Sequence-, structure-, and dynamics-based comparisons of structurally homologous CheY-like proteins

Yi He^a, Gia G. Maisuradze^a, Yanping Yin^a, Khatuna Kachlishvili^a, S. Rackovsky^{a,b}, and Harold A. Scheraga^{a,1}

^aBaker Laboratory of Chemistry and Chemical Biology, Cornell University, Ithaca, NY 14853; and ^bDepartment of Pharmacological Sciences, The Icahn School of Medicine at Mount Sinai, New York, NY 10029

Contributed by Harold A. Scheraga, December 29, 2016 (sent for review October 18, 2016; reviewed by Robert L. Jernigan and Jeffrey Skolnick)

We recently introduced a physically based approach to sequence comparison, the property factor method (PFM). In the present work, we apply the PFM approach to the study of a challenging set of sequences—the bacterial chemotaxis protein CheY, the N-terminal receiver domain of the nitrogen regulation protein NT-NtrC, and the sporulation response regulator Spo0F. These are all response regulators involved in signal transduction. Despite functional similarity and structural homology, they exhibit low sequence identity. PFM sequence comparison demonstrates a statistically significant qualitative difference between the sequence of CheY and those of the other two proteins that is not found using conventional alignment methods. This difference is shown to be consonant with structural characteristics, using distance matrix comparisons. We also demonstrate that residues participating strongly in native contacts during unfolding are distributed differently in CheY than in the other two proteins. The PFM result is also in accord with dynamic simulation results of several types. Molecular dynamics simulations of all three proteins were carried out at several temperatures, and it is shown that the dynamics of CheY are predicted to differ from those of NT-NtrC and Spo0F. The predicted dynamic properties of the three proteins are in good agreement with experimentally determined B factors and with fluctuations predicted by the Gaussian network model. We pinpoint the differences between the PFM and traditional sequence comparisons and discuss the informatic basis for the ability of the PFM approach to detect physical differences between these sequences that are not apparent from traditional alignment-based comparison.

amino acid physical properties | protein fluctuations | all-atom simulations

The investigation of the similarities and differences in the dynamics of sequentially and structurally homologous proteins has a long history (1–14). One of the important computational approaches to this problem involves the identification of conserved residues (1–4) and the investigation of the influence of these conserved residues on the folding mechanism of the proteins. An advantage of this approach is that the influence of conserved residues can be verified by mutation experiments.

More subtle questions are raised by the existence of proteins that are structurally homologous and have similar biological functions but dissimilar amino acid sequences. These molecules are of particular interest because differences in behavior can arise from sequence differences, even though the proteins have almost identical tertiary structures. A central problem then becomes the detection of those sequence characteristics that correlate with observed differences in molecular properties. It is this problem that we address in the present work.

We consider the proteins NT-NtrC, Spo0F, and CheY, which have α/β structures and are known to be response regulators involved in signal transduction (15). All three proteins have about 120 residues and very similar native structures, with pairwise root-mean-square deviations (rmsds) below 3.0 Å (14). However, they exhibit less than 35% pairwise sequence identity. Their secondary structures, determined by the define secondary structure of proteins

(DSSP) algorithm (16), are slightly different, as shown in Fig. S1. Spo0F and CheY have five well-defined α -helices and β -strands, and they exhibit a pairwise rmsd value of about 1.85 Å (14). NT-NtrC not only lacks one α -helix (corresponding to $\alpha 4$ in Spo0F and CheY) and two β -strands ($\beta 4$ and $\beta 5$ in Spo0F and CheY), but also has a slightly larger rmsd (~ 2.50 Å) from both Spo0F and CheY (14) and significantly shorter secondary structural fragments.

Recent investigations, using a Gō model modified to include sequence information (17), suggest that these proteins may have hierarchical folding processes and that formation of certain subdomains is critical to reaching the native state (12, 14). In this model, CheY and NT-NtrC share an N-terminal to C-terminal folding pathway, whereas the folding of Spo0F starts at the center and elongates first to the N terminus and then to the C terminus (12, 14). These folding differences must arise from significant sequence differences (18).

Our approach is twofold. We investigate the interactions and fluctuations of these molecules, in an all-atom representation, in their native states. This can provide information as to how the proteins perform their biological functions (19–22) and also information that can be used to understand their folding processes (23–25). Observed dynamic differences must be encoded in amino acid sequences.

