Application of Computational Molecular Biophysics to Problems in Bacterial Chemotaxis

A Dissertation Presented for the

Doctor of Philosophy

Degree

The University of Tennessee, Knoxville

Davi Ortega

May 2013

DEDICATION

To Milly.

If I can do it, you can do it

ACKNOWLEDGMENTS

This has been a long journey and so many people have contributed in so many different ways that I feel I cannot express all the gratitude they deserve in words. Thus, I will restrict myself to acknowledge some of the key individuals and institutions.

First I would like to acknowledge my advisor Dr. Igor Zhulin for more than his teachings and patience; his mentorship and friendship were crucial for the success of this work. Also, I would like to thank my co-advisor Dr. Jerome Baudry for his guidance in molecular dynamics simulations and his patience in discussing ideas over and over again. I must also acknowledge the other members of my committee for their comments and commitment to my project when I needed: Dr. Mike Guidry, Dr. Lee Riedinger, Dr. Marianne Brienig. I am grateful for the members of the Zhulin lab: Dr. Brian Cantwell, Dr. Kirill Borziak, Dr. Bhanu Rekapalli, Dr. Se-Ran Jun, Harold Shanafield, Amit Upadhyay, Ogun Adebali and Aaron Fleetwood for their teachings, discussions and exchange of ideas that directly and indirectly contributed to this work. In special, I also would like to thank Dr. Kristin Wuichet and Dr. Luke Ulrich for their mentorship and utter commitment in making me a better scientist by improving my analytical skills and critical judgment. In addition, there is a list of collaborators outside the University of Tennessee that must be acknowledge: Dr. Frederick Dahlquist, Dr. Sandy Parkinson and Dr. Grant Jensen for their teachings and collaborative work presented in this dissertation; Dr. Guoya Mo, Kwangwoon Lee for introducing me to a wet lab; and in special Dr. Ariane Briegel who always pushed me to become more precise in my explanations and more accurate in my computational predictions.

I also cannot end this acknowledgment section without showing my gratitude to Dr. Flavio Caldas da Cruz, who guided my first years in science and who heavily influenced my overall approach to scientific problems together with some of my professors and supervisors: Dr. Artemio Scalabrin, Dr. Daniel Pereira, Dr. Chris Oates, Dr. Leo Hollberg, Dr. Richard Fox, Dr. Scott Diddams and Dr. Amir Caldeira. Also, I want to thank several colleagues for their friendship and teachings that were directly involved with my research: Dr. David Figueira, Dr. Daniela Manoel, Dr. Reinaldo Cavasso-Filho, Dr. Yan Le Coq, Dr. Jason Stalnaker, Dr. Tara Fortier, Dr. Wictor Magno, Dr. Luciana Diogenes, Dr. Giovana Nogueira, Dr. Luciano Cruz, Dr. Ronaldo Vescovini, Haroldo Onisto, Marcos Junior, Joao Rodrigues and others. Finally, I would like to thank the Sao Paulo Research Foundation - FAPESP and the National Institute of Health – NIH for the financial support.

ABSTRACT

The combination of physics, biology, chemistry, and computer science constitutes the promising field of computational molecular biophysics. This field studies the molecular properties of DNA, protein lipids and biomolecules using computational methods. For this dissertation, I approached four problems involving the chemotaxis pathway, the set of proteins that function as the navigation system of bacteria and lower eukaryotes.

In the first chapter, I used a special-purpose machine for molecular dynamics simulations, Anton, to simulate the signaling domain of the chemoreceptor in different signaling states for a total of 6 microseconds. Among other findings, this study provides enough evidence to propose a novel molecular mechanism for the kinase activation by the chemoreceptor and reconcile previously conflicting experimental data. In the second chapter, my molecular dynamics studies of the scaffold protein cheW reveals the existence and role of a conserved salt-bridge that stabilizes the relative position of the two binding sites in the chew surface: the chemoreceptor and the kinase. The results were further confirmed with NMR experiments performed with collaborators at the University of California in Santa Barbara, CA. In the third chapter, my colleagues and I investigate the quality of homology modeled structures with cheW protein as a benchmark. By subjecting the models to molecular dynamics and Monte Carlo simulations, we show that the homology models are snapshots of a larger ensemble of conformations very similar to the one generated by the experimental structures. In the fourth chapter, I use bioinformatics and basic mathematical modeling to predict the specific chemoreceptor(s) expressed in vivo and imaged with electron cryo tomography (ECT) by our collaborators at the California Institute of Technology. The study was essential to validate the argument that the hexagonal arrangement of transmembrane chemoreceptors is universal among bacteria, a major breakthrough in the field of chemotaxis.

In summary, this thesis presents a collection of four works in the field of bacterial chemotaxis where either methods of physics or the quantitative approach of physicists were of fundamental importance for the success of the project.

TABLE OF CONTENTS

INTRODUCTION 1

Introduction to protein structure 2

Chemotaxis 4

Escherichia coli's chemotaxis system 7

Molecular Dynamics 9

Bonded and non-bonded interactions 10

The velocity-Verlet integrator 13

Isothermal-isobaric (NPT) ensemble 14

Anton supercomputer 14

Bioinformatics and Genomics 15

Scope of Thesis 16

CHAPTER I SIGNALING MECHANISM OF CHEMORECEPTOR REVEALED BY MICROSECOND MOLECULAR DYNAMICS SIMULATION 17

Abstract 18

Introduction 18

Results 23

Bending properties of chemoreceptor 23

Average of atomic mobility of alpha carbons and order parameter. 26

Phe396 undergo to cis-trans conformational switch during simulations. 28

Cis-trans switch of Phe396-Phe396' is methylation state dependent 32

Phe396 is the most conserved residue in chemoreceptors. 36

Discussion 36

Mechanical properties of the chemoreceptor structure 36

Cis-trans switch of conserved Phe396 suggests novel kinase activation mechanism 40

Conclusion 45

Methods 48

CHAPTER II THE ROLE OF A CONSERVED SALT-BRIDGE IN THE CHEMOTAXIS COMPLEX 53

Abstract 54

Introduction 55

Results 56

Co-evolving class-specific residues in CheW form a short-range salt bridge....56
Mutants E38A and R62A show different stability *in vitro*....61
Backbone Dynamics....61
Direct evidence of the salt-bond between Arg62 and Glu38....67
Interaction between Arg62 and Glu38 affects protein dynamics....70
Local backbone changes are linked to relative position of functionally active sites....70
Discussion....75
Materials and Methods....78
NMR spectroscopy....78
Bioinformatics....79
Molecular Dynamics Protocols....80
Calculation of the frame average RMSD per residue:....81
Order Parameter Calculations:....82
CHAPTER III HOMOLOGY MODELING AND MOLECULAR SIMULATIONS OF CHEW....83
Abstract....84
Introduction....84
Results....87
Discussion....94
Materials and Methods....96
Acknowledgments....99
CHAPTER IV UNIVERSAL ARCHITECTURE OF BACTERIAL CHEMORECEPTOR ARRAYS....101
Abstract....102
Introduction....102
Results and Discussion....103
Conclusion....116
Materials and Methods.....116
Acknowledgements....118
CONCLUSION....119
REFERENCES....123
VITA....139

LIST OF TABLES

Table 1: Standard amino acids, their codes and biochemical groups 3
Table 2: Highly conserved residues in F1 and F7 classes of the CheW protein 59
Table 3: Chemotaxis protein structure entries in the Protein Data Bank. 86
Table 4: RMSD values from Figure 37 for (A) MD and (B) LBMC. 91
Table 5: CheW interactions in the literature. 97
Table 6: Summary of measurements of 13 different bacterial species obtained by ECT 105

LIST OF FIGURES

Figure 1: Main protein structures levels..... 5
Figure 2: Canonical E. coli chemotaxis system. 6
Figure 3: Bonded interactions 11
Figure 4: Scheme of the most common layout of chemoreceptor domains.....21
Figure 5: Profile of averaged of bending angle for each residue layer for each 2 microsecond simulation in three different methylation states: QQQQ (green), QEQE (black) and EEEE (red). 24
Figure 6: Mapping of the QQQQ averaged bending angle profile in the Tsr structure.....25
Figure 7: Profile of r.m.s.f for each residue for each 2 microsecond simulation in three different methylation states: QQQQ (green), QEQE (black) and EEEE (red). 27
Figure 8: Profile of the order parameter (S^2) calculated from each simulation of the chemoreceptor in three different methylation states: QQQQ (green), QEQE (black) and EEEE (red).....29
Figure 9: Temporal analysis of the protein interaction module mobility (r.m.s.d.) shows oscilation between two stable conformations. 30
Figure 10: Interaction between Phe396(blue) and Phe396'(red) in the Tsr X-ray crystal structure (PDB code: 1QU7) [14] 31
Figure 11: Temporal evolution of the $\chi 1$ dihedral angle of the interacting residues Phe396 (above) and Phe396' (below) 33
Figure 12: Time lapse of a section of a chemoreceptor simulations in the wild-type methylation state highlighting the conformation switch of the Phe396(red)-Phe396'(blue) pair.....34
Figure 13: Temporal analysis of the pairwise distances between the helices: C, N, C' and N'…35
Figure 14: Comparison of the distributions of $\chi 1$ dihedral angle in Phe396 during simulations of the chemoreceptor in three states: QQQQ, QEQE and EEEE.....37
Figure 15: Sequence logo of the chemoreceptor protein interaction module. 38
Figure 16: Geometry of the tip of the receptor in cis conformation of the Phe396 (A) and trans conformation of the Phe396 (B). 41
Figure 17: Working model of the Phe396 conformation preference based on molecular dynamics simulations. 42
Figure 18: Predicted geometries of the trimer-of-dimer for cis (A) and trans (B) conformations of the χ_1 dihedral angle of the Phe396.....43

Figure 19: Top view of the protein interaction module of the Tsr trimer of dimers crystal structure. 44
Figure 20: The three distinct regions of the Tsr receptor: adaptation module, flexible bundle and protein interaction module. 46
Figure 21: Schematics of the calculation of the bending angle. 50
Figure 22: Distribution of non-redundant CheW sequences in chemotaxis classes 58
Figure 23 : Highly conserved residues mapped on the CheW secondary structure scheme (A) and the 3D NMR structure (B). 60
Figure 24: Effects of the mutation E38A in the CheW structure. 62
Figure 25: Effects of the mutation R62A in the CheW structure. 63
Figure 26: Backbone amide 15N relaxation parameters for CheW vs. residue number. 65
Figure 27: CPMG dispersion curves for CheW. 66
Figure 28: Temporal evolution of the distance of the side chain charged group centroids 68
Figure 29: The salt-bridge formed by Arg62 and Glu38 in geometry A (top) and geometry B (bottom). 69
Figure 30: Histogram of the distance between alpha carbon from Arg62 and Glu38 in all wild-type simulations combined (black). The frames can be separated by the occurrence of salt-bridges in geometry A (blue) and all other interactions (red). 71
Figure 31: Distributions of the distances between alpha carbons for each conformation between Arg62 and Glu38 in all simulations with WT. 72
Figure 32: Comparison of the histograms of the distances between the alpha carbons of residues 38 and 62 for three alleles: wild-type (black), E38A (purple) and R62A (green). 73
Figure 33: Average of the order parameter calculations for 10 simulations in each simulated allele. Wild-type is in black, E38A in red and R62A in blue. 74
Figure 34: Occurrence of salt-bridge in geometry A between Arg62 and Glu38 improves the stability of the position of the chemoreceptor binding region relative to the kinase binding region. 76
Figure 35: Histogram of the pairwise sequence identities for all against all 1,742 non-redundant CheW sequences selected from the MiST2 database. 88
Figure 36 : Comparison of the RMSDs between the 20 homology models and 20 NMR structures for both *E.coli* and *T. maritima*. 89
Figure 37: RMSD matrices comparing the similarity of each point of the homology-modeled trajectories with each point of the NMR trajectories. 92

Figure 38: Superimposition of the most similar structure in the NMR trajectory (red) with the most similar structure in the homology-modeled trajectory (blue).....93
Figure 39: Histograms of the RMSD values comparing the NMR ensembles and MD/LBMC simulated trajectories. 95
Figure 40: Schematic representation of MCP topology type I. 104
Figure 41: Characteristic appearance of chemoreceptor arrays in vivo. 107
Figure 42: Chemoreceptor arrays in diverse bacteria. 108
Figure 43: Correlation between observed physical length and predicted sequence length. 109
Figure 44: MCP signaling domain length classes observed in the study.....112
Figure 45: The three different domain architectures of MCP topology type I receptors. 113
Figure 46: Universally conserved 12-nm hexagonal arrangement of receptor. 114
Figure 47: Maximum likelihood phylogenetic tree of 406 representative bacteria. 115

INTRODUCTION

Physicists have always combined intricate mathematical models to explain and predict behaviors of different systems in nature: from galaxies to sub-atomic particles. In the last century, physics has contributed to the understanding of many fields of biology. In particular, the study of biomolecules greatly benefited from the methods of physics resulting in the field of molecular biophysics. With the significant increase of available computational power, *in silico* methods became viable to molecular biophysicists. For example, a detailed description of the dynamics of a biomolecule can be obtained by numerically solving Newton's equations of motion for a many-body system. This method, called molecular dynamics simulations (MD) has rapidly evolved in the past few years and is now capable of approaching biological systems of varied time and length scales [1, 2]. Concomitantly, several properties calculated from trajectories produced by MD simulations have been successfully compared to experimentally derived data as a validation of the computational method [3, 4]. However, the advantage of approaching problems in biology from the physics perspective is not restricted to innovative methods. The generally quantitative formalism of a physicist in synergy with the qualitative perspective of a biologist allows for major leaps in the current understanding of life sciences. This multi-part research project seeks to apply this synergy between physics and biology in different problems related to bacterial chemotaxis, primarily focusing on its sophisticated molecular machinery. In addition, bioinformatics methods are used to extrapolate the results from a particular study to a range of similar biological systems, considerably increasing the significance of the discoveries here described.

Introduction to protein structure

At the molecular level, proteins are constructed from building blocks called amino acids. Amino acids are organic compounds formed by a carboxylic acid ($-COOH$) connected to an amine group ($-NH_2$) by a carbon, known as an alpha carbon, which in turn is connected to a hydrogen atom and a side chain. Different amino acids have the same backbone (carboxylic acid, alpha carbon and amine group) but different side chains, or residues. Twenty of these amino acids are specified by the genetic code and are considered standard building blocks of proteins. Each of these amino acids is represented by a unique three letter code or one letter code and can be classified by biochemical characteristics (Table 1). An amide synthesis reaction between any two amino acids builds a peptide bond between them. Repeatedly, this

Table 1: Standard amino acids, their codes and biochemical groups

Amino acids	One letter code	Three letter code	Biochemical group
Alanine	A	Ala	hydrophobic
Arginine	R	Arg	positive charged
Asparagine	N	Asn	polar
Aspartic acid	D	Asp	negative charged
Cysteine	C	Cys	polar
Glutamic acid	E	Glu	negative charged
Glutamine	Q	Gln	polar
Glycine	G	Gly	hydrophobic
Histidine	H	His	positive charged
Isoleucine	I	Iso	hydrophobic
Leucine	L	Leu	hydrophobic
Lysine	K	Lys	positive charged
Methionine	M	Met	hydrophobic
Phenylalanine	F	Phe	hydrophobic
Proline	P	Pro	hydrophobic
Serine	S	Ser	polar
Threonine	T	Tre	polar
Tryptophan	W	Trp	hydrophobic
Tyrosine	Y	Tyr	polar
Valine	V	Val	hydrophobic

process builds elongated amino acid chains also known as polypeptide. Finally, proteins are polypeptides that perform a biological function when folded [5].

From the macromolecular perspective, proteins have four levels of information. The primary structure of a protein is the information related to the amino acid content that forms the protein, also known as the protein sequence. Segments of the peptide sequence of various lengths fold primarily in two basic configurations: alpha-helices and beta-sheets. Information associated to each configuration is called secondary structure. The tertiary structure of a protein is the complete tridimensional fold of the peptide sequence, built from a series of alpha-helices, beta-sheets and unstructured regions. Lastly, the quaternary structure is the association of multiple tertiary structures to form a macromolecular protein complex Figure 1 .

Finally, from the architectural perspective, proteins are divided into domains, which are parts of the protein that fold independently of the others. A protein can have a single domain or multiple domains. Normally, each domain is responsible for a specific task fundamental to the overall function of the protein.

Chemotaxis

Chemotaxis is the phenomena of organisms moving in the environment according to chemical cues. The network of proteins that participates in this process forms the chemotaxis system. The chemotaxis system has been classified as the most complex two component regulatory system (TCS) [6], which, in turn, is a subcategory of the signal transduction cell systems [7]. Generally, the chemotaxis system works by sensing the chemical composition of its surrounding environment and controlling the motility apparatus to navigate the organism accordingly. In other words, it is the navigation control system of prokaryotes and lower eukaryotes. Throughout evolution, bacterial chemotaxis systems have diversified incredibly, both in network structure and components; however, the basic principles of functionality remained the same [8-10]. The core components present in most chemotaxis systems are: the methyl-accepting chemotaxis proteins (MCPs), scaffold protein (CheW), the phosphor-relay histidine kinase (CheA), the CheA response regulator (CheY) (responsible for motor control) and two enzymes: the methyltransferase (CheR) and methylesterase (CheB) (both involved in adaptation mechanisms) [10]. While CheB, CheR and CheY are diffused in the cell, thousands of copies of MCPs, CheAs and CheWs self-assemble into a protein complex generally localized in the cell pole (Figure 2) [11]. By means of whole-cell electron cryo-tomography, it has been

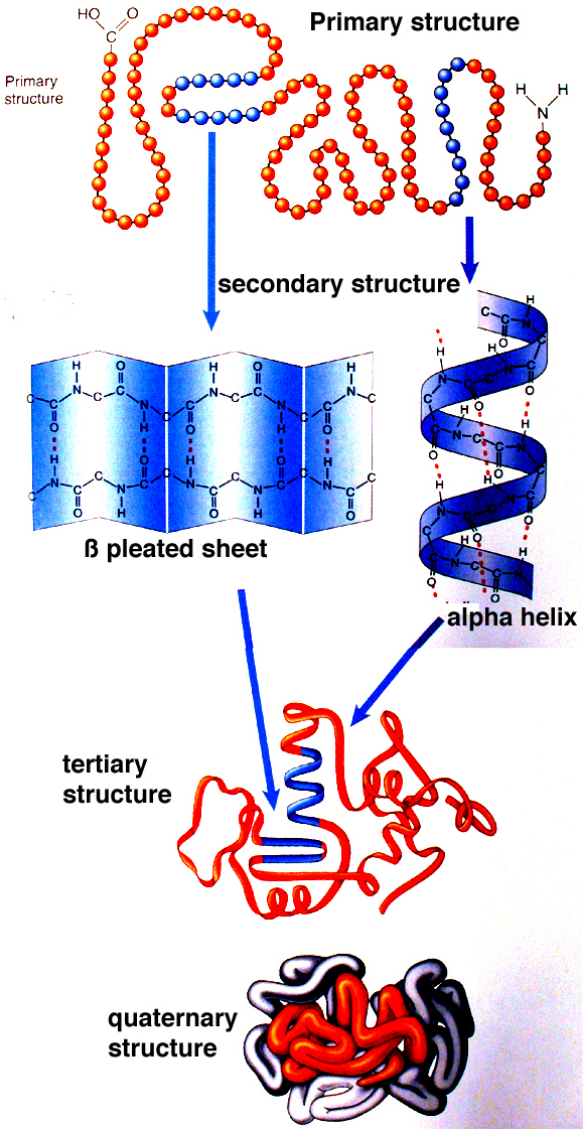


Figure 1: Main protein structures levels

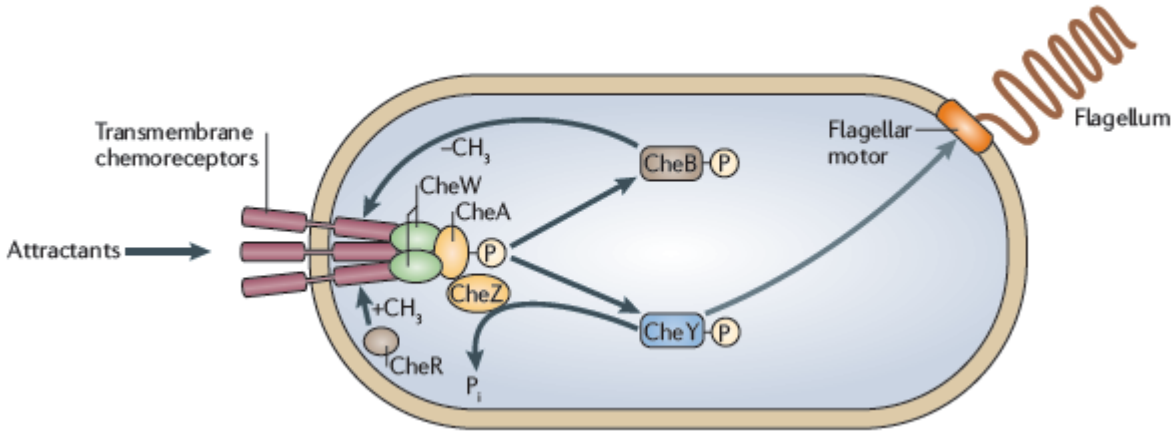


Figure 2: Canonical E. coli chemotaxis system.

Taken from [11]

shown that chemoreceptors in different species representing major bacterial phyla are all arranged into a highly conserved, 12-nm hexagonal array [12]. This rejects the model of hedgerow arrangements for chemoreceptors previously suggested from the crystal structure in *Thermotoga maritima* [13] in favor of the trimer-of-dimers arrangement suggested by the crystal structure of E. coli's receptor Tsr [14]. This result suggests that the basic mechanism and function of receptor clustering is universal among bacterial species and was, thus, conserved during evolution.

Escherichia coli's chemotaxis system

Escherichia coli, which is a model organism in the study of chemotaxis, has a simple chemotaxis network of proteins featuring all the core components plus a phosphotase (CheZ). It has all the defining characteristics of a chemotaxis system: adaptability, signal amplification, dynamic range [15] and robustness [16]. This thesis will further explore the chemotaxis system of *E. coli*.

In *E. coli*'s genome, there is one histidine kinase (CheA), one scaffold protein (CheW) and five MCPs: Tar, Tsr, Trg, Tap and Aer [17]. The two major receptors, Tar and Tsr, are different from the other receptors because they show greater abundance in the cell and have the ability to adapt and function independently of other receptors. On the other hand, the three minor receptors, Trg, Tap, and Aer, depend, to some extent, on major receptors to perform function and adaptation, as well as showing lower abundance in the cell. Thousands of copies of the five chemoreceptors form a mixed forest of trimer-of-dimers [14, 18-22] and, with help from the scaffold protein CheW, bind to a homodimer histidine kinase CheA to form the minimal signaling unit *in vitro* [23]. The stoichiometry of the chemotaxis system in E. coli is still a matter of debate [20]. A membrane bound, ultra-stable[24] macromolecular complex, also called signaling complex, is then formed by arranging many copies of the minimal signaling unit into a large array with hexagonal symmetry [20, 25].

The chemotaxis network in *E.coli* is rather simple compared to other organisms and, therefore, subject to intense study [11]. The qualitative model of E.coli's chemotaxis system implies that the signaling complex oscillates between two conformational states. The "on" receptor conformation promotes the CheA autophosphorylation rate and the "off" conformation disables CheA autophosphorylation. The overall kinase activity, therefore, is the result of the proportion of complexes in the two states [26]. In the presence of attractants, the receptors' conformation equilibrium is shifted to the "off" state; while in the absence of attractants the

receptors' conformation equilibrium is shifted to the "on" state. Shifts in the on-off equilibrium control the flux of phosphoryl groups from CheA to CheY and CheB. CheY-P binds to the motor, increasing the probability of clockwise (CW) rotation, which causes the cell to tumble. Counter-clockwise (CCW) rotation, the behavior in absence of CheY-P, produces smooth swimming. The combination of tumbling and smooth swimming results in a random walk movement of the organism. The phosphatase CheZ hydrolyzes CheY-P, guaranteeing short durations of the tumbling periods.

During its random walk, the cell needs to keep track of the various concentrations of attractants recently experienced in order to bias its net displacement towards higher concentrations. This is done via the use of methylation sites in the MCPs by two enzymes: the methylesterase CheB and the methyltransferase CheR. There are four major methylation sites in the major chemoreceptors of *E. coli*, and they are encoded in the gene (in order) as glutamine, glutamate, glutamine and glutamate, or in short QEQE. This methylation state is, thereby, also called wild-type. The adaptation mechanism is based in constant methylation of the methylation sites by the enzyme CheR and in deamidation of the same methylation sites by the enzyme CheB which is kinase activity dependent [26].

Specifically, the wild-type receptors in absence of attractants balance the constant activity of CheR by producing enough CheB-P (CheB phosphorilated by CheA) to keep the chemoreceptors in the wild-type methylation state. Also, the balance between CheY-P and CheZ in the cell avoids the immediate increase in tumbling and allows a non-bias random walk. As the cell encounters attractants, the receptor undergoes to a series of conformational change, culminating in an overall decrease in kinase activity. This change leads to less CheY-P, which allows for a decrease in the tumbling frequency. It also leads to less CheB-P, and the imbalance between CheB-P and CheR drives the receptor to a more methylated state, counteracting the conformational changes due to ligand binding and ultimately restoring the receptor's equilibrium for tumbling. However, this secondary adaptation pathway occurs in a much slower time scale, allowing the bacteria to effectively drift towards the attractant before adapting, reaching equilibrium, and returning to a non-bias random walk. Lastly, when the cell ceases to encounter attractants, the receptor's conformation shifts to the "on" state, which increase kinase activity and consequently, the concentration of CheB-P and CheY-P. The cell now tumbles frequently (high concentration of CheY-P) and the receptor is driven to a less methylated state, which, in turn, restores the wild-type methylation state and, finally, restores the equilibrium between CheB-P and CheR, as well as CheY-P and CheZ[15]. In summary, the adaptation mechanism of

chemotaxis systems assures responses to gradients rather than absolute concentrations of attractants.

Molecular Dynamics

The intricate molecular mechanisms that govern life are a matter of great interest in biology. However, the deep understanding of such mechanisms is not only related to descriptions of proteins, enzymes, lipids, DNA and ligands in atomic resolution but also to how these entities evolve in time. Despite current advances in microscopy and other techniques, there are many limitations to satisfactorily describing the dynamics of biologically relevant molecules via experimental methods. This creates an opportunity for computational methods to aid experimental techniques in a collaborative way. In the context of atomistic description of biological molecules, a specialized computational technique borrowed from chemistry and physics became popular in the past fifty years: the classical molecular dynamics simulations. In a nutshell, the technique simply uses computers to numerically calculate Newtonian equations of motion for a system with N interacting atoms:

$$m_i \ddot{\vec{r}}_i = -\vec{\nabla} U_{total}(\vec{r}_1, \vec{r}_2, \dots, \vec{r}_N), \quad i = 1,2, \dots, N \tag{1}$$

where m_i and $\vec{r}_i$ are, respectively, the mass and the position of atom i. The total potential energy, $U_{total}(\vec{r}_1, \vec{r}_2, \dots, \vec{r}_N)$, is a function of the spatial distribution of all particles in the system [27]. The solution of such equations is the time evolution of the entire system that can be used to calculate several thermodynamic parameters and even visualize certain molecular mechanisms with atomic precision. However, molecular dynamics techniques are only applicable to molecular biology because the atomic coordinates of several proteins and DNAs have been already resolved experimentally. These atomic coordinates are the starting point of any molecular dynamics simulation. Atomic coordinates derived from X-ray crystallography or NMR experiments are organized in a file, named “pdb file,” which is deposited in a public repository [28].

Unfortunately, MD simulation is a classical approach to molecular problems which would be more accurately described by the solution of the time-dependent Schrödinger’s equation. Current computational techniques and available computational power makes moderately sized problems intractable under the quantum mechanics framework. To classically approach molecular problems in a computationally efficient manner, some approximations must be made:

for example, the Born-Oppenheimer approximation which assumes that electrons move instantaneously with the atomic motion, or the assumption that all electrons are in their ground state at all times of the simulation. While approximations of this nature limit applications of the technique, biomolecules have been successfully treated under the classical framework [3, 29-33].

Bonded and non-bonded interactions

The main success of MD simulations in biomolecules comes from the accurate, computer-friendly, description of the total potential energy U_{total}. In general, the total potential energy of a given atom is decomposed in bonded and non-bonded terms, but in practice, the technical description of it is a matter of choice between the engine and the force field used. All simulations described throughout this thesis use the CHARMM27 force field [34-37]. All engines used in this thesis (NAMD2[27], Desmond [38] and Anton [39]) are able to work with the potential energy description provided by CHARMM27, which separates the total potential energy into seven components:

$$U_{total} = U_{bond} + U_{angle} + U_{UB} + U_{improper} + U_{dihedral} + U_{vdw} + U_{coulomb} \quad (2)$$

the first four terms are intra molecular interactions. They are related to the movements of stretching, bending (angle and Urey-Bradley) and improper torsion angles respectively (**Error! Reference source not found.**). They are quadratic energetic penalties to geometrical deviation from the position of equilibrium and are modeled by the following equations[34] :

$$U_{bond} = \sum_{bonds\ i} k_i^{bond}(r_i - r_{0_i})^2 \quad (3)$$

$$U_{angle} = \sum_{angles\ i} k_i^{angle}(\theta_i - \theta_{0_i})^2 \quad (4)$$

$$U_{UB} = \sum_{UB} k_{ik}^{UB}(r_{ik} - r_{0_{ik}})^2 \quad (5)$$

$$U_{improper} = \sum_{improper\ i} k_i^{improper}(\varphi_i - \varphi_{0_i})^2 \quad (7)$$

where the variables r_i, θ_i, r_{ik}, φ_i are the bond length, bond angle, Urey-Bradley 1,3 distance and the improper torsion angle while r_{0_i}, θ_{0_i}, $r_{0_{ik}}$, φ_{0_i} are equilibrium values and k_i^{bond}, k_i^{angle}, k_{ik}^{UB}, $k_i^{improper}$ force constants. The last intra molecular potential is the energetic adjust for the dihedral angle rotational barrier described by the formula:

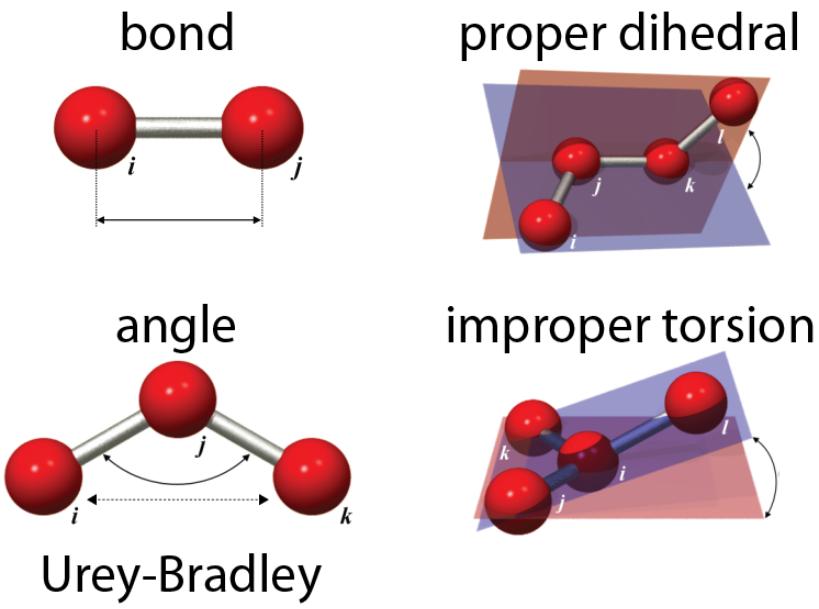


Figure 3: Bonded interactions

$$U_{dihedral} = \sum_{dihedral\,i} k_i^{dihedral}[1 + \cos(\gamma_i n_i - \gamma_{0_i})] \quad (6)$$

here, $\gamma_i$is the dihedral coordinate, n_i determines the periodicity, φ_{0_i} determines the angle where the potential has a minima and $k_i^{dihedral}$ determines its amplitude force constant.

There are two non-bonded terms: van der Walls and Coulombic. The first is described as Lennard-Jones potential 6, 12:

$$U_{vdw} = \sum_{non-bonded} 4\epsilon_{ij}\left[\left(\frac{R_{0,ij}}{r_{ij}}\right)^{12} - \left(\frac{R_{0,ij}}{r_{ij}}\right)^{6}\right] \quad (8)$$

where r_{ij} is the distance between atoms i and j , ϵ_{ij} is the potential depth and $R_{0,ij}$ is the minimal distance allowed between the atoms. Finally, the last term describes the electrostatic potential of the system:

$$U_{coulomb} = \sum_{non-bonded} \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}} \quad (9)$$

where q_i, q_j are the charges of the atoms i and j, r_{ij} the distance between them and ϵ_0 the dielectric constant, which is set to unit in this context.

The accuracy of the force field depends on how accurately the parameters described above portray the interactions between atoms in the simulation. In CHARMM27, the force field parameters have been determined by quantum mechanics calculations and empirical techniques [37]. Recently, a number of studies have shown that simulations with this same force field are used to calculate a number of parameters that have been successfully compared to experimental data [40, 41].

Although the accuracy of force fields allows for long MD simulations, the calculations of the non-bonded components of the potential energy are troublesome. Both van der Walls and Coulombic terms are, in principle, summed over every, single, non-bonded atom of the simulation. To avoid boundary effects, the system is submitted to periodic boundary conditions which makes the calculations of non-bonded terms of the potential energy practically intractable [27, 38]. For that reason, special techniques are used to calculate non-bonded terms of the potential energy.

NAMD2 and Desmond calculate the electrostatic potentials with a technique called Particle mesh Ewald or PME [42]. Desmond and Anton use a similar technique called k-space Gaussian split Ewald. A detailed description of both methods can be found in [43], but basically, the method splits the equation (9) in two parts: (1) a direct sum that decays quickly as distance increases and is limited by a cut-off and (2) a reciprocal sum that would converge slowly if not

for efficient mapping of the charge distribution of the system into a regular mesh before the efficient fast Fourier transform [39]. On the same note, van der Walls computations are usually truncated in space given the fast decay of its potential. Despite the technique described to improve performance, van der Walls and electrostatic forces are responsible for almost 90% of the work load of an MD simulation in a regular, single processor, and both are a current subject of study [39].

The velocity-Verlet integrator

Once the potential energy is calculated, the equations of motion (1) must be evolved in time. There are a couple of different techniques to perform this task, but, the velocity-Verlet integrator [44] is described here, as it is the integrator used in NAMD2, Desmon and Anton.

First, assume that all forces are already calculated for the current time step n and the current position and velocity $(r_i(t=n), v_i(t=n))$ are known. Now, the goal of the integrator is to calculate the next step $(r_i(n+1), v_i(n+1))$ assuming the forces $F(r_i(n))$. The basic velocity-Verlet integrator performs the following instructions:

(1) The velocity is calculated from the actual velocities and the current force for a half-step:

$$v_i\left(n + {}^{1}/_{2}\right) = v_i(n) + m_i^{-1} F\left(r_i(n)\right) * {}^{\Delta t}/_{2} \tag{10}$$

(2) The new positions are calculated using the half-step velocity:

$$r_i(n+1) = r_i(n) + v_i\left(n + {}^{1}/_{2}\right) * \Delta t \tag{11}$$

(3) The force $F(r_i(n+1))$ is computed using the recently calculated position.

(4) Finally the new velocities are computed for the step $n+1$:

$$v_i(n+1) = v_i\left(n + {}^{1}/_{2}\right) + m_i^{-1} F\left(r_i(n+1)\right) * {}^{\Delta t}/_{2} \tag{12}$$

where Δt is the integration time step.

This integrator has desirable accuracy for biomolecular applications and is stable for long time simulations. To improve integration time efficiency, NAMD2, Desmond and Anton employ multiple time stepping [44], also known as RESPA, which stands for reference system propagator algorithm [45] . As the calculation of the potential energy is divided per nature of interaction, RESPA allows that different parts of the potential function can have a different integration time step. Usually, bonded interactions are calculated more often, followed by the short range electrostatics and van der Walls, and less often than the long-range electrostatics forces. While it is desirable to have the largest time-step possible from a performance perspective, it should be small enough to ensure accurate description of the system and stability of the numerical calculations. Hydrogen atoms are very light compared to other particles in the system, therefore, bonds to hydrogen atoms limit the size of the time-step. The most popular way to constrain bonds to hydrogen is by using Lagrange multiplier-based methods. While NAMD uses SHAKE [46], both Desmon and Anton use an updated version M-SHAKE [47]. Finally, the integrator accommodates small modifications, as needed, to keep constraints necessary for specific thermodynamic ensemble.

