# A phenylalanine rotameric switch for signal-state control in bacterial chemoreceptors

Davi R. Ortega[a], Chen Yang[b], Peter Ames[b], Jerome Baudry*[c,d], John S. Parkinson[b] and Igor B. Zhulin*[a,f,e]

[a]Joint Institute for Computational Sciences, University of Tennessee - Oak Ridge National Laboratory, Oak Ridge TN 37861;

[b]Department of Biology, University of Utah, Salt Lake City UT 84112;

[c]Department of Biochemistry and Cellular and Molecular Biology, University of Tennessee, Knoxville TN 37996;

[d]Center for Molecular Biophysics, University of Tennessee - Oak Ridge National Laboratory, Oak Ridge TN 37861;

[f]Department of Microbiology, University of Tennessee, Knoxville TN 37996

[e]Computer Science and Mathematics Division, Oak Ridge National Laboratory, Oak Ridge, TN 37861

*To whom correspondence may be addressed. Email: jbaudry@utk.edu and joulineib@ornl.gov

## Materials and 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 [1]. 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 [1-3]. The structure was embedded in water, tip3p, neutralized and 5 mM of NaCl was added. The total simulation system size was 144,647 atoms (90 x 90 x 182 $Å^3$ ). 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 molecular dynamics engine Desmond 2.4 [4] in the Newton supercomputer at University of Tennessee using 512 nodes for pre-equilibration of the system in NPT ensemble with Berendsen thermostat at 300K constant temperature and 1 atm pressure. The system was then transferred to the 512 node, special-purpose supercomputer, Anton[5] 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 9.1 (Schrodinger, Inc.). The velocities were initialized using Desmond 2.4 prior to be transferred to Anton. Each of the three production simulation was 2μs long. All simulations used CHARMM27[6-9] 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 [10]. The 64 grid points over 220 A length could lead to large errors during the computation of forces. However, this is a hardware limitation of Anton and we increased the cutoff of electrostatics interactions to compensate for this hardware limitation. In fact, the cutoff was not set at random. During the preparation of the system, a short simulation was performed to test the accuracy of force computation on Anton. The program measures the rms force error, defined as the rms error in the force on all particles divided by the rms force. The simulation is only cleared for execution if the relative rms force error is below 0.001, which is considered sufficiently accurate for biomolecular MD simulations on Anton[5]. Consequently, we used the smallest possible cutoff for short range interactions and van der Waals at 16.75 Å , that guarantees best performance with rms force error below 0.001, with sufficient accuracy. The simulations time step was 1 fs and respa scheme 1:1:3 meaning that long-range electrostatic interactions were calculated every third step.

**Local alignment per residue protocol for calculations of the order parameter.**

The current methodology to calculate order parameter assumes that the frames of the simulations have been aligned to a reference frame to avoid coupling between rotational and/or translational movements and the internal motions. 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 overcome this problem we suggest a procedure to minimize the problem of frame alignment in anisotropic structures: local alignment per residue protocol. As the internal correlation function is calculated for each residue, each frame of the simulation is aligned to the reference frame using only a selection of atoms within a certain distance from the target residue. This custom selection of atoms per residue is insensitive to

translational/rotational motions between parts of the structure. Large enough cutoff retrieve the orthodox approach. Here we used 30 Å cutoff. The result is robust to cutoff variations.

**Calculation of the order parameter**. The order parameter is defined as [11-16]:

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

where $C_i(\infty)$ is the internal correlation function when $t \to \infty$. Also, $t$ and $\tau$ 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 1 requires a convergence of $C_i(t)$ as $t$ increases. To verify the convergence, we calculate the correlation function as:

$$C_I(t) = \langle P_2\big(\hat{\mu}(0)\cdot\hat{\mu}(t)\big)\rangle \quad (2)$$

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 [13]. If there is no convergence, the order parameter is considered null.

**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 (Fig. S8).

The calculations were performed using the function “measure inertia” from VMD [17]. 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 [18,19]. 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 deviation of the values in the time series.