We also compare the sequences of the three proteins, using both the property factor method (PFM) (18, 26) and conventional sequence alignment methods (27, 28). In previous systems that we have studied, we have shown that the PFM approach is able to detect differences between sequences that conventional

Significance

We study a set of proteins that exhibit low sequence identity, but high structural homology and functional similarity. It is demonstrated that a physics-based sequence comparison tool, the property factor method, is able to detect differences between the sequences of these proteins that correlate with differences in their structures and dynamics. It is shown that these sequence differences are not detected in this challenging system by conventional alignment methods. This result suggests that a significant amount of the information encoded in protein sequences is not captured by evolutionarily motivated comparison methods.

Author contributions: Y.H., S.R., and H.A.S. designed research; Y.H., G.G.M., Y.Y., and S.R. performed research; Y.H., G.G.M., Y.Y., K.K., S.R., and H.A.S. analyzed data; and Y.H., G.G.M., Y.Y., K.K., S.R., and H.A.S. wrote the paper.

Reviewers: R.L.J., Iowa State University; and J.S., Georgia Institute of Technology.

Conflict of interest statement: In 2014, a WeFold paper [Khouri GA, et al. (2014) WeFold: A competition for protein structure prediction. *Proteins* 82(9):1850–1868] described a hybrid approach generated from several protein structure prediction methodologies of 13 laboratories, including the H.A.S. and J.S. groups, and did not involve any active research collaboration.

¹To whom correspondence should be addressed. Email: has5@cornell.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1621344114/-DCSupplemental.

Table 1. Correlation coefficients between PFM sequence comparisons

Comparison	Correlation coefficient
DS/SC	−0.25
DS/DC	−0.17
DC/SC	0.93

DS, the NT-NtrC/Spo0F comparison; SC, the Spo0F/CheY comparison; DC, the NT-NtrC/CheY comparison.

the DC/SC correlation is very high, reflecting the near identity of the corresponding curves in Fig. 1A.

This situation is reversed in Table 2. The DS/DC correlation is very high, indicating that the similarity between the NT-NtrC/Spo0F and NT-NtrC/CheY comparisons is significant, whereas the correlations between the other Spo0F/CheY comparison and the other two curves are at best marginally significant. Table 3 demonstrates numerically that these results occur because the two sequence comparison methods differ in their treatment of the NT-NtrC/CheY comparison. The correlation between the two curves for this sequence comparison is barely statistically significant. Inspection of the PFM and NW curves for this comparison indicates that this difference arises from the C-terminal region, which the two methods compare very differently.

Structural Comparison Between NT-NtrC, Spo0F, and CheY. Distance matrix analysis was used to investigate the differences in the native structure of these three proteins. It was found that there are significant differences in the C-terminal region of NT-NtrC and Spo0F, compared with CheY, as shown in Fig. 2. Each solid circle in Fig. 2 marks a pairwise distance difference larger than 10 Å for a corresponding pair of residues in two proteins. These results are in resonance with the sequence differences indicated in Fig. 1A.

Fluctuations, Thermostability, and Unfolding. Structural fluctuations are essential for the function of proteins in the native state, as well as for correct folding. As we demonstrated in earlier work (38), analysis of structural fluctuations near the native state of a protein can also provide insight into its dynamical behavior. To examine the fluctuations, thermostability, and unfolding processes of these proteins, all-atom MD simulations starting from the experimental structures were carried out using the CHARMM36 force field. Eight 1-μs MD trajectories for each protein were carried out at 303.15 K and 400 K, and eight 1.5-μs trajectories for each protein were carried out at 450.0 K in a cubic box of three-site transferable intermolecular potential (TIP3P) water, together with counter ions (Na^+ or Cl^-) to ensure charge neutrality. Details of the system setup and simulation parameters are described in *Materials and Methods*. We analyze these simulations using four approaches.

To explore the flexibility of different regions of the molecules at a biologically relevant temperature, root-mean-square fluctuations (rmsfs) were calculated for each residue, as shown in Fig. 3 and Fig. S4. In particular, average values of rmsfs along with standard deviations of the three proteins at each temperature are shown in Fig. 3. Fig. S4A–C, *Top*, at each temperature, illustrates the rmsfs of every trajectory (color curves) and the average rmsfs (thick black curves). Fig. S4A–C, *Bottom* at each temperature is identical to Fig. 3.