Isothermal-isobaric (NPT) ensemble

Molecular dynamics simulations can be performed in different thermodynamic ensembles: microcanonical (NVE), canonical (NVT), isoenthalpic-isobaric and isothermal-isobaric (NPT). From all ensembles, NPT is the one that better simulates experimental conditions since it requires that temperature and pressure of the system remain constant. Conceptually, this requirement is achieved in MD simulations by (1) coupling the system to a thermal bath that exchanges energy with the system in order to keep temperature constant and (2) by scaling the system dimensions to controlling the pressure. Details on how these constraints are implemented in MD simulations can be found in [44, 48, 49].

Anton supercomputer

Anton [39] is a specialized supercomputer for molecular dynamic simulations. The machine has 512 processing nodes, each containing a specialized processor for molecular dynamics calculations. The nodes are connected with a high performance network and the machine runs a specialized molecular dynamics engine to avoid overhead. Unlike other machines, Anton has no operating system. The instructions to solve the Newtonian equations of

motion, the structure file containing the atomic coordinates, the parameters for the force field calculations, the integrator constants, the thermodynamic ensemble and any other requirements are compiled altogether in a front node running Linux, which in turn, submits the job to Anton nodes. The unique combination of hardware and software on Anton reaches performances of 2.5 μs/day compared to 50 ns/day for the same system using Desmond on 512 nodes on the University of Tennessee's supercomputer Newton. Simulations presented in Chapter I of this thesis were performed on Anton using 50,000 hours awarded as a grant.

Bioinformatics and Genomics

The use of bioinformatics in this thesis is limited and primarily enhances the significance of the results obtained by other techniques. The genetic code of a given organism, its DNA, carries the information about what proteins are produced to sustain its life. Protein sequences from various genome projects are annotated and stored in databases such as RefSeq [50], providing a vast amount of genetic information. Bioinformatics provides tools to mine this information by mixing applied mathematics and statistics with computational methods. In fact, bioinformatics alone was able to perform ground-breaking scientific developments such as the neutral theory of molecular evolution [51] and the evolutionary tree built from genome sequence [52]. Here some concepts are introduced for the sake of clarity in the later chapters.

In short, a gene is the stretch of DNA that translates into a protein. Proteins are peptide chains that execute a biological function when folded. A pair of genes is said to be homologous if they are related to each other either by orthology, separated by an event of speciation, or by paralogy, separated by an event of duplication. A protein family is the set of genes where all genes are homologous to each other. Using sequence similarity search algorithms such as BLAST [53] or HMMER [54] , one can search databases for homologous genes. A set of homologous genes can be aligned using multiple sequence alignment software, such as MAFFT [55]. The multiple sequence alignment (MSA) is used to evidence a mutation, a deletion, or an insertion of residues in each protein of the protein family. This allows for a comparison of the rate of side chain exchange in each individual amino acid position during the course of evolution of the protein family. It is a paradigm in bioinformatics that highly conserved positions in a multiple sequence alignment of homologous sequences are likely to be crucial for its biological function [56]. In addition, the conservation of the sequence positions involved in certain tasks indicates that the results of experiments in one gene of the family can be confidently

extrapolated to all other genes in the same family, which enhances the impact of a particular biological finding.

Scope of Thesis

This thesis will portray a series of four interdisciplinary works intersecting physics, biology, chemistry and computer science. The common theme between them is the persistent exploration of problems in molecular chemotaxis in prokaryotes from the viewpoint of a physicist. Throughout the thesis, the importance of understanding the limitations of popularized methods such as molecular dynamics, will become apparent. Chapter 1 will investigate the molecular machinery in a chemoreceptor structure by the use of molecular dynamics simulations performed in a specialized machine. This study, apart from reconciling apparently conflicting experimental data previously reported, proposes a novel mechanism of kinase activation by the receptor that should be conserved across the Bacteria and Achaea domain. Chapter 2 will show a combination of molecular dynamics simulations and bioinformatics to predict the role of a conserved residue for an essential protein in chemotaxis. The prediction is further confirmed by NMR measurements and biochemical experiments done by collaborators. Chapter 3 is a study of the quality of homology modeled structures using molecular dynamics and Monte Carlo simulations with CheW protein as a benchmark. The results of this study will guide future homology modeling on cheW homology models in other model organisms such as *Rhodobacter sphaeroides* and *Bacillus subtilis*. Chapter 4 will provide details of the use of bioinformatics and simple mathematical modeling to recognize which specific chemoreceptors were imaged in a series of cryo-EM tomography of several distant related organisms. The work provided decisive support for the universality of hexagonal packing of the self-assembled chemotaxis array in prokaryotes. Finally, the conclusion will review the work presented in this thesis and future prospects will be discussed.

CHAPTER I SIGNALING MECHANISM OF CHEMORECEPTOR REVEALED BY MICROSECOND MOLECULAR DYNAMICS SIMULATION

This chapter was taken from a manuscript in preparation:

Davi R. Ortega, Jerome Baudry and Igor B. Zhulin, **Signaling Mechanism of Chemoreceptor Revealed by Microsecond Molecular Dynamics Simulation.** Manuscript in preparation.

Conceived and designed experiments: DO JB IZ. Performed experiments: DO. Analyzed the data: DO JB IZ. Wrote the paper: DO JB IZ.

Abstract

Bacterial chemotaxis is a model system for the to study of transmembrane signaling mechanisms. Chemoreceptors typically span the inner membrane as part of a highly ordered complex that includes a cytoplasmic kinase, CheA. Ligand binding by the chemoreceptor periplasmic domain transmits a signal to its distal cytoplasmic tip, which interacts with CheA and modulates its kinase activity. The molecular mechanism of kinase activation and deactivation by the chemoreceptor signaling domain remains largely unknown. Using long all-atom molecular dynamics simulations, we show that there is a shift in equilibrium between two stable conformations of the chemoreceptor kinase-binding region, which is dependent on the chemoreceptor signaling state. The two stable conformations of the kinase- binding region are associated with the cis – trans isomeric state of the $\chi 1$ dihedral angle of Phe396, the most conserved residue in the chemoreceptor superfamily. The wild-type chemoreceptor oscillates equally between the two conformations, whereas each signaling state shifts the equilibrium to a particular conformation state. Our results suggest that the switch between the two conformations is the direct control mechanism of kinase activity in chemotaxis, and that this mechanism is conserved throughout all chemotaxis systems, which are widespread across diverse lineages of Bacteria and Archaea. Our results suggest that long molecular dynamics simulations can be productive in studying molecular mechanisms of signal transduction in other systems.

Introduction

Bacteria navigate their environment via a specialized two-component system, the chemotaxis system, which regulates the flagellar motor. Chemical cues induce conformation changes in

chemoreceptors, also known as methyl-accepting chemotaxis proteins (MCPs), which modulate the kinase's autophosphorylation rate affecting the function of the flagella motor. The fine control of this mechanism leads the organism to swim towards attractants and away from repellents. Chemotaxis systems are vastly diverse with several accessory proteins [10] and interaction network structures [8, 9]. However, the superfamily of chemoreceptors is of central importance in the molecular machinery of all chemotaxis pathways in prokaryotes and it is present in more than 95% of genomes with at least one chemotaxis gene [10, 26]. Multiple copies of these receptors directly interact with the kinase CheA and the scaffold protein CheW forming a large ultra-stable multi-protein complex [24]. Furthermore, electron cryo-tomography images of the chemotaxis protein complex in several organisms show a common hexagonal organization of chemoreceptors arrays despite the idiosyncrasies of each chemotaxis system [12]. The universality of these results supports that signal transduction mechanisms in chemoreceptors might be shared among the majority of chemoreceptor homologs.

The transmembrane serine receptor (Tsr), one of the two major receptors in *Escherichia coli* and *Salmonella typhimurium*, is one of the model chemoreceptors commonly studied to understand the mechanisms of signal transduction in prokaryotes and lower eukaryotes. The other major receptor in *Escherichia coli* and *Salmonella typhimurium*, transmembrane aspartate receptor (Tar), is also target of many studies of signal transduction. Tar and Tsr are expected to behave very similarly and results of the dynamics from one to another are often translated with little or no loss of rigor. Tar and Tsr form homodimers [57] and multiple evidences suggest functional trimeric oligomerization made up of mixed homodimers *in vivo* [14, 22, 58]. The domain architectures of Tar and Tsr consists of a single ligand-binding perisplasmic domain and two cytoplasmic domains: HAMP and signaling domain (Figure 4). The signaling domain of Tsr, which is the focus of this work, is a four-helix bundle [14] that can be further divided in three functional modules: adaptation, coupling and protein interaction [59]. The protein interaction module is highly conserved over extended evolutionary distances due to pressure to maintain interaction with multiple interfaces [60]: homo-dimerization, trimerization and interactions with CheA [61] and CheW [20, 62]. Consequently, the activation of the kinase and the formation of the chemotaxis protein complex are directly linked to this module. The coupling module is the region that connects the adaptation and the protein interaction modules. From all three modules, this is the least conserved and its structure is predicted to be more unstable than the others [60]. In addition, knob truncation experiments established no signal-locking perturbation in the coupling module, also in contrast to the others [59]. However, the center of the coupling

module features a set of conserved glycine residues (G340, G341 and G439), forming the so called glycine hinge [63], that is considered important for proper signaling in Escherichia coli [63]. The glycine hinge is predicted to enable a bending movement of the chemoreceptor. The “bending as a mechanism of signal transduction” hypothesis is further supported by in vivo studies that show relative movement between dimers upon stimulation [26, 63-65]. Finally, the adaptation module lies above the coupling module and contains four methylation sites that are covalently modified by the enzymes CheB and CheR in Escherichia coli. The methylation sites act as a chemical memory that allows the system to adapt to new environments [26]. In Tsr, the four principal methylation sites are: Q297, E304, Q311 and E493 [66], hence the notation for the wild-type methylation state of Tsr: QEQE. The receptor’s ability to activate the kinase is dramatically influenced by the receptor’s methylation state [67, 68]. Mutagenesis studies have shown that the glutamatine side chain mimics the effect of a methyl group addition [69] allowing genetic manipulation of the chemoreceptor’s signaling state. For example, in absence of the enzymes CheB and CheR, Tsr double mutant E304Q/E493Q results in a QQQQ state that mimics all sites being methylated, which locks the receptor in “on” state, activating the kinase independently of presence of attractant. Conversely, the double mutant Q297E/Q311E results in the EEEE state that mimics an unmethylated receptor, locking it in the “off” state, deactivating the kinase.

One periplasmic sensory domain, and two citoplasmic domains, HAMP and the signaling domain. The signaling domain is divided in functional modules: adaptation module, coupling module and signaling module. For clarity only one monomer of the homodimer is shown.

To explain the kinase activation by the receptor, a working model has been proposed based on multiple studies: the yin-yang hypothesis [26, 59, 70, 71]. The model proposes that in presence of attractant, conformation changes would be transmitted to the signaling domain of the chemoreceptor and cause different structural effects in different modules. In presence of attractant, the four helix bundle packing is expected to be weakened in the adaptation module but strengthened in the protein interaction module. This represents the “off” state of the receptor, in which inactivates the CheA kinase. Conversely, in absence of signal, attractant the receptor is in an “on” state and helical packing in the adaptation module is expected to be stable while the protein interaction module to be unstable, promoting kinase activity. It is noteworthy that adaptation to a higher attractant concentration leads to the methylation of the glutamate side chains (changing E to Q), partially neutralizing the negative charges in the adaptation module which would theoretically favor stabilization of the helix. Consequently, adaptation to

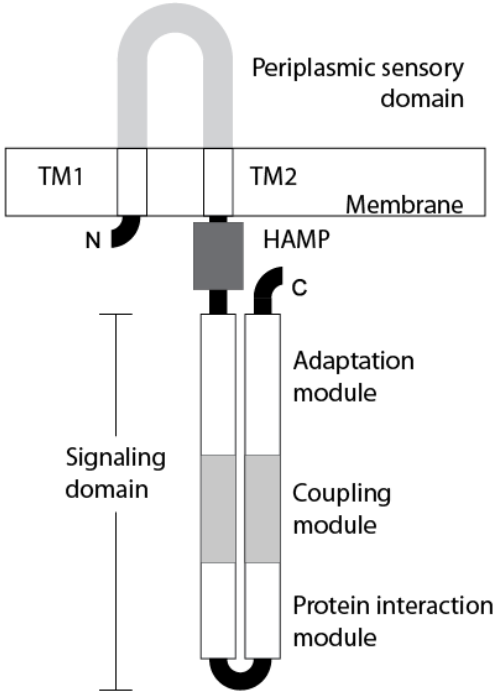


Figure 4: Scheme of the most common layout of chemoreceptor domains.

lower attractant concentration leads to the de-methylation of the glutamine side chains (changing Q to E), increasing the negative charges in the adaptation module and therefore weakening the helical packing by electrostatic repulsion. Thus, adaptation and attractant signals generate opposite effect on the adaptation module (negative feedback) as necessary for proper function of the system [59]. In addition, this model proposes the coupling module as an essential link to transmit the conformation changes between the adaptation module and the protein interaction module. Thus, the authors reject the hypothesis that the middle part of the chemoreceptor would be highly flexible as previously proposed by bioinformatics approach [60] and structural analysis [63], generating debate.

Although the yin-yang model introduces the basic concept of signal transduction in chemoreceptors, the molecular details are still poorly understood [59] despite the multiple experimental techniques applied in the past. Thus, computational methods provide a new avenue of exploration, in particular with molecular dynamics (MD) simulations methods. In the past, the periplasmic and the transmembrane domains of Tar chemoreceptors were successfully subject to different MD techniques addressing the problem of transmembrane signaling transduction[72, 73]. More recently, molecular dynamics coarse grain model was used to gain insights in the general mechanical behavior of the full chemoreceptor bound to the transmembrane [74]. In this work, we use all atoms molecular dynamics (MD) simulations to study in detail the protein dynamics of the signaling domain of Tsr from Escherichia coli in “on”, “off” and wild-type signaling states. We seek to gain insights of the structural and dynamic disparities between the different signaling states shortly after switching. We mimic the “on” and “off” states of the signaling domain of Tsr by varying the methylation state of the chemoreceptor (QQQQ for on, QEQE for wild-type and EEEE for off) via *in silico* mutations.

In molecular dynamics simulations there is a major trade off between the number of simulated atoms versus how long the simulation must be run in order to provide significant information of the system. Consequently, the technical challenges of using MD in this study are twofold: (1) the relatively large size of the signaling domain of the Tsr chemoreceptor and (2) the relatively long timescale when the conformational changes due to changes in methylation states might occur. To overcome this limitation we used Anton [39], a supercomputer specialized in all-atoms molecular dynamics simulations, to simulate 2 μs of physiological time of the signaling domain of Tsr chemoreceptor for each methylation state.

In summary, our data supports a shift in the paradigm that the flexibility of four helix bundle is related to backbone stability or atomic mobility [59, 60]. This reconciles the existence of the flexible bundle with the requirement of backbone stability for signal transduction previously conjectured in the yin yang model. Ultimately, our results suggest a novel mechanism for modulation of the kinase activity by the receptor, which is intimately related to the cis-trans conformation of the highly conserved residue Phe396.

Results

Bending properties of chemoreceptor

The 2 μs simulations of the three methylation states show statistically significant differences ($p < 0.000001$ in two-tailed Kolmogorov-Smirnoff two sample test in pairwise comparison) between the average bending angle in several layers over the period of the simulation (Figure 5). To measure local bending properties in chemoreceptors we pair residues that are equidistant of the center of the harpin turn of the chemoreceptor (residue E391) and call it a layer (see methods). Surprisingly, the regions between layers A307-E475 and I284-S494, including three methylation sites, and the region between A382-A400 and G340-M442, located right below the glycine hinge, are the most inflexible overall. In contrast, the region between layers A314-D464 and L342-E440 is undoubtedly the most variant in all three simulations and includes almost all layers with large average bending angle. These layers are directly above the glycine hinge and span the region between the adaptation and coupling modules (Figure 6).

G280 is the true hinge of the Tsr and is directed linked to signaling properties.

Remarkably, the layers containing the residues forming the glycine hinge did not exibit any especial bending property in comparison to other residues. Conversely, the layer G280-E502 showed the largest average bending angle in all three methylation states: 3.4 ± 0.1 degrees for Tsr_{QQQQ}, 2.9 ± 0.1 degrees Tsr_{QEQE} and 2.9 ± 0.1 degrees Tsr_{EEEE}. In Tar genes, the residue G280 has been mutated to cystein resulting in intradimer disulfide bonds formation [75]. A follow up study in Tar shows that G280 is the only conserved glycine to show kinase

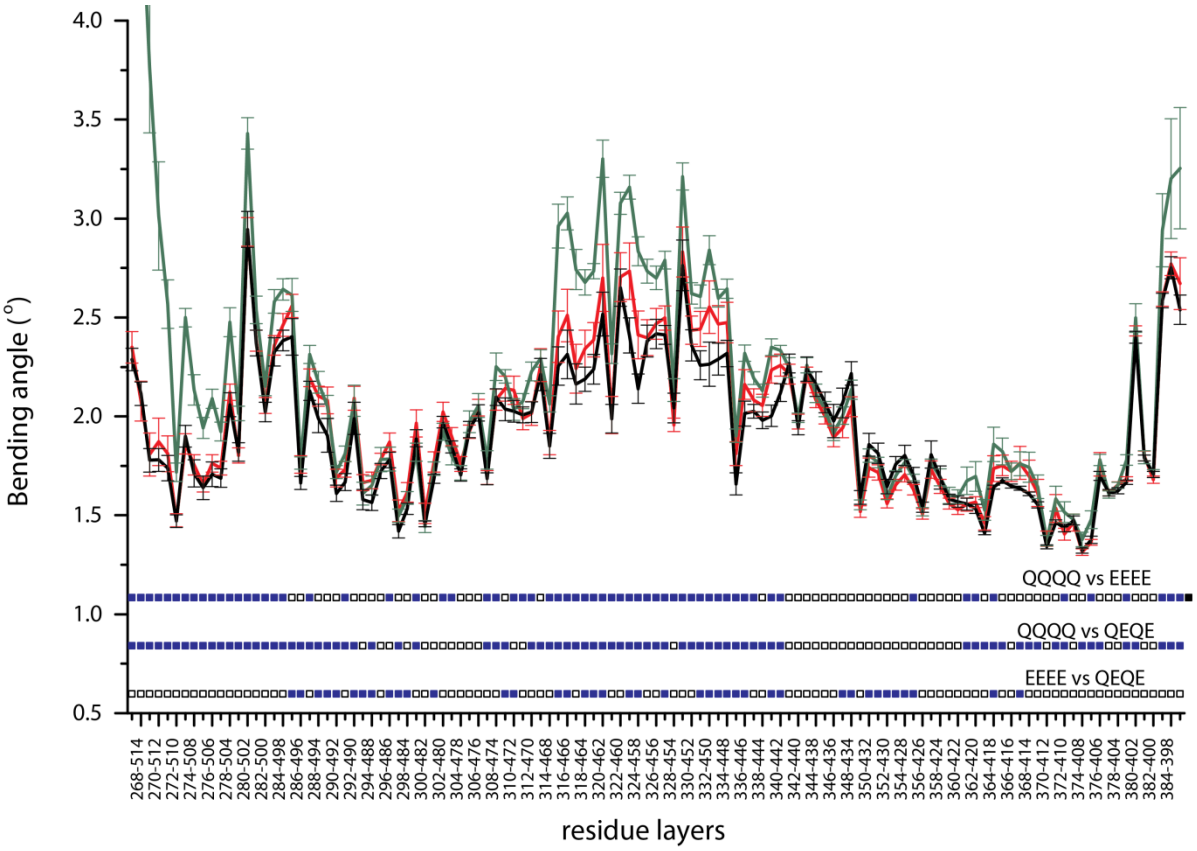


Figure 5: Profile of averaged of bending angle for each residue layer for each 2 microsecond simulation in three different methylation states: QQQQ (green), QEQE (black) and EEEE (red).

Error bars represent the standard deviation of the mean obtained with block jackknife method with N = 10 (see methods). The lines of squares represent the result of Kolmogorov-Smirnoff two sample significance test. Blue squares point the layers with significantly different mean values ($p < 0.000001$) in each pairwise comparison.

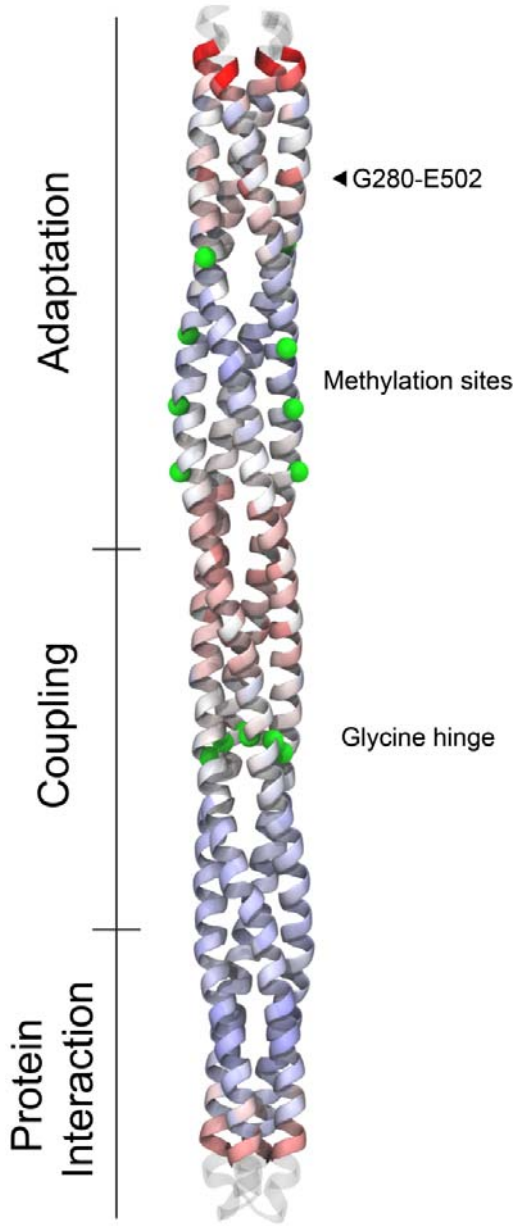


Figure 6: Mapping of the QQQQ averaged bending angle profile in the Tsr structure.

The color indicates high average bending angles (red) and low average bending angles (blue).

activity for three different modifications: G280A, G280C and disulfide formation; even in presence of attractant [63]. Activation of the kinase by G280C disulfide formation makes sense under the yin-yang model that predicts a more stable adaptation module would drive the receptor to “on” state. Nonetheless, lock-on mutant G280A and G280C in reduced state is puzzling. Substitutions to Ala or Cys in a four helix bundle are predicted to destabilize the structure and, if in the adaptation module, would switch the kinase to “off” state based on the yin-yang model.

Average of atomic mobility of alpha carbons and order parameter.

Average atomic mobility, measured by the root mean square fluctuation (r.m.s.f.) of a particular atom, is related to the relative vibrational motion of the atom and in consequence to the Debye-Waller factor, also known as B-factor or the temperature factor. Figure 7 shows the r.m.s.f calculated for each of the 2 μs simulations of the chemoreceptors in all three methylation states. Essentially, different methylation states had slightly altered r.m.s.f along the chemoreceptor structure, especially close to the methylation sites E304 and Q311; however, those differences are not as expressive as the average bending angle. The results also show that the tip of the chemoreceptor and the region between layers A314-D464 and L342-E440, presents high r.m.s.f. in comparison with the rest of the protein. In particular, the residues G457 and L475, a direct neighbor of a conserved glycine G474 [63], demonstrated a particular large mobility among residues on the C-terminus side of the chemoreceptor. In the same line, the residues G280 and its neighbor G283 also show larger atomic mobility compared to its neighbors. The overlay of r.m.s.f and average bending angle data shows an apparent correlation: flexible regions tend to be more dynamic. However, in all three simulations, the region of the glycine hinge, specifically the residues G340, E440 and A443, is particularly mobile compared to all other residues in contrast to the relatively small average bending angle of these residues. It is obvious now that the coupling between r.m.s.f. and bending angle is residual and not purely related to the dynamics. In fact, order parameter S^2 [76] of the bond vector $N^{15} - H$ of the backbone is a measure of dynamics that should be less sensitive to long range motions of slow timescale compared to the backbone dynamics. Calculations of the order parameter for all alpha carbons of the structures show a remarkable similarity between the

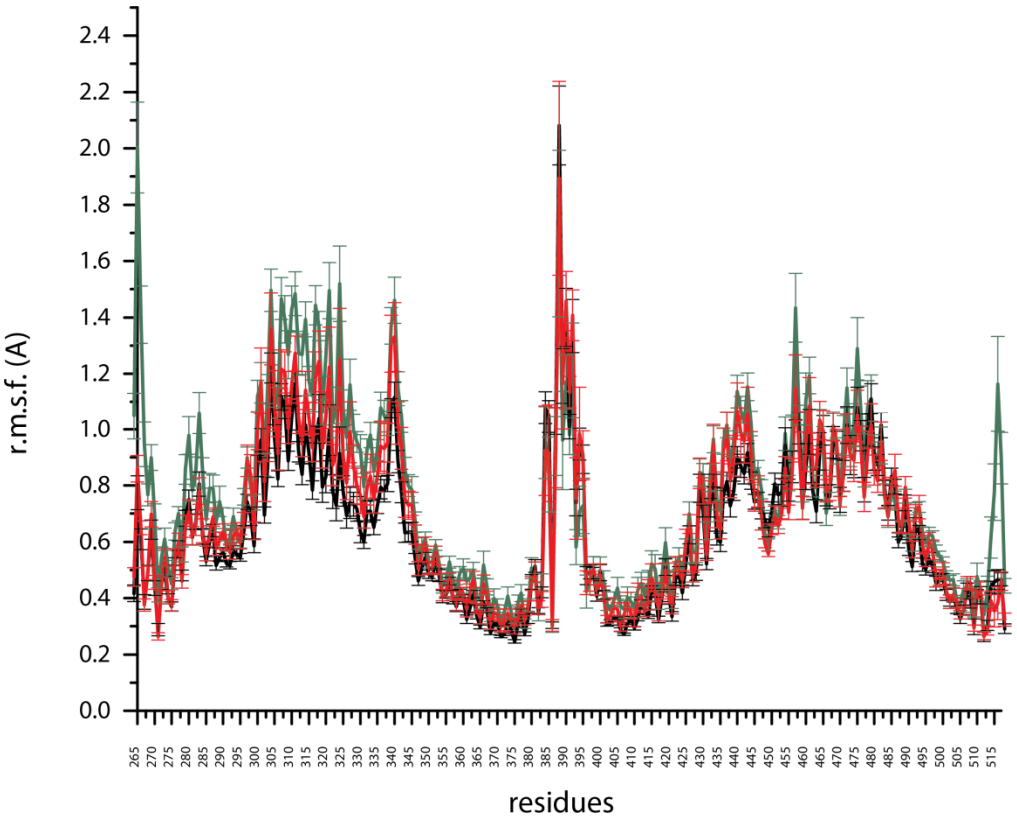


Figure 7: Profile of r.m.s.f for each residue for each 2 microsecond simulation in three different methylation states: QQQQ (green), QEQE (black) and EEEE (red).

Error bars represent the standard deviation of the mean obtained with block jackknife method with N = 10.

simulations in all three methylation states (Figure 8). Overall, the Tsr structure has a relatively rigid backbone $S^2 > 0.9$ with exception for the hairpin turn of the chemoreceptor, as expected.

However, a few isolated residues are more dynamic as exhibited by their lower order parameters. Interestingly, G341 and L342, both part of the layers forming the glycine hinge, are the most dynamic region in the chemoreceptor other than the tip. This resolves the contrast between high r.m.s.f. of the region and the low average bending angle by supporting that the region is unstable, but not necessarily flexible, which in turn argues against the coupling of dynamics and flexibility. Furthermore, residue G280 has an ordinary $S^2 = 0.9$ despite its high flexibility. In addition, residues G426 and G469, both not showing any relevant numbers for r.m.s.f. and bending angle, are also highly dynamical residues. Substitution of G469 by a cysteine causes super activation of the kinase, but maintains wild-type response if in the oxidized state [63]. As previously stated, substitution to cysteine tends to destabilize the helix, locking the kinase on and we hypothesize that disulfide bond formation would rescue the wild-type stability and therefore the signaling properties explaining the otherwise puzzling result. Interestingly, the disulfide formation of the layer G469-G469' in the mutant G469C was reportedly incomplete, which would allow some instability and therefore prevent a complete lock-off state, possibly rescuing the wild-type behavior despite the mutation.

Phe396 undergo to cis-trans conformational switch during simulations.

Despite the insight obtained from the overall mechanistic analysis described above, little to no effect was dependent of the methylation state of the chemoreceptor. However, further analysis of the highly mobile (r.m.s.f) and dynamic (order parameter) protein interacting module indicates otherwise. Analysis of the root mean square deviation (r.m.s.d.) of the protein interaction module in the simulation of the wild-type methylation shows that it oscillates between two different conformations over time (Figure 9). Astonishingly, further investigation of the trajectory for all residues in the protein interaction module indicates that only one residue is responsible for the oscillation of the entire module: the Phe396. Neighboring residues show correlated r.m.s.d. oscillations but in a much lower value, which suggests that they are secondary effects from the movements in the Phe396. During the simulation of the chemoreceptor homodimer, the Phe396 of monomer A interacts with the equivalent residue of monomer B, Phe396'. This interaction is also present in the Tsr X-ray crystal structure (PDB code: 1QU7) [14] as shown in Figure 10.

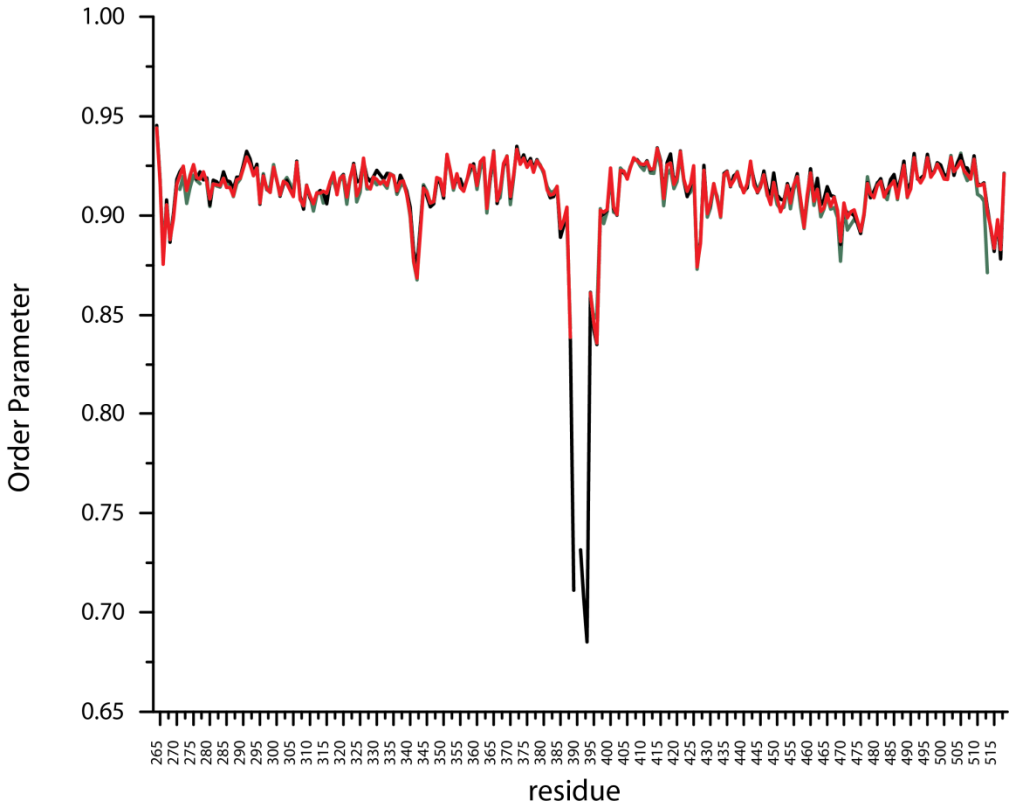


Figure 8: Profile of the order parameter (S^2) calculated from each simulation of the chemoreceptor in three different methylation states: QQQQ (green), QEQE (black) and EEEE (red).

Error bars represent the standard deviation of the mean obtained with block jackknife method with

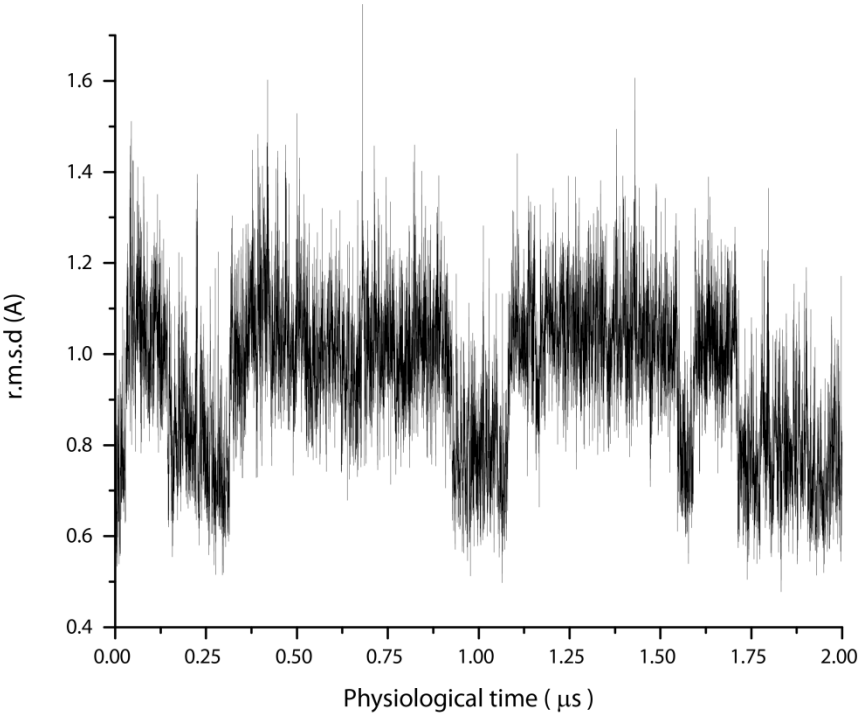


Figure 9: Temporal analysis of the protein interaction module mobility (r.m.s.d.) shows oscilation between two stable conformations.

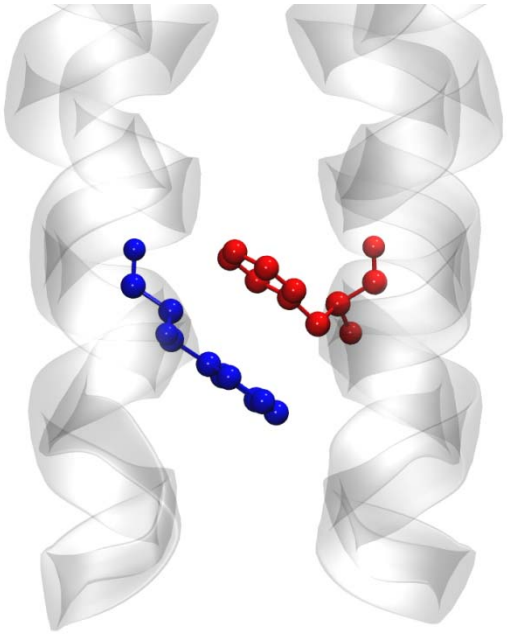


Figure 10: Interaction between Phe396(blue) and Phe396'(red) in the Tsr X-ray crystal structure (PDB code: 1QU7) [14]

Specifically, the oscillation of the r.m.s.d of the protein interaction module is correlated with the position of the X_1 dihedral angle (formed by the dihedral angle of the N, Cα, C$_\beta$ and Cγ atoms) of the residues Phe396 and Phe396' (Figure 11). The Phe396 residue is paired with Val384 to form the 5th layer above the tip. Locally, the cis-trans switch implicates in a "flip" between residues Phe396 and Phe396', where before the switch Phe396 is further away from the tip of the chemoreceptor and after the switch it becomes closer to the tip and vice versa for Phe396'. Analysis of the trajectories shows that the cis-trans switch occurs in approximately 4 ns. Visualization of a switching event in the wild-type chemoreceptors is shown in Figure 12.