**Trajectory analysis**. Dihedral angles and alpha carbon distances were extracted from the trajectories by custom python scripts using the molecular dynamics toolkit MDAnalysis v0.7.5 [20]. Correlation function between the time series of dihedral angle and helix-helix distances (alpha carbon distances) were calculated with “cor” function from R statistical package v2.14.1. In Fig. S7, the result is an average between the absolute correlation values for chain A and B of the receptor. Histograms and time series were plotted with ggplot2 [21] package for R. Figures and movies were made using VMD 1.9.1 [17] that also allowed for visual analysis of the trajectories and RMSD measurements.

**Bioinformatics.** We selected all 12,498 chemoreceptor sequences from complete genomes in the MIST database as in August 2012 [22]. Using HMM models previously published [3], the chemoreceptors were classified and separated in different files according to its heptad classes using HMMER [23]. 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 [24] 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 [25]. 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 [26].

## References

1 Kim, K. K., Yokota, H. & Kim, S.-H. Four-helical-bundle structure of the cytoplasmic domain of a serine chemotaxis receptor. *Nature* **400**, 787-792, (1999).

2 Le Moual, H. & Koshland, D. E., Jr. Molecular evolution of the C-terminal cytoplasmic domain of a superfamily of bacterial receptors involved in taxis. *J Mol Biol* **261**, 568-585, (1996).

3 Alexander, R. P. & Zhulin, I. B. Evolutionary genomics reveals conserved structural determinants of signaling and adaptation in microbial chemoreceptors. *Proc Natl Acad Sci USA* **104**, 2885-2890, (2007).

4 Bowers, K. J. *et al.* in *Proceedings of the 2006 ACM/IEEE conference on Supercomputing* 84 (ACM, Tampa, Florida, 2006).

5 Shaw, D. E. *et al.* Anton, a special-purpose machine for molecular dynamics simulation. *Commun ACM* **51**, 91, (2008).

6 Mayaan, E., Moser, A., MacKerell, A. D., Jr. & York, D. M. CHARMM force field parameters for simulation of reactive intermediates in native and thio-substituted ribozymes. *J Comput Chem* **28**, 495-507, (2007).

7 Buck, M., Bouguet-Bonnet, S., Pastor, R. W. & MacKerell, A. D., Jr. Importance of the CMAP correction to the CHARMM22 protein force field: dynamics of hen lysozyme. *Biophys J* **90**, L36-38, (2006).

8 Mackerell, A. D., Feig, M. & Brooks, C. L. 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* **25**, 1400-1415, (2004).

9 MacKerell, A. D. *et al.* All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins†. *J Phys Chem B* **102**, 3586-3616, (1998).

10 Shan, Y., Klepeis, J. L., Eastwood, M. P., Dror, R. O. & Shaw, D. E. Gaussian split Ewald: A fast Ewald mesh method for molecular simulation. *J Chem Phys* **122**, 54101, (2005).

11 Lipari, G. & Szabo, A. Model-free approach to the interpretation of nuclear magnetic resonance relaxation in macromolecules. 1. Theory and range of validity. *J Am Chem Soc* **104**, 4546-4559, (1982).

12 Chandrasekhar, I., Clore, G. M., Szabo, A., Gronenborn, A. M. & Brooks, B. R. A 500 ps molecular dynamics simulation study of interleukin-1 beta in water. Correlation with nuclear magnetic resonance spectroscopy and crystallography. *J Mol Biol* **226**, 239-250, (1992).

13 Chen, J., Brooks, C. L., 3rd & Wright, P. E. Model-free analysis of protein dynamics: assessment of accuracy and model selection protocols based on molecular dynamics simulation. *J Biomol NMR* **29**, 243-257, (2004).

14 Nederveen, A. J. & Bonvin, A. M. J. J. NMR relaxation and internal dynamics of ubiquitin from a 0.2 μs MD simulation. *J Chem Theory Comput* **1**, 363-374, (2005).

15 Markwick, P. R., Bouvignies, G. & Blackledge, M. Exploring multiple timescale motions in protein GB3 using accelerated molecular dynamics and NMR spectroscopy. *J Am Chem Soc* **129**, 4724-4730, (2007).