From Fig. 3A and Fig. S4A, it can be seen that, at 303.15 K, Spo0F and CheY have smaller fluctuations than NT-NtrC in most regions. An exception is observed in residues approximately 70–80 of CheY. The stability of Spo0F and CheY, especially at low temperature, can be understood in terms of their higher secondary structure content together with the formation of nonpolar cores. In Fig. 3A and Fig. S4A, peaks at residues 61 and 95 of NT-NtrC and residue 76 of CheY highlight flexible regions for these two proteins. The fluctuating regions centered at the above residues are shown in Fig. S5 (blue/red highlights). The flexibility

of these regions is closely related to their biological functions, and our results are in good agreement with previous experimental and computational studies (19, 39–41).

Fig. 3B and Fig. S4B show the fluctuations of the three molecules at 400 K. It can be seen that the C terminus of CheY, including $\alpha 4\beta 5\alpha 5$, has unfolded. However, the N-terminal half of CheY remains structured. In contrast, although fluctuations of $\alpha 4$ in NT-NtrC and Spo0F were also increased in the C-terminal region, the stable β -sheet core, consisting of $\beta 2\beta 1\beta 3\beta 4\beta 5$, holds the C termini of these molecules together in all eight trajectories in our simulations. The N termini of NT-NtrC and Spo0F appear to be more stable than that of CheY at 400 K. At 450 K, all three proteins were unfolded as shown in Fig. 3C and Fig. S4C.

We have calculated the Pearson correlation coefficients R and P values between the average fluctuations of these three proteins at 303.15 K, the biological-functionally relevant temperature. At 303.15 K, $R_{\text{NT-NtrC} - \text{Spo0F}} = 0.55$ ($P = 9.15 \times 10^{-11}$), $R_{\text{NT-NtrC} - \text{CheY}} = 0.49$ ($P = 6.33 \times 10^{-9}$), and $R_{\text{Spo0F} - \text{CheY}} = 0.33$ ($P = 2.68 \times 10^{-4}$). NT-NtrC and Spo0F are more strongly correlated to each other, with respect to fluctuations, than either one is to CheY. The similarity ranking obtained by analyzing fluctuations at 303.15 K agrees with that obtained by sequence analysis, using either conventional methods or the PFM.

We have also compared experimental B factors and those predicted by the GNM to average fluctuations obtained from our all-atom simulations. The average fluctuations obtained using all eight trajectories show good agreement with the B factors reported from experiments and GNM predictions, as shown in Fig. S6.

We next examine the unfolding processes of all three proteins. The total average number of native contacts formed by each residue in all snapshots of a trajectory was calculated. Two different residue separation cutoffs (three and five residues) were used to define native contacts. Presumably, the residues that form the largest number of native contacts are those that contribute most to the stabilization of the native structure during the unfolding processes.

As shown in Fig. 4, NT-NtrC has a fairly even distribution of native contacts across the whole chain, and it shares with Spo0F a stable region in the C-terminal half that includes $\alpha 3\beta 4$, as shown by the red curves. The rest of the C-terminal region is similar to that of CheY. In contrast, the native contacts in CheY are concentrated in the N-terminal half, with significantly fewer in the C-terminal half, especially when using the five-residue cutoff (red line). The lack of a stabilized $\alpha 3\beta 4$ in CheY makes it difficult to form stable interactions between the N- and C-terminal halves. These results for CheY suggest the formation of a stable core in the N-terminal half during unfolding, which is consistent with the N-to-C folding mechanism proposed in previous publications (12, 14).

It was suggested in previous work (12, 14) that the nonpolar clusters play key roles in the stabilization of these proteins. We therefore identified the residues that form the largest number of native contacts in the unfolding processes. To filter out contacts formed by neighboring residues, we identified contacts involving residues with a minimum separation of 5 residues. We highlighted the 10 residues in each protein that have the largest averaged number of native contacts during the unfolding processes (Fig. S7). It can be seen that, for NT-NtrC and Spo0F, these are

Table 2. Correlation coefficients between NW sequence comparisons

Comparison	Correlation coefficient
DS/SC	−0.23
DS/DC	0.8
DC/SC	0.16

DS, the NT-NtrC/Spo0F comparison; SC, the Spo0F/CheY comparison; DC, the NT-NtrC/CheY comparison.