To further study the effects of the cis-trans switch in the overall structure, the distances between the alpha carbons of Phe396, Phe396', Val384 and Val384' were measured during the simulation. Val384 and Phe396 are found in the same layer of the four helix bundle. The first is in the N-terminus helix (N) and the second is in the C-terminus helix (C) of each monomer. Herein, Phe396/Phe396' and Val384/Val384' are described as C/C' and N/N', respectively, according to the helices in which they are found. Therefore, the relative position of the residues towards each other can be also interpreted as the distances between the four alpha helices composing the four helix bundle. Surprisingly the only deformation occurring during the cis-trans switch is the distances between the helix C and N' and between the helix C'-N, all other distances remains constant during the simulation. A summary of the temporal evolution of these measurements is shown on Figure 13.

Distances between helices were measured as the alpha carbons distances between residues Phe396 from the helix C, Val384 from the helix N, Phe396' from the helix C' and Val394' from helix N'. Over the course of the simulation, only distances between C-N' and N-C' varies as the cis-trans C1 dihedral angle. This suggests that the consequence of the cis-trans oscillation of Phe396 to the overall structure is the variation of the relative position of the helices C and N' and the relative position of the helices N-C'.

Cis-trans switch of Phe396-Phe396' is methylation state dependent

In the wild-type simulation, the Phe396 oscillates between the two stable conformations cis and trans. In contrast, the switch between conformations is slightly less frequent in the simulation of the chemoreceptor in the "on" state, QQQQ, revealing a preferencial trans conformation of the Phe396. Simulations of the chemoreceptor in the "off" state, EEEE, also

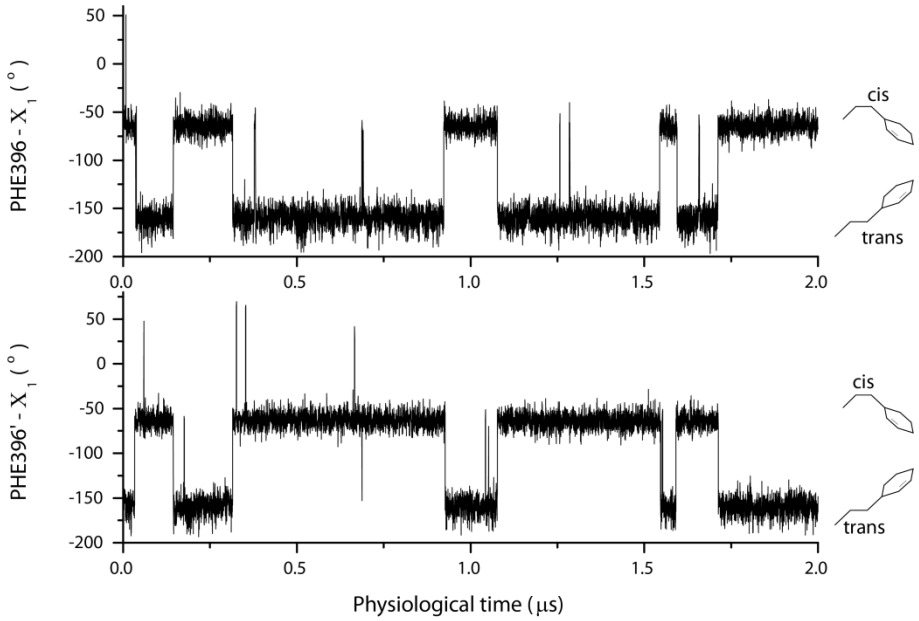


Figure 11: Temporal evolution of the X1 dihedral angle of the interacting residues Phe396 (above) and Phe396' (below)

Comparison between the temporal evolution of the X1 dihedral angle and the r.m.s.d of the entire protein module Figure 9 shows that the switching between cis-trans is correlated to major configuration changes in the protein interaction region in the chemoreceptor

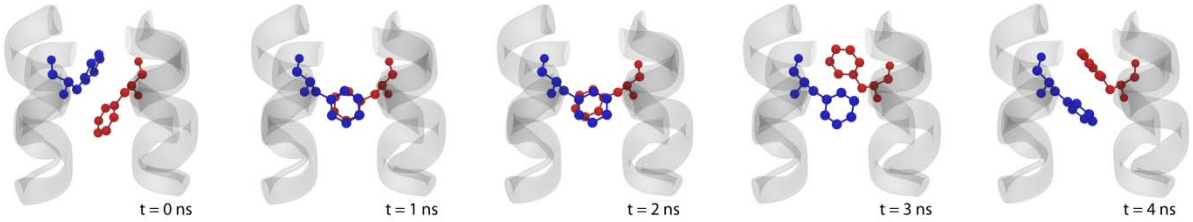


Figure 12: Time lapse of a section of a chemoreceptor simulations in the wild-type methylation state highlighting the conformation switch of the Phe396(red)-Phe396'(blue) pair.

Phe396 starts in a cis conformation (right) and then switches to a trans conformation (left) over the course of 4ns.

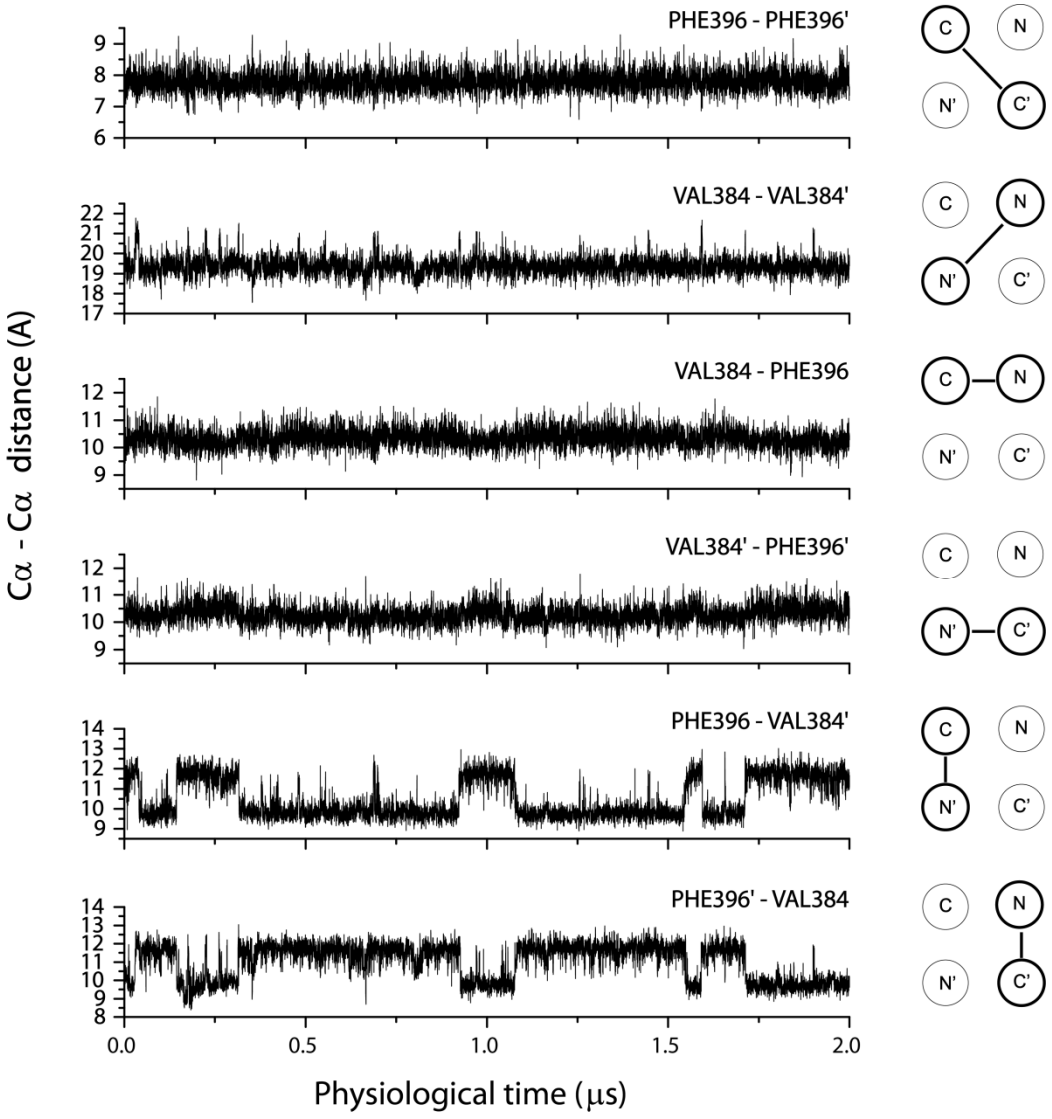


Figure 13: Temporal analysis of the pairwise distances between the helices: C, N, C' and N'.

reveal a preferential conformation; however, it is a cis conformation preference in contrast to the QQQQ state simulations. This stark divide is evident in the distribution of the χ_1 dihedral angles observed during the simulations (Figure 14).

Phe396 is the most conserved residue in chemoreceptors.

One of the central paradigms of molecular evolutionary biology is that residues conserved over long evolutionary distances are important for the function of the protein, either by contributing to enzymatic activity, serving as docking site for an interacting protein or even to maintain the tertiary structure [56, 77]. The protein interaction module of the chemoreceptor is one of the most conserved sequences in nature, and this conservation is attributed to the multi-faceted nature of its interactions [60]. Analysis of 7,809 non-redundant sequences of chemoreceptor fetched from complete genomes in the MIST database in August 2012, reveals the Phe396 as the most conserved residue from in the chemoreceptor protein family (Figure 15). This suggests that if the Phe396 cis-trans switching is in fact the mechanism of kinase activation by the chemoreceptor, then the mechanism is conserved in all bacterial chemotaxis systems.

Discussion

Mechanical properties of the chemoreceptor structure

Previously, functional modules of the chemoreceptor signaling domain were identified based on features of the multiple sequence alignment [60]. Here we show that the previously described flexible bundle [60], which was later defined as the coupling module [59], shows large average bending angles, although our experiments define new boundaries for this module. We find the largest bending angles between the layers A314-D464 and L342-E440 in contrast with the much larger region previously predicted [60] (Figure 5 and Figure 6). The adaptation module that serves as a substrate for the adaptation enzymes, cheB and cheR, is one of the most rigid regions of the chemoreceptor together with the protein interaction module where CheA and CheW are suppose to bind to the chemoreceptor [26] . Remarkably, while local stability is not required for protein-protein interaction [78], we found that bending properties of the four helix bundle seems correlate with the chemoreceptor's binding sites. Therefore, since this intermediate region of the receptor has no special properties but the high average bending angle, we agree with the flexible bundle nomenclature.

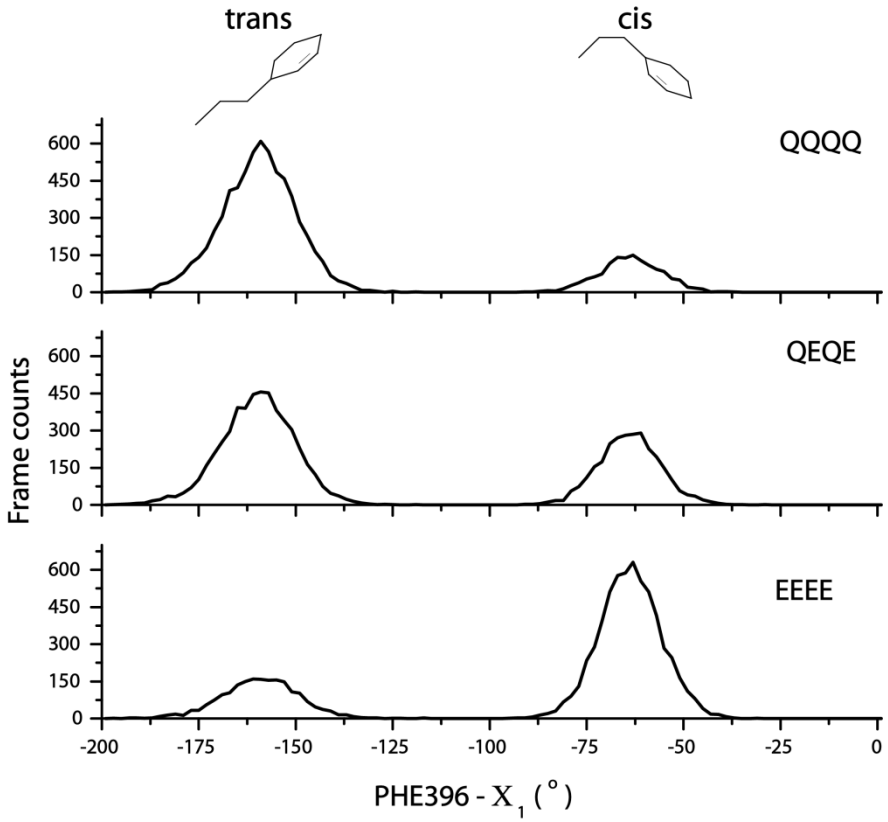


Figure 14: Comparison of the distributions of X1 dihedral angle in Phe396 during simulations of the chemoreceptor in three states: QQQQ, QEQE and EEEE.



Figure 15: Sequence logo of the chemoreceptor protein interaction module.

The Phe396 (designated by an asterisk) is the most conserved residue in the entire family of chemoreceptors.

In addition, the results presented here do not support a bending role for the glycine hinge and suggest the region should be termed as glycine bundle. The conservation of these glycine residues together with the unique high plasticity of their backbone, relative low order parameter and high r.m.s.f., still supports that the glycine bundle is important for the proper function of the chemoreceptor, but not as a highly bendable hinge. In contrast, the layer G280-E502 appears to be the true hinge of the chemoreceptor having the highest average bending angle in all three simulations. These residues have not being attributed to have a bending role before since it is present in the adaptation module that was thought to be rigid. In addition, previous methodologies applied to measure bending angle were subject to cumulative bending of adjacent layers [63, 74] misleading the measurement.

The combination of average bending angle, atomic mobility and backbone dynamics, explains the apparent conflicting result of mutants targeting G280 [63] and the yin-yang model [59]. We speculate that substitutions in position G280 could compromise the bending properties of the layer causing an increase in stability and consequently locking the kinase in on state. This hypothesis is also in line with the idea that replacing the glycine with a disulfide bond would lead to loss of bending properties which in turn stabilizes the local helical packing. In addition it is noteworthy that this relationship is not symmetrical: increasing stability in neighboring residues does not necessarily affect bending capabilities of a given layer; however impairing bending capabilities of a given layer must contribute to stabilization of the region. Another factor to consider is that residue E502, which is a member of the layer G280-E502, is also known to be a minor methylation site [66]. Changes from Glu to Gln can alter the local bending properties of the layer given the large number of charged residues in the adaptation module [79] serving as a fine tune to receptor adaptability. In light of these results, we show that there is an intrinsic connection between receptor bending properties, helical stability and kinase activation. However, it does not necessarily mean that highly flexible regions are unstable and largely mobile. In other words, the relationship between these three mechanical properties seems to depend on several factors such as the biochemical composition of the local environment. Also, no substantial change in any of the three mechanical properties measured here could be correlated to differences in methylation states.

Cis-trans switch of conserved Phe396 suggests novel kinase activation mechanism

We identified two stable conformations of the Phe396-Phe396' interacting pair (Figure 10). Our results show that switching from one conformation to another takes approximately 4ns, however the switching event is rather rare (Figure 12). The only significant consequence of the Phe396 conformational switch in the overall structure is the variation between the relative position of the helices C and N' and between the helices N and C' (Figure 13). Surprisingly, the Phe396 conformational switch does not alter the other intra-monomer helix distances (C-N, C'-N') nor the inter-monomer distances of the equivalent helices (C-C' and N-N'). Figure 16 shows a representation of the two configurations of the four helices depending on the conformation of Phe396. If monomer A (C and N) is exactly symmetrical to monomer B (C' and N') then the two configurations from Figure 16 are the same if monomer A and monomer B are switched and rotated 180°. We work with the hypothesis that monomer A and B cannot be perfectly symmetrical given the chaotic nature of biological systems and that the two conformations are indeed not equivalent. Thus, our results show that during the simulation of the chemoreceptor in the wild-type methylation state, the pair visits the two conformations equally suggesting energetic equilibrium between the two conformations in this methylation state. Interestingly, the equilibrium is shifted in different directions for different methylation states Figure 14 and Figure 17 .

However, recent experimental evidences show that the quaternary structure of the chemoreceptor, the trimer of dimers [12], is necessary for successful modulation of the kinase activity [23]. An important question is what does these two conformations mean for the trimer of dimer context? Figure 18 shows a prediction of two conformations assumed by the trimer of dimer in case of all Phe396 in cis and trans state (panels A and B, respectively). Analysis of the Tsr trimer of dimer crystal structure further supports our predictions (Figure 19). The structure's snapshot shows the trimer with the Phe396 in cis state and provides supports for the predicted conformation for the Phe396 trans state as well. Surprisingly, the residues from the helix C' are not in close contact with the residues from the neighboring helix N which facilitates the ejection of the helix C' in trans. In this case, we also speculate that the ejection of C' helix would allow for the three helices N to come closer together. In this scenario, this hypothesis can be tested with crosslinking experiments on residues V398 and R394 for chemoreceptors in "on" and "off" signaling state.

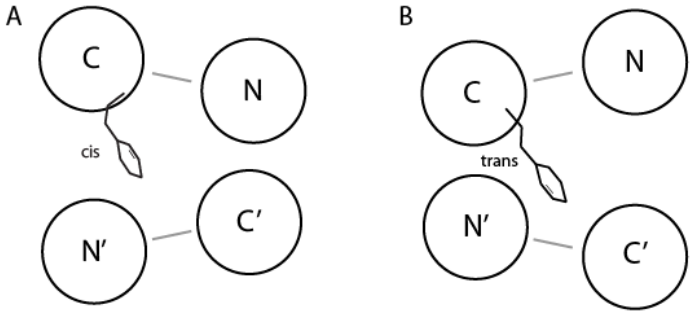


Figure 16: Geometry of the tip of the receptor in cis conformation of the Phe396 (A) and trans conformation of the Phe396 (B).

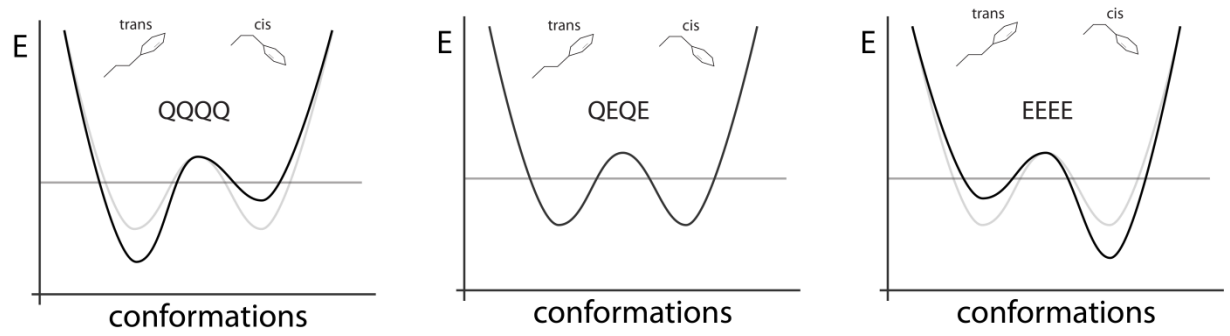


Figure 17: Working model of the Phe396 conformation preference based on molecular dynamics simulations.

The potential energy is shown in the y-axis. In the wild-type, both conformations are equally preferable (central panel). In the QQQQ and EEEE methylation states (left and right panels, respectively) the wild-type Phe396 equilibrium conformation (gray) shifts to a preferential conformation (black)

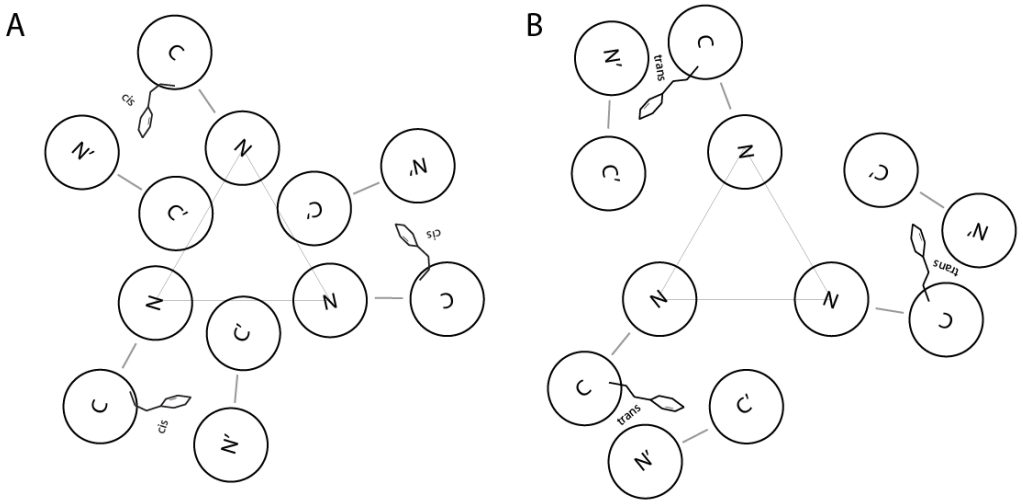


Figure 18: Predicted geometries of the trimer-of-dimer for cis (A) and trans (B) conformations of the X_1 dihedral angle of the Phe396.

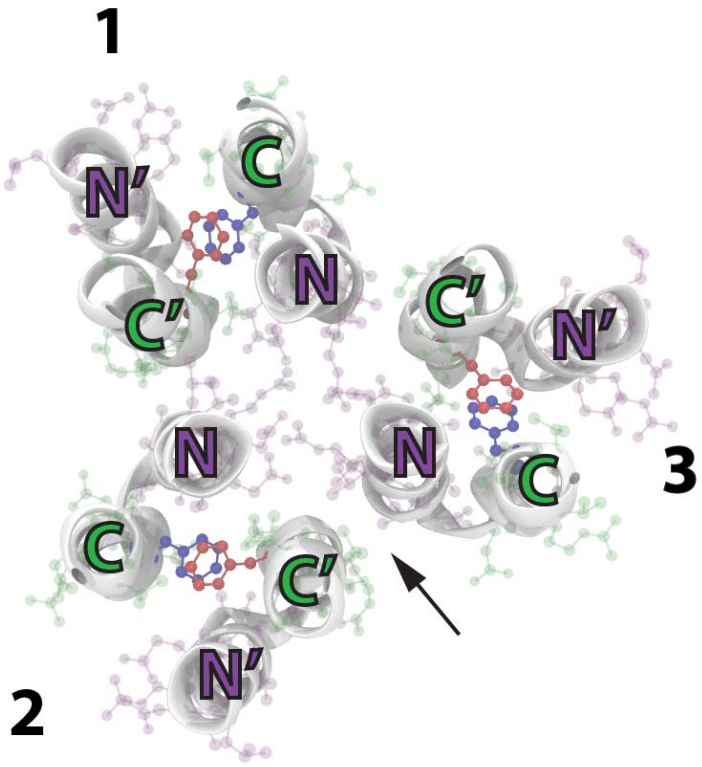


Figure 19: Top view of the protein interaction module of the Tsr trimer of dimers crystal structure.

Residues from the N – terminus (purple) are responsible for the trimerization contacts in only one monomer of each the dimer (1, 2 and 3). Remarkably, the residues from the C – terminus (green) do not seem to make a strong connection to residues from the neighbor dimer helix N (black arrow). This provides further support that for the two states conformation of the trimer of dimers suggested in Figure 18.

Conclusion

Using a specialized supercomputer, Anton, we performed three 2μs simulations of the full signaling domain of the chemoreceptor Tsr from the *Escherichia coli*. Each simulation had the chemoreceptor in a different methylation state that mimic three distinct signaling states: “on” (QQQQ), “off” EEEE, and wild-type (QEQE), which is an equilibrium between “on” and “off” states, as assumed by a two-state model of chemotaxis [26]. This analysis provided means to use structural features common in all simulations to redefine the boundaries between that were previously characterized by sequence analysis [60] (Figure 20.)

The adaptation module begins at the start of the chemoreceptor signaling domain and ends at the layer A314-R468. This region contains the methylation sites responsible for adaptation Q297, E304, Q311 and E493, in Tsr *E. coli*). It is remarkably rigid in terms of average bending angle. In addition, this region contains the layer G280-E502 that shows remarkable bending capabilities and is the true hinge of the chemoreceptor. This finding reconciles the conflicting results of the *in vitro* data on kinase activity of various mutants [63]. Specifically, G280A and G280C in reduced and oxidized state which lock the kinase in “on” state instead of lock the kinase in “off” state as expected by the yin-yang model [59]. We hypothesize that any mutation in this layer affects its special bending properties resulting in a lock-on behavior of the receptor. In other words, we believe the mutations that were originally thought to destabilize the helix have the opposite effect and jam the chemoreceptor hinge, stabilizing the helix. The high conservation of this glycine in close homologs also supports the importance of the position for the function of the chemoreceptor [63].

The coupling module, more appropriately called here the flexible bundle, links the adaptation module and the protein interaction module. We show that this region is highly flexible as predicted by Alexander and Zhulin [60] and yet fairly stable as conjectured by the yin-yang model [59], reconciling the conflicting evidence. Here, we introduce the concept that the bending capabilities and backbone dynamics are not always correlated. Furthermore, our results show that measurements of atomic mobility are a combination of these two measurements and possibly other structural movements such as torsion and stretching, which were not explored in this work. The boundary between the flexible bundle and the protein interaction module is the so called glycine hinge, more precisely the layer G341-T441. However, our results indicate that the region containing the glycines G340, G341 and G439, forming the glycine hinge, does not bend more than other regions of the chemoreceptor. On the other hand, this region exhibits high

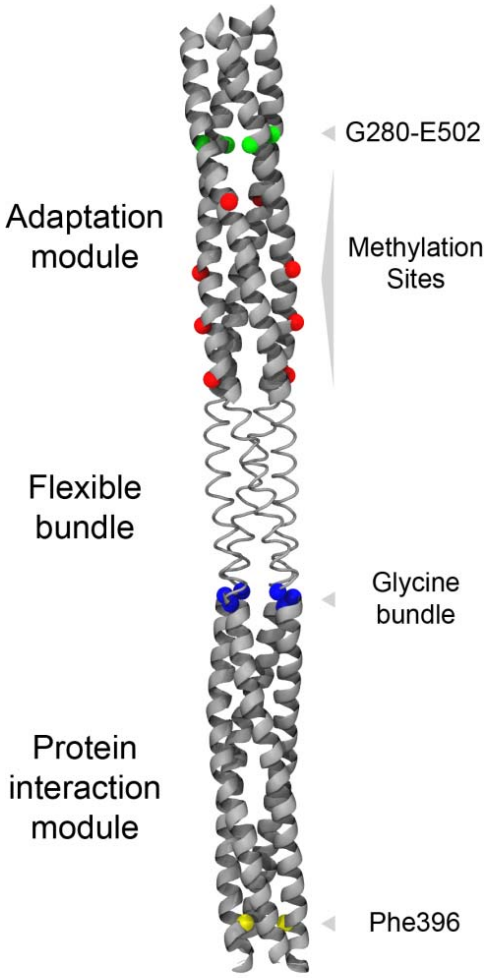


Figure 20: The three distinct regions of the Tsr receptor: adaptation module, flexible bundle and protein interaction module.

Features of interest are highlighted: the G280-E502 the hinge of the receptor (green), the methylation sites (red), the glycine bundle (blue) and the central residue in the model for kinase activation the Phe396 (yellow).

atomic mobility given the plasticity of their backbones as shown in lower order parameter relative to the rest of residues. Therefore, we propose to rename this region the glycine bundle based on our evidence that the previously described “hinge-like” observations[63] are related to cumulative bending in adjacent layers and not to the region itself.

The protein interacting module can be clearly divided in two regions. The “upper” part of the protein interaction module, roughly from the layer L380-E402 to the glycine bundle, is well characterized structurally by our methods. Our results indicate that the region is inflexible to bending and has a highly stable backbone (with the exception of residue G426). In contrast, the “lower” part of the protein interaction module, below the layer L380-E402 is extremely dynamic and hints towards a more complex role than simply steady protein interaction.

We further investigated the most conserved residue in the extensive chemoreceptor family, Phe396, which is located in the center of the dynamic part of the protein interaction module. Our results show that during the simulations the residue transitions between two stable and distinct conformations: the cis and trans conformations of its χ_1 dihedral angle. We found that over the course of the simulation of the receptor in the wild-type methylation state, the Phe396 assumes both cis and trans conformations equally. In contrast, Phe396 exhibits opposing preferential conformation states in QQQQ and EEEE simulations (trans and cis, respectively). The switch from cis to trans of Phe396 significantly affects the overall geometry of the dynamic part of the protein interaction module by altering the distances between helix C and N’ and C’ and N Figure 16 . In the context of trimer-of-dimers this switch would putatively change the topology of the allegedly interacting sites of CheW and CheA regulatory domain. Since each methylation state represents a different signaling state, we suggest a novel molecular mechanism of the CheA kinase regulation where the cis-trans switch of the Phe396 is directly responsible for the modulation of kinase activity.

In our hypothesis, wild-type chemoreceptors have an equal likelihood of existing in either conformation, which has a net effect of an equal ratio of “on”:“off” chemoreceptor along with and active:inactive CheA kinases, accordingly. Changes in the adaptation module, either by (de)methylation of the adaptation sites or by change in concentration of the attractant, alter the energy landscape of the dynamic part of the protein interaction module making one of the conformations energetically more favorable than the other Figure 17. This in turn changes the ratio of “on”:“off” chemoreceptors and active:inactive CheA kinases, and therefore, each

conformation is associated to a different signaling state. How changes in adaptation module alter the energy landscape of the protein interaction module is still unknown.

Overall, the results presented here support a novel mechanism of kinase activation, only revealed by use of computational methods of molecular dynamics simulations. As the switch occurs in 4 ns, the changes in conformation are rather rare. Ideally, the extension of such simulations is desired although the results here are enough to provide a testable hypothesis for the mechanism of kinase activation.

Methods

Simulation system. The X-ray crystal structure of the Tsr chemoreceptor in QQQQ methylation state deposited in the Protein Data Bank (PDB code: 1QU7) is not fully resolved. However, the authors built a complete model based on the X-ray crystal structure and cross linking data [14]. Water molecules trapped in the 1QU7 were transferred to the model, total of 120. The model was truncated at the residues 263 to 519, the coordinates around the limits of the signaling domain [14, 60, 80]. The structure was embedded in water, tip3p, neutralized and 0.05mM of NaCl was added. The total simulation system size was 144,647 atoms (90x90x182 $Å^3$). To keep the receptor in place during the simulations we added a 50 kcal/mol/$Å^2$ restrain in the backbone of the residues 263 and 519 and one 25 kcal/mol/$Å^2$ in the backbone of the residues 264 and 518.

Simulations. We performed a 50ns simulation with the engine Desmond in the Newton supercomputer at University of Tennessee using 512 nodes for pre-equilibration of the system in NPT ensemble with Berendsen at 300K constant temperature and 1 atm pressure. The system was then transferred to the 512 node, special-purpose supercomputer, Anton where a one 1μs simulation was performed to assure equilibration of the entire structure. Copies of the last frame of this simulation were mutated to change the methylation states of the structure: Q304E and Q493E to build QEQE structure and Q297E, Q304E, Q311E and Q493E to build EEEE structure. Waters and ions were added as needed to restore minor changes in density and neutralize the system. Local minimization was performed for 8 steps in the recently mutated side chains on Maestro. The velocities were initialized on Desmond prior to be transferred to Anton. Each of the three production simulation was 2μs long. All simulations used CHARMM27

forcefield, NPT ensemble, 300K, 1 atm and Berendsen integrator. Long range electrostatics interactions used Gaussian split Ewald with a 64 x 64 x 64 FFT mesh. Short range and van der Waals interactions were cut off at 16.75 Å. The simulations time step was 1 fs and respa 1:1:3 meaning that long-range interactions were calculated every third step.

Calculation of the average bending angle. To measure local bending properties in chemoreceptors we pair equidistant residues of the center of the harping turn of the chemoreceptor (residue E391) and call it a residue layer. For example the 10^{th} residue from the center of the harping turn E391 towards the N-terminus is the residue N381 which is paired to the 10^{th} residue towards the C-terminus G401 to for the layer E391-G401. The angle between the largest component of the principal axis of inertia calculated for the alpha carbons of the four layers above the target layer and below the target layer is then denoted bending angle (Figure 21). The calculations was performed using the function “measure inertia” from VMD [81]. This strategy aims to minimize coupling between other movements such as shear, torsion or stretching that might appear as bending, as well as misleading measurements by cumulative bending of adjacent layers in a given frame, as occurred in [74]. A time series of the bending angle was extracted for each layer and averaged over time for each production simulation. The error bars are the standard error calculated by the jackknife procedure [82]. More specifically, by the delete-a-group procedure as explained in reference [83] using N = 10. We used the Kolmogoroff-Simirnoff two sample test as the test of significance since the test is sensitive to shape and average of the distributions compared.

Local alignment per residue protocol for calculations of the r.m.s.f and order parameter.

In molecular dynamics, both r.m.s.f and order parameter calculations need a reference frame. R.m.s.f calculations can be performed against any frame of the simulation or even to an average structure of the entire ensemble assuming that the system is ergotic. The order parameter, as it is a measurement of the asymptotic value of a correlation function has to be performed against the initial frame (see methods below). The current methodology to calculate these values assume that the frames of the simulations have been aligned to the reference frame to avoid coupling between rotational and/or translational movements to the dynamics being measured. This procedure works well for globular proteins but it fails in the case of multidomain structures and/or largely anisotropic structures such as the chemoreceptor. To

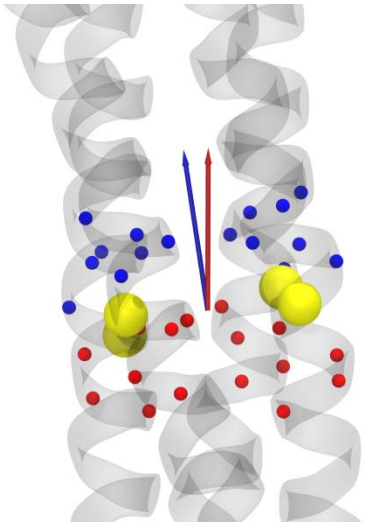


Figure 21: Schematics of the calculation of the bending angle.

The target layer is represented by the yellow spheres. The largest component of the principal axis of inertia of the alpha carbons in the superior 4 layers (blue spheres) is compared to the one calculated from the lower 4 layers (red spheres) as represented by the colored arrows. The angle between the arrows is called the bending angle

overcome this problem we suggest a procedure much more computationally intensive but that eliminates the problem of frame alignment in anisotropic structures that we call: local alignment per residue protocol. As the value is calculated for each residue, each frame is aligned to the reference frame using only atoms within a certain threshold from the target residue. This custom selection of atoms filters long range movements of parts of the structure that might take place during the simulation. In both R.m.s.f and order parameter the cutoff chosen was of 30 Å. We generally call this procedure as local alignment per residue protocol.

Calculation of the r.m.s.f. The r.m.s.f. here is calculated for the alpha carbon of each residue as the formula:

$$R.m.s.f(i) = \frac{1}{T}\sum_t^T \langle \vec{r}_i(t) - \overrightarrow{r_{i_{ave}}} \rangle^2 \quad 10$$

where T is the total number of frames, $\vec{r}_i(t)$ is the position of the alpha carbon of residue i in the frame t and $\overrightarrow{r_{i_{ave}}} = \langle \vec{r}_i(t) \rangle$ is the average position of the alpha carbon of residue i. The error bar is the standard error also calculated by delete-a-group jackknife procedure as in the average bending angle.

Calculation of the order parameter. The order parameter is defined as [76, 84-88]:

$$S^2 = C_I(\infty) = \frac{1}{T^2}\sum_{t=0}^{T/2}\sum_{\tau=0}^{T/2} P_2\big(\hat{\mu}(\tau)\cdot\hat{\mu}(t+\tau)\big) \quad 11$$

where $C_i(\infty)$ is the internal correlation function when $t \to \infty$. Also, t and τ scans over the sequence of frames, $\hat{\mu}$ is the unit vector pointing along the backbone ^{15}N-H bond. $P_2(x) = \left(\frac{3x^2}{2} - \frac{1}{2}\right)$ is the second Legendre polynomial. The equation X requires a convergence of $C_i(t)$ as t increases. To verify the convergence, we calculate the correlation function as:

$$C_I(t) = \langle P_2\big(\hat{\mu}(0)\cdot\hat{\mu}(t)\big)\rangle \quad 12$$

then we define C_{tail} as the average of the values of the last 0.5 ns of the correlation function. Convergence is assumed if $|C_I(\infty) - C_{tail}| < 0.005$ as proposed before [85]. If there is no convergence, the order parameter is considered null.

