Entropic Folding Pathway of Human Epidermal Growth Factor Explored by Disulfide Scrambling and Amplified Collective Motion Simulations[†,‡]

Zhiyong Zhang,[§] Paul C. Boyle,[§] Bao-Yuan Lu,[||] Jui-Yoa Chang,[||] and Willy Wriggers*[,§]

School of Health Information Sciences, University of Texas Health Science Center at Houston, 7000 Fannin Street, Suite 600, Houston, Texas 77030, and Research Center for Protein Chemistry, Brown Foundation Institute of Molecular Medicine, University of Texas Health Science Center at Houston, 1825 Pressler Street, Houston, Texas 77030

Received July 26, 2006; Revised Manuscript Received October 26, 2006

ABSTRACT: Epidermal growth factor (EGF) regulates cell proliferation and differentiation by binding to the EGF receptor (EGFR) extra-cellular domains. Human EGF is a small, single-chain protein comprising three distinct loops (A, B, and C), which are connected by three disulfide bridges (Cys6−Cys20, Cys14−Cys31, and Cys33−Cys42). These disulfide bridges are essential for structural stability and biological activity. EGF was extensively studied by disulfide scrambling, an experimental technique for the conformational entrapment of intermediate states, which allows us to study the folding pathway of proteins containing disulfide bonds. The experimental results showed that there is a major 2-disulfide intermediate (denoted EGF-II) and that the native disulfide bonding pattern is less prevalent in one of the mutants. In this article, we investigated for the first time the solution conformations of wild-type EGF, EGF-II, and the mutant S9C through extensive molecular dynamics (MD) simulations in water using both the standard MD technique and a recently developed amplified-collective-motion (ACM) sampling method. Compared to standard MD simulations, we achieved a much more enhanced sampling by the ACM simulations, and the structures were sufficiently relaxed to estimate configurational entropies. The simulation results suggest a predominantly entropic folding pathway governed by the disorder of three functional loop regions. Although EGF-II exhibits two native disulfide bonds (Cys14−Cys31 and Cys33− Cys42), its large configurational entropy inhibits a direct transition to the native structure in the folding process. When Ser9 is mutated into Cys, a non-native disulfide bridge Cys9− Cys20 is slightly more favorable than the native Cys6−Cys20 because a less constrained *N*-terminus affords larger entropy. Isomers that are functionally less active also exhibit a more localized dynamics of the functional loop regions, which may suggest a possible mechanism for the modulation of EGF activity.

Human epidermal growth factor (EGF[1]) is a single-chain polypeptide with 53 residues, and the corresponding receptor (EGFR) is a trans-membrane protein (*1*) comprising 1186 residues. EGF induces dimerization of EGFR by binding to the extracellular region of the receptor, which leads to a 2:2 EGF−EGFR complex (*2*, *3*) (Figure 1a). A recent experimental structure (*4*) supports a receptor-mediated mechanism for the receptor dimerization (*3*). An EGF molecule binds to the extracellular domains (L1 and L2) of a receptor molecule, which induces conformational changes of EGFR so that its dimerization surface in the S1 domain is exposed. After dimerization, the cytoplasmic tyrosine kinase domains in the two EGFR molecules are close enough for autophosphorylation, which will activate the intrinsic tyrosine kinase activity. Then the EGFR tyrosine kinase triggers numerous downstream signaling pathways to regulate cell proliferation and differentiation (*5*, *6*).

The presently known experimental structures of human EGF are very similar (*4*, *7*, *8*), also the EGFs from different species such as murine (*9*−*11*) have essentially the same fold as human EGF. The primary structure of EGF comprises three distinct loops (A, B, and C loops), which are divided by three disulfide bridges (Cys6−Cys20, Cys14−Cys31, and Cys33−Cys42), as shown in Figure 1b. The *N*-terminal A-loop (residues 6−19) is fastened by the disulfide bond Cys6−Cys20 containing an α-helical fragment. The so-called B-loop (residues 20−31) comprises a two-stranded antiparallel β-sheet, and the C-loop (residues 33−42), constrained

[†] This work is supported in part by grants from the NIH (1R01GM62968), Human Frontier Science Program (RGP0026/2003), and Alfred P. Sloan Foundation (BR-4297) (to W.W.), a training fellowship from the Keck Center Pharmacoinformatics Training Program of the Gulf Coast Consortia (NIH Grant No. 1 R90 DK071505-01) (to Z.Z.) as well as the financial support from the Protein Institute, Inc. and the Robert Welch Foundation (to J.Y.C.).