16 Trbovic, N., Kim, B., Friesner, R. A. & Palmer, A. G., 3rd. Structural analysis of protein dynamics by MD simulations and NMR spin-relaxation. *Proteins* **71**, 684-694, (2008).

17 Humphrey, W., Dalke, A. & Schulten, K. VMD: visual molecular dynamics. *J Mol Graph* **14**, 33-38, 27-38, (1996).

18 Hall, B. A., Armitage, J. P. & Sansom, M. S. Mechanism of bacterial signal transduction revealed by molecular dynamics of Tsr dimers and trimers of dimers in lipid vesicles. *PLoS Comput Biol* **8**, e1002685, (2012).

19 Coleman, M. D., Bass, R. B., Mehan, R. S. & Falke, J. J. Conserved glycine residues in the cytoplasmic domain of the aspartate receptor play essential roles in kinase coupling and on-off switching. *Biochemistry* **44**, 7687-7695, (2005).

20 Michaud-Agrawal, N., Denning, E. J., Woolf, T. B. & Beckstein, O. MDAnalysis: A toolkit for the analysis of molecular dynamics simulations. *J Comput Chem* **32**, 2319-2327, (2011).

21 Wickham, H. *ggplot2: elegant graphics for data analysis*. (Springer New York, 2009).

22 Ulrich, L. E. & Zhulin, I. B. The MiST2 database: a comprehensive genomics resource on microbial signal transduction. *Nucleic Acids Res* **38**, D401-407, (2010).

23 Finn, R. D., Clements, J. & Eddy, S. R. HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res* **39**, W29-37, (2011).

24 Finn, R. D. *et al.* Pfam: clans, web tools and services. *Nucleic Acids Res* **34**, D247-251, (2006).

25 Katoh, K. & Toh, H. Recent developments in the MAFFT multiple sequence alignment program. *Brief Bioinform* **9**, 286-298, (2008).

26 Crooks, G. E., Hon, G., Chandonia, J. M. & Brenner, S. E. WebLogo: a sequence logo generator. *Genome Res* **14**, 1188-1190, (2004).

**Table S1:** Sequence conservation within the chemoreceptor protein interaction region (as defined by Alexander and Zhulin 2007) ranked by entropy calculated with Weblogo. Phe396 is the most conserved residue in the chemoreceptor family. Multiple sequence alignment is available as Supplementary Data set S1.