Bioinformatics. We selected all 12,498 chemoreceptor sequences from complete genomes in the MIST database as in August 2012 [89]. Using HMM models previously published [60], the chemoreceptors were classified and separated in different files according to its heptad classes

using HMMER [90]. From this set, 2,312 sequences were excluded from our analysis by not matching any of the heptad classes. For each file, the MCPsignal PFAM model [91] was used to only select the region of the protein matching the PFAM definition of the signaling domain. Each file was independently aligned using MAFFT [55]. To avoid bias, we excluded sequences 98% identical. Also, 46 sequences were removed for the reason of being incomplete in the region of interest. Finally, the MSA of each heptad class was manually trimmed to include only the closest 4 heptads from the hairpin turn from the N-terminus and the C-terminus, total of 8 heptads or 57 residues. In Tsr number the region selected is from D363 to S419. The sequence logo with the information content, which in turn indicates the amino acid distribution of each position of the MSA was built using the software Weblogo [92].

CHAPTER II THE ROLE OF A CONSERVED SALT-BRIDGE IN THE CHEMOTAXIS COMPLEX

This chapter was taken from a manuscript in preparation:

Davi Ortega, Guoya Mo, Kwangwoon Lee, Hongjun Zhou, Jerome Baudry, Frederick Dahlquist, Igor Zhulin. **Investigating structural properties of CheW with molecular dynamics and NMR**. Manuscript in preparation

Conceived and designed experiments: DO JB FD IZ. Performed experiments: DO GM KL HZ. Analyzed the data: DO GM KL HZ JB FD IZ. Wrote the paper: DO GM KL JB FD IZ.

Abstract

The prokaryotic chemotaxis system is one of the best studied signal transduction pathways in nature. Recently, a comprehensive evolutionary study revealed that the birth of a single component, CheW, led to the divergence of the chemotaxis system from simpler two component systems. CheW increases the binding affinity between the receptors and the kinase in the chemotaxis complex, which supports its role as scaffold and possibly the main promoter of chemotaxis lattice formation. However, *in vitro* and *in vivo* experiments targeting a highly conserved position Arg62 in *E. coli* suggest that CheW may play a more complex and dynamic role in chemotaxis given its null phenotype, despite showing only small changes in binding affinity with the kinase and the chemoreceptor. Here we show that the Arg62 form a salt-bridge interaction with Glu38 that is of fundamental importance for the signaling mechanism in bacterial chemotaxis. By means of a total of ~ 2.7 μs of multiple molecular dynamics simulations we establish the salt-bridge formation between Arg62 and Glu38 in wild-type CheW. We found that disruption of this interaction affects the structure of the first sub-domain which, in turn, affects the overall stability between the binding sites for the chemoreceptor (second sub-domain) and the kinase (first sub-domain). NMR experiments shows that the mutation R62A only introduces local changes in CheW structure but relaxation dispersion analysis suggests that the mutation will increase the dynamics of the second sub-domain. Taking these results together, we provide a significant step towards a better understanding of the kinase activation by the chemoreceptor and place CheW protein not as a simple scaffold protein but as a component with an active role in signal transduction.

Introduction

Signal transduction is a fundamentally important process for all living organisms. While the majority of signal transduction events in prokaryotes are carried out by relatively simple one- and two-component systems [6, 93, 94], the pathway that controls chemotaxis and other cellular functions in bacteria and archaea [10, 11] appears to be as complex as some eukaryotic receptor-kinase cascades. The *Escherichia coli* chemotaxis pathway employs dedicated chemoreceptors that are anchored in the membrane and detect signals from both outside and inside the cell [26]. Chemoreceptors relay this information to the CheA histidine kinase, which then communicates the information to its cognate response regulator CheY. In a phosphorylated form, the CheY protein binds to flagellar motors causing a change in the direction of its rotation, thus converting the initial signal detected by chemoreceptors into a behavioral response – a change in the swimming direction. This pathway also employs receptor modifying enzymes CheB and CheR as well as a CheZ phosphatase that, acts on CheY [7].

The key features of this remarkable system include high sensitivity, wide dynamic range, signal integration, memory, and precise adaptation [67, 95-98] all of which are consequences of a highly ordered arrangement of chemoreceptors and kinases at the cell pole [12, 95, 99]. The geometry of a hexagonal array with a lattice spacing of 20 nm is conserved over long evolutionary distances [12], indicating the importance of relevant positions between interacting members of the complex. In addition to chemoreceptors and the CheA kinase, this complex also contains the CheW protein, which is usually referred to as a docking, coupling or adaptor protein [100-102].

Three structures of CheW have been resolved for three organisms: *Thermotoga maritima* [100], *Escherichia coli* [103] and *Thermoanaerobactor tengcongensis* [104]. The CheW fold is composed by two five-stranded β-barrel connected by a hydrophobic core [100]. CheW is needed for proper kinase activation by the chemoreceptor [105] since it is necessary to the formation of the chemoreceptor-CheW-CheA complex [106]. Also, overexpression of CheW leads to impairment of chemotaxis [62] since it disrupts trimer formation in chemoreceptors by blocking trimer contacts in the receptor [102, 107, 108]. The binding sites of the kinase CheA and the chemoreceptor have been mapped on CheW in a series of publications [102, 107, 109-111]. Although the overall results of binding assays suggest a modest role of simple scaffold for CheW, the chemotaxis inhibitor mutant R62H challenges that perspective. While only

moderately affecting in vitro binding affinity for both the receptor and the kinase, the mutant has null phenotype for chemotaxis [110]. This astonishing result suggests that CheW might have a bigger role in chemotaxis system yet to be uncovered [110].

In this study, we address this problem by using a combination of comparative sequence analysis, NMR and computational molecular biophysics, and analyzing results in the context of available structural, biochemical and genetic data. This approach revealed an evolutionary conserved salt bridge on the surface of CheW that is responsible for maintaining a specific geometry within the signaling complex. Overall, our results show that CheW is not just molecular "glue" but a highly dynamic protein that might have a "pro-active" role in signal transduction.

Results

Co-evolving class-specific residues in CheW form a short-range salt bridge

Residues in proteins that are conserved over long evolutionary distance play the most critical roles in their structure. Because positional conservation is strongly affected by functional diversification during the evolution of the protein family, it is important to avoid entangled information in search of function-specific residues. CheW is a single-domain protein; however this domain, also termed CheW, is present in some other multi-domain chemotaxis proteins, such as CheA [112] and CheV [113]. However, diversification within the CheW protein family may not be limited to its presence in functionally distinct proteins. The chemotaxis signal transduction pathway originated early in the evolution of bacteria and diversified into many distinct classes, in which a repertoire of interacting proteins can be quite different [10]. For example, in F1 class exemplified by *Bacillus subtilis*, CheW protein interacts with chemoreceptors that are structurally different from those in F7 class exemplified by *Escherichia coli* [60]. Furthermore, within a genomic dataset, protein sequences in each class are unequal in both numbers and phylogenetic relatedness, which further complicates analysis. In order to identify residues that are only critical to the function of the CheW protein, we collected all sequences of proteins with a single CheW domain from MIST database [89]. CheW sequences larger than 238 amino acids were also discarded to decrease the chances of selecting poorly annotated proteins containing the CheW domain as opposed to the scaffold protein CheW

studied here. Also, sequences with less than 138 amino acids, which is the length of the CheW model in PFAM, were removed from the dataset to avoid truncated sequences. Both of these length-based filters do not affect the overall results and are placed here for the sake of clarity. Finally, sequences with more than 98% identity were deleted from the dataset to avoid bias from the different number of genomes per organism. We have assigned the resultant sequences to chemotaxis classes and found that F1 and F7 are the most abundant classes that are also comparable in size (Figure 22). Therefore, we have performed further comparative sequence analysis on CheW-F1 and CheW-F7 subsets only.

Earlier analysis of CheW sequences indicated it is a poorly conserved protein [114]. Therefore, it was not surprising to discover that among the five most conserved positions in each class only two glycine residues are common to both classes (Table 2). Conservation of a glycine residue usually indicates its unique structural role, either by allowing sharp turns and bends, or by its placement in space constraint environment [63]. Indeed, Gly63 is located at a critical turn on the CheW tertiary structure and Gly57 is present in a beta sheet bend (Figure 23). A strikingly unexpected find, however, was a nearly absolute conservation of two charged residues (Arg62 and Glu38 in *E. coli* sequence) in the F7 class (Table 2). We therefore focused our investigation on the properties of CheW proteins that belong to F7 class, especially *E. coli* CheW protein is from the F7 class.

Arg62 and Glu38 are in close proximity in the tertiary structure (Figure 23). Interestingly, both Arg62 and Glu38 (along with some other residues) have been implicated as functionally important in previous experimental studies with the *E. coli* protein. Mutations in Glu38 reduce the binding constant between CheW and the Tar chemoreceptor making it a possible candidate for the receptor binding site [111]. Mutations in residues in close proximity of Arg62 decrease the binding affinity between CheW and CheA; however, mutations in Arg62 itself do not considerably affect binding affinities for either CheW or CheA while significantly impairing chemotaxis [111]. Thus, defining the role of this conserved residue remains a challenge despite the fact that it has been approached by different experimental techniques [102, 109, 110]. Physical proximity and the opposite charge suggest that Arg62 and Glu38 residues are likely to interact. Furthermore, the highest level of evolutionary conservation suggest this interaction is critical to protein function. Because “self-interactions” between residues in a protein molecule are likely to contribute to protein dynamics, we first attempted to examine the role of Arg62 and Glu38 residues by using NMR.

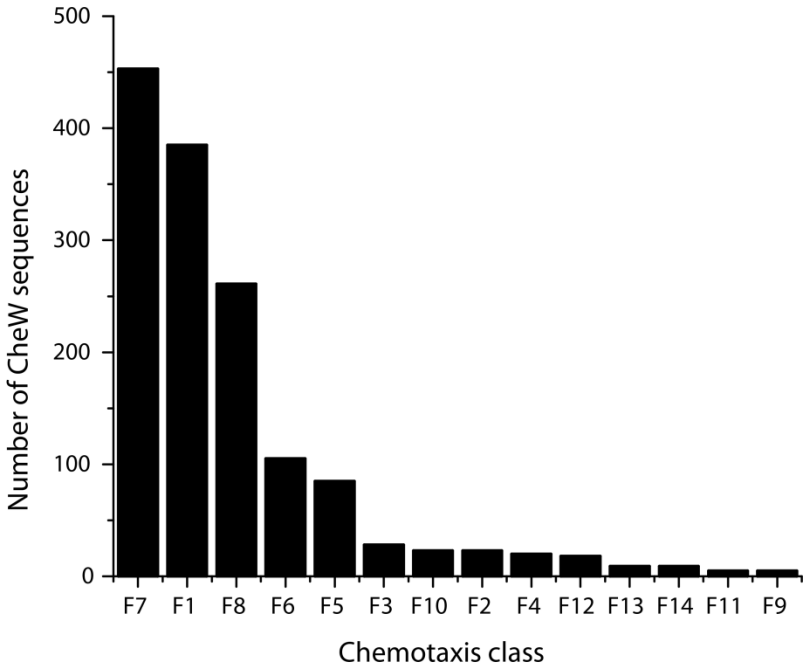


Figure 22: Distribution of non-redundant CheW sequences in chemotaxis classes

Table 2: Highly conserved residues in F1 and F7 classes of the CheW protein

F1		F7	
Residue	Identity (%)	Residue	Identity (%)
Gly57	100	Gly57	99.8
Pro49	98.7	**Arg62**	99.8
Val102	97.1	**Glu38**	99.3
Gly63	94.0	Gly63	99.1
Phe22	93.8	Gly99	99.1

Residue numbers are given in reference to the E. coli CheW protein.

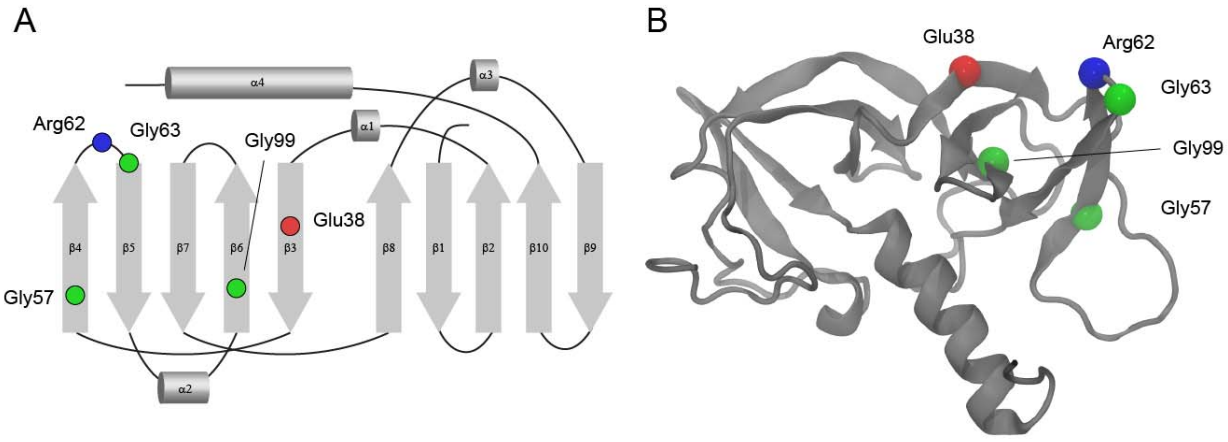


Figure 23 : Highly conserved residues mapped on the CheW secondary structure scheme (A) and the 3D NMR structure (B).

Mutants E38A and R62A show different stability *in vitro*

^{15}N labeled CheW (1 mM) and its mutant CheW R62A (1.5 mM) were uniformly examined for their relaxation properties. The wild-type CheW backbone chemical shift assignments were obtained from previously published results (BMRB accession No. 15322) [103]. Of the 154 published assignments, 123 were transferred to our wild-type CheW ^{15}N-HSQC spectrum. The remaining assignments (20%) were not transferred due to the overlap of certain resonances or the weak intensity of the resonances in our condition.

To see how the mutation affects CheW structure, we compared the ^{15}N-HSQC spectrum of E38A and R62A CheW to the wild-type CheW spectrum, and the residues with significant chemical shift perturbations were mapped onto the structure of CheW (Figure 24 and Figure 25). The results showed that the E38A mutation caused a global structural perturbation, while the R62A mutation only caused local structural perturbation. This is probably because the residue Glu38 is located in the middle of β3, which is critical for structure stability, or proper fold. Residue Arg62, on the other hand, is located in the turn that connects β4 and β5, and a single residue mutation in a flexible turn does not often affect the protein global structure. Residues of the R62A mutant that showed chemical shift perturbation with respect to the wild-type CheW greater than 20 Hz were 40-43, 49, 51, 55, 57, 60, 61, 63, 64, 66, 68-70, 73, 76, 78, 86, 88, and 101. These residues are mainly located in β4-β5, C-terminus of the β-sheet containing Glu38 (β3), and residues in close proximity to these limited regions. For this reason, the subsequent experiments focused on the R62A mutant only. Taken together, these results suggested that, while the global structure does not appear to be significantly affected by this mutation, the absence of the interaction between Glu38 and Arg62 might lead to a slightly altered local arrangement of the 5 β-strands in the second subdomain of CheW.

Backbone Dynamics

To further investigate the significance of the interaction between Glu38 and Arg62, we measured the relaxation parameters of the backbone ^{15}N nuclei in both wild-type and R62A CheW. The average longitudinal relaxation rate R_1 was 1.299 s^{-1} for WT and 1.295 s^{-1} for the mutant (Figure 26a). The average transverse relaxation rate R_2 was 14.62 s^{-1} for WT and 15.38 s^{-1} for R62A mutant (Figure 26b). The average R_2/R_1 value was 11.36 for WT and 12.02 for the mutant. The slight increase in R_2 in the R62A mutant, despite the almost identical in R_1 values,

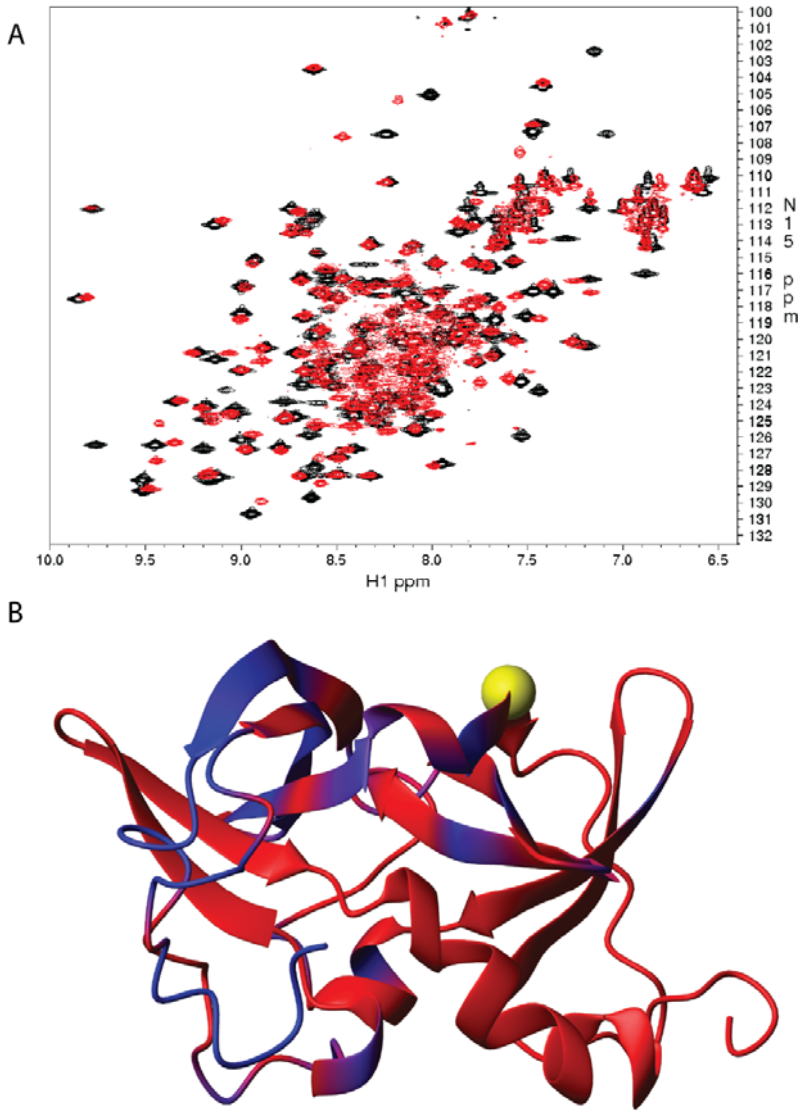


Figure 24: Effects of the mutation E38A in the CheW structure.

A) Superposition of ^{1}H-^{15}N HSQC spectra of Wild-type CheW (black) and the mutant CheW E38A (red). B) The chemical shift perturbation between wild-type- and E38A CheW color-mapped onto the CheW structure (PDB code 2HO9). The red color indicates larger chemical shift difference and blue color showed smaller differences. The mutation site E38 is shown in yellow.

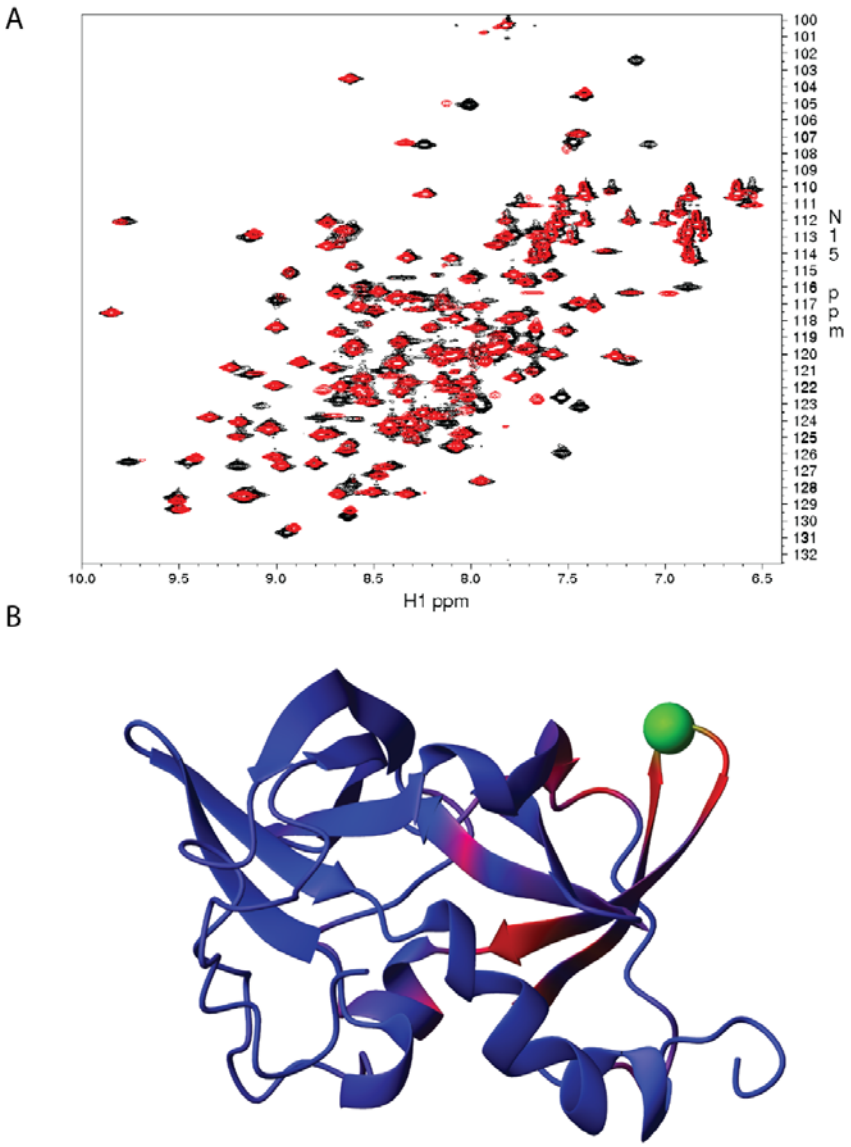


Figure 25: Effects of the mutation R62A in the CheW structure.

A) Superposition of ^{1}H-^{15}N HSQC spectra of Wild-type CheW (black) and the mutant CheW R62A (red). The chemical shift perturbation between wild-type- and E38A CheW shown in (a) was color-mapped onto the CheW structure (PDB code 2HO9). The red color indicates larger chemical shift difference and blue color showed smaller differences. The mutation site R62 is shown in green.

To address such possibilities, further analysis of the dynamics was performed by using the MODELFREE program to provide information on the internal and overall motions of both constructs (8-12). The ^{15}N R_1, R_2 and ^{1}H-^{15}N NOE data were fitted to a single isotropic rotational diffusion model with τ_m fixed at 10.98 ns for WT and 11.6 ns for the mutant, and internal parameters were optimized. The generalized order parameter, S^2, reflects amide HN bond vector motions on the picoseconds-to-nanosecond time scale, and it ranges from 0 (unrestricted motion) to 1 (completely restricted motion). Generally, order parameter values greater than 0.85 are associated with rigid regions of the protein and values below 0.8 are associated with flexible regions [115-120]. The order parameter S^2 obtained from the isotropic model shows that the majority of the backbone amides are rigid, the loops and the turns connecting the β-sheets shows some degree of freedom, and the N- and C- termini of CheW are highly flexible (Figure 26c). The order parameter information for the C-terminal end of loop 1 is missing, probably due to its increased mobility on the microsecond-to-millisecond time scale, which causes resonance broadening and a lack of data for this region.

Relaxation dispersion analysis

During the model-free analysis, an apparent chemical exchange term, R_{ex} was found to make a significant contribution to achieving adequate fit of the ^{15}N relaxation data. This suggested that there were conformational motions in CheW on the microsecond-to-millisecond time scale. For accurate characterization of the R_{ex} term, a series of Carr-Purcell-Meiboom-Gill (CPMG) [121-123] relaxation dispersion experiments were performed on both ^{15}N labeled wild-type and R62A CheW [117, 124, 125]. The R_{ex} terms from R_1/R_2/NOE fitting are similar to the results from the separate CPMG measurements below.

The phenomenological transverse relaxation rate constant, R_{ex}, represents the damping constant due to exchange between sites. The value of R_{ex} results from the chemical shift difference between two exchange sites (Φ_{ex}) and the reduced lifetime of the exchange sites (τ_{ex}) [125]. The differences between R_{ex} measured at τ_{cp} values ranging from 20 ms to 1 ms for wild-type and R62A CheW are shown in Figure 26e, and the differences between these two constructs are shown in Figure 26f. The relaxation dispersion curves for residues Ile65, Thr86, Ser164 are shown in Figure 27, and they are typical of those residues exhibiting τ_{cp} dependent dispersion. The residues that have ΔR_{ex} greater than 5 s^{-1} are Ile65, Phe75, Ser76, Gln77,

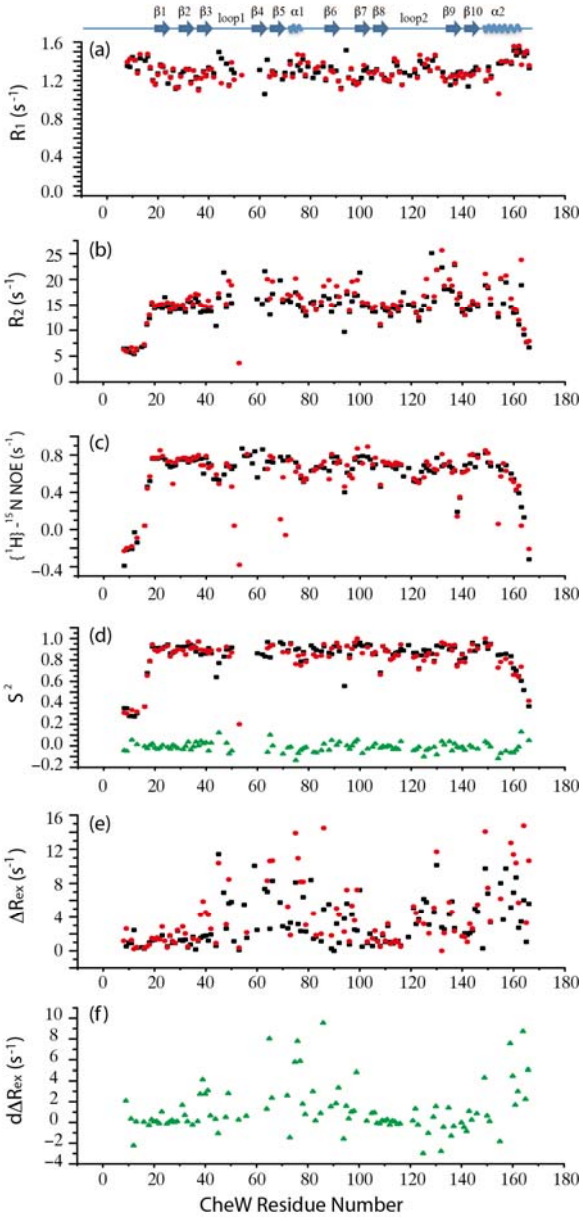


Figure 26: Backbone amide 15N relaxation parameters for CheW vs. residue number.

The black squares represent wild-type CheW, the red circles represent R62A mutant, and the green triangles represent the difference between these two constructs. Approximate location of secondary structural elements is shown at the top: (a) the longitudinal relaxation rate R_1; (b) the transverse relaxation rate R_2; (c) [4]-^{15}N NOE; (d) the extracted order parameter S^2 (e) the differences phenomenological transverse relaxation rate constant $\Delta R_{ex} = R_{ex}(20ms) - R_{ex}(1ms)$; (f) the differences between the ΔR_{ex} in (e) $\Delta R_{ex}(R62A) - \Delta R_{ex}(WT)$.

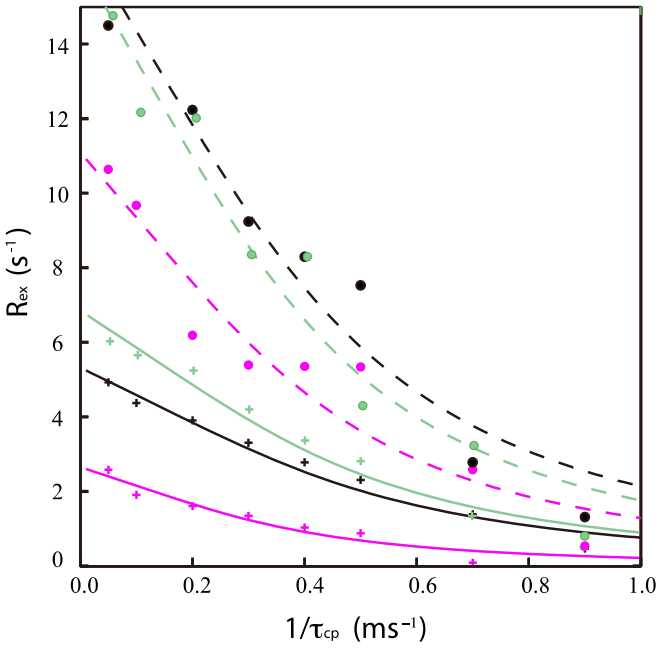


Figure 27: CPMG dispersion curves for CheW.

Values of R_{ex} are plotted versus $1/\tau_{cp}$ for Ile65 (magenta), Thr86 (black), Ser164 (green). Solid lines for WT (+) and dash lines for R62A (•) are best fits to eq. 14.

Thr86, Leu159, Ser164, and their Φ_{ex} and τ_{ex} extracted from R_{ex} data are given in Table 1. In these two CheW constructs, the majority of the backbone ^{15}N spins showed no significant differences in relaxation rate constant. However, some residues located in β4 and β5, and the loops and the two helical regions showed larger R_{ex} values. Furthermore, the single mutation of Arg62 increased the R_{ex} value in some residues on loop1, α1, α2, β4 and β10, indicating that there are increased conformational exchange motions in microsecond-to-millisecond time scale. This suggested that the mutation of R62, which disrupted the interaction between E38 and R62, will destabilize the second subdomain of the CheW structure.

Unfortunately, some critical dynamics information in the C-terminus of loop1 and β4 are missing, thereby limiting further insights into the impact of the interaction between E38 and R62 on the structure, dynamics, and function. To address this limitation, molecular dynamics (MD) was used to complement the NMR studies.

Direct evidence of the salt-bond between Arg62 and Glu38

Because of all the limitations, we decided to study the dynamics of residues Glu38 and Arg62 by using multiple molecular dynamics (MD) simulations. Using the NMR solution structure of the CheW protein from *E. coli* [103] we performed ten independent MD simulations of 90 ns each with a total of 450 thousand frames after an equilibration period of 30 ns (Materials and Methods). According to [126] a salt bridge between Arg and Glu can be inferred from an *in-silico* model if the pair of residues meet the following criteria: i) the side chain charged group centroids are within 4.0 A of each other and ii) at least one pair of carbonyl oxygen and side chain nitrogen atoms are within a 4.0 A. In our simulations, 84% of the frames meet both criteria and 11% meet only the latter, indicating N-O bridge formation. In only 5% of the frames these criteria not met. The temporal evolution of the distance of the side chain charged group centroids is show in Figure 28.

In further analysis, we found there are only two representative distinct salt bridge geometries. We define these geometries as follows: geometry A requires both atoms NH1 and NH2 are within 4.0 A from a two distinct oxygen atoms in the glutamate side chain, and geometry B requires both atoms NE and NH2 are within 4.0 A from a different oxygen atom in the glutamate side chain, (Figure 29). In 64% of all frames the residues are in geometry A and in 20% they are in geometry B.

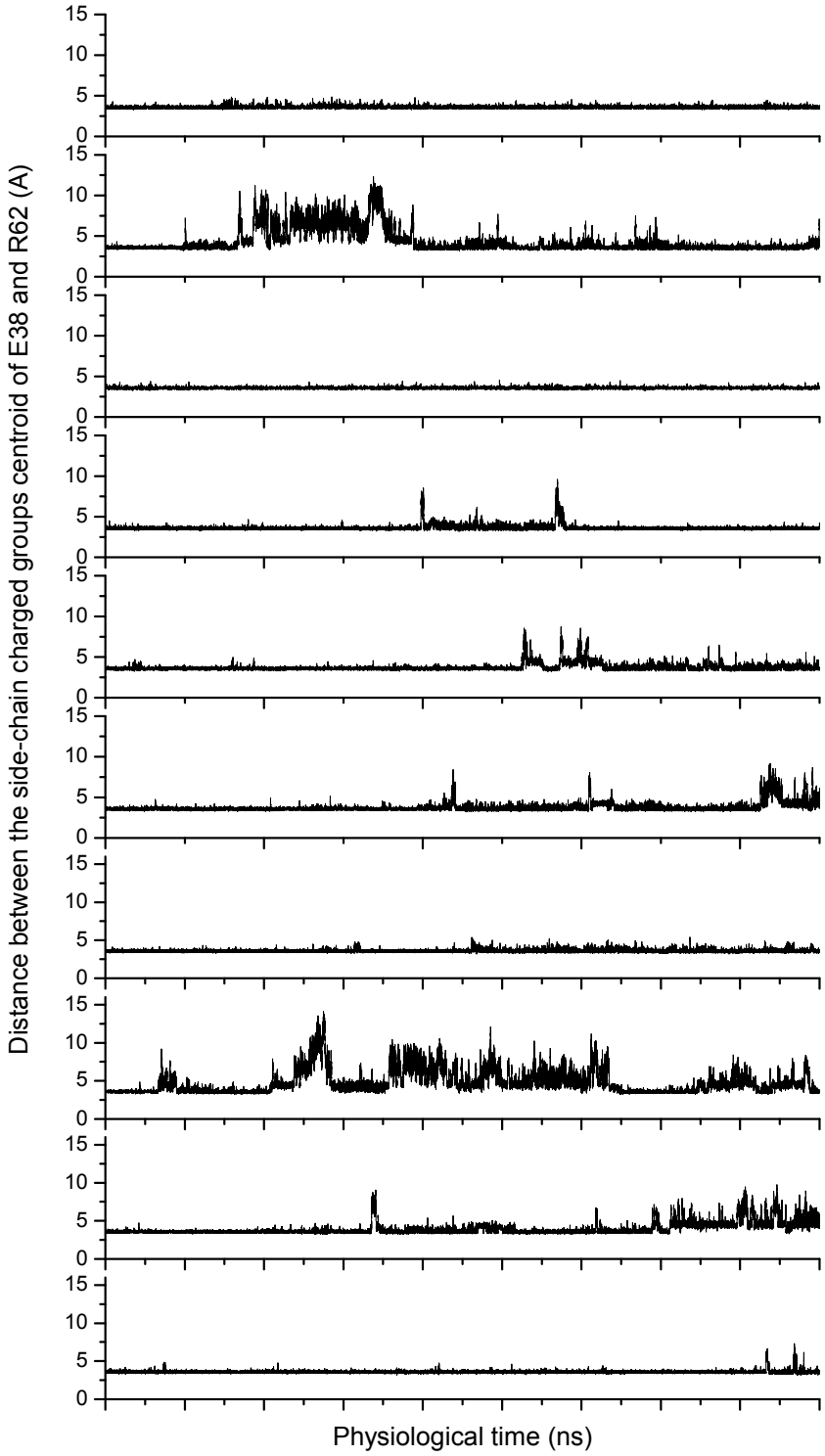


Figure 28: Temporal evolution of the distance of the side chain charged group centroids

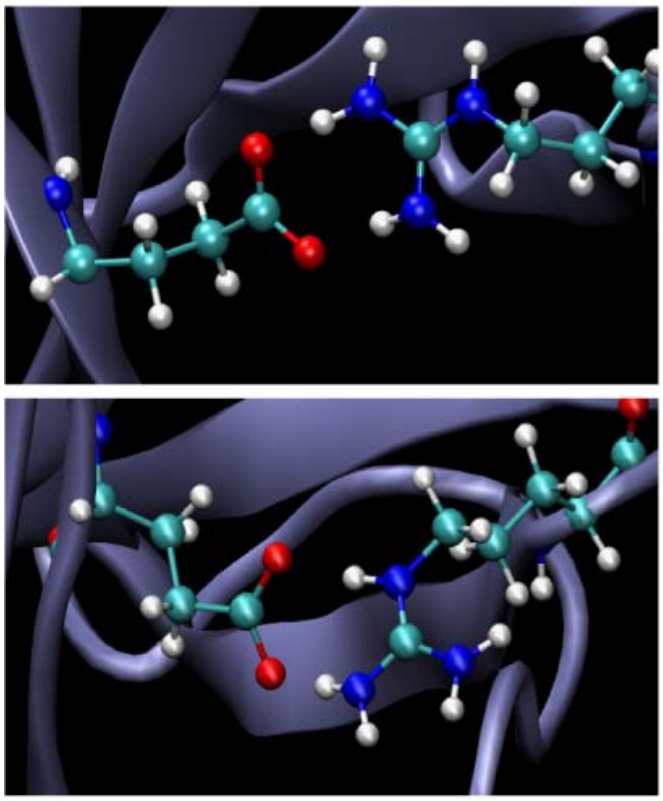


Figure 29: The salt-bridge formed by Arg62 and Glu38 in geometry A (top) and geometry B (bottom).