[‡] This article is dedicated to the memory of Paul C. Boyle, M.S., born Dec 13, 1979 in Killybegs, Ireland, died Nov 29, 2004 in Houston, Texas as the result of an accident. P.C.B. carried out the initial work of this article as a graduate student at the School of Health Information Sciences.

* Corresponding author. Tel: (713) 500-3961. Fax: (713) 500-3907. E-mail: wriggers@biomachina.org.

[§] School of Health Information Sciences.

[||] Research Center for Protein Chemistry.

[1] Abbreviations: EGF, epidermal growth factor; EGFR, EGF receptor; N-WT, wild-type EGF; EGF-II, 2-disulfide folding intermediate; EGF-IIIA and EGF-IIIB, two 3-disulfide scrambled EGF structures; S9C, a mutant of EGF with Ser9/Cys9 substitution; n-S9C, less favorable isomer of S9C; N-S9C, more favorable isomer of S9C; MD, molecular dynamics; ACM, amplified collective motions; MM-GBSA, molecular mechanics generalized Born surface area; rmsd, root mean square deviation; SASA, solvent-accessible surface area; EDA, essential dynamics analysis.

10.1021/bi0615083 CCC: $33.50
Published on Web 11/30/2006

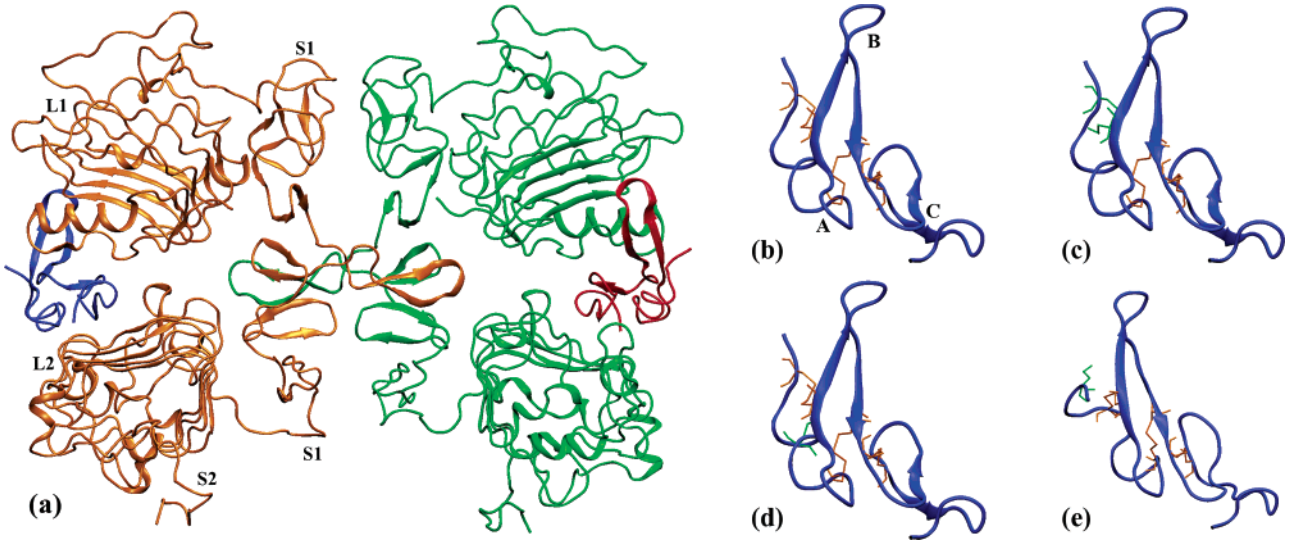

FIGURE 1: Cartoon renderings of the human EGF−EGFR complex and the four EGF molecules used for the simulation. (a) 2:2 Human EGF−EGFR complex. The two EGF chains are colored in blue and red. The two corresponding EGFR chains (extra-cellular region) are colored in orange and green. The L1, S1, L2, and S2 domains in EGFR are indicated. (b) Wild-type EGF (N-WT). The three disulfide bonds are represented by bonds and colored in orange. The three loops (A, B, and C loops) are indicated. (c) Folding intermediate (EGF-II). Similar to b, but Cys6 and Cys20, which do not form disulfide bonds, are colored in green. (d) Less favorable S9C isomer (n-S9C). Similar to b, but Cys9, which does not form a disulfide bond, is colored in green. (e) Favorable S9C isomer (N-S9C). Similar to b, but Cys6, which does not form a disulfide bond, is colored in green. Figures 1, 4, and 6−8 were created with VMD (*35*).