Table 3. Correlation coefficients between PFM and NW comparisons

Comparison	Correlation coefficient
DS (PFM)/DS (NW)	0.83
SC (PFM)/SC (NW)	0.69
DC (PFM)/DC (NW)	−0.31

DS, the NT-NtrC/Spo0F comparison; SC, the Spo0F/CheY comparison; DC, the NT-NtrC/CheY comparison.

predominantly nonpolar residues in β -strands. (All 10 residues are nonpolar in NT-NtrC, and 9 of 10 are nonpolar in Spo0F). Moreover, as shown in Fig. S7, there is good correspondence of the stabilization cores of NT-NtrC and Spo0F, including $\beta 1\beta 3\beta 4$, whereas the stabilization core of CheY is located in the N-terminal half and includes $\beta 1\alpha 1\beta 2\beta 3$.

We also identified the stabilization cores of the three proteins from their fluctuations at 303.15 K and 400 K. A superposition technique was used in which the structures of interest are anchored at the most invariant region. In contrast to the calculation of the rmsf, which requires only one round of structure superposition for all frames in a trajectory, core identification requires new structure superpositions for all frames at each iteration, when an atom with large fluctuation is removed across all of the frames (42–44). To accelerate these calculations, only C $^{\alpha}$ carbon atoms were considered in this process. The stabilization cores of each protein obtained from CHARMM simulations at both 303.15 K and 400 K are shown in Fig. 5. In addition to the cores reported in Fig. 5, the order of residues identified during each round, when identifying the stabilization core, is shown in Fig. S8.

All three proteins have a common stabilization core that includes $\beta 1\beta 3\beta 4$, at both 303.15 K and 400 K. Although helix packing against the $\beta 1\beta 3\beta 4$ core is different at 303.15 K for each of the three proteins when examining only the behavior of the 30 most stable residues, there is a common $\beta 1\alpha 1\beta 2$ core, in addition to the $\beta 1\beta 3\beta 4$ core, when we include the 60 most stable residues. NT-NtrC has almost identical stabilization cores at 303.15 K and 400 K, demonstrating the high thermostability of this core (Fig. 5 A and B). The stabilization core of Spo0F is almost the same at the two temperatures when the 30 most stable residues are considered, except for an expansion of the stabilization core to

$\beta 4$. When we include the 60 most stable residues, however, $\alpha 5$ is part of the stabilization core at 303.15 K, but is replaced by $\alpha 4$ at 400 K (Fig. 5 C and D). In contrast, the stabilization core of CheY not only shifted from the C-terminal to the N-terminal half between 303.15 K and 400 K, but we also find that the most stable 30-residue stabilization core also has different helix packing against the $\beta 2\beta 1\beta 3$ plane (Fig. 5 E and F).

A detailed comparison of the motions of the proteins in their native states and along unfolding pathways is shown in Fig. S9 (Supporting Information, *Motions of the CheY-Like Proteins*).

The results of these studies of unfolding processes and the identification of stabilization cores demonstrate the interplay between helical content and the nonpolar core in stabilizing the protein and the very different behavior of CheY in this regard from the other two proteins. Recent experimental studies on CheY highlighted the role of $\alpha 2$, $\alpha 3$, and $\alpha 4$ in the stabilization of native structure (36). It is clear from our native contact analysis of the unfolding processes of CheY (shown in Fig. 4) and fluctuation core analysis (Fig. 5) that $\alpha 2$, which is a part of the stabilization core, plays a significant role in stabilizing the native structure, whereas $\alpha 3$ and $\alpha 4$ clearly make a lesser contribution to stabilization.

Discussion and Conclusions

Kidera et al. (45, 46), Rackovsky (47), and Mirny et al. (2, 3) have remarked that protein structure similarity must be encoded in and should be detected using the physical properties of the amino acids in the relevant protein sequences. Although the three proteins we have examined herein have very similar secondary-structure arrangements and tertiary structures, pairwise physical-property-based sequence comparisons and all-atom simulations indicate that NT-NtrC and Spo0F share a higher degree of sequence and dynamical similarity than either one does to CheY, especially at 303.15 K.