In simulations with wild-type structure, 64% of frames have the residues Arg62 and Glu38 in salt-bridge formation in geometry A, and 20% in geometry B.

Interaction between Arg62 and Glu38 affects protein dynamics

The interaction between Arg62 and Glu38 has consequences to the overall protein structure. To understand the contribution of the salt bridge to the local backbone stability of the protein, we analyzed the distance between alpha carbons of the involved residues in all 10 simulations. We found that the salt-bridge in geometry A maintains the distance between the alpha carbons of residues Glu38 and Arg62 at 12.3 ± 0.3 A. All other conformations assumed by the ion pair, including the salt-bridge in geometry B and the N-O bridge, slightly shift the distance between the alpha carbons and increase their relative motion range (Figure 30). The independent distribution of each conformation is shown on Figure 31. This result indicates that the stability of the relative distance between the backbone atoms of Glu38 and Arg62 is compromised if the residues do not form a salt-bridge in geometry A. To further validate this finding, we created two *in silico* mutations where Glu38 and Arg62 were replaced with Ala and ten, independent simulations of 90 ns were carried out for each mutant. Both mutations intrinsically forbid salt-bridge formation. We noticed that, in both mutants, the distance between the backbone atoms, in specific alpha carbons, was not restricted, as in the wild-type, with the salt-bridge in geometry A (Figure 32).

This result leads to the conclusion that the formation of the salt-bridge between residues Glu38 and Arg62 in a specific geometry maintains a stable, relative position between their correspondent backbone atoms.

Local backbone changes are linked to relative position of functionally active sites

Arg62 is located close to a proposed CheA binding site, and Glu38 is located within a proposed receptor binding site. Consistent changes in alpha carbon fluctuations calculated for each allele show an increase in motion of the receptor binding site relative to the kinase binding site. Local changes in backbone positioning relative to these sites were seen in all frames where the interaction between Arg62 and GluE38 was not maintained in a specific geometry. This appears to be the most critical consequence of the disruption of the interaction for the protein structure in the performed simulations. As revealed by the analysis of the order parameter, this local change in backbone mobility could not be linked to overall protein dynamics changes in the pico- to nanosecond time scale (Figure 33).

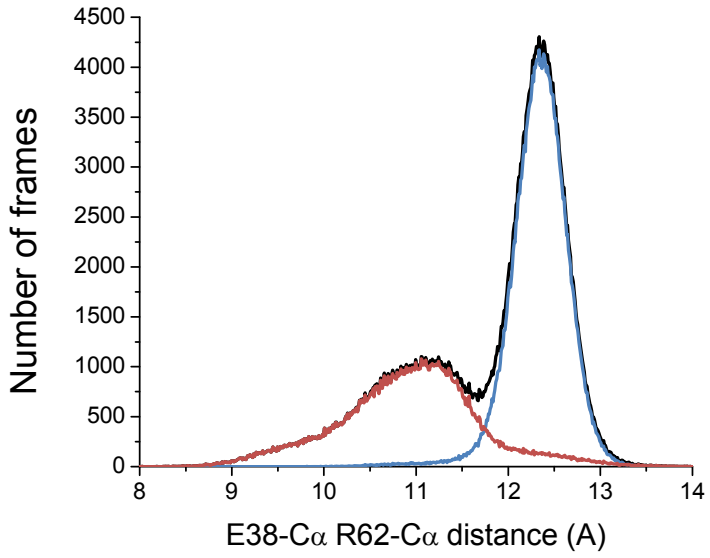


Figure 30: Histogram of the distance between alpha carbon from Arg62 and Glu38 in all wild-type simulations combined (black). The frames can be separated by the occurrence of salt-bridges in geometry A (blue) and all other interactions (red).

The latter includes frames with geometry B and frames with no interaction between the residues Arg62 and Glu38. The salt-bridge in geometry A is solely responsible for the peak of stability of alpha carbon distance in 12.3 ± 0.3 A.

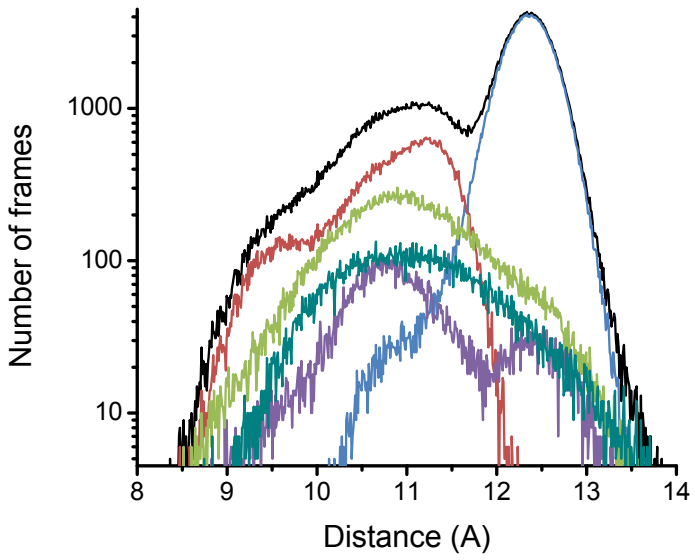


Figure 31: Distributions of the distances between alpha carbons for each conformation between Arg62 and Glu38 in all simulations with WT.

In black is the sum of all conformations, salt bridge in geometry A is in blue, salt bridge in geometry B is in red, salt bridge in other salt-bridge geometries is in purple, N-O bridge in light green and finally longer range in dark green. Note the log scale for easy display of the less populated conformations.

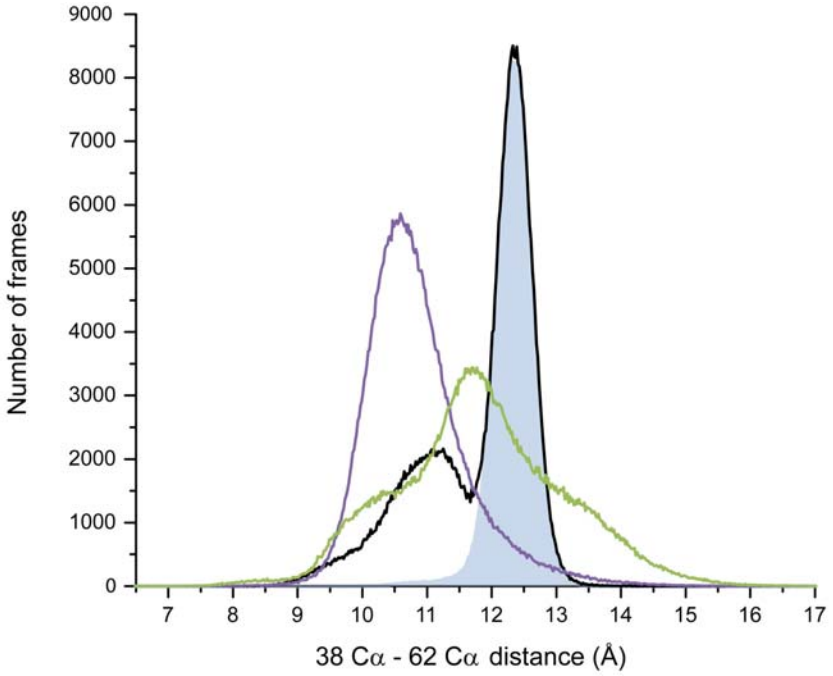


Figure 32: Comparison of the histograms of the distances between the alpha carbons of residues 38 and 62 for three alleles: wild-type (black), E38A (purple) and R62A (green).

Both mutants show an increase in instability (broader peaks) with respect to the wild-type salt-bridge with geometry A (light blue shade).

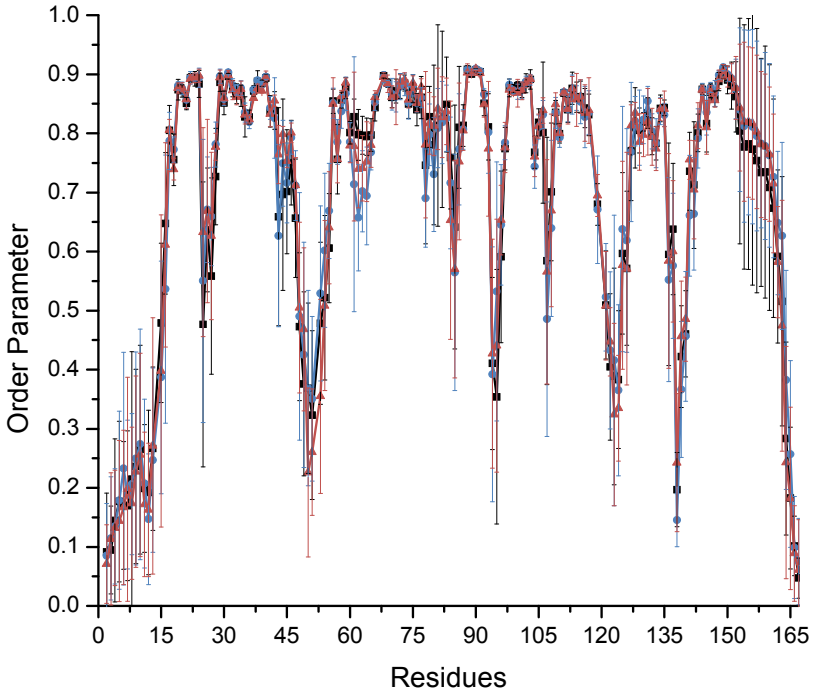


Figure 33: Average of the order parameter calculations for 10 simulations in each simulated allele. Wild-type is in black, E38A in red and R62A in blue.

Only local changes around position 62 and loop regions are prone of consistent changes in dynamics. Error bars represent the standard deviation of the calculations for the 10 simulation in each allele.

We hypothesize that the main evolutionary pressure in the class CheW-F7 is to maintain this key intra-protein interaction between the two residues that are also a part of two different protein-protein interfaces within the signaling complex.

To further support this hypothesis, we analyzed the difference in frame averaged root mean square deviation (RMSD) per residue between frames collected from the simulations of the R62A mutant in comparison with those of the wild-type (salt bridge in geometry A). Mutations in position Glu38 disrupt the interaction of CheW with the receptor [111], therefore we have not performed the same analysis for the mutant E38A. To measure the relative fluctuation of the receptor binding region related to the CheA binding region, we aligned the frames using only the backbone atoms of the residues Ile55 and Val68, which is a proposed CheA binding site [102, 111]. The frames with the salt bridge in geometry A were selected separately from each simulation, and the final frame averaged RMSD per residue value is an average of the values independently calculated for each simulation. Since only 64% of the frames from the wild-type simulation had a salt bridge in geometry A, we randomly selected 64% of the frames from all ten R62A simulations. The frame averaged RMSD per residue was calculated in the same fashion as described for the wild-type. Overall, the R62A mutant protein was more dynamical than the wild-type (higher frame averaged RMSD per residue) (Figure 34A). However, considering the fluctuation of the results from each simulation, only few residues were significantly more dynamical in R62A than in the wild-type (p-value < 0.00002). Satisfactorily, more than half of these residues were found on the chemoreceptor binding region (Figure 34B), further suggesting that the most important consequence of disrupting the salt bridge between Glu38 and Arg62 is the increase in fluctuation of the relative position between the kinase and receptor binding sites on the CheW surface.

Discussion

The results presented here provide a compelling explanation for the strong evolutionary pressure on residues Arg62 and Glu38 of the scaffold protein CheW. We found that both residues are highly conserved among all CheW proteins. In particular, they are invariant in all currently available CheW sequences from the most populated chemotaxis class, F7. While Glu38 was previously suggested to participate in the interaction with chemoreceptors [111], previous studies failed to propose a role for Arg62, although this residue was recognized as highly conserved and shown to be critical for chemotaxis [110]. Using NMR measurements and MD simulations, we now show that Arg62 and Glu38 form a stable salt-bridge with a specific

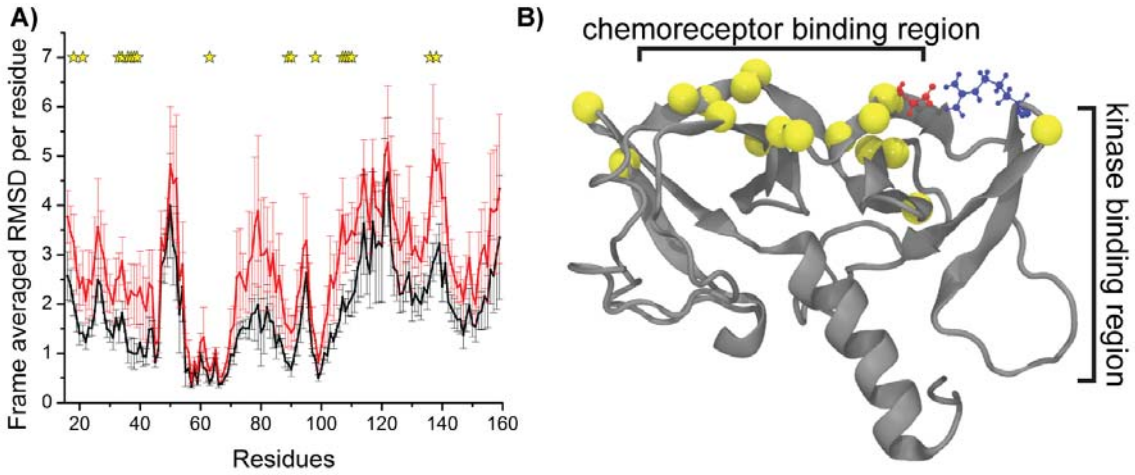


Figure 34: Occurrence of salt-bridge in geometry A between Arg62 and Glu38 improves the stability of the position of the chemoreceptor binding region relative to the kinase binding region.

A) Frame averaged root mean square deviation (RMSD) per residue was calculated for frames from wild-type simulations with salt-bridge in geometry A between Arg62 and Glu38 (black) and also for frames from R62A simulations (red). The error bars represent standard error of the mean. B) Cartoon representation of the CheW structure and residues Arg62 (blue) and Glu38(red). Residues presenting significant difference ($P < 0.0002$) in RMSD are marked in the plot (yellow star) and mapped in the structure (yellow balls). Spectacularly, all residues lay in the chemoreceptor binding region.

geometry. Simulations with the mutant R62A show that disruption of the salt bridge does not lead to overall structural and dynamic compromises of the protein. However, it resulted in a detectable loss in stability of the relative position of the chemoreceptor and the kinase binding regions. NMR experiments determined that mutant R62A only perturb the CheW structure locally, agreeing with the MD results. However, the chemical shifts experienced by several residues in the mutant E38A, in contrast to only local changes of the same mutant in the molecular dynamics simulations, suggests that E38A is important for the proper fold of CheW. *In-silico* mutants do not experiencing the folding process. Furthermore, the NMR relaxation dispersion experiment suggests slow, local motion around the region of Arg62 in the mutant R62A. It may be inferred that immediate changes in stability observed by the MD simulations will lead to major motions in microsecond-to-millisecond time scale. Finally, despite the differences in the environment conditions between the NMR experiments and the MD simulations, both mutant R62A and wild-type showed a relative good agreement in order parameter, a long stand validation of molecular dynamics methodology by NMR quantitative experiment.

Our results, combined with previous experimental data, lead to the unambiguous conclusion that the Arg62 and Glu38 interaction is fundamental to CheW function as a dedicated scaffold in the chemotaxis complex. We speculate that the maintenance of stability between the chemoreceptor and kinase binding regions of CheW might play a role in signal transduction between the chemoreceptor trimer of dimers and the kinase.

The role of CheW in bridging chemoreceptors and the histidine kinase in the chemotaxis signaling complex is reminescent to that of SH2 and SH3 domain-containing scaffold proteins that connect G-proteins and tyrosine kinases to downstream enzymes in eukaryotic signal transduction schemes. A traditional view on scaffold proteins as a molecular “glue” [107] is challenged by observations that proteins are conformationally dynamic and exhibit functional promiscuity [127]. Our results demonstrate that the CheW protein, which participates in a unique signaling complex in bacteria, is not just static “glue” that keeps the elements of the complex together. It is a highly dynamic protein and, thus, might play a role in signal transduction. For example, transmitting conformational changes in chemoreceptors and the kinase can be achieved by changing stability between corresponding binding regions of CheW that we have detected.

By implementing a computational approach that combines bioinformatics and molecular simulation with in-vitro NMR techniques, we were able to find a key determinant – an evolutionary conserved salt bridge on the surface of the protein – and to propose a mechanism for such control. As more genomic and structural information becomes available, our approach is directly applicable to similar biological systems.

Materials and Methods

NMR spectroscopy

All the NMR data was collected at 30°C with a Varian Inova 600 MHz spectrometer equipped with a four-channel (^{1}H, ^{13}C, ^{15}N, and ^{2}H) cryoprobe and Z-axis pulsed field gradients. NMR data was analyzed with the nmrPipe package and ANSIG3.3 [128, 129]. The backbone ^{1}H and ^{15}N sequential assignments were taken from a previous publication [103]. All NMR samples were in 30 mM Tris-HCl (pH 7.3), 30 mM NaCl, 0.2% sodium azide in 90% H_2O and 10% D_2O.

The longitudinal relaxation time T_1 (or inverse rates R_1), transverse relaxation time (or inverse rates R_2), and the ^{1}H-^{15}N NOE factor of backbone amide ^{15}N nuclei in CheW were measured using inverse-detected two-dimensional (2D) experiments [115-119]. R_1 relaxation rates were determined from spectra recorded with delays of 11, 55, 110, 220, 330, 440, 660, 880, and 1210 msec. R_2 relaxation rates were determined from spectra recorded with delays of 16.5, 33, 49.5, 66, 82.6, 99.1, 115.6, 132.1, and 148.6 msec. A recycle delay of 1.5 sec was used in both R_1 and R_2 measurements. Relaxation rates R_1 and R_2 were extracted by fitting the peak intensities with a single-exponential decay function. The ^{1}H-^{15}N NOE factor was taken as the ratio of the peak intensities with and without proton saturation during 3 sec of the total 8 sec recycle delay period [119, 120]. Further analysis of the dynamics data was performed by using the MODELFREE program [76, 116, 118, 130] to provide information on the internal and overall motions of CheW. The ^{15}N R_1, R_2 and ^{1}H-^{15}N NOE values were fitted into a single isotropic rotational diffusion model described by the overall correlation time τ_m. The model contains a contribution from fast internal motions described by the order parameter S^2 and the correlation time τ_e and from additional exchange broadening (R_{ex}) on the time scale of μsec to msec. In the

data fitting procedure, τ_m was fixed at 11.0 ns for WT and 11.6 ns for the mutant, and internal motional parameters were optimized [76, 115-118, 124, 125, 130].

For more accurate characterization of the chemical exchange contribution (R_{ex}) to the transverse relaxation rate constant, a series of modified Carr-Purcell-Meiboom-Gill (CPMG) relaxation dispersion experiments were performed [121-123]. The total CPMG period was kept constant at 80.0 msec while the delay τ_{cp} was varied for a total of 9 values ranging from 1.0 ms to 20.0 ms. The ΔR_{ex} term, with a base value at the fastest spin-echo rate or the shortest τ_{cp}= 1 ms, can be extracted by the following equation:

$$\Delta R_{ex} = -12.5 \ln (I/I_0) \quad 13$$

where I is the peak intensity at τ_{cp} and I_0 is the peak intensity with τ_{cp}= 1 msec. The value of R_{ex} is contributed by the chemical shift difference between two exchange sites (Φ_{ex}) and the reduced lifetime of the exchange sites (τ_{ex}):

$$R_{ex} = \Phi_{ex}\, \tau_{ex} [1-2\tau_{ex}/\tau_{cp}) \tanh(\tau_{cp}/2\tau_{ex})] \quad 14$$

in which $\Phi_{ex} = (\omega_1 - \omega_2)^2 p_1 p_2$; and p_i and ω_i are the population and Larmor frequency for the nuclear spin at site i, respectively, and τ_{ex} is the reduced lifetime of the exchanging sites [125].

Bioinformatics

We collected all 3738 CheW proteins sequences from MIST database [89] using its definition of CheW in August 2012 from draft and complete genomes. HMMER 2.0 [54] was used to classify the CheW according to its chemotaxis classification (flagelar only) determined by Wuichet and Zhulin [10]. 1185 sequences were uncategorized, attributed to alternative systems (Acf) or type IV pili systems (Tfp) and were discarded from this analysis. To avoid truncated sequences, we removed sequences with less than 138 amino acids, the length of the pfam model. We recognize that CheW domain is also part of other proteins with large portions with no hit to pfam models (data not shown).To avoid contamination by such proteins, we selected only sequences shorter than 238 amino acids long, since the scaffold CheW presumably contains only one CheW domain. The number 238 comes from adding 100 amino acids, which is the average size of a protein domain, in addition to the 138 amino acids covered

by the pfam model. In this step, 296 sequences were deleted for being too large and 72 were deleted for being too short. The length-based filter applied here does not significantly change the results and was used to significantly improve readability of the multiple sequence alignment built. CheWs sequences were separated according to their chemotaxis classes in different files. Each file was subjected to multiple sequence alignment using algorithm L-INS-I from the package MAFFT [55]. To avoid bias, the distribution by redundant genomes for each class, sequences with more than 98% identity were removed from the dataset. The final dataset contained 1429 sequences. Figure 22 shows the distribution of these sequences in natural chemotaxis classes. These sequences were used to plot Figure 22 and the sets for class F1 and F7 were used to calculate the identities presented in Table 2.

Structures and simulation system.

The atomic coordinates of *E. coli* CheW were obtained from the NMR structure deposited on PDB (PDB code: 2HO9) [103]. There are 20 frames in the PDB file, and the frame with the lowest alpha carbon RMSD against the average of all frames was selected. This structure contains 2576 atoms (including hydrogen). The structure was solvated with Solvate plug-in of VMD [81]. Using the Autoionize plug-in of VMD, 30 mM of NaCl was added to the system and randomly placed ions were used to neutralize the simulation box. The final system contains 36193 atoms in a 64 x 91 x 70 A simulation cell.

CheW structure equilibrates after ~30 ns (data not shown). We arbitrarily select a frame at 40 ns to be the starting point for production simulations of the wild-type protein. To build the mutant R62A and E38A, we used the Mutate plug-in of VMD and changed arginine to alanine and glutamate to alanine in two, independent copies of the same frame, which was selected as the starting point for the wild-type protein simulations. To ensure that all three simulations systems (wild-type, mutant R62A and E38A) were similar, only 200 steps of energy minimization were applied to each of the two simulation cells after the mutation.

Molecular Dynamics Protocols

Equilibration: The simulation system was subjected to 1000 steps of standard energy minimization, followed by 280 ns simulation in NPT ensemble. All simulations were performed with NAMD2 [27] using CHARMM22 [37] force fields for proteins and TIP3P model for water

[131]. The temperature was held constant at 298 K using Langevin dynamics in non-hydrogen atoms with a damping coefficient of 5 ps -1. The pressure was also held constant at 1 atm using a Nose-Hoover Langevin piston [49] with a period of 100 fs and a decay time of 50 fs. The integration time step was set to 2 fs under a multiple time stepping scheme [132], with bonded and non-bonded interactions calculated at every step and long range electrostatics interactions calculated at every other step. For the description of the long range forces, van der Waals forces had a cutoff of 12 Å and a switching function started at 10 Å to ensure smoothness. Electrostatic interactions were calculated using particle mesh Ewald (PME) with a grid point density of over 1/Å.

Production: For the wild-type and each mutant, ten, 90 ns long, independent simulations were produced. In each simulation, atom velocities were reinitialized, guaranteeing independence between each run. The same simulation settings described in the equilibration section were used. The computation was performed using 512 nodes in the Newton Cluster at The University of Tennessee-Knoxville with a performance of ~33 ns/day. Temperature and pressure were constant throughout all simulations (data not shown).

Calculation of the frame average RMSD per residue:

To calculate the frame average RMSD per residue, we executed the following procedure: (1) Select all frames from all simulations with the wild-type structure in which Arg62 and Glu38 form a salt-bridge in geometry A. The frames were kept separated according to its parent simulation producing ten sets of frames. (2) For each one of the ten sets of frames, the RMSD per residue is calculated for each frame against the initial frame common to all simulations. (3) The RMSDs per residue were averaged over the number of frames in each set independently.

The RMSDs were calculated using the VMD tcl command “rmsd” and all atoms were taken in consideration. The values plotted in Figure 34A are means of the frame average RMSD per residue for all ten values of the ten sets. The error bars represent standard error of the mean.

The same procedure was executed for the simulations with the mutant R62A structure. However, to produce the ten sets of frames, the same amount of frames selected from the wild-type simulations (64%) were randomly selected from all ten simulations and later grouped in ten

sets according to the parent simulation. Statistical significance was calculated using two-tailed t-tests for each residue independently.

Order Parameter Calculations:

We calculate the order parameter defined by Lipari and Szabo [76]. We use a discrete version of the equation 3 in [84]:

$$S^2 = \frac{1}{T^2}\sum_{t=0}^{T/2}\sum_{\tau=0}^{T/2} P_2\left(\hat{\mu}(\tau)\cdot\hat{\mu}(t+\tau)\right) \quad 15$$

where t and τ scan over the sequence of frames, $\hat{\mu}$ is the unit vector pointing along the backbone N-H bond. $P_2(x) = \left(\frac{3x^2}{2} - \frac{1}{2}\right)$ is the second Legendre polynomial.

General protocol for frame alignments:

CheW contains several loops. In our simulation, these loops were very flexible and alignment of the frames was rather poor, which dramatically affected the results of the order parameter calculations. It is then important to align the frames using only the most stable regions of the molecule. The residues with the lowest RMSF per residue calculated from the production part of the initial 280 ns simulation were selected for the alignment. The cutoff was determined by the 75th percentile of the distribution of the calculated RMSF for each residue. As a result, only residues with less than 4.87 Å RMSF were used to align the frames for order parameter calculations: 15 to 43 46 47 53 to 61 64 to 81 85 to 119 126 to 136 140 to 156.

CHAPTER III Homology Modeling and Molecular Simulations of CheW

This chapter was taken from a manuscript in preparation:

Derek J. Cashman; Davi R. Ortega; Igor B. Zhulin; Jerome Y. Baudry

Homology Modeling and Molecular Simulations of CheW, Nature of a Key Dynamic Component of the Chemotaxis Signaling. Manuscript in preparation

Conceived and designed experiments: DC DO IZ JB. Performed experiments: DC DO. Analyzed the data: DC DO IZ JB. Wrote the paper: DC DO IZ JB .

Abstract

Homology models of the *E. coli* and *T. maritima* chemotaxis protein CheW were constructed to quantify the quality of structural predictions and their usefulness in chemotaxis: i) a homology model of *E. coli* CheW was constructed using the *T. maritima* NMR structure as a template, and ii) a model of *T. maritima* CheW was constructed using the *E. coli* NMR structure as a template. The conformational space accessible to the homology models and to the NMR structures was investigated using molecular dynamics and Monte Carlo simulations. The results show that even though homology models of cheW sequences and their corresponding experimental structures may be structurally different, they explore, through their dynamics, similar fractions of their accessible conformational space. This shows that homology models of cheW (and potentially, of other proteins) should be seen as snapshots of an otherwise large ensemble of structures similar to that of experimental structures. The understanding of the structural and functional properties of chemotaxis complexes benefits from the description of the dynamics of predictive homology models.

Introduction

Bacterial chemotaxis is a model system for understanding the fundamental principles of signal transduction in biological systems. The core signaling complex in chemotaxis consists of chemoreceptors and the histidine kinase, CheA, that are linked by the coupling protein, CheW. Chemoreceptors detect various extracellular and intracellular stimuli and modulate CheA activity, which transduces the signals to the flagellar apparatus via its cognate response

regulator, CheY[26, 133]. In many organisms, the signaling complex assembles into organized arrays at the cell poles, where chemoreceptors cooperatively regulate kinase activity [12, 134]. This high-order structure is critical for signal amplification, the remarkable sensitivity of the system, and its precise adaptation [99, 135]. Although the general concepts involved in the chemotaxis pathway are understood, the details of the molecular mechanism are still a matter of intensive research [11, 15]. Thus, an atomic description and complete molecular analysis of the chemotaxis components is fundamental to address this challenging topic.

There are currently 34 atomic structures deposited in the Protein Data Bank for the protein elements of the chemotaxis signaling complex formed by the five domain kinase CheA, the scaffold protein CheW and the chemoreceptor, which is also known as methyl-accepting chemotaxis protein (MCP) (Table 3). All of these structures are for proteins in *Thermotoga maritima*, a thermo-stable protein, and *Escherichia coli*, a model organism for chemotaxis. However, other model organisms, such as *Rhodobacter sphaeroides* [11, 136] still and *Bacillus subtilis* [11, 136], still do not have resolved three-dimensional structures available for any of their chemotaxis proteins. The modest number of available 3D structures compared to the vast number of chemotaxis proteins that have been sequenced is dramatic. For instance, there are 3,738 CheW protein sequences from draft and complete genomes in the Microbial Signal Transduction (MiST2) database as of August 2012 [89]. This is in sharp contrast to four structures resolved for three organisms. To translate this avalanche of sequence data into structural knowledge, it becomes necessary to use *in silico* approaches to build molecular models.

Homology modeling is the approach of choice to build such models when only the sequence of a protein and at least one structure of a homologous protein are known. This technique has been used extensively in a wide variety of applications, including analyzing ligand binding sites [137, 138], substrate specificity[139], docking and scoring involved in rational drug design [140], generating ensembles for docking [141], generating and analyzing binding sites for protein-protein interactions [142], as well as providing starting models in X-ray crystallography [143] and NMR spectroscopy [144]. In homology modeling, the higher the sequence identity between the protein sequence to be modeled (also called the “target”) and the protein template, the higher the quality of the model [145]. Empirically, sequence identity levels of less than ~30% between the template and the target proteins have been found to lead to homology models of poor quality. Thus, proteins in this range of sequence identity are often referred to as being in the “Twilight Zone” of homology modeling [146]. This is an issue for cheW, where proteins of

Table 3: Chemotaxis protein structure entries in the Protein Data Bank.

All structures are derived from X-ray crystallography, except the underlined structures, which are derived from NMR spectroscopy.

Structural availability summary	CheA					CheW	MCP
	P1	P2	P3	P4	P5		
Escherichia coli		1FWP 1EAY 1A00 1FFG 1FFS 1FFW				2HO9	3ZX6 1QU7
Thermotoga maritima	1TQG 2LD6	1U0S	1B3Q	1B3Q 1I58 1I59 1I5A 1I5B 1I5C 1I5D 2CH4 3UR1	1B3Q 2CH4 3UR1	1K0S 2CH4 3UR1	2CH7 3G67 3G6B 3UR1
Thermoanaerobacter tengcongensis						2QDL	
Salmonella typhimurium	1I5N						

similar function have often low or very low sequence identity. As shown in Figure 35, virtually all sequences exhibit less than 30% sequence identity to each other, and most of them actually exhibit sequence identity of less than 15%. However, there are many examples of proteins in different families having very low sequence identity yet sharing an overall similar fold [147-155]. This is especially true for proteins that share similar functions, such as the known sequences of CheW.

In the case of chemotaxis proteins, and in particular of CheW, where known sequences outnumber known structures by orders of magnitude, homology modeling is the only technique available to develop a structural rationale to function. But how accurate would such models be, which have to be based on low sequence identities and high function and conservation? Even if overall structures are correctly predicted, would specific details in the structures be missed that are important for the protein's function? Generally, what level of confidence can the field of chemotaxis have in such computationally predicted cheW structures? The present work aims at addressing these issues by quantifying the quality of structural predictions and the extent to which structural predictions can explain and rationalize the function of the corresponding proteins.

In the present work, homology models of *E. coli* cheW are built using the NMR structure of *T. maritima* CheW (PDB code: 1K0S) as a template. Similarly, homology models of *T. maritima* CheW are built using the *E. coli* NMR structure as a template (PDB code: 2HO9). The residue identity between the sequences of these two proteins is low – 25.8% – which places these proteins in the aforementioned "twilight zone" of homology modeling. Molecular dynamics (MD) and Monte Carlo (MC) simulations are used to assess the structural variations of these homology models, and to assess the similarities and differences of the models compared to the experimental NMR structures of the corresponding proteins. This identifies the regions of the CheW structures that can be modeled with high confidence and therefore explains their corresponding functional role.

Results

Comparison of the RMSDs between the starting, static 20 homology models and 20 NMR structures (Figure 36) show fairly large values, up to 8 Å for *E. coli* and up to 6 Å for *T. maritima*. Hence, the homology models are overall significantly different than the experimental

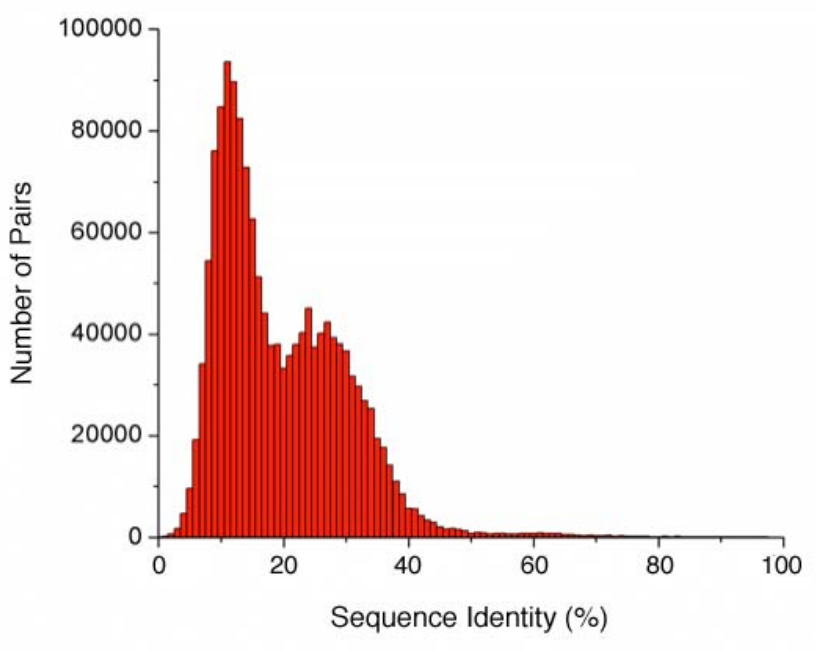


Figure 35: Histogram of the pairwise sequence identities for all against all 1,742 non-redundant CheW sequences selected from the MiST2 database.

