alkali and halide monovalent ion parameters for use in explicitly solvated biomolecular simulations. *Journal of Physical Chemistry B*, 112(30):9020–9041.

[43] Joung, S. and Cheatham, T. E. (2009). Molecular dynamics simulations of the dynamic and energetic properties of alkali and halide ions using water-model-specific ion parameters. *Journal of Physical Chemistry B*, 113(40):13279–13290.

[44] Juers, D. H. and Matthews, B. W. (2001). Reversible lattice repacking illustrates the temperature dependence of macromolecular interactions. *Journal of Molecular Biology*, 311(4):851–862.

[45] Karan, R., Capes, M. D., and DasSarma, S. (2012). Function and biotechnology of extremophilic enzymes in low water activity. *Aquat. Biosyst.*, 8(1):4.

[46] Kashefolgheta, S. and Vila Verde, A. (2017). Developing force fields when experimental data is sparse: AMBER/GAFF-compatible parameters for inorganic and alkyl oxoanions. *Phys. Chem. Chem. Phys.*, 19(31):20593–20607.

[47] Kastenholz, M. A. and Hünenberger, P. H. (2004). Influence of Artificial Periodicity and Ionic Strength in Molecular Dynamics Simulations of Charged Biomolecules Employing Lattice-Sum Methods. *Journal of Physical Chemistry B*, 108(2):774–788.

[48] Kastenholz, M. A. and Hünenberger, P. H. (2006). Computation of methodology-independent ionic solvation free energies from molecular simulations. II. the hydration free energy of the sodium cation. *Journal of Chemical Physics*, 124(22).

[49] Kennedy, S. P., Ng, W. V., Salzberg, S. L., Hood, L., and DasSarma, S. (2001). Understanding the adaptation of Halobacterium species NRC-1 to its extreme environment through computational analysis of its genome sequence. *Genome Research*, 11(10):1641–1650.

[50] Kirkwood, J. G. (1935). Statistical mechanics of fluid mixtures. *The Journal of Chemical Physics*, 3(5):300–313.

[51] Kirkwood, J. G. (1951). The statistical mechanical theory of solutions. I. *The Journal of Chemical Physics*, 19(6):774–777.

[52] Klasczyk, B. and Knecht, V. (2010). Kirkwood-Buff derived force field for alkali chlorides in simple point charge water. *Journal of Chemical Physics*, 132(2):24109.

[53] Klimovich, P. V., Shirts, M. R., and Mobley, D. L. (2015). Guidelines for the analysis of free energy calculations HHS Public Access. *J Comput Aided Mol Des*, 29(5):397–411.

[54] Krüger, P., Schnell, S. K., Bedeaux, D., Kjelstrup, S., Vlugt, T. J. H., and Simon, J.-M. (2013). Kirkwood-buff integrals for finite volumes. *J. Phys. Chem. Lett.*, 4(2):235–238.

[55] Kuntz Jr., I. D. (1971). Hydration of macromolecules. IV. Polypeptide conformation in frozen solutions. *Journal of the American Chemical Society*, 93(2):516–518.

[56] Kusalik, P. G. and Patey, G. N. (1987). The thermodynamic properties of electrolyte solutions: Some formal results. *J. Chem. Phys.*, 86(9):5110–5116.

[57] Lanyi, J. K. (1974). Salt-dependent properties of proteins from extremely halophilic bacteria. *Bacteriological Reviews*, 38(3):272–290.

[58] Lazaridis, T. (1998). Inhomogeneous fluid approach to solvation thermodynamics. 1. Theory. *Journal of Physical Chemistry B*, 102(18):3531–3541.

[59] Lebowitz, J. L. and Percus, J. K. (1961). Long-range correlations in a closed system with applications to nonuniform fluids. *Phys. Rev.*, 122(6):1675–1691.

[60] Lee, S. B. and Kim, K.-J. (1995). Effect of water activity on enzyme hydration and enzyme reaction rate in organic solvents. *J. Ferment. Bioeng.*, 79(5):473 – 478.

[61] Li, P., Roberts, B. P., Chakravorty, D. K., and Merz, K. M. (2013). Rational design of particle mesh ewald compatible lennard-jones parameters for +2 metal cations in explicit solvent. *Journal of Chemical Theory and Computation*, 9(6):2733–2748.

[62] Lin, B. and Pettitt, B. M. (2011). Note: On the universality of proximal radial distribution functions of proteins. *J. Chem. Phys.*, 134(10):106101–.

[63] Madern, D., Ebel, C., and Zaccai, G. (2000). Halophilic adaptation of enzymes. *Extremophiles*, 4(2):91–98.

[64] Madern, D. and Zaccai, G. (2004). Molecular adaptation: The malate dehydrogenase from the extreme halophilic bacterium Salinibacter ruber behaves like a non-halophilic protein. *Biochimie*, 86(4-5):295–303.

[65] Maier, J. A., Martinez, C., Kasavajhala, K., Wickstrom, L., Hauser, K. E., and Simmerling, C. (2015). ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *J. Chem. Theory Comput.*, 11(8):3696–37132.

[66] Makarov, V. A., Andrews, B. K., and Pettitt, B. M. (1998). Reconstructing the protein-water interface. *Biopolymers*, 45(7):469–478.

[67] Mammen, M., Choi, S. K., and Whitesides, G. M. (1998). Polyvalent interactions in biological systems: Implications for design and use of multivalent ligands and inhibitors. *Angewandte Chemie - International Edition*, 37(20):2754–2794.

[68] Mark, P. and Nilsson, L. (2001). Structure and dynamics of the tip3p, spc, and spc/e water models at 298 k. *J. Phys. Chem. A*, 105(43):9954–9960. J. Phys. Chem. A.

[69] Mevarech, M., Frolow, F., and Gloss, L. M. (2000). Halophilic enzymes: Proteins with a grain of salt. *Biophysical Chemistry*, 86(2-3):155–164.