We emphasize that the PFM does not compare sequences based on an arbitrarily selected subset of the physical properties of the amino acids. Rather, as we note in *Materials and Methods*, the approach arises from a factor analysis of all available physical property sets for the 20 amino acids and leads to a complete, orthonormal numerical representation of the physics of protein sequences. This representation carries 86% of the variance of the entire set of available physical properties (45, 46). We have further shown (48) that the representation cannot be simplified,

Fig. 3. (A–C) Average values of rmsfs, and associated standard deviations, of eight MD trajectories for NT-NtrC, Spo0F, and CheY at 303.15 K (A), 400 K (B), and 450 K (C). The solid lines at the bottom of each panel correspond to β -strands (black) and α -helices (red).

Fig. 4. Average total number of native contacts formed by each residue at 450 K, with standard deviations. The lines were obtained by counting the native contacts formed between residues separated by a minimum of three (black) or five (red) residues. Note that the β -strands are shown as gray bars and α -helices are shown as black bars at the top of each panel.

because any deviation from the full property factor representation results in a loss of physical information. No factor encodes information about any of the others, and all encode roughly equal amounts. It follows that the sequence differences between the three proteins detected in the present work are not ascribable to specific physical properties—the more so as they are differences between large fragments of sequence. Rather, they are detected in the full space of properties. Because conventional alignment

methods, including advanced methods based on multiple alignment, are based on evolutionary, rather than physical, comparison criteria, the information under consideration is not directly accessible to them.

Materials and Methods

Sequence Comparison Using Physical Property Factors. The PFM was discussed in earlier work (18, 26). The PFM algorithm compares sequences based solely on the physical properties of the 20 amino acids, with no ancillary evolutionary assumptions. Kidera et al. (45) derived 10 property factors that describe the amino acids physically, based on an exhaustive statistical analysis of previous experimental and theoretical results. These property factors form a complete, orthonormal, physics-based representation of the amino acids and therefore avoid issues of incompleteness and correlation associated with representations based only on a few selected physical properties.

Using these physical property factors, a protein sequence of length N can be represented by an $N \times 10$ matrix. With a fast normalized cross-correlation (FNCC) algorithm (49), the correlation score of a pair of matrices can be calculated, and the resulting normalized cross-correlation score provides a quantitative measure of the similarity of any pair of protein sequences in terms of their physical properties. The application of PFM to sequence comparison and homolog detection was reported earlier (26).

All-Atom Simulation Details. The solution NMR structure of NT-NtrC [Protein Data Bank (PDB) ID: 1DC7] (39) and crystal structures of Spo0F (PDB ID: 1SRR) (50) and CheY (PDB ID: 3CHY) (51) are used in the all-atom MD simulations.

Eight 1- μ s long MD trajectories at 303.15 K and 400 K, and 1.5- μ s MD trajectories at 450 K were carried out for each protein, using the CHARMM36 force field and the Groningen Machine for Chemical Simulations (GROMACS) package (52–54), version 5.1.2. Because of the stability of Spo0F at high temperature, 1.5- μ s trajectories were used to explore the unfolding process of all three proteins. Each protein was placed in a cubic box of TIP3P water together with counter ions (Na^+ or Cl^-) to neutralize the whole system. The water box contains at least a 10-Å water/counter-ion layer around each protein at 303.15 K, a minimum 20-Å water/counter-ion layer around each protein at 400 K, and a minimum 40-Å water/counter-ion layer around each protein at 450 K. Long-range electrostatics were calculated using the particle-mesh Ewald (PME) algorithm (55, 56). Periodic boundary conditions were applied in all directions. Each system of protein, water, and counter ion was prepared using CHARMM-GUI (57, 58), which generates a series of GROMACS inputs for subsequent MD simulations.

To generate equilibrated starting structures for the MD simulations, after placing each protein in a water box with counter ions, steepest-descent minimization was carried out, followed by a 25-ps MD simulation with a time step of 1 fs, to heat the whole system from 1 K to the desired temperature. All bonds with hydrogen atoms are converted to constraints with the algorithm LINear Constraint Solver (LINCS) (59), using the default parameters of the GROMACS package. The equilibrated structures obtained from the above steps were used for subsequent production runs. A Nose-Hoover temperature thermostat (60, 61) was used to maintain the temperature. The time step is 2 fs, and snapshots were taken every 100 ps.