| Residue Number | Entropy | Counts | Identity (%) |
|---|---|---|---|
| F396 | 3.2359 | 7805 | 99.949 |
| N381 | 3.1058 | 7797 | 99.846 |
| Q374 | 3.0404 | 7598 | 97.298 |
| N376 | 2.8706 | 7494 | 95.966 |
| T375 | 2.7197 | 7637 | 97.797 |
| E402 | 2.7024 | 7729 | 98.976 |
| I368 | 2.6978 | 7381 | 94.519 |
| V399 | 2.6942 | 7789 | 99.744 |
| E385 | 2.6939 | 7695 | 98.540 |
| V398 | 2.6263 | 7684 | 98.399 |
| G393 | 2.6123 | 7794 | 99.808 |
| G390 | 2.5754 | 7753 | 99.283 |
| I371 | 2.5738 | 7148 | 91.535 |
| R388 | 2.572 | 7394 | 94.686 |
| G395 | 2.5685 | 7738 | 99.091 |
| I417 | 2.5368 | 7029 | 90.012 |
| A382 | 2.4888 | 7801 | 99.898 |
| A400 | 2.4665 | 7772 | 99.526 |
| A389 | 2.4415 | 7744 | 99.168 |
| V403 | 2.3994 | 7024 | 89.947 |
| L406 | 2.3969 | 7802 | 99.910 |
| L378 | 2.394 | 7798 | 99.859 |
| A386 | 2.3883 | 7630 | 97.708 |
| A387 | 2.3643 | 7634 | 97.759 |
| A379 | 2.3255 | 7481 | 95.800 |
| A407 | 2.2993 | 7408 | 94.865 |
| L380 | 2.288 | 7631 | 97.721 |
| R404 | 2.2638 | 6329 | 81.048 |
| V384 | 2.2549 | 5480 | 70.175 |
| A397 | 2.1672 | 7227 | 92.547 |
| E391 | 2.1605 | 6520 | 83.493 |
| R394 | 2.079 | 5515 | 70.624 |
| A372 | 2.0423 | 6728 | 86.157 |
| I377 | 1.8934 | 5847 | 74.875 |
| S410 | 1.8108 | 3681 | 47.138 |
| I364 | 1.7073 | 4508 | 57.728 |
| A413 | 1.6859 | 4796 | 61.416 |
| A383 | 1.6602 | 5639 | 72.212 |
| R409 | 1.6101 | 4462 | 57.139 |
| E416 | 1.576 | 5189 | 66.449 |
| A414 | 1.548 | 3709 | 47.496 |
| Q392 | 1.4438 | 2826 | 36.189 |
| V367 | 1.1379 | 3855 | 49.366 |
| I365 | 1.1176 | 2366 | 30.298 |
| F373 | 1.0742 | 2248 | 28.787 |
| G401 | 1.0322 | 3224 | 41.286 |
| Q408 | 0.9874 | 2436 | 31.195 |
| A411 | 0.9447 | 3236 | 41.439 |
| G370 | 0.8434 | 2147 | 27.494 |
| N405 | 0.6712 | 1928 | 24.689 |
| D369 | 0.6551 | 1755 | 22.474 |
| D363 | 0.6099 | 1708 | 21.872 |
| R415 | 0.5896 | 2367 | 30.311 |
| S366 | 0.5531 | 1536 | 19.670 |
| K418 | 0.5235 | 1670 | 21.386 |
| Q412 | 0.4477 | 1173 | 15.021 |
| S419 | 0.4234 | 1169 | 14.970 |

## SI Figures

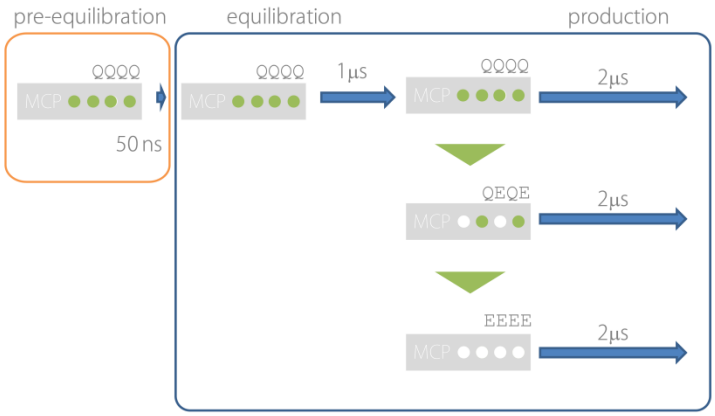


**Figure. S1.** Scheme of simulations performed in this study.

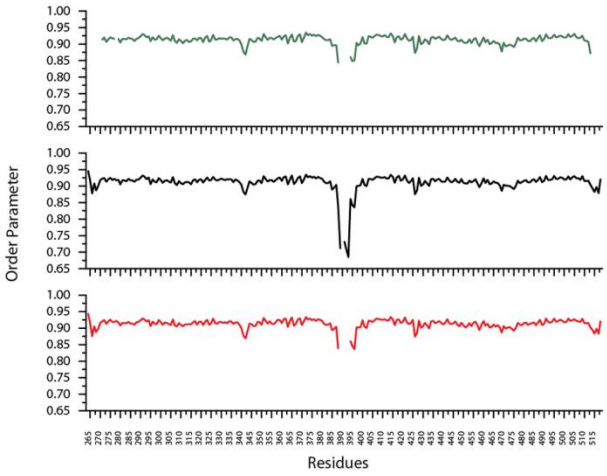


**Figure. S2.** Order parameter profile for the Tsr signaling domain calculated from the trajectories of the molecular dynamics simulations in three different methylation states: QQQQ (red), QEQE (black) and EEEE (green). Null values indicate positions with no convergence of the internal correlation function (see SI Materials and Methods).