[70] O'Donoghue, P. and Luthey-Schulten, Z. (2003). On the Evolution of Structure in Aminoacyl-tRNA Synthetases. *Microbiology and Molecular Biology Reviews*, 67(4):550–573.

[71] Ortega, G., Diercks, T., and Millet, O. (2015). Halophilic Protein Adaptation Results from Synergistic Residue-Ion Interactions in the Folded and Unfolded States. *Chemistry and Biology*, 22(12):1597–1607.

[72] Paliwal, H. and Shirts, M. R. (2011). A benchmark test set for alchemical free energy transformations and its use to quantify error in common free energy methods. *Journal of Chemical Theory and Computation*, 7(12):4115–4134.

[73] Patriksson, A. and Van Der Spoel, D. (2008). A temperature predictor for parallel tempering simulations. *Phys. Chem. Chem. Phys.*, 10(15):2073–2077.

[74] Paul, S., Bag, S. K., Das, S., Harvill, E. T., and Dutta, C. (2008). Molecular signature of hypersaline adaptation: insights from genome and proteome composition of halophilic prokaryotes. *Genome Biology*, 9(4):R70.

[75] Peters, M. B., Yang, Y., Wang, B., Füsti-Molnár, L., Weaver, M. N., and Merz, K. M. (2010). Structural survey of zinc-containing proteins and development of the zinc AMBER force field (ZAFF). *Journal of Chemical Theory and Computation*, 6(9):2935–2947.

[76] Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C., and Ferrin, T. E. (2004). UCSF Chimera - A visualization system for exploratory research and analysis. *Journal of Computational Chemistry*, 25(13):1605–1612.

[77] Pitzer, K. S. (1991). *Activity Coefficients in Electrolyte Solutions*. CRC Press, London, second edition.

[78] Pundak, S. and Eisenberg, H. (1981). Structure and Activity of Malate Dehydrogenase from the Extreme Halophilic Bacteria of the Dead Sea. *European Journal of Biochemistry*, 470:463–470.

[79] Qvist, J., Ortega, G., Tadeo, X., Millet, O., and Halle, B. (2012). Hydration dynamics of a halophilic protein in folded and unfolded states. *Journal of Physical Chemistry B*, 116(10):3436–3444.

[80] Ramshaw, J. D. (1980). Functional derivative relations for a finite non-uniform molecular fluid in the canonical ensemble. *Mol. Phys.*, 41(1):219–227.

[81] Roberts, E., Eargle, J., Wright, D., and Luthey-Schulten, Z. (2006). MultiSeq: Unifying sequence and structure data for evolutionary analysis. *BMC Bioinformatics*, 7:382.

[82] Robinson, R. A. and Stokes, R. H. (1949). Tables of osmotic and activity coefficients of electrolytes in aqueous solution at 25 °c. *Trans. Faraday Soc.*, 45(7):612–624.

[83] Robinson, R. A. and Stokes, R. H. (2002). *Electrolyte Solutions*. Dover Publications, INC., Mineola, New York, second edition.

[84] Robustelli, P., Piana, S., and Shaw, D. E. (2018). Developing a molecular dynamics force field for both folded and disordered protein states. *Proceedings of the National Academy of Sciences*, 115(21):E4758–E4766.

[85] Rocklin, G. J., Mobley, D. L., Dill, K. A., and Hünenberger, P. H. (2013). Calculating the binding free energies of charged species based on explicit-solvent simulations employing lattice-sum methods: An accurate correction scheme for electrostatic finite-size effects. *Journal of Chemical Physics*, 139(18).

[86] Russell, R. B. and Barton, G. J. (1992). Multiple protein sequence alignment from tertiary structure comparison: Assignment of global and residue confidence levels. *Proteins: Structure, Function, and Bioinformatics*, 14(2):309–323.

[87] Ryckaert, J.-P., Ciccotti, G., and Berendsen, H. J. (1977). Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *Journal of Computational Physics*, 23(3):327–341.

[88] Shirts, M. R. and Chodera, J. D. (2008). Statistically optimal analysis of samples from multiple equilibrium states. *Journal of Chemical Physics*, 129(12).

[89] Siglioccolo, A., Paiardini, A., Piscitelli, M., and Pascarella, S. (2011). Structural adaptation of extreme halophilic proteins through decrease of conserved hydrophobic contact surface. *BMC Structural Biology*, 11:50.

[90] Simonson, T. (1993). Free energy of particle insertion an exact analysis of the origin singularity for simple liquids. *Molecular Physics*, 80(2):441–447.

[91] Smiatek, J. (2017). Aqueous ionic liquids and their effects on protein structures: An overview on recent theoretical and experimental results. *Journal of Physics Condensed Matter*, 29(23).

[92] Stevens, M. J. and Rempe, S. L. (2016). Ion-Specific Effects in Carboxylate Binding Sites. *Journal of Physical Chemistry B*, 120(49):12519–12530.

[93] Sugita, Y. and Okamoto, Y. (1999). Replica-exchange molecular dynamics method for protein folding. *Chemical Physics Letters*, 314(1-2):141–151.

[94] Sun, Y. and Petersen, P. B. (2017). Solvation shell structure of small molecules and proteins by IR-MCR spectroscopy. *J. Phys. Chem. Lett.*, 8(3):611–614.

[95] Swendsen, R. H. and Wang, J. S. (1986). Replica Monte Carlo simulation of spin-glasses. *Physical Review Letters*, 57(21):2607–2609.

[96] Tadeo, X., López-Méndez, B., Trigueros, T., Laín, A., Castaño, D., and Millet, O. (2009). Structural basis for the aminoacid composition of proteins from halophilic archea. *PLoS Biology*, 7(12):e1000257.

[97] Tang, F., Ohto, T., Hasegawa, T., Xie, W. J., Xu, L., Bonn, M., and Nagata, Y. (2018). Definition of Free O-H Groups of Water at the Air-Water Interface. *Journal of Chemical Theory and Computation*, 14(1):357–364.