Fig. 5. The stabilization cores (red) of NT-NtrC, Spo0F, and CheY overlap with their experimental structures (light gray). (A) For NT-NtrC at 303.15 K, the 60-residue stabilization core (Left) and the 30-residue stabilization core (Right). (B) For NT-NtrC at 400 K, the 60-residue stabilization core (Left) and the 30-residue stabilization core (Right). (C) For Spo0F at 303.15 K, the 60-residue stabilization core (Left) and the 30-residue stabilization core (Right). (D) For Spo0F at 400 K, the 60-residue stabilization core (Left) and the 30-residue stabilization core (Right). (E) For CheY at 303.15 K, the 60-residue stabilization core (Left) and the 30-residue stabilization core (Right). (F) For CheY at 400 K, the 60-residue stabilization core (Left) and the 30-residue stabilization core (Right).

ACKNOWLEDGMENTS. This research was supported by National Institutes of Health Grant GM-14312 and National Science Foundation (NSF) Grant MCB-10-19767. Computational resources were provided by (i) the Argonne Leadership Computing Facility at Argonne National Laboratory, which is supported by the Office of Science of the US Department of Energy under Contract DE-AC02-06CH11357; (ii) Interdisciplinary Center of Mathematical and Computer Modeling, University of Warsaw; (iii) our 956-processor

Beowulf cluster at the Baker Laboratory of Chemistry, Cornell University, (iv) our 692-processor Beowulf cluster at the Faculty of Chemistry, University of Gdańsk; and (v) the Academic Computer Centre in Gdańsk (Centrum Informatyczne Trójmiejskiej Akademickiej Sieci Komputerowej). This research was also supported by an allocation of advanced computing resources provided by the NSF (www.nics.tennessee.edu/) and by the NSF through TeraGrid resources provided by the Pittsburgh Supercomputing Center.