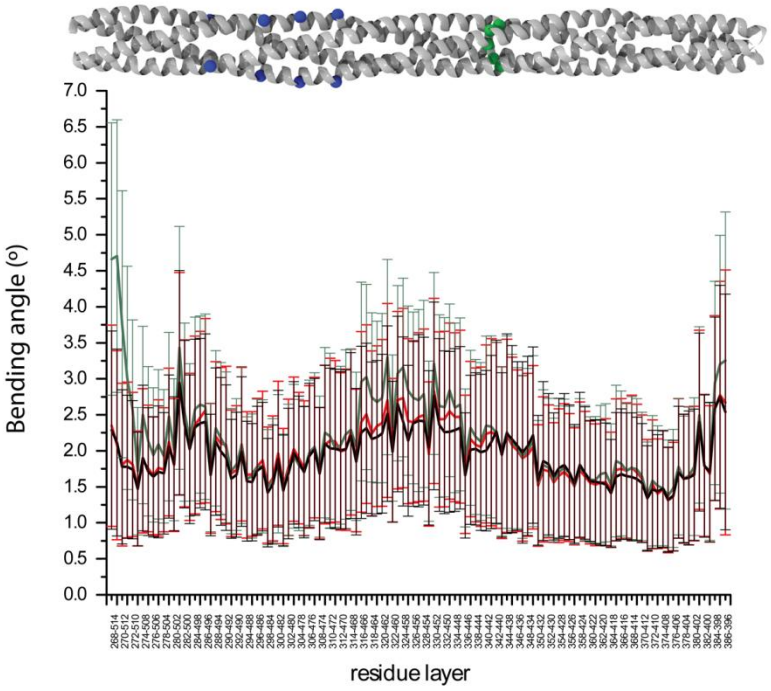


**Figure. S3.** Average bending angle along the structure measured from the frames of all three simulations: QQQQ(red), QEQE(black) and EEEE(green). Error bars are standard deviation of the mean.

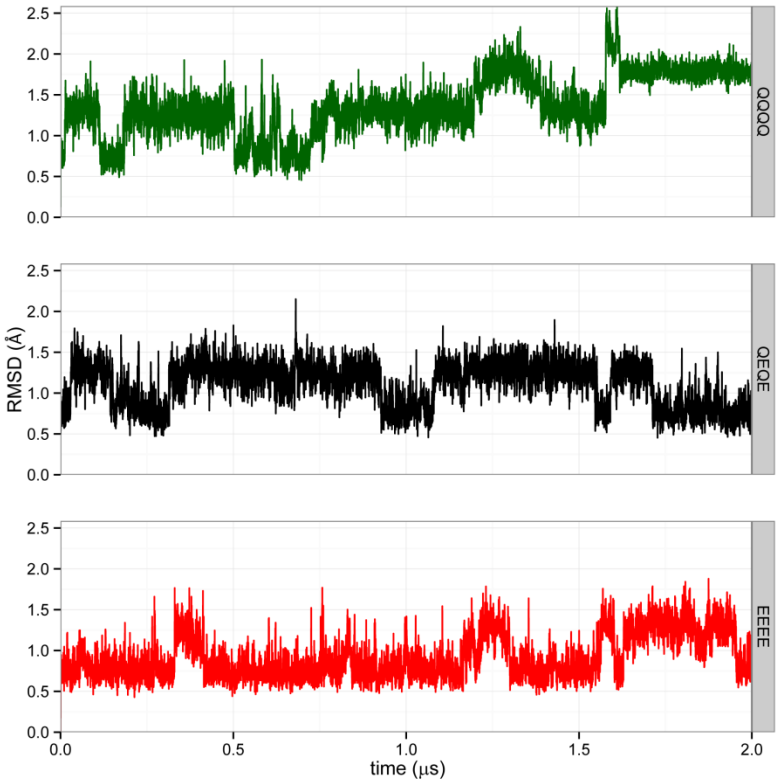


**Figure. S4.** Root-mean-square-deviation (RMSD) of each frame against the first frame in all simulations. Analysis of the RMSD over time shows that in all three signaling states this region oscillates between two conformations.