[98] Tehei, M., Franzetti, B., Wood, K., Gabel, F., Fabiani, E., Jasnin, M., Zamponi, M., Oesterhelt, D., Zaccai, G., Ginsburg, M., and Ginsburg, B. Z. (2007). Neutron scattering reveals extremely slow cell water in a Dead Sea organism. *Proceedings of the National Academy of Sciences*, 104(3):766–771.

[99] Tehei, M., Madern, D., Pfister, C., and Zaccai, G. (2002). Fast dynamics of halophilic malate dehydrogenase and BSA measured by neutron scattering under various solvent conditions influencing protein stability. *Proceedings of the National Academy of Sciences*, 98(25):14356–14361.

[100] Teilum, K., Olsen, J. G., and Kragelund, B. B. (2011). Protein stability, flexibility and function. *Biochimica et Biophysica Acta - Proteins and Proteomics*, 1814(8):969–976.

[101] Tesi, M. C., Janse Van Rensburg, E. J., Orlandini, E., and Whittington, S. G. (1996). Monte Carlo study of the interacting self-avoiding walk model in three dimensions. *Journal of Statistical Physics*, 82(1-2):155–181.

[102] Tessier, P. M. and Lenhoff, A. M. (2003). Measurements of protein self-association as a guide to crystallization. *Curr. Opin. Biotechnol.*, 14(5):512 – 516.

[103] Theobald, D. L. and Wuttke, D. S. (2008). Accurate structural correlations from maximum likelihood superpositions. *PLoS Comput Biol*, 4(2):e43.

[104] Van Der Spoel, D., Lindahl, E., Hess, B., Groenhof, G., Mark, A. E., and Berendsen, H. J. (2005). GROMACS: Fast, flexible, and free. *Journal of Computational Chemistry*, 26(16):1701–1718.

[105] Van Der Wielen, P. W., Bolhuis, H., Borin, S., Daffonchio, D., Corselli, C., Giuliano, L., D'Auria, G., De Lange, G. J., Huebner, A., Varnavas, S. P., Thomson, J., Tamburini, C., Marty, D., McGenity, T. J., and Timmis, K. N. (2005). The enigma of prokaryotic life in deep hypersaline anoxic basins. *Science*, 307(5706):121–123.

[106] van Gunsteren, W. F. and Berendesen, J. C. (1988). A Leap-frog Algorithm for Stochastic Dynamics. *Mol. simul.*, 1(3):173–185.

[107] Vila Verde, A., Santer, M., and Lipowsky, R. (2016). Solvent-shared pairs of densely charged ions induce intense but short-range supra-additive slowdown of water rotation. *Phys. Chem. Chem. Phys.*, 18(3):1918–1930.

[108] Vinogradov, S. N. and Linnell, R. H. (1971). *Hydrogen bonding*. Van Nostrand Reinhold, New York.

[109] Vrbka, L., Vondrášek, J., Jagoda-Cwiklik, B., Vácha, R., and Jungwirth, P. (2006). Quantification and rationalization of the higher affinity of sodium over potassium to protein surfaces. *Proceedings of the National Academy of Sciences*, 103(42):15440–15444.

[110] Wang, J., Wolf, R. M., Caldwell, J. W., Kollman, P. A., and Case, D. A. (2004). Development and testing of a general amber force field. *J. Comput. Chem.*, 25(9):1157–1174.

[111] Wang, P. H., Best, R. B., and Blumberger, J. (2011). Multiscale simulation reveals multiple pathways for H_2 and O_2 transport in a [NiFe]-hydrogenase. *Journal of the American Chemical Society*, 133(10):3548–3556.

[112] Warden, A. C., Williams, M., Peat, T. S., Seabrook, S. A., Newman, J., Dojchinov, G., and Haritos, V. S. (2015). Rational engineering of a mesohalophilic carbonic anhydrase to an extreme halotolerant biocatalyst. *Nature Communications*, 6:1–10.

[113] Werber, M. M. and Mevarech, M. (1978). Purification and characterization of a highly acidic 2Fe-ferredoxin from Halobacterium of the Dead Sea. *Archives of Biochemistry and Biophysics*, 187(2):447–456.

[114] Xiao, L. and Honig, B. (1999). Electrostatic contributions to the stability of hyperthermophilic proteins. *Journal of Molecular Biology*, 289(5):1435–1444.

[115] Zaccai, G., Cendrin, F., Haik, Y., Borochov, N., and Eisenberg, H. (1989). Stabilization of halophilic malate dehydrogenase. *Journal of Molecular Biology*, 208(3):491–500.

[116] Zaks, A. and Klibanov, A. M. (1988). The effect of water on enzyme action in organic media. *The Journal of biological chemistry*, 263(17):8017–8021.

[117] Zhang, H., Jiang, Y., Cui, Z., and Yin, C. (2018a). Force field benchmark of amino acids. 2. partition coefficients between water and organic solvents. *J. Chem. Inf. Model.*, 58(8):1669–1681.

[118] Zhang, H. Y., Yin, C. H., Jiang, Y., and van der Spoel, D. (2018b). Force field benchmark of amino acids: I. hydration and diffusion in different water models. *J. Chem. Inf. Model.*, 58(5):1037–1052.

[119] Zwanzig, R. W. (1954). High-Temperature Equation of State by a Perturbation Method. I. Nonpolar Gases. *Journal of Chemical Physics*, 22(8):1420–1426.

Appendix A

Supporting information

Table A.1 Distance between K^+ and the indicated carboxylate oxygens of acidic amino acids in an halophilic ferredoxin (pdb ID 1DOI), from crystallography and from simulation. Related to Fig. 2.3 of the main text.