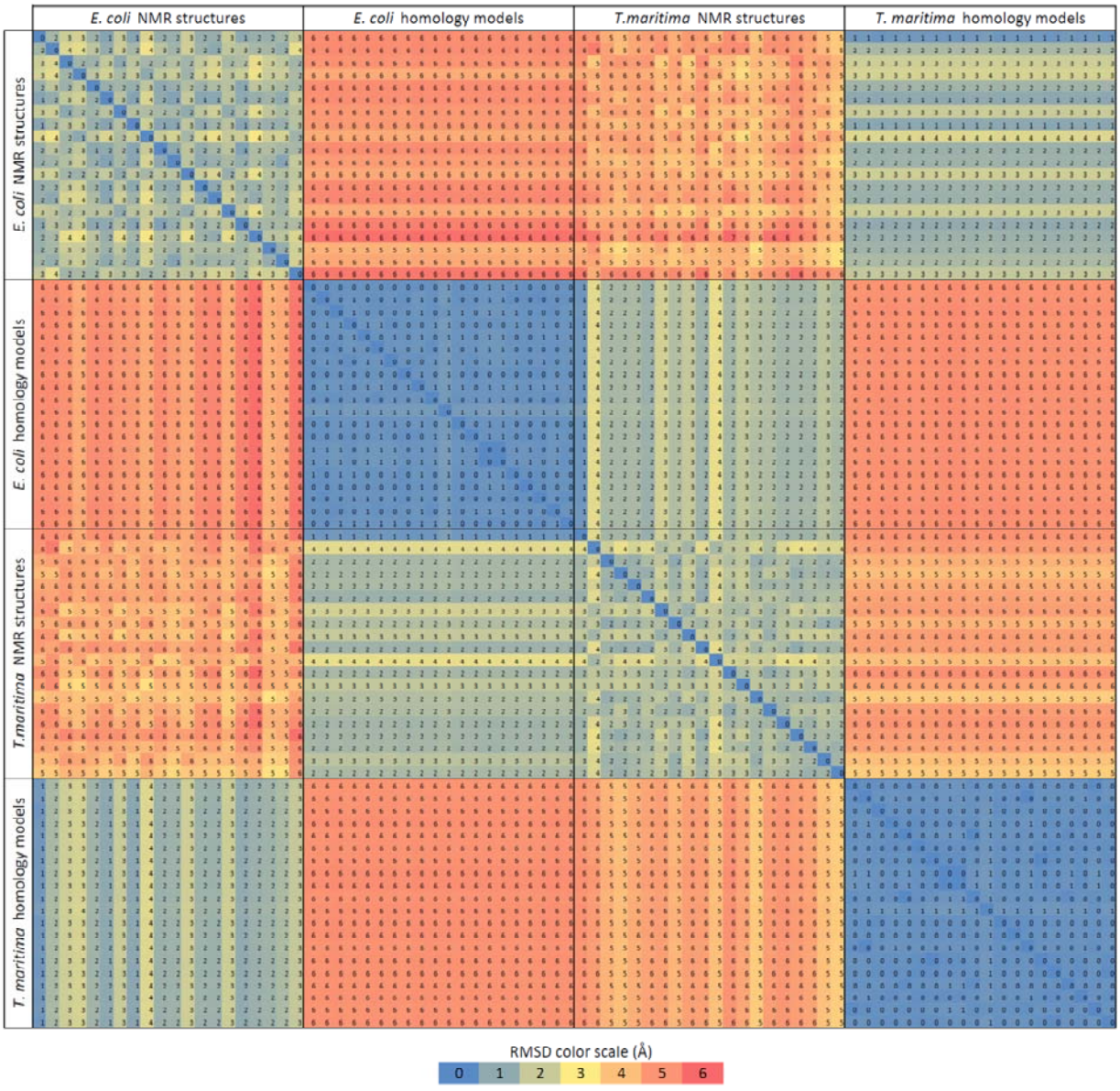


Figure 36 : Comparison of the RMSDs between the 20 homology models and 20 NMR structures for both *E.coli* and *T. maritima*.

NMR structures of the corresponding species. In contrast, the RMSD values between the experimental static NMR sub-structures are no larger than 4 Å for both *E. coli* and *T. maritima*.

Prior to the RMSD calculation of each pair, the structures were aligned taking in consideration the backbone atoms of the residues that can be aligned without gap in the protein cheW from *E. coli* and *T. maritime* pairwise alignment. The selected residues for E. coli are: 7 to 72 74 to 120 123 to 151 154 to 161 while for T. maritime were all residues but 151. The RMSD values calculated for the same set of residues used in the alignment were calculated using the function “measure rmsd” from VMD [81].

Molecular dynamics and Monte Carlo simulations of selected homology models and NMR structures (see Materials and Methods) allow the sampling of structural variations that are thermodynamically accessible at room temperature. RMSDs between each of the 25,000 structures from the MD simulation of the selected NMR models and each of the 25,000 structures from the MD simulation of the selected homology model were calculated as described in the Materials and Methods section. The results are summarized in **Error! Reference source not found.**, and, in the case of the α/β consensus residues, displayed on Figure 37. The “all residues” RMSD values between the starting homology model and the corresponding NMR structure are high: 9.1 Å for *E. coli* and 6.8 Å for *T. maritima* (**Error! Reference source not found.**). However, the core structural elements of the proteins exhibit a lower RMSD when the termini are excluded, and the RMSD values are further for the α/β consensus residues ranging between 3.0 Å for *E. coli* and 4.4 Å for *T. maritima* (**Error! Reference source not found.**). This indicates that about half to two-thirds of the relatively high overall RMSD difference between the homology model and NMR structures is due to differences in the mostly terminal flexible parts of the protein, with the structural core being better modeled by the homology approach. Importantly, **Error! Reference source not found.** also indicates that the molecular dynamics trajectories as well as the Monte Carlo trajectories produce structural variations of the starting NMR and homology model structures that were much closer to each other than the static starting models, as low as 0.8 Å (for *E. coli*) and 1.5 Å (for *T. maritima*) for the α/β consensus residues. The corresponding structures are superimposed in Figure 38 that shows that the core structures are highly similar in NMR and homology model trajectories. Yet there is very little overlap seen in the N-terminal and C-terminal regions, or in some of the internal loops. Significant differences are also visible in the large loop located near the top of each structure in Figure 38 (corresponding to residues 43-54 in *E. coli* and residues 37-48 in *T. maritima*).

Table 4: RMSD values from Figure 37 for (A) MD and (B) LBMC.

A	**E. Coli CheW**			**T. Maritima CheW**		
RMSD vs. NMR Trajectories	**All Residues**	**Residues 17-157 Only**	**α/β Consensus Residues Only**	**All Residues**	**Residues 10-147 Only**	**α/β Consensus Residues Only**
Starting values	9.1	6.2	3.0	6.8	5.8	4.4
Lowest values	5.2	3.0	0.8	3.4	2.8	1.5
Highest values	10.4	6.3	3.4	6.0	5.1	2.7
Average values	8.5	5.3	2.1	5.3	4.5	2.4
B	**E. Coli CheW**			**T. Maritima CheW**		
RMSD vs. NMR Trajectories	**All Residues**	**Residues 17-157 Only**	**α/β Consensus Residues Only**	**All Residues**	**Residues 10-147 Only**	**α/β Consensus Residues Only**
Starting values	9.1	6.2	3.0	6.8	5.8	4.4
Lowest values	4.9	3.9	1.1	3.9	3.5	1.1
Highest values	11.4	6.8	2.9	9.1	6.8	6.3
Average values	8.1	5.2	1.9	6.0	4.9	3.4

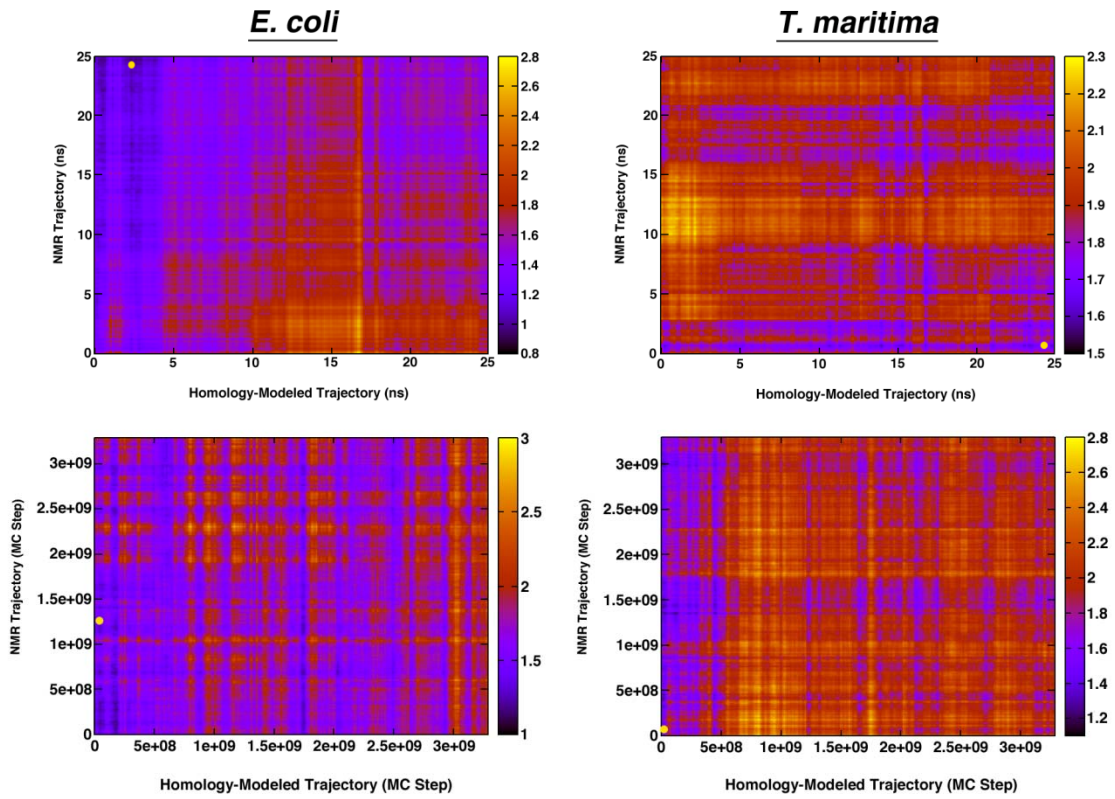


Figure 37: RMSD matrices comparing the similarity of each point of the homology-modeled trajectories with each point of the NMR trajectories.

The RMSD is calculated using the α/β consensus residues only, and is indicated by color, according to the color-coded scale to the right of each matrix. The top two matrices are for the MD simulations, and the bottom two are for the LBMC simulations. The small yellow dots in each graph indicate the point where the RMSD between both structures in the ensemble is the lowest.

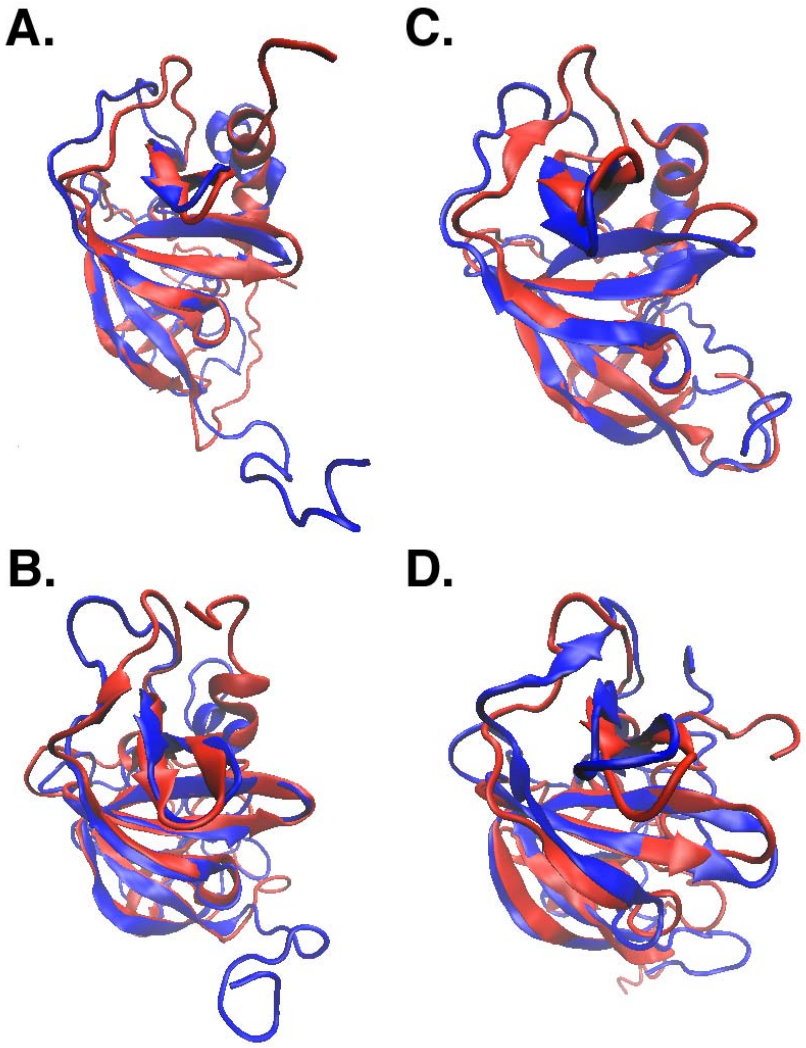


Figure 38: Superimposition of the most similar structure in the NMR trajectory (red) with the most similar structure in the homology-modeled trajectory (blue)

(A): *E. coli* MD simulation (RMSD = 0.8 Å), (B) *E. coli* LBMC simulation (RMSD = 1.0 Å), (C) *T. maritima* MD simulation (RMSD = 1.5 Å), and (D) *T. maritima* LBMC simulation (RMSD = 1.1 Å).

Figure 39 shows the distributions of RMSD values found comparing the structures in the NMR and homology model simulations (as for the instance shown in Figure 37 for the α/β consensus residues). In all of the cases represented in Figure 39, the distribution representing the 20 homology models vs. each other (blue) consistently exhibit the lowest range of RMSD values, indicating that they are relatively structurally close to each other. The RMSD range of the 20 NMR structures vs. each other (green) consistently exhibit RMSD values shifted toward higher values than the homology models, indicating that the NMR structures collectively describe more configurational space than the homology models. This configurational space is different when calculating RMSDs between the homology models and the NMR structures (red). As shown in **Error! Reference source not found.**, the NMR and homology models are closer to each other in the case of the core residues than when comparing all residues, and further so in the case of the α/β consensus residues. The purple and cyan distributions show that the range of conformation sampled in molecular dynamics and LBMC simulations are slightly different in the case of simulations of the selected NMR model and of the selected homology model. Importantly, the red distribution in Figure 39.E. shows that, in the case of the MD simulation of the *E. coli* homology model and of the *E. coli* NMR structure, the configurational space sampled is leading to RMSD values that can be lower than that exhibited between NMR structures. In other words, a static, NMR model and a static homology model, differing by approximately 3 Å RMSD (black line in Figure 39E.) can, when sampling the configurational space thermodynamically allowed to them, find themselves closer to each other at RMSD values less than 1 Å than individual NMR structures of the same protein.

Discussion

These results show that homology models of cheW can be constructed using other cheW templates of low sequence similarities. When comparing such homology models to their corresponding experimental structures, the limits of homology modeling become apparent: homology models are closer to their template structure than to their target structures. However, although homology models and NMR models of CheW may be overall different from each other, the sampling of structural space accessible by these models using molecular dynamics or Monte Carlo simulations significantly improves the agreement between predicted and

experimental models of the same protein. Experimental and predicted models may be closer to each other than some experimental NMR structures are from each other. This suggests that

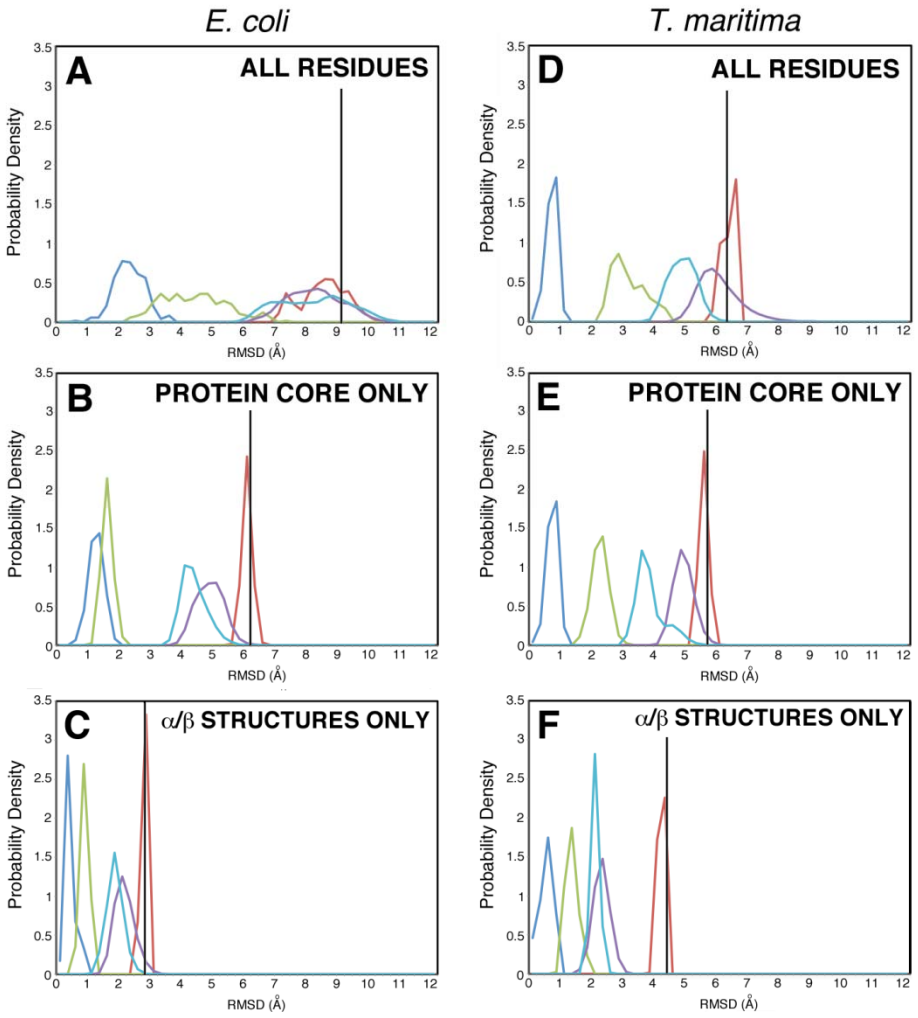


Figure 39: Histograms of the RMSD values comparing the NMR ensembles and MD/LBMC simulated trajectories.

Blue represents the histogram for the RMSD values of the 20 homology models versus each other; Green, the histogram of the RMSD values of the 20 NMR structures versus each other; Red, the histogram of the RMSD values of the 20 homology models versus 20 NMR structures; Purple, the histogram of the RMSD values of every structure of the homology model simulation versus every structure of the NMR simulation, using LBMC; Cyan, the histogram of the RMSD values of every structure of the homology model simulation versus every structure of

the NMR simulation, using MD. The vertical black line indicates the starting RMSD value for the homology model/NMR structure pair that was simulated.
whenever possible, homology models should not be seen as “the” most correct possible model, but rather as a starting point, and they should be subjected to MD or Monte Carlo simulations to identify the range of structures thermodynamically accessible.

However, even this improvement brought by including the models’ structural variations has its limits in the modeling of CheW. Not surprisingly, the N-terminal and C-terminal regions of CheW are too flexible to be correctly modeled by homology. This is also the case of the flexible loop from residues 38 to 48. What does this mean in terms of confidence of the model when it comes to translate CheW *structure* into CheW *function*? Table 5 shows the CheW residues proposed to be involved in protein/protein interactions in the chemotaxis complexes of *E. coli* and *T. maritima.* The CheW residues in the α/β consensus region, for which the “dynamics-improved” homology models are in very good agreement with the experimental structure, contain many residues that are involved with the chemoreceptor proteins. This suggests that homology modeling of the CheW sequences done here is likely to be a good structural basis to predict the interacting regions between CheW and MCP. This is less so however for the proposed interactions between CheW and the kinase protein of the chemotaxis complex, as the residues proposed to be involved in these interactions belong to the more structurally variable regions that are not so well predicted by homology modeling. The dynamic loop between residues 43-54 in *E. coli* and residues 37-48 in *T. maritima,* that is involved in the interaction between CheW and the CheA P5 domain[13, 20], is also more difficult to model predictively. However, the present simulations indicate that these regions are relatively flexible at room temperature and the structures observed in NMR experiments may not correspond to the structures of the same regions when in contact with another protein in the functional chemotaxis complexes.

Materials and Methods

Multiple sequence alignments of cheW. CheW sequences were retrieved from complete genomes in the August 2012 release of the MiST2 database[89]. The sequences were then pruned using the CheW domain definition from the Pfam [156] model PF01584 with HMMER3 [54] and 2,240 sequences with a single hit to the Pfam model were selected. A multiple sequence alignment was generated using linsi from the package MAFFT [55]. Sequences

exhibiting 98% identity were deleted to avoid redundant sequences. The final dataset contained 1,742 sequences that were re-aligned using linsi. The pairwise identity I_{ij} between the

Table 5: CheW interactions in the literature.

For *E. coli*, the residues included were: 17-19, 22-24, 27-30, 36-39, 57-61, 64-69, 87-93, 96-102, 104-105, 109-111, 133-135, 142-144, and 154-160 (57 residues, or 34.1%, out of the total 167 residues in the protein). For T. maritima, the residues included were: 12-17, 22-26, 30-34, 51-55, 58-63, 65-69, 80-84, 92-95, 97-103, 127, 132, 134-135, and 139-147 (61 residues, or 40.4%, out of the total 151 residues in the protein).

Protein Partner	Organism	Residues	Reference
Chemoreceptor	*T. maritima*	14 27 39 98 99	[107]
	E. coli	37 62	[102]
	E. coli	38 87	[111]
	E. coli	38 62 99 88 86 108	[109]
Kinase	*E. coli*	46 48 59 60 64	[102]
	E. coli	45 46 56 158	[111]

sequences i and j was calculated for $i = 1,2,3 \ldots,1741$ and $j = i + 1, i + 2, \ldots,1742$. All I_{ij} values were binned in 1% bins and displayed in the histogram format.

Homology modeling of *E. coli* and *T. maritima* cheW structures. For the modeling of *E. coli* CheW, the sequence was obtained from the UniProt database (Entry ID: P0A964) [157], and modeled based on the *T. maritima* CheW structure as the template obtained from the Protein Data Bank [158], PDB ID: 1K0S [100]. Similarly, the *T. maritima* CheW protein was modeled from its sequence (UniProt Entry ID: Q56311) using the *E. coli* CheW structure as the template (PDB ID: 2HO9) [103]. The program MOE, version 2010 (Chemical Computing Group, Inc., Montreal, Quebec, Canada), was used to build 20 homology models of both the *E. coli* and *T. maritima* CheW proteins. The C-terminal and N-terminal outgap modeling and automatic disulfide bond detection options were enabled in MOE. The models generated were scored based on Coulomb and Generalized Born interaction energies [159], and the top scoring homology model was selected for simulation using molecular dynamics and Monte-Carlo simulations.

Molecular Dynamics Simulation. The dynamics of the selected CheW homology models was investigated using all-atom molecular dynamics (MD) simulations, including the top scoring homology model generated for *E. coli* and *T. maritima* CheW. In addition, the dynamics of the first NMR structure (which is also the most thermodynamically favorable) of the PDB entries for the corresponding proteins (1K0S and 2HO9) was also simulated. Each protein was solvated using periodic boundary conditions with 8,067 TIP3P water molecules [27, 131]. The molecular dynamics program NAMD2 version 2.7 [27] was used with the CHARMM22 all-atom force field at a simulated temperature of 300 K. The integration step was set to 2 fs and all of the distances in the system involving hydrogen atoms were constrained to equilibrium values. All simulated systems were initially energy minimized using the conjugate gradient algorithm for 2,000 steps. After initial energy minimization, the systems were gradually heated in an equilibration procedure from 100 K to 300 K, in incremental steps of 50 K for 100 ps at a time, for a total of 500 ps. This was followed by a production run of 50 ns. The similarities and differences between two given structures were quantified by calculating the root-mean-square deviation (RMSD)

between these structures. Three different RMSDs were calculated that focus on different structural subsets of cheW: i) using the backbone heavy atoms of all residues, ii) using the backbone heavy atoms of only the "protein core", i.e. excluding residues 1-16 and 158-167 in *E. coli*, or residues 1-9 and 148-151 in *T. maritima*, and iii) using only the backbone heavy atoms of only the α/β consensus residues, i.e.; residues having either a α-helix or β-pleated sheet structure in all 20 sub-structures of the NMR models. For *E. coli*, this included residues 17-19, 22-24, 27-30, 36-39, 57-61, 64-69, 87-93, 96-102, 104-105, 109-111, 133,-135, 142-144, and 154-160 (57 residues, or 34.1%, out of the total 167 residues in the protein). For *T. maritima*, the residues included were 12-17, 22-26, 30-34, 51-55, 58-63, 65-69, 80-84, 92-95, 97-103, 127, 132, 134-135, and 139-147 (61 residues, or 40.4%, out of the total 167 residues in the protein). RMSDs were calculated between all 25,000 structures generated in the molecular dynamics simulations of the selected homology model, and all 25,000 structures generated in the molecular dynamics simulations of the selected NMR model.

Monte Carlo Simulations. The same systems that were used in the MD simulations were used in Monte-Carlo simulations using the LBMC method [160]. All simulations were run using an equilibration phase of 3×10^8 Monte Carlo MC steps, followed by a total of 3×10^9 MC steps. The simulation temperature was chosen to be slightly below the unfolding temperature, based on 13 short simulations of 3×10^8 MC steps, i.e. of $k_BT/\varepsilon = 0.7$, where ε is the depth of the Gō potential [161, 162]. Frames were saved every 10^5 MC steps. Trial moves consisted of swapping three consecutive peptide planes per step and/or changing the corresponding Ψ angles[160] with an acceptance rate of approximately 20-25% ; setting the fraction of local moves to 10% and the fraction of Ψ-only moves to 30%. RMSDs were calculated using the same numbers of residues as with the MD simulations (above), and for the LBMC ensembles, 33,000 structures were generated in the molecular dynamics simulations of the selected homology model and NMR model.

Acknowledgments

This work was supported by NIH grant 2R01 GM072285-06 to IBZ and JB. The authors gratefully acknowledge the Newton HPC Program at the University of Tennessee – Knoxville as

well as the UT/ORNL Center for Molecular Biophysics for computational resources. We thank Dr. Barry Bruce and Aaron Fleetwood for helpful discussions.

CHAPTER IV Universal architecture of bacterial chemoreceptor arrays

This chapter was originally published by Ariane Briegel and Davi R. Ortega, et al in Proceedings of National Academy of Science:

Ariane Briegel, Davi R. Ortega, Elitza I. Tocheva, Kristin Wuichet, Zhuo Li, Songye Chen, Axel Müller, Cristina V. Iancu, Gavin E. Murphy, Megan J. Dobro, Igor B. Zhulin and Grant J. Jensen (2009). **Universal architecture of bacterial chemoreceptor arrays.** Proc Natl Acad Sci U S A 106(40): 17181-17186.

Author's contribution: AB and GJ designed and led the project. AB, ET, ZL, SC, AM, CI, GM and MD collected tomograms. AB analyzed the tomograms. AB, DO, IZ, and ET performed the bioinformatics and length correlation analyses. AB, DO, IZ, ET, and GJ wrote the paper.

Abstract

Chemoreceptors are key components of the high-performance signal transduction system that controls bacterial chemotaxis. Chemoreceptors are typically localized in a cluster at the cell pole, where interaction among the receptors in the cluster is thought to contribute to the high sensitivity, wide dynamic range and precise adaptation of the signaling system. Previous structural and genomic studies have produced conflicting models, however, for the arrangement of the chemoreceptors in the clusters. Using whole-cell electron cryotomography, here we show that chemoreceptors of different classes and in many different species representing major bacterial phyla are all arranged into a highly conserved, 12-nm hexagonal array consistent with the proposed "trimer of dimers" organization. The various observed lengths of the receptors confirm current models for the methylation, flexibility, signaling, and linker subdomains in vivo. Our results suggest that the basic mechanism of transmembrane signaling in chemotaxis is universal among bacterial species and was thus structurally conserved during evolution.

Introduction

Most motile prokaryotes rely on a chemosensory system to control their movement toward favorable environmental conditions [7]. This process of chemotaxis depends on transmembrane chemoreceptors called methyl-accepting chemotaxis proteins (MCPs). MCPs can be classified by topology type [163] and signaling domain class [60]. Topology type 1 MCPs have large periplasmic ligand-binding domains [163] and an elongated cytoplasmic region consisting of a HAMP (histidine kinase, adenylyl cyclases, methyl-binding proteins, and phosphatases) domain

followed by a signaling domain, which in turn is composed of "methylation," "flexible bundle," and "signaling" subdomains [26, 60] Figure 40. MCPs cluster together with other chemotaxis proteins including CheA and CheW in large arrays at the cell pole [95, 99, 134, 164, 165].

Because MCPs act cooperatively, their arrangement and interactions within the arrays is critical to their function. Based on the crystal structure of the Tsr receptor from *E. coli* [14] as well as cross-linking and other studies [166, 167], it seems clear now that the basic functional unit in that organism is a "trimer of receptor dimers." It was further proposed that in *E. coli*, trimers of receptor dimers form a hexagonal array with a lattice spacing of 20 nm [168]. A subsequent ECT study showed that overexpresed Tsr chemoreceptors in *E. coli* pack into a hexagonal lattice with a center-to-center spacing of 7.5 nm [169-172]. In these overexpression strains the receptors surprisingly form a ‘zipper-like’ double layer, where large invaginations of the inner membrane allow the cytoplasmic tips of one layer to interact with the cytoplasmic tips of a second, facing layer. This arrangement was later proposed to represent the activated form of the receptors [173]. Adding further complication, MCPs from *T. maritima* crystalized as rows of dimers [13]. This structure, combined with pulsed ESR and crystallographic studies of a CheA-CheW dimer, led to a third, "hedgerows of dimers" model for the architecture of chemoreceptor arrays [13]. Finally, through direct imaging of intact *Caulobacter crescentus* cells, we [174] and others [169] showed that the chemoreceptors in that organism are arranged in a hexagonal lattice whose 12 nm spacing suggested that trimers of receptor dimers occupied each 3-fold symmetric vertex. While the MCPs of *E. coli* and *C. crescentus* all belong to the same signaling domain class (36H), those from *T. maritima* belong to a different class (44H) [60]. Thus, based on structural and bioinformatics data, it was unclear whether receptors from different MCP classes and organisms clustered similarly, or if not, how many architectures there might be. By imaging wild-type cells in nearly-native states, here we show that the chemoreceptors of diverse species from six different signaling domain classes are all arranged into a highly conserved, 12-nm hexagonal array consistent with a single "trimer of receptor dimers" functional unit at each vertex.

Results and Discussion

Position of chemoreceptor arrays within cells. In order to visualize the arrangement of chemoreceptors in diverse bacteria, we selected thirteen distantly related organisms, which together possess receptors from all seven major signaling domain classes [60] (Table 6) and recorded nearly 700 electron cryotomograms of intact, frozen-hydrated cells. Previous

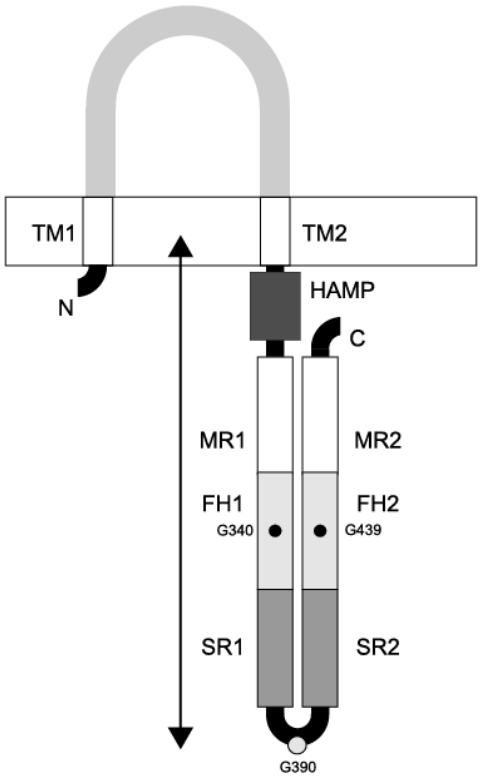


Figure 40: Schematic representation of MCP topology type I.

Two transmembrane regions (TM1 and TM2) anchor the receptor in the membrane. The extracellular (periplasmic) ligand-binding domain is shown in light gray. The cytoplasmic portion of the receptor consists of the HAMP domain, linker, and the signaling domain, which, in turn, is comprised of 3 sub-domains: methylation regions (MR1 and MR2, white), the flexible bundle sub-domain (FH1 and FH2, light gray), and the signaling sub-domain (SR1 and SR2, dark gray). In the *E. coli* Tsr receptor, Gly-340 and Gly-439 comprise the glycine ‘‘hinge’’ in the flexible bundle sub-domain, and the receptor hairpin is Gly-390. The ‘‘cytoplasmic’’ sequence length plotted on Fig. 3 was measured from the middle of TM2 to Gly-390 in the hairpin.

Table 6: Summary of measurements of 13 different bacterial species obtained by ECT

Bacteria	Phylum	Average cell diameter x length (µm)	MCP class (no. of receptors)	Location	Distance from IM to base plate	Lattice	Surface area (nm^2), estimated no. of receptors
Escherichia coli	Gamma-proteobacteria	0.5 – 1.3 x 2-5	36H (**4**,1)	Mainly polar	22 nm	12 nm	~ 53 k, ~5200
Caulobacter crescentus	Alpha-proteobacteria	0.4-0.7 x 0.9-2.2	36H (**9**,7), 38H (**1**), Unc (1)	Polar, Convex side	31 nm	12 nm	~ 17 k, ~ 1700
Thermotoga maritima	Thermotogae	0.5-1 x >1.5	44H (**6**), Unc (1)	polar	25 nm	12 nm	~ 97 k, ~ 9400
Vibrio cholera	Gamma-proteobacteria	0.8 – 0.9 x 1.5	40H (**32**,10), 44H (1), 36H (2), 24H (2), Unc (2)	Polar, Convex side	25 nm	12 nm	~ 121 k, ~11800
Magnetospirillum magneticum	Alpha-proteobacteria	0.5 x 2–10	38H (**39**,23), 40H (**2**), 44H (1), Unc (7)	Polar	28 nm	12 nm	~ 12 k, ~ 1200
Helicobacter hepaticus	Epsilon-proteobacteria	0.3 x 1.5 – 4	28H (**4**,1), 40H (**3**), Unc (1)	Polar 'cap'	24 nm	12 nm	~112 k, ~ 10900
Campylobacter jejuni	Epsilon-proteobacteria	0.4 x 0.5–5	28H (**4**), 40H (**1**,1), 24H (3), Unc (1)	Polar 'cap'	24 nm	12 nm	~144 k, ~ 14000
Rhodobacter sphaeroides	Alpha-proteobacteria	0.7–0.8 x1.3–1.4	34H (**7**,1), 36H (2), Unc (2)	Polar	21 nm	12 nm	~22 k, ~ 2200
Borrelia burgdorferi	Spirochaetes	0.2 x variable length (> 10)	34H (**2**), 48H (**1**), Unc (3)	Sub-polar	27 nm	No lattice observed	NA
Listeria monocytogenes	Firmicutes	0.5 x > 1.5	44H (**1**), 24H (1)	Polar	26 nm	No lattice observed	~30 k, ~ 2900
Acetonema longum	Firmicutes	0.3 x variable length (> 10)	44H (**20**,2), Unc (10)	Sub-polar	26 nm	12 nm	~ 51 k, ~5000
Treponema primitia	Spirochaetes	0.4 x3-8	48H (**10**,1), 40H (**1**), Unc (**1**,2)	Polar	28 nm	12 nm	~ 15 k, ~ 1500
Halothiobacillus neapolitanus	Gamma-proteobacteria	0.4 –0.5 x1.6–1.7	40H (**4**)	Polar	24 nm	12 nm	~ 31 k, ~ 3000

immunolabeling [175] and correlated light and electron microscopy studies [174] had already established that in *E. coli* and *C. crescentus*, chemoreceptor arrays can be recognized as clusters of thin, pillar-like densities extending from the inner membrane to a prominent "base plate" 20-30 nm below. Satisfyingly, similar structures were seen in all thirteen organisms imaged here (Figure 41 and Figure 42), but their locations within the cell varied. As in *E. coli*, the chemoreceptor arrays in *Magnetospirillum magneticum, Rhodobacter sphaeroides, Treponema primitia, Thermotoga maritima*, and *Listeria monocytogenes* were polar. In contrast, the arrays in *Helicobacter hepaticus* and *Campylobacter jejuni* formed a polar "collar" completely surrounding the tip of the cell, but with a gap at the apex occupied by the flagellar motor. As in *C. crescentus* [174], the arrays in *Vibrio cholerae* were polar but consistently localized to the convex side of the crescent-shaped cells. Receptor arrays in *Acetonema longum* and *Borrelia burgdorferi* were typically subpolar but inconsistently positioned. The array in *A. longum* was found, for instance, to range from immediately adjacent to the pole to nearly a micron away. Although cytoplasmic MCP arrays have been reported in *R. sphaeroides* [176-178], none were observed here.