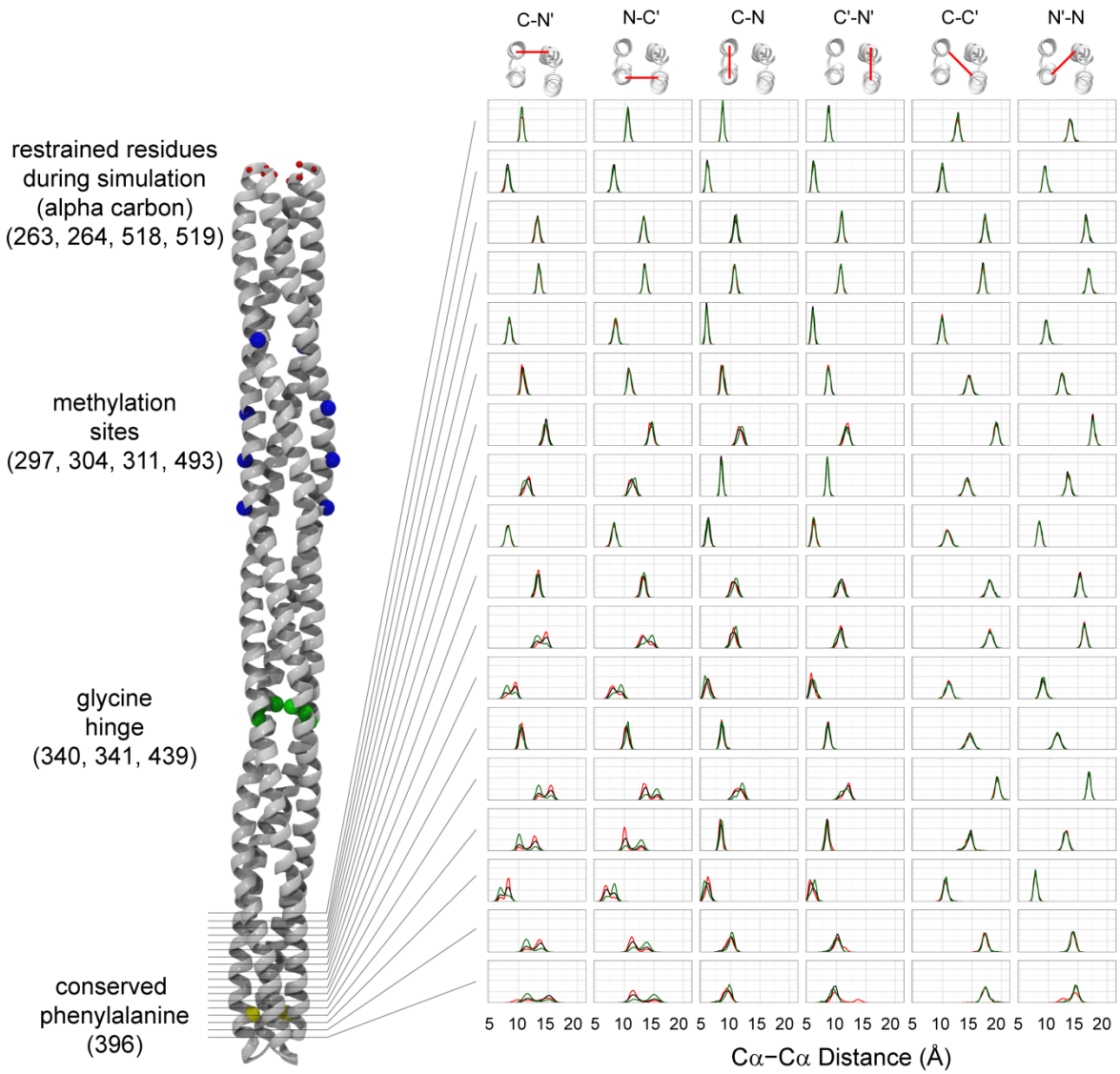


**Figure. S5.** Distances between helices of the Tsr structure measured at a single-pair level for each methylation state: QQQQ (green), QEQE(black) and EEEE(red). Note the bottom left panels that show a methylation state dependent bimodal distribution indicating two stable conformations of the tip of the receptor up to ~ 1.5 nm from the harping turn.