		$f_{R_{\min,K^+O}}$					
		1	1.04	1.06	1.08	1.09	1.1
Prot. site[a]	$r^b_{cryst.}$ (Å)	r^c_{sim} (Å)					
12ASP_OD1	3.01	2.678	2.833	2.859	2.942	2.993	2.846
81ASP_OD2	2.97	2.671	2.832	2.848	2.894	2.822	2.75
107ASP_OD1	3.08	2.68	2.834	2.893	2.946	2.985	2.976
109ASP_OD1	3.12	2.688	2.853	2.888	2.887	2.834	2.813
110GLU_OE2	2.8	2.674	2.807	2.911	2.836	2.808	2.832

[a] The protein site is identified by the residue number, residue name and oxygen name.

[b] $r_{cryst.}$: distances reported in ref. 24 for the 1DOI crystal structure.

[c] $r_{sim.}$: position of the first peak of the RDFs calculated using molecular dynamics simulations of the same protein at room temperature, solvated in 1 mol/dm^3 KCl and with constrained backbone, using the indicated parameter values for the interaction between K^+ and the oxygen carboxylates.

A.1 Water-protein hydrogen bonds per residue type

Fig. A.1 shows the average number of hydrogen bonds donated by water molecules to the protein, averaged separately for different types of residues. Acidic residues accept far more hydrogen bonds than any other residue type. The number of hydrogen bonds accepted by any given residue depends weakly (if at all) on protein identity and on the concentration of KCl in the bulk.

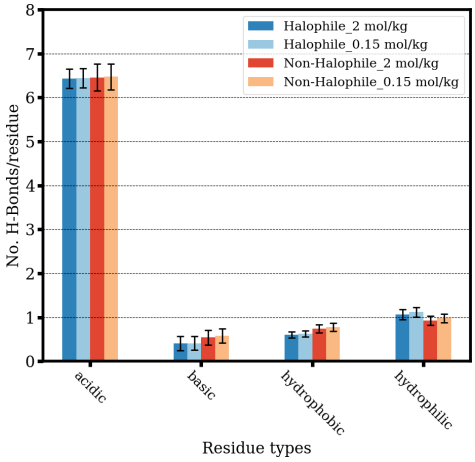
A.2 Cumulative number of potassium ions as a function of distance to the protein surface

Fig. A.2 shows the number of potassium ions within any given distance to the protein surface. The largest differences between halophilic proteins and their mesophilic counterpart are seen for the ferredoxin (Fig. A.2a) and the beta-lactamase (Fig. A.2d). These large differences reflect not only their different amino acid composition, but also their different size (in the case of ferredoxin) and shape (beta-lactamase)

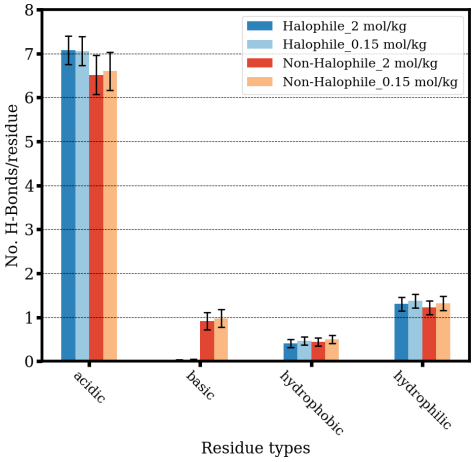
A.3 Mean square displacement of water and of K^+

SI Figs. A.3 and A.4 show the MSD of water or K^+ ions belonging to the first hydration shell of each ferredoxin protein, simulated at high and low KCl concentration, calculated as described in the main text. The MSD of water and of K^+ calculated for the other proteins (results not shown) are qualitatively similar to these. For $t < 10$ ps, both water and K^+ are clearly in the subdiffusive regime. The diffusive regime is only seen for $t = [10, 50]$ ps; beyond this time, the particles move supradiffusively for a short time interval. For $t > 1$ ns the MSD saturates, reflecting the fact that this MSD is calculated using coordinates wrapped back to the main simulation cell.

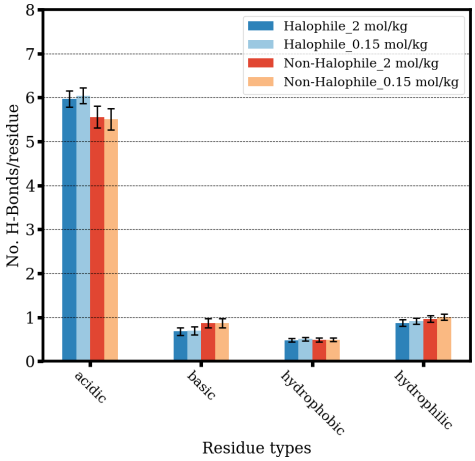
In Figs. A.5 and A.6 we show the diffusion coefficients, D, of water and of potassium for all proteins at $b_{\mathrm{KCl}} = 0.15$ mol/kg. Details of the calculation of D are given in the main text.