Receptor lengths. Despite the similarity of the arrays, the distance between the base plate and the inner membrane varied between species (Table 6). Measured values ranged from 21 nm in *R. sphaeroides* to 31 nm in *C. crescentus*, but were constant within each species. Because distinct periplasmic densities were observed above the arrays in nearly all the cells, and only topology type 1 MCPs have large periplasmic domains, we infer that at least the majority of the MCPs composing these arrays were of topology type I. In 7 of the organisms imaged (*E. coli, V. cholerae, H. neapolitanus, A. longum, L. monocytogenes, R. sphaeroides* and *T. maritima*) all the topology type I MCPs present in their respective genomes belong to a single (but different for each organism) signaling domain class (Table 6). When the observed distance between the inner membrane and base plate was plotted against the number of cytoplasmic residues in the corresponding receptor sequences, there was a strong correlation with a slope of 0.142 nm per residue (Figure 43a). Because all the MCPs shown in Figure 43a contain a single HAMP domain, and its size is constant [179, 180], its presence should not affect the slope. The remarkable match of the observed slope with the rise per residue seen in the coiled coil crystal structure of a *T. maritima* receptor's signaling domain (0.145 nm/residue) [13] therefore confirms that the methylation, flexible bundle, and signaling subdomains in all the receptors shown are also coiled coils *in vivo*.

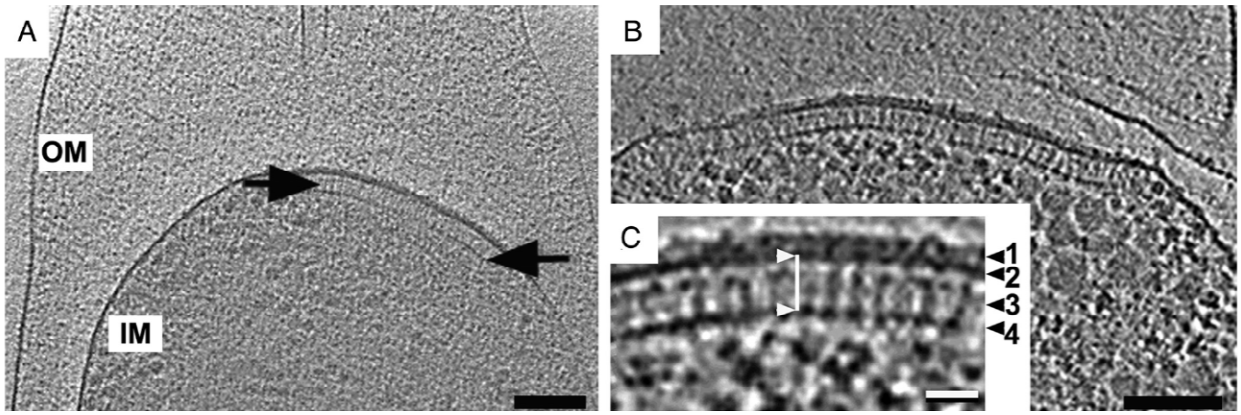


Figure 41: Characteristic appearance of chemoreceptor arrays in vivo.

(A) A 55-nm-thick tomographic slice through a T. maritima cell pole (signaling domain class 44H). Typical features like the inner (IM) and outer membrane (OM) and the enclosed extended periplasm are clearly visible. The arrows indicate the location of the chemoreceptor array within the inner membrane and densely packed cytoplasm. Bar: 100 nm. (B) A 3-nm-thick tomographic slice through the pole of a T. maritima cell treated with Polymyxin B. The reduced cytoplasmic crowding clarifies chemoreceptor features compared to those in untreated cells. Bar 100 nm.

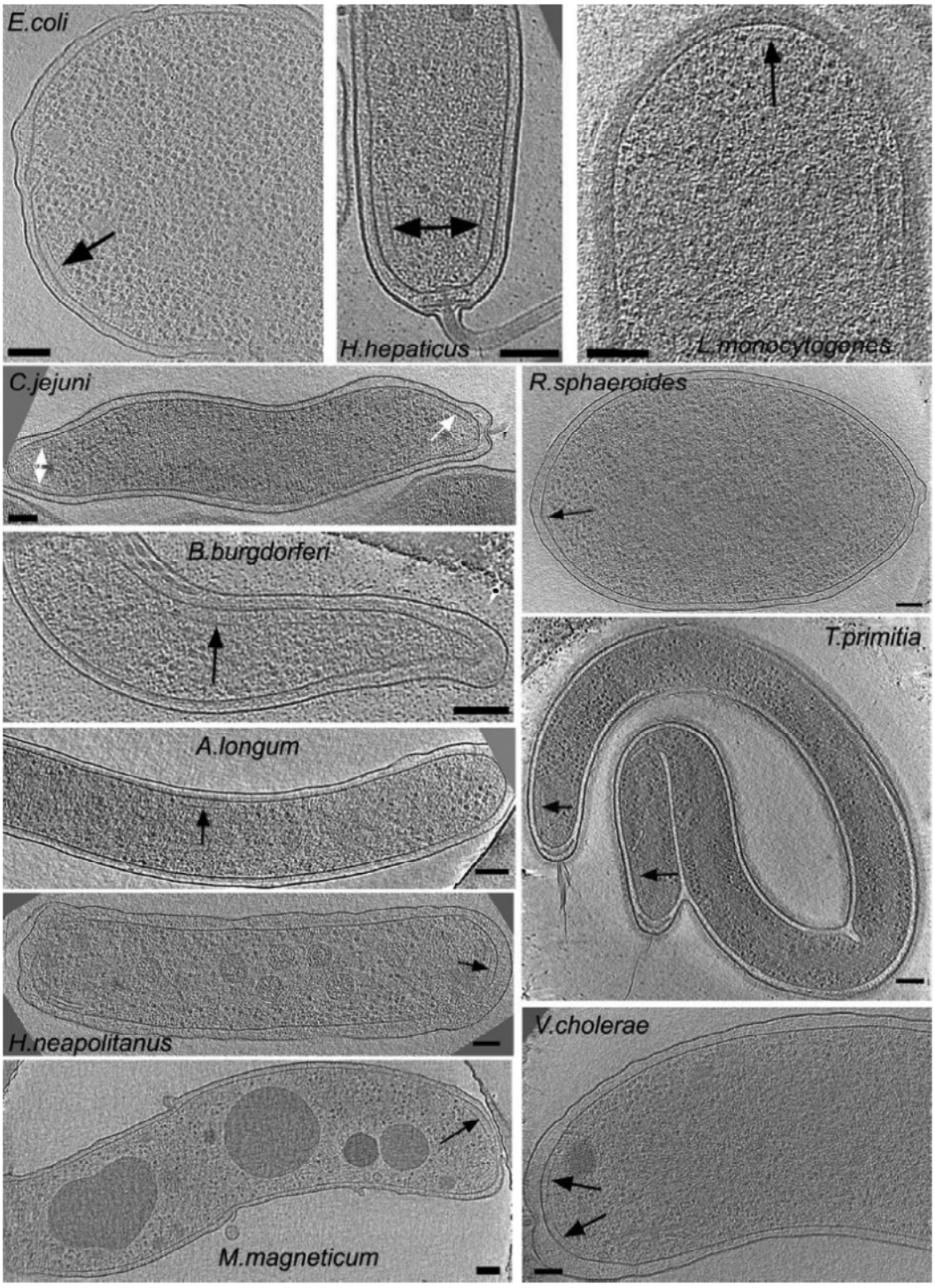
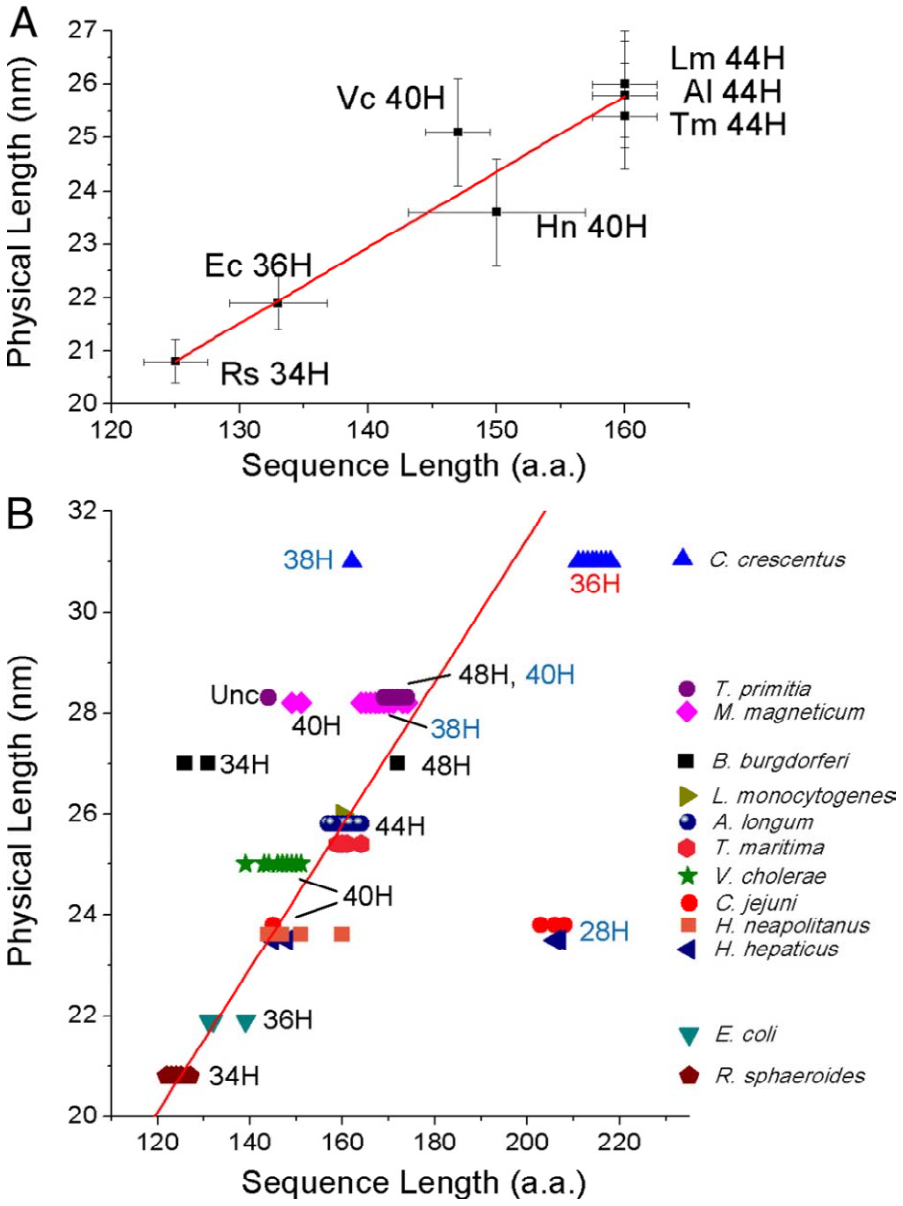


Figure 42: Chemoreceptor arrays in diverse bacteria.

Tomographic slices through cells of 11 different species illuminating the varied location but consistent appearance of the arrays. (T. maritima and C. crescentus are not shown, because they are available in Figure 41 and ref [174] respectively.) (Scale bars: 100 nm.)

Figure 43: Correlation between observed physical length and predicted sequence length.

(A) Organisms possessing a single class of topology type I receptors. Physical length and sequence length were measured as described in *Materials and Methods*. The sequence length is an average of all topology type I MCPs in the given genome. Vertical bars indicate SD of measurements from different cryo-tomograms and positions within the array, horizontal bars indicate the larger of the SD of the various MCP sequence lengths present in the genome or the estimated uncertainty in the position of the transmembrane region (~ 5 residues, see Materials and Methods). The line is a least squares fit whose slope confirms that the cytoplasmic domains of the receptors form extended coiled coils. Al, *A. longum*; Ec, *E. coli*; Hn, *H. neapolitanus*; Lm, *L. monocytogenes*; Rs, *R. sphaeroides*; Tm, *T. maritima*; and Vc, *V. cholera*. (B) All topology type I MCPs in all 13 organisms imaged. Each MCP sequence in each organism is represented by a symbol, color- and shape-coded by organism (Right). All the MCPs of a particular organism appear at the same height on the graph (the measured distance between the inner membrane and base plate layer), even though it is not known which were actually imaged. MCPs of particular signaling domain classes cluster closely, and are labeled with the color of the label itself (e.g., 34H, 36H) indicating whether the receptors of that class are typical (black), contain extra linkers (blue), or contain both extra linkers and a second HAMP domain (red; see Figure 44). The sequence lengths of typical receptors (i.e., those without extra linkers and HAMP domains) are seen to progress steadily with class number across the graph from left to right. Receptors with additional linkers or a second HAMP domain (blue and red labels) appear further to the right than expected because of their extra residues. The Unc label represents an MCP that does not correspond to a known length class, but was given a sequence length measurement as described in Materials and Methods. The graph shows that within the organisms that possess 2 classes of receptors (*C. jejuni*, *H. hepaticus*, *B. burgdorferi*, *T. primitia*, *M. magneticum*, and *C. crescentus*), only one class matches the trend line found in A, suggesting that it was the receptor class forming the arrays.


A
Physical Length (nm)
Sequence Length (a.a.)
Lm 44H
Al 44H
Tm 44H
Vc 40H
Hn 40H
Ec 36H
Rs 34H
B
Physical Length (nm)
Sequence Length (a.a.)
38H
36H
Unc
40H
48H, 40H
38H
34H
48H
44H
40H
28H
36H
34H
C. crescentus
T. primitia
M. magneticum
B. burgdorferi
L. monocytogenes
A. longum
T. maritima
V. cholerae
C. jejuni
H. neapolitanus
H. hepaticus
E. coli
R. sphaeroides

The genomes of the other 6 organisms imaged (*C. jejuni*, *H. hepaticus*, *B. burgdorferi*, *T. primitia*, *M. magneticum*, and *C. crescentus*), each contain topology type I MCPs belonging to 2 different signaling domain classes and some of their MCPs possess linkers and/or an additional HAMP domain (Figure 44 and Figure 45). It was unclear therefore which class of MCPs were being expressed and forming the arrays. When the observed lengths were plotted against the number of cytoplasmic residues in all the topology type I receptors, however, only one candidate signaling class within each organism matched the trend line (Figure 43b). In the cases of *B. burgdorferi* and *T. primitia*, for instance, the sequence length of the class 48H MCPs fell compellingly well on the trend line but the class 34H (*B. burgdorferi*) and unclassified (*T. primitia*) receptors did not. In the cases of *C. jejuni* and *H. hepaticus*, receptors of the 40H class fell on the trend line, but those of class 28H, which contain long (~95 residue) extra linkers between their HAMP and signaling domains (Figure 45), did not. In the case of *M. magneticum*, while its class 40H receptors did not fit the trend line, surprisingly, its class 38H receptors did because they contain an extra linker (of ~30 residues). Finally, in the case of *C. crescentus*, neither the class 36H nor 38H receptors matched the trend line well. However, close inspection of the sequences revealed that the class 36H receptors also contain second HAMP domain. Because a HAMP domain is expected to be approximately 4 nm shorter than a (presumably) α-helical linker of the same number of residues [179], if this deficit is taken into account, the observed length of the class 36H receptors also matches the trend line well. Our interpretations are therefore that (i) in the single growth condition used for each particular species, the arrays were composed of receptors from a single predominant signaling class that could be identified by the observed distance between the inner membrane and base plate, (ii) the methylation, flexible bundle, and signaling subdomains present in all 13 organisms are in fact coiled coils, and (iii) the linkers are α-helical *in vivo*. Assuming this is correct, our data contained images of 5 major signaling domain classes (44H, 40H, 38H, 36H and 34H) and one minor signaling domain class (48H).

Lattice arrangement. Eleven of the species imaged here presented clear "top" views of the arrays (those of *L. monocytogenes* and *B. burgdorferi* were inaccessible, see Figure 46Figure 45). Surprisingly, they all revealed the same ~12 nm honeycomb-like hexagonal arrangement observed previously in *C. crescentus*. We conclude that throughout the entire wide range of species and receptor classes imaged here (including WT *E. coli* and *T. maritima,* for which MCP crystal structures and alternative models exist, as well as organisms from 6 diverse taxonomic groups that span the bacterial kingdom (Figure 47), trimers of receptor dimers pack at the

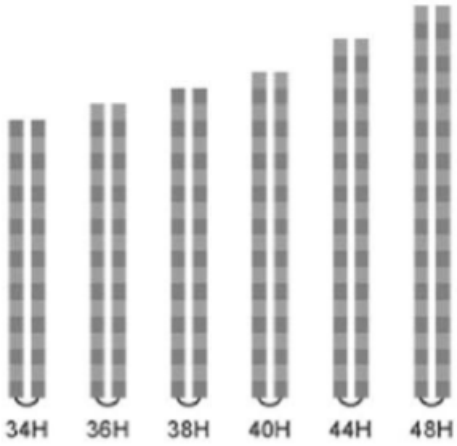


Figure 44: MCP signaling domain length classes observed in the study.

Light and dark gray blocks correspond to heptads (7 aa) for each class (i.e., there is a total of 34 heptads in the 34H MCP signaling domain).

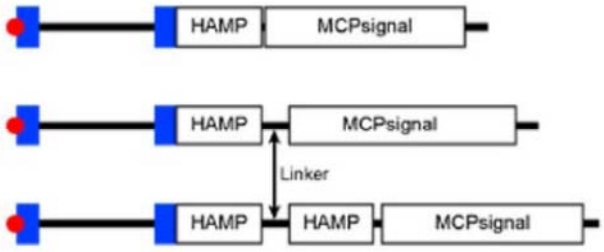


Figure 45: The three different domain architectures of MCP topology type I receptors.

Red circles, predicted signal peptides; blue rectangles, predicted transmembrane regions; HAMP, predicted HAMP domains; linker, predicted linker region; MCPsignal, predicted signaling domains. The domain architecture at the top is most common. Receptor classes with the top, middle, and bottom domain architectures are identified by black, blue, and red labels in Figure 43b, respectively.

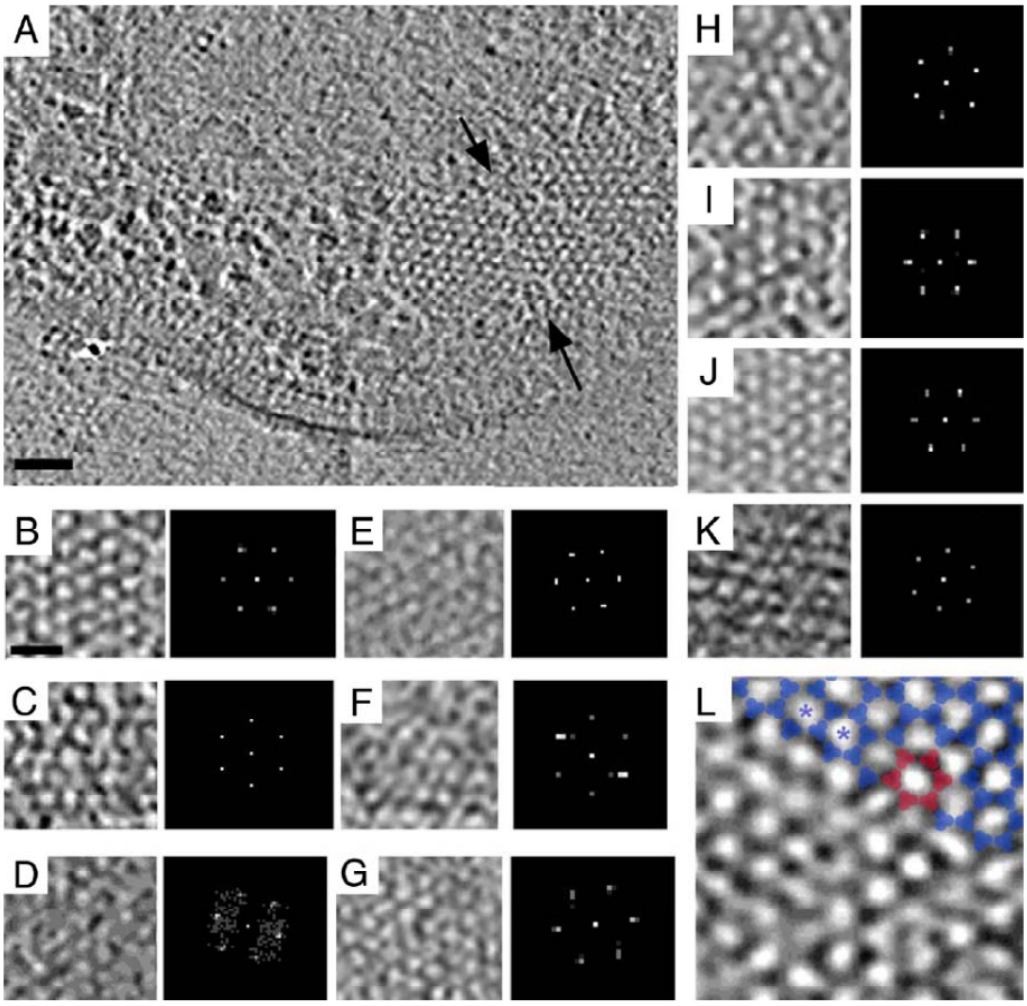


Figure 46: Universally conserved 12-nm hexagonal arrangement of receptor.

(A) Top" view of a chemoreceptor array (black arrows) in *T. maritima* (signaling domain class 44H). Bar: 50 nm. (B–K) Top views (Left) and power spectra (Right) of receptor arrays all reveal the same ~12-nm hexagonal lattice. B, *T.maritima*; C, *A. longum*; D, *C. jejuni*; E, *H. hepaticus*; F, *M. magneticum*; G, *H. neapolitanus*; H, *R. sphaeroides*; I, *E. coli*; J, *V. cholerae*; K, *T. primitia*. (Scale bars: 25 nm; power spectra enlarged.) (L) Trimer of dimers (blue) fit into the vertices of the hexagonal lattice in a chemoreceptor array (V. cholerae). Six trimers of dimers (red) enclose one hexagon. The spacing from the center of one hexagon to the center of an adjacent one is consistently 12 nm (blue asterisks).

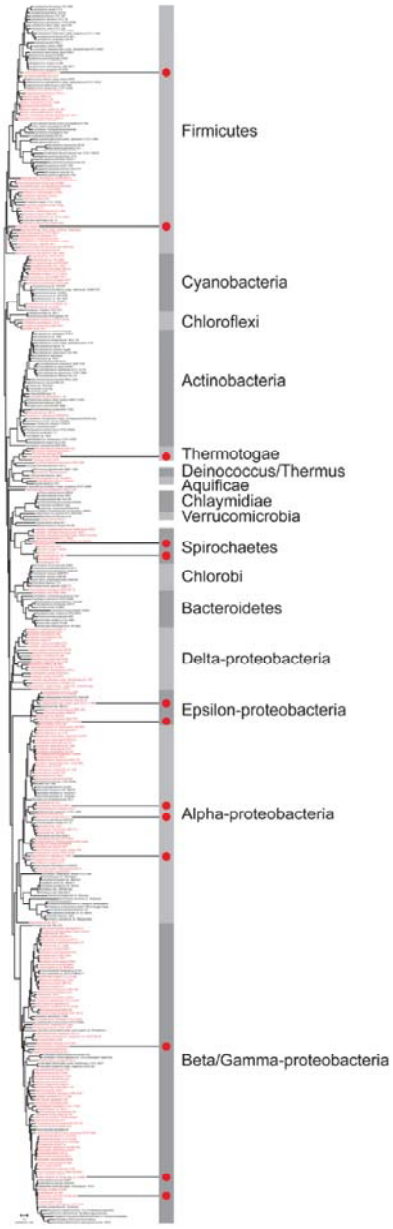


Figure 47: Maximum likelihood phylogenetic tree of 406 representative bacteria.

Organisms that encode one or more MCPs within their genomes are shown in red. Phyla are shown at the right (proteobacteria are further subdivided into classes) with the exception of unique organisms that are sole representatives of their phyla. Red circles and their associated lines show the placement of organisms examined in this study. The Firmicutes clade was used to root the tree, because it is currently considered to be the oldest phylum [181, 182].

vertices of a 12 nm hexagonal lattice. In all the arrays we observed, the honeycomb-like lattice was clearest just above the base plate but deteriorated as it rose towards the inner membrane. These observations support the notion that the major architectural contacts occur near the signaling subdomain of chemoreceptors [26]. While the basic arrangement of all the arrays was clearly hexagonal, none of the arrays were perfectly regular, supporting the idea that the degree of local order could reflect activation and/or regulation [169]. The size of the arrays, and thus the estimated number of receptors, varied by an order of magnitude (from ~1,200 in *M. magneticum* to ~14,400 in *C. jejuni* (Table 6)), without obvious correlation to the cell size or bacterial taxonomy.

Conclusion

Tightly coupled, communicating chemoreceptor arrays are thought to enable the main features of the signaling mechanism: heightened sensitivity [183], signal gain [96], cooperativity [67, 135] and adaptation [98, 184]. The universal hexagonal architecture and secondary structure of chemoreceptor arrays that we observed in diverse bacterial species implies, therefore, that the trimer-of-dimers arrangement and the underlying signaling mechanism are preserved over long evolutionary distances. The main features of the signaling mechanism that are being revealed in *E. coli* are therefore likely to be applicable to other bacterial species. This is important because while chemotaxis is critical to both pathogenic [185] and symbiotic [186] interactions of bacteria with higher organisms, the molecular details of this fascinating system can at present only be studied in a few model organisms.

Materials and Methods.

Strains, sample preparation, EM data collection, and image processing. Bacterial strains (*Caulobacter crescentus* CB15N, *Escherichia coli* RP 437, *Thermotoga maritima* MSB8/DSM 3109, *Vibrio cholerae* TRH7000, *Magnetospirillum magneticum* sp. AMB-1, *Helicobacter hepaticus* ATCC 51449, *Campylobacter jejuni* ATCC 29428, *Rhodobacter sphaeroides* NCIB 8253, *Borrelia burgdorferi* B31 cells ATCC 35210, *Listeria monocytogenes* strain 10403S (serotype 1/2a), *Acetonema longum* APO-1 DSM 6540, *Treponema primitia* strain ZAS-2, *Halothiobacillus neapolitanus* C2 ATCC 23641) were grown in standard media. To flatten the thickest cell types slightly, *E. coli* cells were incubated with 462 iu/ml PenicillinG for 60 min at 30

°C [187] and *T. maritima* cells were treated with 1 mg/ml PolymyxinB for 10 hours on ice. Cultures were plunge-frozen across EM grids as described [188]. Standard EM tilt-series were collected on 300 kV electron cryomicroscopes and three-dimensional reconstructions were calculated as described [189, 190].

MCP sequences and classification. MCP sequences from the complete genomes (*E. coli, C. crescentus, T. maritima, V. cholerae, M. magneticum, H. hepaticus, C. jejuni, R. sphaeroides, B. burgdorferi, L. monocytogenes*) were downloaded from the MIST database [191]. For the draft genomes (*A. longum*, *T. primitia* and *H. neapolitanus),* contigs were subjected to the GeneMark gene finding program [192] to obtain the translated sequences. MCPs were then identified in translated proteins using the MCPsignal domain model (Pfam database [156] accession number PF00015) and the HMMER software package [54]. The final set contained 223 MCPs from 13 genomes (Table 6). MCPs were assigned to signaling classes and membrane topology types as previously described [60, 193]. Sequences which did not match any established signaling class were left unclassified ("unc" in Table 6).

Physical and sequence lengths measurements. Because of the well-understood point-spread-function in electron cryotomography (including a final low-pass filter), the edges of objects appear less sharp in tomograms than they really are. The exact positions of the top and bottom of the inner membrane or CheAW base plate are therefore difficult to ascertain. The location of their midplanes are, however, highly reliable, since the point-spread-function only smooths (and does not shift) peaks. The distance between the peaks (midplanes) of the inner membrane and CheAW base plate was therefore used as an estimate of the physical length of the cytoplasmic portion of the MCPs. Likewise, the center of transmembrane regions can be more reliably predicted from sequence than the edges, and neither is exact because transmembrane helices likely drift up and down a few residues within the fluid bilayer. The sequence length of the "cytoplasmic" domains was therefore taken to be the number of amino acids from the middle of TM2 to the conserved glycine at the tip of the hairpin (Gly390 in the Tsr protein of *E. coli*)(Figure 40). While it is not yet known exactly where the tip of the hairpin is with respect to the midplane of the CheAW base plate, because whatever discrepancy that might exist is likely to be the same for all the receptors, it should not have affected the slope of the correlation between physical and sequence lengths across different receptor classes.

Acknowledgements

The authors wish to thank Eric Matson, Reinhard Rachel, Kevin Bruhn, Gordon Cannon, Alan Barbour, Sarkis Mazmanian, Jeanette Beatty, Maria Sandkvist, Dianne Newman and John S. Parkinson for bacterial strains; Howard Berg for the penicillin treatment protocol for *E. coli*; Jane H. Ding for computational support; Alasdair McDowall and Jian Shi for EM support; and Kristin Wuichet, Roger Alexander, Luke Ulrich and Bhanu Rekapalli for assistance and helpful suggestions. This work was supported in part by NIH grants R01 AI067548 and P50 GM082545 to GJJ and R01 GM72285 to IBZ as well as the Howard Hughes Medical Institute, the Beckman Institute at Caltech, and gifts to Caltech from the Gordon and Betty Moore Foundation and Agouron Institute.

CONCLUSION

In this thesis, different problems in chemotaxis were approached with computational molecular biophysics. In the first three chapters, molecular dynamics techniques were applied to fundamentally diverse problems: the molecular mechanism of cell signaling, protein structure and homology modeling. The last chapter is a classic example of how the scientific training of a physicist is useful in other areas such as biology even without using physics methods. Regardless, the four projects resulted in four novel findings in the fields of chemotaxis and protein structure.

In chapter one, the use of long molecular dynamics simulations reconciles apparently conflicting arguments in the chemoreceptor's dynamic structural properties from two different methods: bioinformatics and crosslinking experiments. The first predicts the center portion of the chemoreceptor to be a flexible bundle while the other predicts a stable structure. We show that both are correct and that bending properties are not strictly related to backbone dynamics in four helix bundles. In addition, this study provides enough evidence to redefine the boundaries of the functional modules of chemoreceptors. Ultimately, and most importantly, the results of this work suggest a novel model for the modulation of the kinase activity by the chemoreceptor. This is the first molecular mechanism proposed to explain the activation of the kinase by conformational changes in the receptor's protein interaction region.

In the second chapter, the confirmation of a salt-bridge formation between two highly conserved residues in the cheW protein family explains a ten year mystery in the field of chemotaxis: if cheW is only a scaffold, as previously thought, why does the mutant R62H have null phenotype and yet manage to keep the wild-type binding affinities with the chemoreceptor and the kinase? Here, molecular dynamics and NMR experiments are combined to show that the aforementioned residue participates in a salt-bridge interaction critical for the proper function of chemotaxis in the organism.

The third chapter addresses a major problem in the field of structural biology. The information obtained by sequencing techniques dramatically outnumbers the number of atomic coordinates resolved for protein structures by X-ray crystallography and NMR. Molecular dynamics is used to access the quality of a common technique used to bridge this information gap: homology modeling. The results show that although the models are structurally close to the template, ensemble sampling techniques such as molecular dynamics or Monte Carlo simulations improves the general quality of the model.

Finally, in chapter four, mathematical modeling and bioinformatics were used to demonstrate that the selection of organisms showing hexagonal pack of transmembrane

chemoreceptors arrays is in fact a broad survey of evolutionarily distant chemoreceptors. This result was central to the claim that the hexagonal packing observed is universal in bacteria.

The field of molecular dynamics simulations is only in its early developments and popularization among young scientists. With the increase in computer power and advances in specialized hardware specific for molecular dynamics, it is expected a higher influx of scientific discoveries using the method. Experimental techniques are clearly still necessary to lay the groundwork prior to *in silico* studies, but the significance of high-throughput computational methods is already undeniable.

Future Aims

The future of computational biophysics relies on the advances of computational methods and technologies yet to come such as new algorithms, faster devices and new architectures. At the current stage, mid-size single molecule molecular dynamics needs specialized machines or large supercomputers to produce a meaningful, statistically tractable, ensemble. However, in the near future, longer and larger simulations will become more feasible and that opens an entire pathway for the field of chemotaxis.

The ultimate goal of simulating the entire chemotaxis complex anchored to a lipid bilayer should be pursued, given the outcomes. Presented here, a complete picture of the individual conformational changes that form the signaling cascade of the chemotaxis system will have an unprecedented impact in health sciences. Chemotaxis is not only a model system for several other cell signaling mechanisms, but it is also essential for understanding infection by pathogenic bacteria and therefore a target for drug design.

However, multiple advances are necessary to successfully simulate the entire chemotaxis complex. First, homology modeling and molecular dynamics simulations are necessary to confidently build missing structures and capture the native states of proteins to be assembled in the complex. Second, bioinformatics and structural biology will be necessary to correctly dock the proteins in the complex. Third and most importantly, the assembled system will have to undergo long (microseconds to miliseconds) molecular dynamics. The complete chemotaxis complex system could be as large as 100 million atoms and simulations of the order of milliseconds will be needed to effectively gain insight in the molecular machinery of the entire system at once. For that, new technologies in hardware and software will be necessary

Conversely, there are some topics that can be promptly addressed. The use of molecular dynamics to produce a reliable, complete, all-atoms structural model of the

chemoreceptor is one example. The link between the HAMP domain and the signaling domain remains subject of research and molecular dynamics can help to provide some answers. The combination of bioinformatics, homology modeling and molecular dynamics can drive major advances to the complete understanding of chemotaxis in the near future.

REFERENCES

1. Lindorff-Larsen, K., et al., *How Fast-Folding Proteins Fold.* Science, 2011. **334**(6055): p. 517-520.
2. Schulz, R., et al., *Scaling of Multimillion-Atom Biological Molecular Dynamics Simulation on a Petascale Supercomputer.* Journal of Chemical Theory and Computation, 2009. **5**(10): p. 2798-2808.
3. Maragakis, P., et al., *Microsecond Molecular Dynamics Simulation Shows Effect of Slow Loop Dynamics on Backbone Amide Order Parameters of Proteins†.* The Journal of Physical Chemistry B, 2008. **112**(19): p. 6155-6158.
4. Caballero-Manrique, E., et al., *A Theory of Protein Dynamics to Predict NMR Relaxation.* 2007. **93**(12): p. 4128-4140.
5. Branden, C.-I. and J. Tooze, *Introduction to Protein Structure*1999: Garland Publishing.
6. Wuichet, K., B.J. Cantwell, and I.B. Zhulin, *Evolution and phyletic distribution of two-component signal transduction systems.* Current Opinion in Microbiology, 2010. **13**(2): p. 219-225.
7. Wadhams, G.H. and J.P. Armitage, *Making sense of it all: bacterial chemotaxis.* Nat Rev Mol Cell Biol, 2004. **5**(12): p. 1024-1037.
8. Rao, C.V., J.R. Kirby, and A.P. Arkin, *Design and Diversity in Bacterial Chemotaxis: A Comparative Study in Escherichia coli and Bacillus subtilis.* PLoS Biol, 2004. **2**(2): p. e49.
9. Hamadeh, A., et al., *Feedback Control Architecture and the Bacterial Chemotaxis Network.* PLoS Comput Biol, 2011. **7**(5): p. e1001130.
10. Wuichet, K. and I.B. Zhulin, *Origins and Diversification of a Complex Signal Transduction System in Prokaryotes.* Sci. Signal., 2010. **3**(128): p. ra50-.
11. Porter, S.L., G.H. Wadhams, and J.P. Armitage, *Signal processing in complex chemotaxis pathways.* Nat Rev Micro, 2011. **9**(3): p. 153-165.
12. Briegel, A., et al., *Universal architecture of bacterial chemoreceptor arrays.* Proc Natl Acad Sci U S A, 2009. **106**(40): p. 17181-6.
13. Park, S.-Y., et al., *Reconstruction of the chemotaxis receptor-kinase assembly.* Nat Struct Mol Biol, 2006. **13**(5): p. 400-407.
14. Kim, K.K., H. Yokota, and S.-H. Kim, *Four-helical-bundle structure of the cytoplasmic domain of a serine chemotaxis receptor.* Nature, 1999. **400**(6746): p. 787-792.
15. Hazelbauer, G.L. and W.-C. Lai, *Bacterial chemoreceptors: providing enhanced features to two-component signaling.* Current Opinion in Microbiology, 2010. **13**(2): p. 124-132.