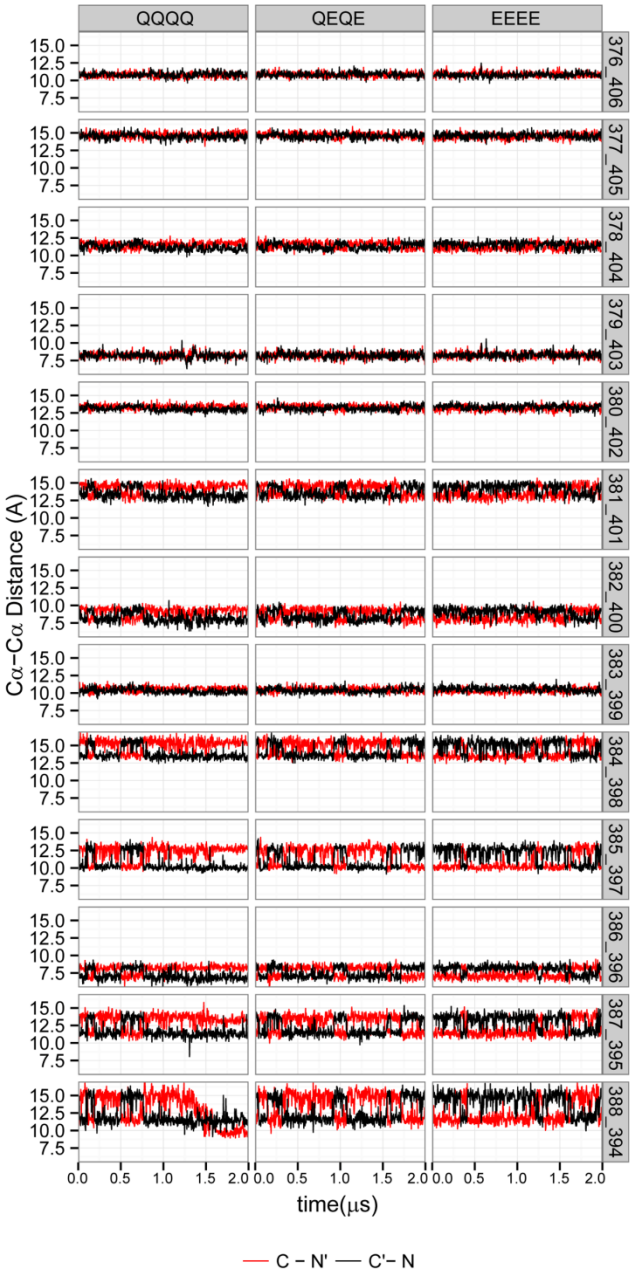


**Figure. S6.** Time evolution of the distances between helices C – N' (red) and C' – N (black) for several layers in the protein interaction region.