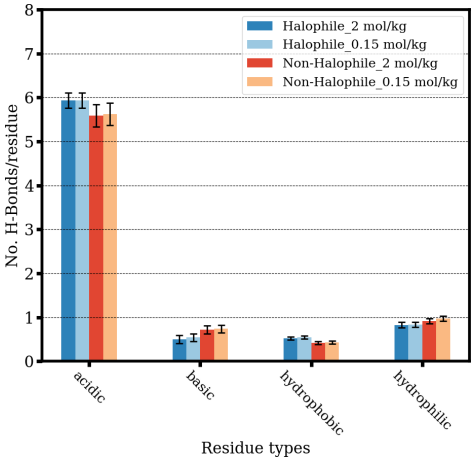
(a) 1DOI_1FRD protein pair



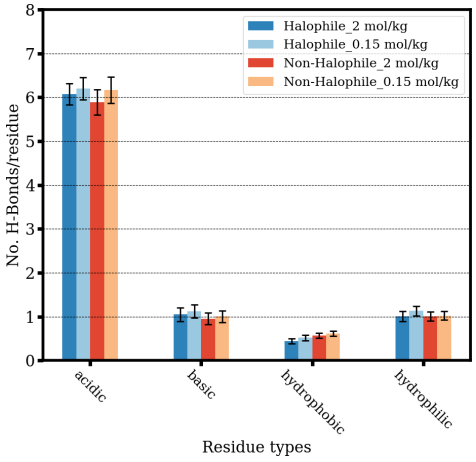
(b) 2KAC_1HZ6 protein pair



(c) 4CNX_1V9E protein pair

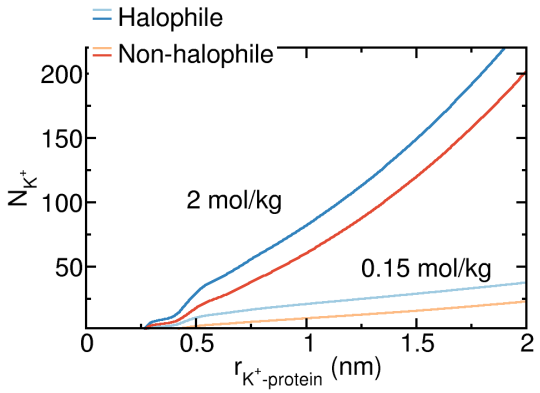


(d) 3WRT_1ZKJ protein pair

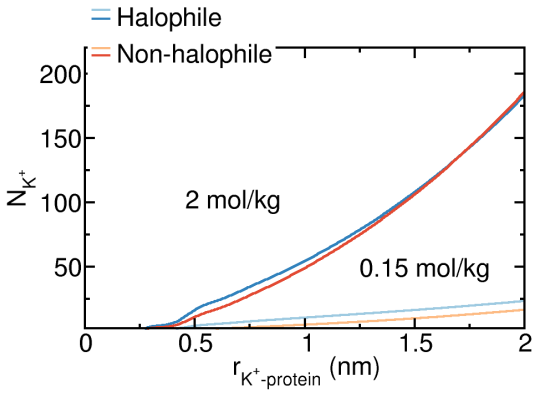


(e) 2ITH_2L28 protein pair

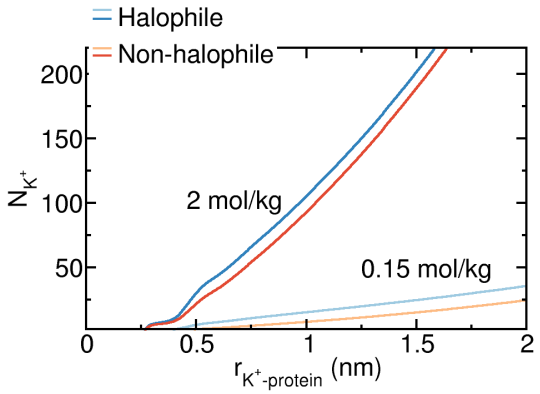
Fig. A.1 Number of water-protein hydrogen-bonds per residue, averaged over the indicated amino acid types (acidic, basic, hydrophobic, and hydrophilic (=polar, non-charged), for the indicated halophilic-mesophilic proteins, identified by their pdb IDs.