- Shakhnovich E, Abkevich V, Pitsyn O (1996) Conserved residues and the mechanism of protein folding. *Nature* 379(6560):96–98.
- Mirny LA, Abkevich VI, Shakhnovich EI (1998) How evolution makes proteins fold quickly. *Proc Natl Acad Sci USA* 95(9):4976–4981.
- Mirny LA, Shakhnovich EI (1999) Universally conserved positions in protein folds: Reading evolutionary signals about stability, folding kinetics and function. *J Mol Biol* 291(1):177–196.
- Pitsyn OB, Ting KL (1999) Non-functional conserved residues in globins and their possible role as a folding nucleus. *J Mol Biol* 291(3):671–682.
- Nishimura C, Prytulla S, Dyson HJ, Wright PE (2000) Conservation of folding pathways in evolutionarily distant globin sequences. *Nat Struct Biol* 7(8):679–686.
- Ortiz AR, Skolnick J (2000) Sequence evolution and the mechanism of protein folding. *Biophys J* 79(4):1787–1799.
- Gunasekaran K, Eyles SJ, Hagler AT, Gierasch LM (2001) Keeping it in the family: Folding studies of related proteins. *Curr Opin Struct Biol* 11(1):83–93.
- Ivankov DN, et al. (2003) Contact order revisited: Influence of protein size on the folding rate. *Protein Sci* 12(9):2057–2062.
- Friel CT, Capaldi AP, Radford SE (2003) Structural analysis of the rate-limiting transition states in the folding of Im7 and Im9: Similarities and differences in the folding of homologous proteins. *J Mol Biol* 326(1):293–305.
- Scott KA, Bates S, Hooton KA, Clarke J (2004) The folding of spectrin domains I: Wild-type domains have the same stability but very different kinetic properties. *J Mol Biol* 344(1):195–205.
- Zarrine-Afsar A, Larson SM, Davidson AR (2005) The family feud: Do proteins with similar structures fold via the same pathway? *Curr Opin Struct Biol* 15(1):42–49.
- Hills RD, Jr, Brooks CL, 3rd (2008) Subdomain competition, cooperativity, and topological frustration in the folding of CheY. *J Mol Biol* 382(2):485–495.
- Wensley BG, Gärtner M, Choo WX, Bates S, Clarke J (2009) Different members of a simple three-helix bundle protein family have very different folding rate constants and fold by different mechanisms. *J Mol Biol* 390(5):1074–1085.
- Hills RD, Jr, et al. (2010) Topological frustration in β - α -repeat proteins: Sequence diversity modulates the conserved folding mechanisms of $\alpha/\beta/\alpha$ sandwich proteins. *J Mol Biol* 398(2):332–350.
- Allweiss B, Dostal J, Carey KE, Edwards TF, Freter R (1977) The role of chemotaxis in the ecology of bacterial pathogens of mucosal surfaces. *Nature* 266(5601):448–450.
- Kabsch W, Sander C (1983) Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* 22(12):2577–2637.
- Karanicolas J, Brooks CL, 3rd (2002) The origins of asymmetry in the folding transition states of protein L and protein G. *Protein Sci* 11(10):2351–2361.
- Scheraga HA, Rackovsky S (2014) Homolog detection using global sequence properties suggests an alternate view of structural encoding in protein sequences. *Proc Natl Acad Sci USA* 111(14):5225–5229.
- Hu X, Wang Y (2006) Molecular dynamic simulations of the N-terminal receiver domain of NtrC reveal intrinsic conformational flexibility in the inactive state. *J Biomol Struct Dyn* 23(5):509–518.
- Knaggs MH, Salisbury FR, Jr, Edgell MH, Fetrow JS (2007) Insights into correlated motions and long-range interactions in CheY derived from molecular dynamics simulations. *Biophys J* 92(6):2062–2079.
- Peters GH (2009) The effect of Asp54 phosphorylation on the energetics and dynamics in the response regulator protein Spo0F studied by molecular dynamics. *Proteins* 75(3):648–658.
- Paul M, Hazra M, Barman A, Hazra S (2014) Comparative molecular dynamics simulation studies for determining factors contributing to the thermostability of chemotaxis protein “CheY”. *J Biomol Struct Dyn* 32(6):928–949.
- Cote Y, Senet P, Delarue P, Maisuradze GG, Scheraga HA (2012) Anomalous diffusion and dynamical correlation between the side chains and the main chain of proteins in their native state. *Proc Natl Acad Sci USA* 109(26):10346–10351.
- Maisuradze GG, Liwo A, Senet P, Scheraga HA (2013) Local vs global motions in protein folding. *J Chem Theory Comput* 9(7):2907–2921.
- Cote Y, Maisuradze GG, Delarue P, Scheraga HA, Senet P (2015) New insights into protein (Un)folding dynamics. *J Phys Chem Lett* 6(6):1082–1086.
- He Y, Rackovsky S, Yin Y, Scheraga HA (2015) Alternative approach to protein structure prediction based on sequential similarity of physical properties. *Proc Natl Acad Sci USA* 112(16):5029–5032.
- Smith TF, Waterman MS (1981) Identification of common molecular subsequences. *J Mol Biol* 147(1):195–197.
- Durbin R, Eddy S, Krogh A, Mitchison G (1998) *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids* (Cambridge Univ Press, Cambridge, UK).
- Rader AJ, Chennubhotla C, Yang L-W, Bahar I (2006) The Gaussian Network Model: Theory and applications. *Theoretical Model Analysis: Theory and Applications to Biological and Chemical Systems*, Mathematical and Computational Biology Series, eds Qui C, Bahar I (Chapman & Hall/CRC, Boca Raton, FL), pp 41–63.