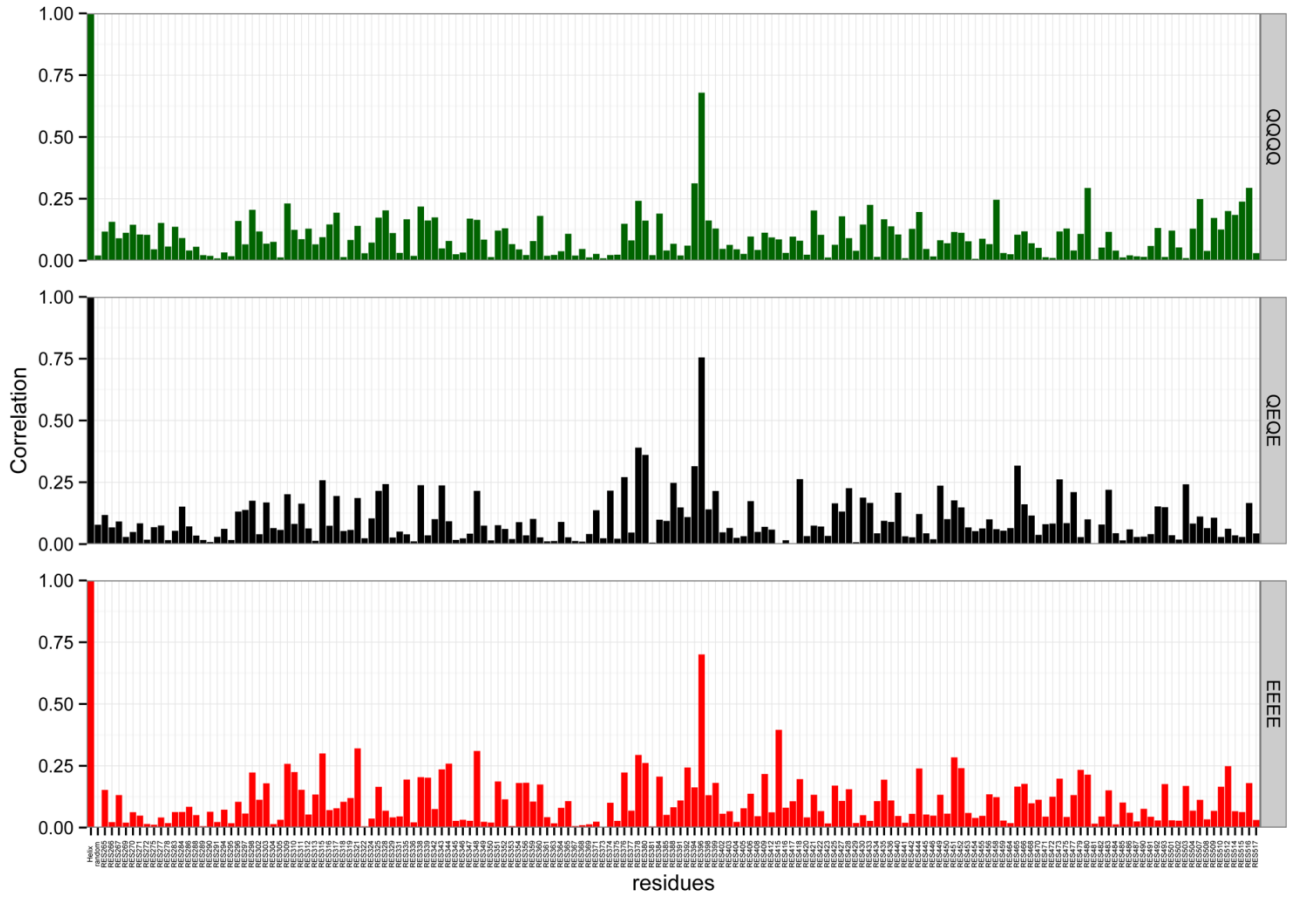


**Figure S7.** Correlation of the distance between the helices C – N' and the dihedral angle $\chi_1$ measured for all residues in the structure. The highest peak corresponds to Phe396.

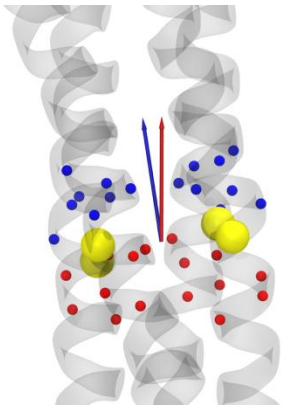

**Figure S8.** Visualization of the technique used to measure the bending angle along the structure for each layer. The principal axis of inertia (arrows) of the four alpha carbons above (blue) and below (red) the layer were calculated. The bending angle is defined as the angle between the two vectors.