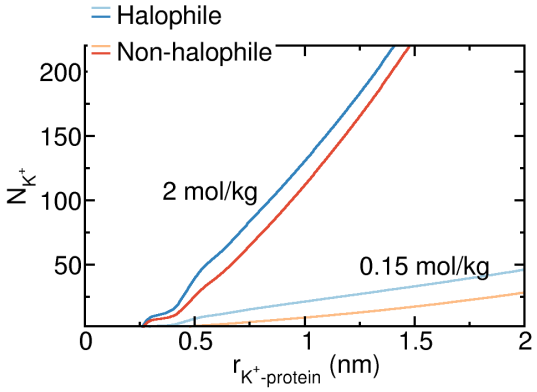
(a) 1DOI_1FRD protein pair



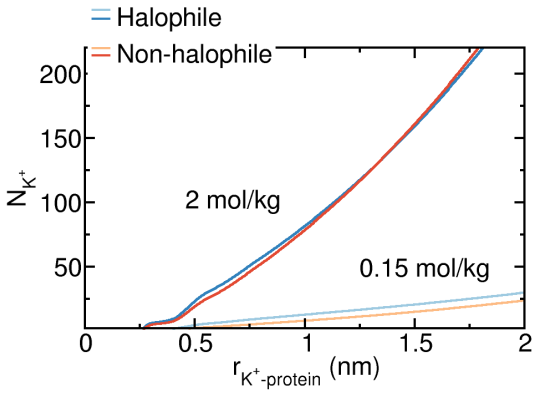
(b) 2KAC_1HZ6 protein pair



(c) 4CNX_1V9E protein pair

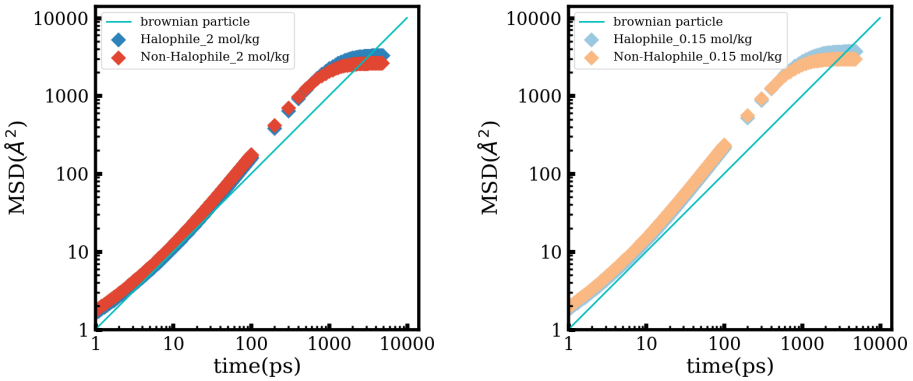


(d) 3WRT_1ZKJ protein pair



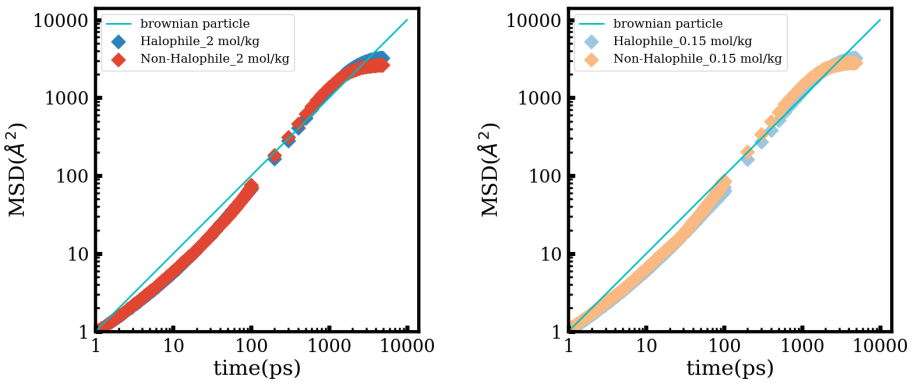
(e) 2ITH_2L28 protein pair

Fig. A.2 Cumulative number of potassium ions as a function of the distance to the heavy atoms defining the protein surface, from simulations at the indicated concentration of KCl.



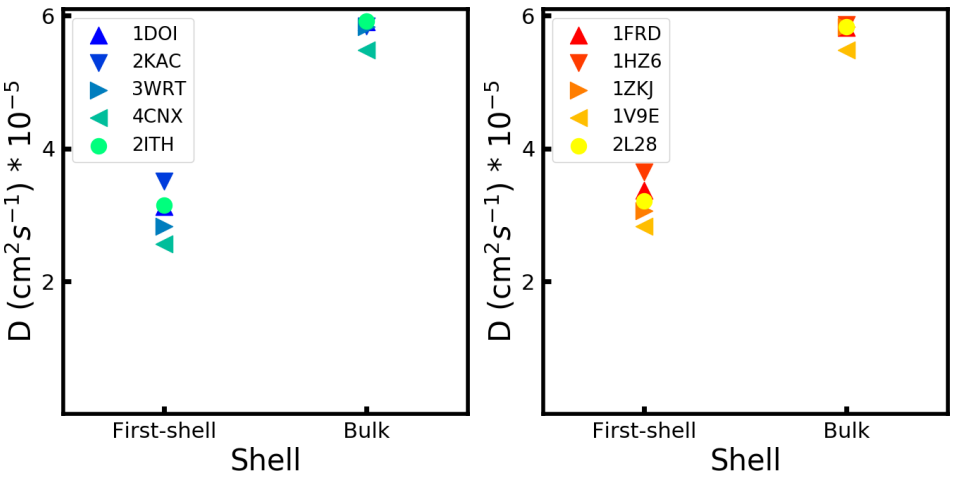
(a) $b_{\mathrm{KCl}} = 2$ mol/kg. (b) $b_{\mathrm{KCl}} = 0.15$ mol/kg.

Fig. A.3 MSD of the subpopulation of water molecules that belong to the first hydration shell of the halophilic or non-halophilic ferredoxin at $t = 0$, for the indicated salt concentration. The light blue line illustrates the diffusive limit of a Brownian particle with $D = 1/6$ Å^2/ps.



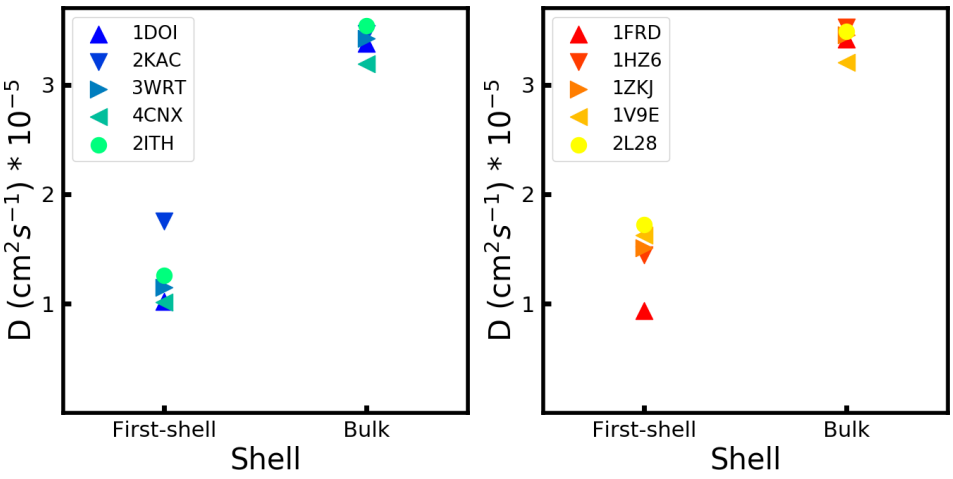
(a) $b_{\mathrm{KCl}} = 2$ mol/kg. (b) $b_{\mathrm{KCl}} = 0.15$ mol/kg.

Fig. A.4 MSD of the subpopulation of K^+ ions that belong to the first hydration shell of the halophilic or non-halophilic ferredoxin at $t = 0$, for the indicated salt concentrations. The light blue line illustrates the diffusive limit of a Brownian particle with $D = 1/6$ Å^2/ps.



(a) Halophiles (b) Non-Halophiles

Fig. A.5 Diffusion coefficients of: (First-shell) water molecules that belong to the first hydration shell of the indicated proteins at $t = 0$, simulated at $b_{\mathrm{KCl}} = 0.15$ mol/kg; (Bulk) all water molecules in the same simulation.



(a) Halophiles (b) Non-Halophiles

Fig. A.6 Diffusion coefficients of: (First-shell) potassium ions that belong to the first solvation shell of the indicated proteins at $t = 0$, simulated at $b_{\mathrm{KCl}} = 0.15$ mol/kg; (Bulk) all potassium ions in the same simulation.