16. Alon, U., et al., *Robustness in bacterial chemotaxis.* Nature, 1999. **397**(6715): p. 168-171.
17. Grebe, T.W. and J. Stock, *Bacterial chemotaxis: The five sensors of a bacterium.* Current Biology, 1998. **8**(5): p. R154-R157.
18. Gosink, K.K., Y. Zhao, and J.S. Parkinson, *Mutational Analysis of N381, a Key Trimer Contact Residue in Tsr, the Escherichia coli Serine Chemoreceptor.* Journal of Bacteriology, 2011. **193**(23): p. 6452-6460.
19. Massazza, D.A., J.S. Parkinson, and C.A. Studdert, *Cross-Linking Evidence for Motional Constraints within Chemoreceptor Trimers of Dimers.* Biochemistry, 2010. **50**(5): p. 820-827.
20. Briegel, A., et al., *Bacterial chemoreceptor arrays are hexagonally packed trimers of receptor dimers networked by rings of kinase and coupling proteins.* Proceedings of the National Academy of Sciences, 2012. **109**(10): p. 3766-3771.
21. Studdert, C.A. and J.S. Parkinson, *Crosslinking snapshots of bacterial chemoreceptor squads.* Proceedings of the National Academy of Sciences of the United States of America, 2004. **101**(7): p. 2117-2122.
22. Ames, P., et al., *Collaborative signaling by mixed chemoreceptor teams in Escherichia coli.* Proceedings of the National Academy of Sciences, 2002. **99**(10): p. 7060-7065.
23. Li, M. and G.L. Hazelbauer, *Core unit of chemotaxis signaling complexes.* Proceedings of the National Academy of Sciences, 2011. **108**(23): p. 9390-9395.
24. Erbse, A.H. and J.J. Falke, *The Core Signaling Proteins of Bacterial Chemotaxis Assemble To Form an Ultrastable Complex.* Biochemistry, 2009. **48**(29): p. 6975-6987.
25. Liu, J., et al., *Molecular architecture of chemoreceptor arrays revealed by cryoelectron tomography of Escherichia coli minicells.* Proceedings of the National Academy of Sciences, 2012. **109**(23): p. E1481–E1488.
26. Hazelbauer, G.L., J.J. Falke, and J.S. Parkinson, *Bacterial chemoreceptors: high-performance signaling in networked arrays.* Trends in Biochemical Sciences, 2008. **33**(1): p. 9-19.
27. Phillips, J.C., et al., *Scalable molecular dynamics with NAMD.* Journal of Computational Chemistry, 2005. **26**(16): p. 1781-1802.
28. Bernstein, F.C., et al., *The protein data bank: A computer-based archival file for macromolecular structures.* Archives of Biochemistry and Biophysics, 1978. **185**(2): p. 584-591.

29. Freddolino, P.L., et al., *Ten-Microsecond Molecular Dynamics Simulation of a Fast-Folding WW Domain.* 2008. **94**(10): p. L75-L77.
30. Klein, M.L. and W. Shinoda, *Large-Scale Molecular Dynamics Simulations of Self-Assembling Systems.* Science, 2008. **321**(5890): p. 798-800.
31. Sanbonmatsu, K.Y. and C.S. Tung, *High performance computing in biology: Multimillion atom simulations of nanoscale systems.* Journal of Structural Biology, 2007. **157**(3): p. 470-480.
32. Showalter, S.A. and R. Brüschweiler, *Validation of Molecular Dynamics Simulations of Biomolecules Using NMR Spin Relaxation as Benchmarks: Application to the AMBER99SB Force Field.* Journal of Chemical Theory and Computation, 2007. **3**(3): p. 961-975.
33. Freddolino, P.L., et al., *Molecular Dynamics Simulations of the Complete Satellite Tobacco Mosaic Virus.* Structure, 2006. **14**(3): p. 437-449.
34. Mayaan, E., et al., *CHARMM force field parameters for simulation of reactive intermediates in native and thio-substituted ribozymes.* J. Comput. Chem., 2007. **28**(2): p. 495-507.
35. Buck, M., et al., *Importance of the CMAP Correction to the CHARMM22 Protein Force Field: Dynamics of Hen Lysozyme.* 2006. **90**(4): p. L36-L38.
36. Mackerell, A.D., M. Feig, and C.L. Brooks, *Extending the treatment of backbone energetics in protein force fields: Limitations of gas-phase quantum mechanics in reproducing protein conformational distributions in molecular dynamics simulations.* Journal of Computational Chemistry, 2004. **25**(11): p. 1400-1415.
37. MacKerell, A.D., et al., *All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins†.* The Journal of Physical Chemistry B, 1998. **102**(18): p. 3586-3616.
38. Bowers, K.J., et al., *Scalable algorithms for molecular dynamics simulations on commodity clusters*, in *Proceedings of the 2006 ACM/IEEE conference on Supercomputing*2006, ACM: Tampa, Florida. p. 84.
39. Shaw, D.E., et al., *Anton, a special-purpose machine for molecular dynamics simulation.* Commun. ACM, 2008. **51**(7): p. 91-97.
40. Lindorff-Larsen, K., et al., *Systematic Validation of Protein Force Fields against Experimental Data.* PLoS ONE, 2012. **7**(2): p. e32131.

41. Piana, S., K. Lindorff-Larsen, and David E. Shaw, *How Robust Are Protein Folding Simulations with Respect to Force Field Parameterization?* Biophysical journal, 2011. **100**(9): p. L47-L49.
42. Darden, T., D. York, and L. Pedersen, *Particle mesh Ewald: An N [center-dot] log(N) method for Ewald sums in large systems.* The Journal of Chemical Physics, 1993. **98**(12): p. 10089-10092.
43. Shan, Y., et al., *Gaussian split Ewald: A fast Ewald mesh method for molecular simulation.* The Journal of Chemical Physics, 2005. **122**(5): p. 054101.
44. Grubmüller, H., et al., *Generalized Verlet Algorithm for Efficient Molecular Dynamics Simulations with Long-range Interactions.* Molecular Simulation, 1991. **6**(1-3): p. 121-142.
45. Tuckerman, M.E., B.J. Berne, and G.J. Martyna, *Molecular dynamics algorithm for multiple time scales: Systems with long range forces.* The Journal of Chemical Physics, 1991. **94**(10): p. 6811-6815.
46. Ryckaert, J.-P., G. Ciccotti, and H.J.C. Berendsen, *Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes.* Journal of Computational Physics, 1977. **23**(3): p. 327-341.
47. Kräutler, V., W.F. van Gunsteren, and P.H. Hünenberger, *A fast SHAKE algorithm to solve distance constraint equations for small molecules in molecular dynamics simulations.* Journal of Computational Chemistry, 2001. **22**(5): p. 501-508.
48. Quigley, D. and M.I.J. Probert, *Langevin dynamics in constant pressure extended systems.* The Journal of Chemical Physics, 2004. **120**(24): p. 11432-11441.
49. Feller, S., et al., *Constant pressure molecular dynamics simulation: The Langevin piston method.* The Journal of Chemical Physics, 1995. **103**(11): p. 4613-4621.
50. Pruitt, K.D., et al., *NCBI Reference Sequences: current status, policy and new initiatives.* Nucleic Acids Research, 2009. **37**(suppl 1): p. D32-D36.
51. Kimura, M., *The Neutral Theory of Molecular Evolution*. The neutral theory of molecular evolution1983: Cambridge University Press.
52. Klotz, L.C., et al., *Calculation of evolutionary trees from sequence data.* Proceedings of the National Academy of Sciences, 1979. **76**(9): p. 4516-4520.
53. Altschul, S.F., et al., *Basic local alignment search tool.* Journal of Molecular Biology, 1990. **215**(3): p. 403-410.
54. Eddy, S.R., *Profile hidden Markov models.* Bioinformatics, 1998. **14**(9): p. 755-763.

55. Katoh, K. and H. Toh, *Recent developments in the MAFFT multiple sequence alignment program.* Briefings in Bioinformatics, 2008. **9**(4): p. 286-298.
56. Higgs, P. and T. Attwood, *Bioinformatics and Molecular Evolution*2005: Wiley-Blackwell.
57. Milligan, D.L. and D.E. Koshland, *Site-directed cross-linking. Establishing the dimeric structure of the aspartate receptor of bacterial chemotaxis.* Journal of Biological Chemistry, 1988. **263**(13): p. 6268-6275.
58. Falke, J.J., *Cooperativity between bacterial chemotaxis receptors.* Proceedings of the National Academy of Sciences, 2002. **99**(10): p. 6530-6532.
59. Swain, K.E., M.A. Gonzalez, and J.J. Falke, *Engineered socket study of signaling through a four-helix bundle: evidence for a yin-yang mechanism in the kinase control module of the aspartate receptor.* Biochemistry, 2009. **48**(39): p. 9266-77.
60. Alexander, R.P. and I.B. Zhulin, *Evolutionary genomics reveals conserved structural determinants of signaling and adaptation in microbial chemoreceptors.* Proceedings of the National Academy of Sciences, 2007. **104**(8): p. 2885-2890.
61. Miller, A.S., et al., *CheA Kinase of Bacterial Chemotaxis: Chemical Mapping of Four Essential Docking Sites†.* Biochemistry, 2006. **45**(29): p. 8699-8711.
62. Liu, J.D. and J.S. Parkinson, *Role of CheW protein in coupling membrane receptors to the intracellular signaling system of bacterial chemotaxis.* Proceedings of the National Academy of Sciences of the United States of America, 1989. **86**(22): p. 8703-8707.
63. Coleman, M.D., et al., *Conserved Glycine Residues in the Cytoplasmic Domain of the Aspartate Receptor Play Essential Roles in Kinase Coupling and On−Off Switching†.* Biochemistry, 2005. **44**(21): p. 7687-7695.
64. Vaknin, A. and H.C. Berg, *Physical Responses of Bacterial Chemoreceptors.* Journal of Molecular Biology, 2007. **366**(5): p. 1416-1423.
65. Vaknin, A. and H.C. Berg, *Osmotic stress mechanically perturbs chemoreceptors in Escherichia coli.* Proceedings of the National Academy of Sciences of the United States of America, 2006. **103**(3): p. 592-596.
66. Rice, M.S. and F.W. Dahlquist, *Sites of deamidation and methylation in Tsr, a bacterial chemotaxis sensory transducer.* Journal of Biological Chemistry, 1991. **266**(15): p. 9746-53.
67. Li, G. and R.M. Weis, *Covalent Modification Regulates Ligand Binding to Receptor Complexes in the Chemosensory System of Escherichia coli.* Cell, 2000. **100**(3): p. 357-365.

68. Bornhorst, J.A. and J.J. Falke, *Evidence That Both Ligand Binding and Covalent Adaptation Drive a Two-State Equilibrium in the Aspartate Receptor Signaling Complex.* J. Gen. Physiol., 2001. **118**(6): p. 693-710.
69. Dunten, P. and D.E. Koshland, *Tuning the responsiveness of a sensory receptor via covalent modification.* Journal of Biological Chemistry, 1991. **266**(3): p. 1491-1496.
70. Lai, W.-C., B.D. Beel, and G.L. Hazelbauer, *Adaptational modification and ligand occupancy have opposite effects on positioning of the transmembrane signalling helix of a chemoreceptor.* Molecular Microbiology, 2006. **61**(4): p. 1081-1090.
71. Falke, J.J. and G.L. Hazelbauer, *Transmembrane signaling in bacterial chemoreceptors.* Trends in Biochemical Sciences, 2001. **26**(4): p. 257-265.
72. Hall, B.A., J.P. Armitage, and M.S.P. Sansom, *Transmembrane Helix Dynamics of Bacterial Chemoreceptors Supports a Piston Model of Signalling.* PLoS Comput Biol, 2011. **7**(10): p. e1002204.
73. Park, H., W. Im, and C. Seok, *Transmembrane Signaling of Chemotaxis Receptor Tar: Insights from Molecular Dynamics Simulation Studies.* Biophysical journal, 2011. **100**(12): p. 2955-2963.
74. Hall, B.A., J.P. Armitage, and M.S.P. Sansom, *Mechanism of Bacterial Signal Transduction Revealed by Molecular Dynamics of Tsr Dimers and Trimers of Dimers in Lipid Vesicles.* PLoS Comput Biol, 2012. **8**(9): p. e1002685.
75. Danielson, M.A., R.B. Bass, and J.J. Falke, *Cysteine and Disulfide Scanning Reveals a Regulatory α-Helix in the Cytoplasmic Domain of the Aspartate Receptor.* Journal of Biological Chemistry, 1997. **272**(52): p. 32878-32888.
76. Lipari, G. and A. Szabo, *Model-free approach to the interpretation of nuclear magnetic resonance relaxation in macromolecules. 1. Theory and range of validity.* Journal of the American Chemical Society, 1982. **104**(17): p. 4546-4559.
77. Pazos, F., et al., *Correlated mutations contain information about protein-protein interaction.* Journal of Molecular Biology, 1997. **271**(4): p. 511-523.
78. Jones, S. and J.M. Thornton, *Principles of protein-protein interactions.* Proceedings of the National Academy of Sciences of the United States of America, 1996. **93**(1): p. 13-20.
79. Starrett, D.J. and J.J. Falke, *Adaptation mechanism of the aspartate receptor: electrostatics of the adaptation subdomain play a key role in modulating kinase activity.* Biochemistry, 2005. **44**(5): p. 1550-60.

80. Le Moual, H. and J.D.E. Koshland, *Molecular Evolution of the C-terminal Cytoplasmic Domain of a Superfamily of Bacterial Receptors Involved in Taxis.* Journal of Molecular Biology, 1996. **261**(4): p. 568-585.
81. Humphrey, W., A. Dalke, and K. Schulten, *VMD: Visual molecular dynamics.* Journal of Molecular Graphics, 1996. **14**(1): p. 33-38.
82. QUENOUILLE, M.H., *NOTES ON BIAS IN ESTIMATION.* Biometrika, 1956. **43**(3-4): p. 353-360.
83. Philippopoulos, M. and C. Lim, *Molecular Dynamics Simulation ofE. coliRibonuclease H1in Solution: Correlation with NMR and X-ray Data and Insights into Biological Function.* Journal of Molecular Biology, 1995. **254**(4): p. 771-792.
84. Chandrasekhar, I., et al., *A 500 ps molecular dynamics simulation study of interleukin-1[beta] in water : Correlation with nuclear magnetic resonance spectroscopy and crystallography.* Journal of Molecular Biology, 1992. **226**(1): p. 239-250.
85. Chen, J., C.L. Brooks, and P.E. Wright, *Model-free Analysis of Protein Dynamics: Assessment of Accuracy and Model Selection Protocols Based on Molecular Dynamics Simulation.* Journal of Biomolecular NMR, 2004. **29**(3): p. 243-257.
86. Nederveen, A.J. and A.M.J.J. Bonvin, *NMR Relaxation and Internal Dynamics of Ubiquitin from a 0.2 μs MD Simulation.* Journal of Chemical Theory and Computation, 2005. **1**(3): p. 363-374.
87. Markwick, P.R.L., G. Bouvignies, and M. Blackledge, *Exploring Multiple Timescale Motions in Protein GB3 Using Accelerated Molecular Dynamics and NMR Spectroscopy.* Journal of the American Chemical Society, 2007. **129**(15): p. 4724-4730.
88. Trbovic, N., et al., *Structural analysis of protein dynamics by MD simulations and NMR spin-relaxation.* Proteins: Structure, Function, and Bioinformatics, 2008. **71**(2): p. 684-694.
89. Ulrich, L.E. and I.B. Zhulin, *The MiST2 database: a comprehensive genomics resource on microbial signal transduction.* Nucleic Acids Research, 2010. **38**(suppl 1): p. D401-D407.
90. Finn, R.D., J. Clements, and S.R. Eddy, *HMMER web server: interactive sequence similarity searching.* Nucleic Acids Research, 2011. **39**(suppl 2): p. W29-W37.
91. Finn, R.D., et al., *Pfam: clans, web tools and services.* Nucleic Acids Research, 2006. **34**(suppl 1): p. D247-D251.
92. Crooks, G.E., et al., *WebLogo: A Sequence Logo Generator.* Genome Research, 2004. **14**(6): p. 1188-1190.

93. Stock, A.M., V.L. Robinson, and P.N. Goudreau, *TWO-COMPONENT SIGNAL TRANSDUCTION.* Annual Review of Biochemistry, 2000. **69**(1): p. 183-215.
94. Ulrich, L.E., E.V. Koonin, and I.B. Zhulin, *One-component systems dominate signal transduction in prokaryotes.* Trends in microbiology, 2005. **13**(2): p. 52-56.
95. Gestwicki, J.E. and L.L. Kiessling, *Inter-receptor communication through arrays of bacterial chemoreceptors.* Nature, 2002. **415**(6867): p. 81-84.
96. Sourjik, V. and H.C. Berg, *Receptor sensitivity in bacterial chemotaxis.* Proceedings of the National Academy of Sciences of the United States of America, 2002. **99**(1): p. 123-127.
97. Sourjik, V. and H.C. Berg, *Functional interactions between receptors in bacterial chemotaxis.* Nature, 2004. **428**(6981): p. 437-441.
98. Endres, R.G. and N.S. Wingreen, *Precise adaptation in bacterial chemotaxis through "assistance neighborhoods".* Proceedings of the National Academy of Sciences, 2006. **103**(35): p. 13040-13044.
99. Bray, D., M.D. Levin, and C.J. Morton-Firth, *Receptor clustering as a cellular mechanism to control sensitivity.* Nature, 1998. **393**(6680): p. 85-88.
100. Griswold, I.J., et al., *The solution structure and interactions of CheW from Thermotoga maritima.* Nat Struct Biol, 2002. **9**(2): p. 121-5.
101. Bhatnagar, J., et al., *Structure of the Ternary Complex Formed by a Chemotaxis Receptor Signaling Domain, the CheA Histidine Kinase, and the Coupling Protein CheW As Determined by Pulsed Dipolar ESR Spectroscopy.* Biochemistry, 2010. **49**(18): p. 3824-3841.
102. Underbakke, E.S., Y. Zhu, and L.L. Kiessling, *Protein Footprinting in a Complex Milieu: Identifying the Interaction Surfaces of the Chemotaxis Adaptor Protein CheW.* Journal of Molecular Biology, 2011. **409**(4): p. 483-495.
103. Li, Y., et al., *Solution structure of the bacterial chemotaxis adaptor protein CheW from Escherichia coli.* Biochem Biophys Res Commun, 2007. **360**(4): p. 863-7.
104. Yao, W., L. Shi, and D.-C. Liang, *Crystal structure of scaffolding protein CheW from thermoanaerobacter tengcongensis.* Biochemical and Biophysical Research Communications, 2007. **361**(4): p. 1027-1032.
105. Borkovich, K.A., et al., *Transmembrane signal transduction in bacterial chemotaxis involves ligand-dependent activation of phosphate group transfer.* Proceedings of the National Academy of Sciences, 1989. **86**(4): p. 1208-1212.

106. Gegner, J.A., et al., *Assembly of an MCP receptor, CheW, and kinase CheA complex in the bacterial chemotaxis signal transduction pathway.* 1992. **70**(6): p. 975-982.
107. Vu, A., et al., *The Receptor–CheW Binding Interface in Bacterial Chemotaxis.* Journal of Molecular Biology, 2012. **415**(4): p. 759-767.
108. Cardozo, M.J., et al., *Disruption of chemoreceptor signalling arrays by high levels of CheW, the receptor–kinase coupling protein.* Molecular Microbiology, 2010. **75**(5): p. 1171-1181.
109. Liu, J.D. and J.S. Parkinson, *Genetic evidence for interaction between the CheW and Tsr proteins during chemoreceptor signaling by Escherichia coli.* J Bacteriol, 1991. **173**(16): p. 4941-51.
110. Boukhvalova, M.S., F.W. Dahlquist, and R.C. Stewart, *CheW Binding Interactions with CheA and Tar.* Journal of Biological Chemistry, 2002. **277**(25): p. 22251-22259.
111. Boukhvalova, M., R. VanBruggen, and R.C. Stewart, *CheA Kinase and Chemoreceptor Interaction Surfaces on CheW.* Journal of Biological Chemistry, 2002. **277**(26): p. 23596-23603.
112. Bilwes, A.M., et al., *Structure of CheA, a Signal-Transducing Histidine Kinase.* 1999. **96**(1): p. 131-141.
113. Alexander, R.P., et al., *CheV: CheW-like coupling proteins at the core of the chemotaxis signaling network.* Trends in microbiology, 2010. **18**(11): p. 494-503.
114. Alexandre, G. and I.B. Zhulin, *Different Evolutionary Constraints on Chemotaxis Proteins CheW and CheY Revealed by Heterologous Expression Studies and Protein Sequence Analysis.* Journal of Bacteriology, 2003. **185**(2): p. 544-552.
115. Kay, L.E., D.A. Torchia, and A. Bax, *Backbone dynamics of proteins as studied by nitrogen-15 inverse detected heteronuclear NMR spectroscopy: application to staphylococcal nuclease.* Biochemistry, 1989. **28**(23): p. 8972-8979.
116. Clore, G.M., et al., *Analysis of the backbone dynamics of interleukin-1.beta. using two-dimensional inverse detected heteronuclear nitrogen-15-proton NMR spectroscopy.* Biochemistry, 1990. **29**(32): p. 7387-7401.
117. Palmer, A.G., M. Rance, and P.E. Wright, *Intramolecular motions of a zinc finger DNA-binding domain from Xfin characterized by proton-detected natural abundance carbon-13 heteronuclear NMR spectroscopy.* Journal of the American Chemical Society, 1991. **113**(12): p. 4371-4380.

118. Mandel, A.M., M. Akke, and I.I.I.A.G. Palmer, *Backbone Dynamics ofEscherichia coliRibonuclease HI: Correlations with Structure and Function in an Active Enzyme.* Journal of Molecular Biology, 1995. **246**(1): p. 144-163.
119. Zhou, H., et al., *Long-Range Structural and Dynamical Changes Induced by Cofactor Binding in DNA Methyltransferase M.HhaI‡.* Biochemistry, 2007. **46**(24): p. 7261-7268.
120. Griswold, I.J. and F.W. Dahlquist, *The dynamic behavior of CheW from Thermotoga maritima in solution, as determined by nuclear magnetic resonance: implications for potential protein-protein interaction sites.* Biophysical Chemistry, 2002. **101-102**: p. 359-373.
121. Carr, H.Y. and E.M. Purcell, *Effects of Diffusion on Free Precession in Nuclear Magnetic Resonance Experiments.* Physical Review, 1954. **94**(3): p. 630-638.
122. Meiboom, S. and D. Gill, *Modified Spin-Echo Method for Measuring Nuclear Relaxation Times.* Review of Scientific Instruments, 1958. **29**(8): p. 688-691.
123. Loria, J.P., M. Rance, and A.G. Palmer, *A Relaxation-Compensated Carr−Purcell−Meiboom−Gill Sequence for Characterizing Chemical Exchange by NMR Spectroscopy.* Journal of the American Chemical Society, 1999. **121**(10): p. 2331-2332.
124. Abragam, A., *Principle of Nuclear Magnetism*1961, Oxford: Clarendon Press.
125. Woessner, D.E., *Nuclear Spin Relaxation in Ellipsoids Undergoing Rotational Brownian Motion.* The Journal of Chemical Physics, 1962. **37**(3): p. 647-654.
126. Kumar, S. and R. Nussinov, *Salt bridge stability in monomeric proteins.* Journal of Molecular Biology, 1999. **293**(5): p. 1241-1255.
127. Bécavin, C., et al., *Improving the efficiency of multidimensional scaling in the analysis of high-dimensional data using singular value decomposition.* Bioinformatics, 2011. **27**(10): p. 1413-1421.
128. Kraulis, P.J., *ANSIG: A program for the assignment of protein 1H 2D NMR spectra by interactive computer graphics.* Journal of Magnetic Resonance (1969), 1989. **84**(3): p. 627-633.
129. Delaglio, F., et al., *NMRPipe: A multidimensional spectral processing system based on UNIX pipes.* Journal of Biomolecular NMR, 1995. **6**(3): p. 277-293.
130. Lipari, G. and A. Szabo, *Model-free approach to the interpretation of nuclear magnetic resonance relaxation in macromolecules. 2. Analysis of experimental results.* Journal of the American Chemical Society, 1982. **104**(17): p. 4559-4570.
131. Jorgensen, W.L., et al., *Comparison of simple potential functions for simulating liquid water.* The Journal of Chemical Physics, 1983. **79**(2): p. 926-935.

132. Schlick, T., et al., *Algorithmic challenges in computational molecular biophysics.* J. Comput. Phys., 1999. **151**(1): p. 9-48.
133. Sourjik, V. and J.P. Armitage, *Spatial organization in bacterial chemotaxis.* The EMBO journal, 2010. **29**(16): p. 2724-33.
134. Maddock, J.R. and L. Shapiro, *Polar location of the chemoreceptor complex in the Escherichia coli cell.* Science, 1993. **259**(5102): p. 1717-1723.
135. Sourjik, V., *Receptor clustering and signal processing in E. coli chemotaxis.* Trends in microbiology, 2004. **12**(12): p. 569-576.
136. Szurmant, H. and G.W. Ordal, *Diversity in Chemotaxis Mechanisms among the Bacteria and Archaea.* Microbiology and Molecular Biology Reviews, 2004. **68**(2): p. 301-319.
137. Sali, A., et al., *Three-dimensional models of four mouse mast cell chymases. Identification of proteoglycan binding regions and protease-specific antigenic epitopes.* The Journal of biological chemistry, 1993. **268**(12): p. 9023-34.
138. Matsumoto, R.S., A.; Ghildyal, N.; Karplus, M.; Stevens, R.L., *Packaging of Proteases and Proteoglycans in the Granules of Mast Cells and Other Hematopoietic Cells: A Cluster of Histidines on Mouse Mast Cell Protease 7 Regulates Its Binding to Heparin Serglycin Proteoglycans.* J. Biol. Chem., 1995. **270**: p. 19524-19531.
139. Caputo, A.J., M.N.G.; Powers, J.C.; Hudig, D.; Bleackley, R.C., *Conversion of the substrate specificity of mouse proteinase granzyme B.* Nature Structural & Molecular Biology, 1994. **1**(6): p. 364-367.
140. Ring, C.S., et al., *Structure-based inhibitor design by using protein models for the development of antiparasitic agents.* Proceedings of the National Academy of Sciences of the United States of America, 1993. **90**(8): p. 3583-7.
141. Novoa, E.M.d.P., L.R.; Barril, X.; Orozco, M., *Ensemble Docking from Homology Models.* Journal of Chemical Theory and Computation, 2010. **6**: p. 2547-2557.
142. Droupadi, P.R., J.M. Varga, and D.S. Linthicum, *Mechanism of allergenic cross-reactions--IV. Evidence for participation of aromatic residues in the ligand binding site of two multi-specific IgE monoclonal antibodies.* Molecular immunology, 1994. **31**(7): p. 537-48.
143. Carson, M.B., C.E.; DeLucas, L.J.; Narayana, S.V.L., *Structure of a Novel Serine Protease Compared to Homology Models.* Acta Cryst., 1994. **D50**: p. 889-899.
144. Sutcliffe, M.J., C.M. Dobson, and R.E. Oswald, *Solution structure of neuronal bungarotoxin determined by two-dimensional NMR spectroscopy: calculation of tertiary*

structure using systematic homologous model building, dynamical simulated annealing, and restrained molecular dynamics. Biochemistry, 1992. **31**(11): p. 2962-70.

145. Baker, D. and A. Sali, *Protein Structure Prediction and Structural Genomics.* Science, 2001. **294**(5540): p. 93-96.
146. Doolittle, R.F., *Of Urfs and Orfs: Primer on how to analyze derived amino acid sequences.*1986, Mill Valley, CA: University Science Books.
147. Bork, P., L. Holm, and C. Sander, *The immunoglobulin fold. Structural classification, sequence patterns and common core.* Journal of molecular biology, 1994. **242**(4): p. 309-20.
148. Rougon, G.H., O., *New insights into the diversity and*

function of neuronal immunoglobulin superfamily molecules. Following An. Rev. Neurosci., 2003. **26**: p. 207-238.

149. Smith, D.K.X., H., *Sequence profiles of immunoglobulin and immunoglobulin-like domains.* J. Mol. Biol., 1997. **274**: p. 530-545.
150. Williams, A.F.B., A.N., *The immunoglobulin superfamily—domains for cell surface recognition.* Ann. Rev. Immunol., 1988. **6**: p. 381-405.
151. Farber, G.K. and G.A. Petsko, *The evolution of alpha/beta barrel enzymes.* Trends Biochem. Sci., 1990. **15**: p. 228-234.
152. Murphy, M.E., P.F. Lindley, and E.T. Adman, *Structural comparison of cupredoxin domains: domain recycling to construct proteins with novel functions.* Prot. Sci., 1997. **6**: p. 761-770.
153. Nelson, D.R. and H.W. Strobel, *On the membrane topology of vertebrate cytochrome P-450 proteins.* The Journal of biological chemistry, 1988. **263**(13): p. 6038-50.
154. Nelson, D.R. and H.W. Strobel, *Secondary structure prediction of 52 membrane-bound cytochromes P450 shows a strong structural similarity to P450cam.* Biochemistry, 1989. **28**(2): p. 656-60.
155. Hasemann, C.A., et al., *Structure and function of cytochromes P450: a comparative analysis of three crystal structures.* Structure, 1995. **3**(1): p. 41-62.
156. Finn, R.D., et al., *The Pfam protein families database.* Nucleic Acids Research, 2008. **36**: p. D281-288.
157. Consortium, T.U., *Reorganizing the protein space at the Universal Protein Resource (UniProt).* Nucleic Acids Research, 2012. **40**(D1): p. D71-D75.

158. Bernstein, F.C., et al., *The Protein Data Bank. A computer-based archival file for macromolecular structures.* European journal of biochemistry / FEBS, 1977. **80**(2): p. 319-24.
159. Labute, P., *The generalized Born/volume integral implicit solvent model: estimation of the free energy of hydration using London dispersion instead of atomic surface area.* Journal of computational chemistry, 2008. **29**(10): p. 1693-8.
160. Mamonov, A.B., et al., *General library-based Monte Carlo technique enables equilibrium sampling of semi-atomistic protein models.* The journal of physical chemistry. B, 2009. **113**(31): p. 10891-904.
161. Taketomi, H., Y. Ueda, and N. Gō, *Studies on Protein Folding, Unfolding and Fluctuations by Computer Simulation. I. The effect of specific amino acid sequence represented by specific inter-unit interactions.* Int. J. Pept. Protein Res., 1975. **7**(6): p. 445-459.
162. Ueda, Y., H. Taketomi, and N. Go, *Studies of protein folding, unfolding, and fluctuations by computer simulation, II. 3-dimensional lattice model for lysozyme.* Biopolymers, 1978. **17**: p. 1531-1538.
163. Zhulin, I.B., *The superfamily of chemotaxis transducers: from physiology to genomics and back.* Adv. Microb. Physiol. , 2001. **45**: p. 157-198.
164. Boukhvalova, M.S., F.W. Dahlquist, and R.C. Stewart, *CheW binding interactions with CheA and Tar: Importance for chemotaxis signaling in E. coli* J Biol Chem, 2002. **277**: p. 22251-22259.
165. Lybarger, S.R. and J. Maddock, *Polarity in Action: Asymmetric Protein Localization in Bacteria.* J Bacteriol, 2001. **183**(11): p. 3261-3267.
166. Boldog, T., et al., *Nanodiscs separate chemoreceptor oligomeric states and reveal their signaling properties.* Proc Natl Acad Sci U S A, 2006. **103**(31): p. 11509-14.
167. Studdert, C.A. and J.S. Parkinson, *In vivo crosslinking methods for analyzing the assembly and architecture of chemoreceptor arrays.* Methods Enzymol, 2007. **423**: p. 414-431.
168. Shimizu, T.S., et al., *Molecular model of a lattice of signalling proteins involved in bacterial chemotaxis.* Nat Cell Biol, 2000. **2**(11): p. 792-796.
169. Khursigara, C.M., X. Wu, and S. Subramaniam, *Chemoreceptors in Caulobacter crescentus: trimers of receptor dimers in a partially ordered hexagonally packed array.* J Bacteriol, 2008. **190**(20): p. 6805-6810.

170. Lefman, J., et al., *Three-dimensional electron microscopic imaging of membrane invaginations in Escherichia coli overproducing the chemotaxis receptor Tsr.* J Bacteriol, 2004. **186**(15): p. 5052-5061.
171. Weis, R.M., et al., *Electron microscopic analysis of membrane assemblies formed by the bacterial chemotaxis receptor Tsr.* J Bacteriol, 2003. **185**(12): p. 3636-3643.
172. Zhang, P., et al., *Direct visualization of receptor arrays in frozen-hydrated sections and plunge-frozen specimens of E. coli engineered to overproduce the chemotaxis receptor Tsr.* J Microsc, 2004. **216**: p. 76-83.
173. Wolanin, P.M., et al., *Self-assembly of receptor / signaling complexes in bacterial chemotaxis.* Proc Natl Acad Sci U S A, 2006. **103**(39): p. 14313-14318.
174. Briegel, A., et al., *Location and architecture of the Caulobacter crescentus chemoreceptor array.* Molecular Microbiology, 2008. **69**(1): p. 30-41.
175. Zhang, P., et al., *Direct visualization of Escherichia coli chemotaxis receptor arrays using cryo-electron microscopy.* Proceedings of the National Academy of Sciences, 2007. **104**(10): p. 3777-3781.
176. Porter, S.L., G.H. Wadhams, and J.P. Armitage, *Rhodobacter sphaeroides: complexity in chemotactic signalling.* Trends Microbiol, 2008. **16**(6): p. 251-260.
177. Wadhams, G.H., et al., *TlpC, a novel chemotaxis protein in Rhodobacter sphaeroides, localizes to a discrete region in the cytoplasm.* Mol Microbiol, 2002. **46**(5): p. 1211-1221.
178. Wadhams, G.H., et al., *Targeting of two signal transduction pathways to different regions of the bacterial cell.* Mol Microbiol, 2003. **50**(3): p. 763-770.
179. Hulko, M., et al., *The HAMP Domain Structure Implies Helix Rotation in Transmembrane Signaling.* Cell, 2006. **126**(5): p. 929-940.
180. Aravind, L. and C.P. Ponting, *The cytoplasmic helical linker domain of receptor histidine kinase and methyl-accepting proteins is common to many prokaryotic signalling proteins.* FEMS Microbiology Letters, 1999. **176**(1): p. 111-116.
181. Ciccarelli, F.D., et al., *Toward Automatic Reconstruction of a Highly Resolved Tree of Life.* Science, 2006. **311**(5765): p. 1283-1287.
182. Gupta, R.S., *The natural evolutionary relationships among prokaryotes.* Crit Rev Microbiol, 2000. **26**: p. 111-131.
183. Duke, T.A.J. and D. Bray, *Heightened sensitivity of a lattice of membrane receptors.* Proc Natl Acad Sci U S A, 1999. **96**(18): p. 10104-10108.
184. Li, M. and G.L. Hazelbauer, *Adaptational assistance in clusters of bacterial chemoreceptors.* Mol Microbiol, 2005. **56**(6): p. 1617-1626.

185. Butler, S.M. and A. Camilli, *Both chemotaxis and net motility greatly influence the infectivity of Vibrio cholerae.* Proc Natl Acad Sci U S A, 2004. **101**(14): p. 5018-5023.
186. Miller, L.D., et al., *The major chemotaxis gene cluster of Rhizobium leguminosarum bv. viciae is essential for competitive nodulation.* Mol Microbiol, 2007. **63**: p. 348-362.
187. Eisenbach, M. and J. Adler, *Bacterial cell envelopes with functional flagella.* J Biol Chem, 1981. **256**(16): p. 8807-8814.
188. Iancu, C.V., et al., *Electron cryotomography sample preparation using the Vitrobot.* Nat Protoc, 2007. **1**: p. 2813-2819.
189. Mastronarde, D.A., *Dual-axis tomography: An approach with alignment methods that preserve resolution.* J Struct Biol, 1997. **120**: p. 343-352.
190. Hegerl, R., *The EM program package: a platform for image processing in biological electron microscopy.* J Struct Biol, 1996. **116**: p. 30-34.
191. Ulrich, L.E. and I.B. Zhulin, *MiST: a microbial signal transduction database.* Nucleic Acids Res, 2007. **35**: p. D386-90.
192. Besemer, J. and M. Borodovsky, *GeneMark: web software for gene finding in prokaryotes, eukaryotes and viruses.* Nucleic Acids Research, 2005. **33**: p. W451-454.
193. Wuichet, K., R.P. Alexander, and I.B. Zhulin, *Comparative genomic and protein sequence analyses of a complex system controlling bacterial chemotaxis.* Methods Enzymol, 2007. **422**: p. 1-31.

VITA

Davi Ortega was born in São José do Rio Preto, SP, Brazil. He did his undergraduate studies at Unicamp in Campinas, SP, Brazil. He was once a guest researcher at NIST in Boulder, CO working with calcium optical atomic clocks. He moved to Knoxville, TN to work on his PhD. in computational molecular biophysics. Davi likes to play music, to drink beer with friends, to play chess; he also likes new technology and politics. This year seems to be a good one: he recorded an album with his band Siricutico, achieved the blue-belt level at jiu-jitsu and is likely to get his PhD. Good year!