- Yang LW, et al. (2006) oGNM: Online computation of structural dynamics using the Gaussian Network Model. *Nucleic Acids Res* 34(Web Server issue):W24–W31.
- Best RB, Hummer G, Eaton WA (2013) Native contacts determine protein folding mechanisms in atomistic simulations. *Proc Natl Acad Sci USA* 110(44):17874–17879.
- MacKerell AD, et al. (1998) All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J Phys Chem B* 102(18):3586–3616.
- Mackerell AD, Jr, Feig M, Brooks CL, 3rd (2004) 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. *J Comput Chem* 25(11):1400–1415.
- Best RB, et al. (2012) Optimization of the additive CHARMM all-atom protein force field targeting improved sampling of the backbone ϕ , ψ and side-chain $\chi(1)$ and $\chi(2)$ dihedral angles. *J Chem Theory Comput* 8(9):3257–3273.
- Huang J, MacKerell AD, Jr (2013) CHARMM36 all-atom additive protein force field: Validation based on comparison to NMR data. *J Comput Chem* 34(25):2135–2145.
- Nobrega RP, et al. (2014) Modulation of frustration in folding by sequence permutation. *Proc Natl Acad Sci USA* 111(29):10562–10567.
- Henikoff S, Henikoff JG (1992) Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci USA* 89(22):10915–10919.
- Cote Y, Senet P, Delarue P, Maisuradze GG, Scheraga HA (2010) Nonexponential decay of internal rotational correlation functions of native proteins and self-similar structural fluctuations. *Proc Natl Acad Sci USA* 107(46):19844–19849.
- Kern D, et al. (1999) Structure of a transiently phosphorylated switch in bacterial signal transduction. *Nature* 402(6764):894–898.
- Volkman BF, Lipson D, Wemmer DE, Kern D (2001) Two-state allosteric behavior in a single-domain signaling protein. *Science* 291(5512):2429–2433.
- De Carlo S, et al. (2006) The structural basis for regulated assembly and function of the transcriptional activator NtrC. *Genes Dev* 20(11):1485–1495.
- Gerstein M, Chothia C (1991) Analysis of protein loop closure. Two types of hinges produce one motion in lactate dehydrogenase. *J Mol Biol* 220(1):133–149.
- Schmidt R, Gerstein M, Altman RB (1997) LPFC: An Internet library of protein family core structures. *Protein Sci* 6(1):246–248.
- Grant BJ, Rodrigues APC, ElSawy KM, McCammon JA, Caves LSD (2006) Bio3d: An R package for the comparative analysis of protein structures. *Bioinformatics* 22(21):2695–2696.
- Kidera A, Konishi Y, Oka M, Ooi T, Scheraga HA (1985) Statistical analysis of the physical properties of the 20 naturally occurring amino acids. *J Protein Chem* 4(1):23–55.
- Kidera A, Konishi Y, Ooi T, Scheraga HA (1985) Relation between sequence similarity and structural similarity in proteins. Role of important properties of amino acids. *J Protein Chem* 4(5):265–297.
- Rackovsky S (1993) On the nature of the protein folding code. *Proc Natl Acad Sci USA* 90(2):644–648.
- Scheraga HA, Rackovsky S (2016) Global informatics and physical property selection in protein sequences. *Proc Natl Acad Sci USA* 113(7):1808–1810.
- Levis JP (1995) Fast template matching. *Vis Interface* 95:120–123.
- Madhusudan, et al. (1996) Crystal structure of a phosphatase-resistant mutant of sporulation response regulator Spo0F from *Bacillus subtilis*. *Structure* 4(6):679–690.
- Volz K, Matsumura P (1991) Crystal structure of *Escherichia coli* CheY refined at 1.7-Å resolution. *J Biol Chem* 266(23):15511–15519.
- Berendsen HJC, van der Spoel D, van Drunen R (1995) GROMACS: A message-passing parallel molecular dynamics implementation. *Comput Phys Commun* 91(1–3):43–56.
- Pall S, Abraham MJ, Kutzner C, Hess B, Lindahl E (2015) *Tackling Exascale Software Challenges in Molecular Dynamics Simulations with GROMACS*, Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) (Springer, Cham, Switzerland), pp 3–27.
- Abraham MJ, et al. (2015) Gromacs: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* 1–2:19–25.
- Darden T, York D, Pedersen L (1993) Particle mesh Ewald: An N-log(N) method for Ewald sums in large systems. *J Chem Phys* 98(12):10089.
- Essmann U, et al. (1995) A smooth particle mesh Ewald method. *J Chem Phys* 103:8577–8593.
- Jo S, Kim T, Iyer VG, Im W (2008) CHARMM-GUI: A web-based graphical user interface for CHARMM. *J Comput Chem* 29(11):1859–1865.
- Lee J, et al. (2016) CHARMM-GUI input generator for NAMD, GROMACS, AMBER, OpenMM, and CHARMM/OpenMM simulations using the CHARMM36 additive force field. *J Chem Theory Comput* 12(1):405–413.
- Hess B, Bekker H, Berendsen HJC, Fraaije JGEM (1997) LINCS: A linear constraint solver for molecular simulations. *J Comput Chem* 18(12):1463–1472.
- Nose S (1984) A unified formulation of the constant temperature molecular dynamics methods. *J Chem Phys* 81(1):511.
- Hoover WG (1985) Canonical dynamics: Equilibrium phase-space distributions. *Phys Rev A Gen Phys* 31(3):1695–1697.
- Frauenfelder H, Sligar SG, Wolynes PG (1991) The energy landscapes and motions of proteins. *Science* 254(5038):1598–1603.
- Brooks CL, 3rd, Onuchic JN, Wales DJ (2001) Statistical thermodynamics. Taking a walk on a landscape. *Science* 293(5530):612–613.