A.4 Ion pairing in potassium acetate and sodium acetate solutions

Fig. A.7 shows that contact ion pairs (CIP) are more abundant in $NaCH_3COO$ than in KCH_3COO. In both cases, however, the solution is dominated by solvent shared ion pairs (SIP).

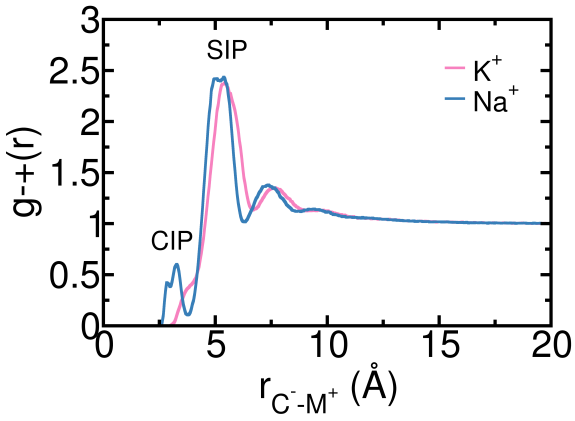


Fig. A.7 Radial distribution function between the metal ion (M^+) and the carboxylate carbon (C^-), from simulations of aqueous solutions of $NaCH_3COO$ with molality 0.5 mol/kg, or KCH_3COO at the same concentration. The parameters for the interaction between carboxylate and water and between carboxylate and Na^+ are from ref. 46; the interaction between carboxylate and K^+ is modelled using the optimized parameter shown in Table 2.1 of the main text. The remaining parameters are from GAFF and TIP3P water.

A.5 Water activity in NaCl and KCl solutions

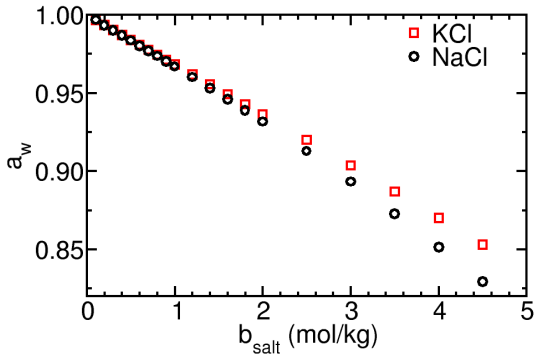


Fig. A.8 Experimentally determined water activity (a_w) in solutions of NaCl or KCl with the indicated molality, *b*. From ref. 82.

A.6 Free energy of mutations studied in the Chapter 4

Table A.2 Free energies (kcal/mol) of mutation for halophilic protein ferredoxin. The standard error of the mean (SEM) is reported for each value. $\Delta G_{\mathrm{X}a\mathrm{Y}} = \Delta G_{\mathrm{X}a\mathrm{X}^0} + \Delta G_{\mathrm{X}^0a\mathrm{Y}^0} + \Delta G_{\mathrm{Y}^0a\mathrm{Y}}$; see Scheme 4.10.

	b_{KCl} = 2 mol/kg					b_{KCl} = 0.15 mol/kg				
mutation	$-\Delta G_{\mathrm{Y}^0a\mathrm{Y}}$ (±0.1)	$-\Delta G_{\mathrm{X}a\mathrm{X}^0}$ (±0.6)	$\Delta G_{\mathrm{X}^0a\mathrm{Y}^0}$ (±0.02)	$\Delta G_{\mathrm{X}a\mathrm{Y}}$ [a] (±0.5)	distance [b]	$-\Delta G_{\mathrm{Y}^0a\mathrm{Y}}$ (±0.1)	$-\Delta G_{\mathrm{X}a\mathrm{X}^0}$ (±0.6)	$\Delta G_{\mathrm{X}^0a\mathrm{Y}^0}$ (±0.04)	$\Delta G_{\mathrm{X}a\mathrm{Y}}$ (±0.5)	distance (Å)
(D12)...D13N	60.76736	-137.79309	-0.55357	76.47216	5.63	63.24650	-140.28636	-0.56827	76.47159	5.96
(N12)...D13N	59.74974	-138.17301	-0.50017	77.9231	5.53	62.57828	-141.54472	-0.60791	78.35853	5.94
(E26)...D29N	56.26149	-132.62702	-0.64023	75.7253	6.36	57.82612	-133.13112	-0.09011	75.21489	5.19
(Q26)...D29N	56.11569	-133.70953	-0.60941	76.98443	7.74	56.45076	-132.36193	-0.09076	75.82041	5.41
(D81)...D107N	63.03267	-141.40526	-1.55017	76.82242	4.36	61.32781	-139.49030	-1.58711	76.57538	5.32
(N81)...D107N	62.65884	-141.83911	-1.64680	77.53347	5.58	61.34764	-140.34335	-1.78861	77.2071	5.23
(D83)...D81N	63.83078	-139.56623	0.32738	76.06283	10.97	66.46505	-142.27500	-0.85744	74.95251	6.04
(N83)...D81N	63.70651	-139.70081	0.25179	76.24609	9.86	66.10691	-143.31791	-0.93470	76.2763	4.84
(D109)...E110Q	57.22122	-135.08841	-0.59288	77.27431	5.53	57.41629	-134.64142	-0.33888	76.88625	7.65
(N109)...E110Q	56.89173	-134.29882	-0.75998	76.64711	6.74	57.98058	-136.46016	-0.10906	78.37052	8.50

[a] Combination of all thermodynamic cycles' free energy, and multiplying the values of $-\Delta G_{\mathrm{Y}^0a\mathrm{Y}}$, and $-\Delta G_{\mathrm{X}a\mathrm{X}^0}$ part by a negative to fit the scheme 4.10 with the associated value of standard error of the mean.
[b] The average distance between the terminal carbons of the side-chain of the mutating residue, D or E, and that of the neighboring residue, D, E, N, Q.The average is taken from the $-\Delta G_{\mathrm{X}^0a\mathrm{Y}^0}$ step of the mutation simulation at λ zero.

Table A.3 Free energy of mutation of Aspartic acid to Asparagine for halophilic protein L.

	b_{KCl} = 2 mol/kg					b_{KCl} = 0.15 mol/kg				
mutation	$-\Delta G_{\mathrm{Y}^0a\mathrm{Y}}$ (±0.1)	$-\Delta G_{\mathrm{X}a\mathrm{X}^0}$ (±0.6)	$\Delta G_{\mathrm{X}^0a\mathrm{Y}^0}$ (±0.04)	$\Delta G_{\mathrm{X}a\mathrm{Y}}$ (±0.5)	distance	$-\Delta G_{\mathrm{Y}^0a\mathrm{Y}}$ (±0.1)	$-\Delta G_{\mathrm{X}a\mathrm{X}^0}$ (±0.6)	$\Delta G_{\mathrm{X}^0a\mathrm{Y}^0}$ (±0.04)	$\Delta G_{\mathrm{X}a\mathrm{Y}}$ (±0.5)	distance (Å)
(E2)...E3Q	60.94070	-140.71040	-0.54365	79.22605	9.95	57.32818	-141.04686	-0.46305	83.25563	8.93
(Q2)...E3Q	61.10848	-139.17715	-0.35378	77.71489	9.97	57.41838	-141.39776	-0.36833	83.61105	9.58
(E28)...E32Q	65.31491	-142.93726	-0.33877	77.28358	7.87	66.00519	-141.92388	-0.46811	75.45058	8.07
(Q28)...E32Q	65.48402	-143.10207	-0.20677	77.41128	7.55	65.05565	-143.16430	-0.65581	77.45284	9.12
(E41)...E42Q	64.44933	-142.49745	-0.32551	77.72261	8.85	63.36859	-142.30051	0.02803	78.95995	7.67
(Q41)...E42Q	63.29743	-142.96200	-0.47503	79.18954	8.71	63.41466	-143.84089	-0.30592	80.12031	8.56
(D38)...E41Q	71.77454	-149.83710	-0.93222	77.13034	5.22	70.41324	-148.54546	-0.43475	77.69747	6.62
(N38)...E41Q	70.65305	-147.07521	-0.95782	75.46434	4.73	69.17052	-146.88446	-0.69658	77.01736	5.94
(E42)...D43N	57.20438	-135.56766	-0.67833	77.68495	6.17	57.07769	-134.68398	-0.43039	77.1759	6.28
(Q42)...D43N	56.67696	-134.85054	-0.46103	77.71255	6.00	57.89925	-137.23697	-0.34339	78.99433	5.76
(E23)...E21Q	60.58391	-136.00383	-1.04906	74.37086	8.57					
(Q23)...E21Q	60.04999	-134.23224	-0.54084	73.64141	7.87					
(E61)...D50N	59.29745	-138.22341	-0.93365	77.99231	6.96					
(Q61)...D50N	58.93284	-137.70456	-1.10447	77.66725	6.21					
(E28)...E27Q	66.99108	-142.59274	-0.22245	75.37921	6.30					
(Q28)...E27Q	62.82825	-140.17936	-0.17533	77.17578	7.15					

Table A.4 Free energy of mutation of Aspartic acid to Asparagine for halophilic protein Dihydrofolate reductase

$b_{\mathbf{KCl}}$ = 2 mol/kg					
mutation	$-\Delta G_{Y^0 a Y}$ ($\pm$0.1)	$-\Delta G_{X a X^0}$ ($\pm$0.6)	$\Delta G_{X^0 a Y^0}$ ($\pm$0.04)	$\Delta G_{X a Y}$ ($\pm$0.5)	distance
(D18)...E20Q	56.77630	-130.71713	-0.52123	73.4196	5.95
(N18)...E20Q	56.41874	-133.81749	-0.75389	76.64486	9.88
(D54)...D55N	56.58246	-136.05554	-0.73275	78.74033	7.23
(N54)...D55N	55.64233	-135.87124	-0.70413	79.52478	6.46
(E133)...D135N	66.45264	-145.07489	-0.34306	78.27919	7.90
(Q133)...D135N	66.52757	-143.00993	-0.31854	76.16382	8.11
(D135)...E138Q	69.22918	-146.38492	-0.30062	76.85512	5.78
(N135)...E138Q	68.29959	-146.02664	-0.27535	77.4517	5.26
(E144)...D146N	57.65902	-131.85082	-0.35306	73.83874	6.11
(Q144)...D146N	57.69196	-133.70504	-0.23425	75.77883	5.53

Table A.5 Free energy of mutation of Aspartic acid to Asparagine for unfolded halophilic protein L.

$b_{\mathbf{KCl}}$ = 2 mol/kg					
mutation	$-\Delta G_{Y^0aY}$ ($\pm$0.1)	$-\Delta G_{XaX^0}$ ($\pm$0.6)	$\Delta G_{X^0aY^0}$ ($\pm$0.04)	ΔG_{XaY} ($\pm$0.5)	distance
(E2)...E3Q	61.22625	-140.86904	-0.14001	79.50278	9.38
(Q2)...E3Q	62.13673	-141.91212	-0.27358	79.50181	7.44
(E21)...E23Q	56.39029	-137.91575	-0.06632	81.45914	7.50
(Q21)...E23Q	57.07932	-137.98829	-0.97050	79.93847	8.21
(E41)...E42Q	64.36035	-139.94819	0.04432	75.63216	6.16
(Q41)...E42Q	63.92586	-140.09838	-0.30096	75.87156	5.86
(E41)...D43N	61.78881	-135.53975	-0.27822	73.47272	10.64
(Q41)...D43N	62.42679	-133.52279	-0.46981	70.62619	9.62
(D43)...E46Q	61.07022	-139.94670	-0.16233	78.71415	5.96
(N43)...E46Q	60.45539	-136.06756	0.12897	75.74114	6.80