by the third disulfide bond Cys33−Cys42, forms part of a second antiparallel β-sheet. The three loops in EGF interact extensively with three specific sites in EGFR (Figure 1a). The A-loop and Arg41 (in the C-loop) interact with site 2 in domain L2 of EGFR. The B-loop binds to site 1 in domain L1, and the *C*-terminal region near Arg45 (including some residues in the C-loop) interacts with site 3 in domain L2.

The three disulfide bonds are very important for the conformational stability and biological activity of EGF. A fully reduced and denatured EGF molecule can refold by disulfide oxidation to spontaneously form the native conformation, a folding process that has been investigated by the technique of disulfide scrambling (*12*, *13*). Disulfide scrambling induces the conformational entrapment of intermediate states, which reveals information about the folding pathway and the conformational stability of the native and scrambled proteins. Along its oxidative folding pathway, EGF accumulates very fast as a single stable 2-disulfide kinetic trap EGF-II (*13*), which contains two native disulfide bonds (Cys14−Cys31 and Cys33−Cys42) without the *N*-terminal one (Cys6−Cys20). However, it is difficult for EGF-II to connect Cys6 and Cys20 directly to form the A-loop. In most cases, EGF-II will reach the native structure through 3-disulfide scrambled isomers, such as EGF-IIIA (Cys6−Cys42, Cys14−Cys33, and Cys20−Cys31) or EGF-IIIB (Cys6−Cys14, Cys20−Cys31, and Cys33−Cys42), by disulfide reshuffling. However, the formation of Cys6−Cys20 is essential for the biological activity of EGF. A synthetic analogue of murine EGF, (Abu6, 20)mEGF (Abu here means amino-butyric acid), has been investigated by NMR spectroscopy (*10*). The results indicate that the loss of disulfide bond Cys6−Cys20 does not affect the global fold of mEGF, but it significantly decreases the binding affinity with EGFR.

To further understand the nature of the kinetic trap of EGF-II, Chang and co-workers have prepared three EGF mutants, each with a single Ser/Cys mutation at one Ser residue (Ser2, Ser4, and Ser9) (*14*) (Figure 2). These mutated Cys2, Cys4, and Cys9 are allowed to compete with Cys6 for the association with Cys20, forming different disulfide bonds during oxidative folding. In the mutant S2C and S4C, the native disulfide bond Cys6−Cys20 is still favored over both Cys2−Cys20 and Cys4−Cys20 (Figure 2a). However, the case is different for the mutant S9C. With 87% probability, a non-native disulfide bond Cys9−Cys20 is formed, whereas the probability to form the native disulfide bond Cys6−Cys20 is only 13% (Figure 2a), which means a non-native disulfide bond is thermodynamically more stable than the native one by a free-energy difference of 1.1 kcal/mol.

In this article, we investigate for the first time the conformational stability, dynamics, and biological activity of wild-type (N-WT) EGF (Figure 1b), the folding intermediate EGF-II (Figure 1c), and the most interesting mutant S9C by molecular dynamics (MD) simulations. S9C has two isomers explored in the simulations: n-S9C (with a disulfide bond Cys6−Cys20, Figure 1d) and N-S9C (with a disulfide bond Cys9−Cys20, Figure 1e). We are concerned mainly with two questions: (1) why EGF-II would choose a more complicated pathway to reach the native structure instead of a simpler one and (*2*) why N-S9C is more favorable than n-S9C.

Following the first MD simulation of a protein 29 years ago (*15*) there were many successful applications in the study of structure and function of biological macromolecules (*16*, *17*). However, the standard MD technique faces a major limitation because of the so-called sampling problem (*18*). On the typical nanosecond to microsecond time scales one can only sample a limited number of functional states of the protein, leaving many other states to be poorly sampled. This limitation, thereby, reduces the predictive power of the simulations. Here, we used a recently developed amplified-collective-motion (ACM) method (*19*) to improve the conformational sampling of EGF molecules. ACM utilizes the low-frequency normal modes obtained with a coarse-grained model (*20*) to guide the atomic simulations. Applications to

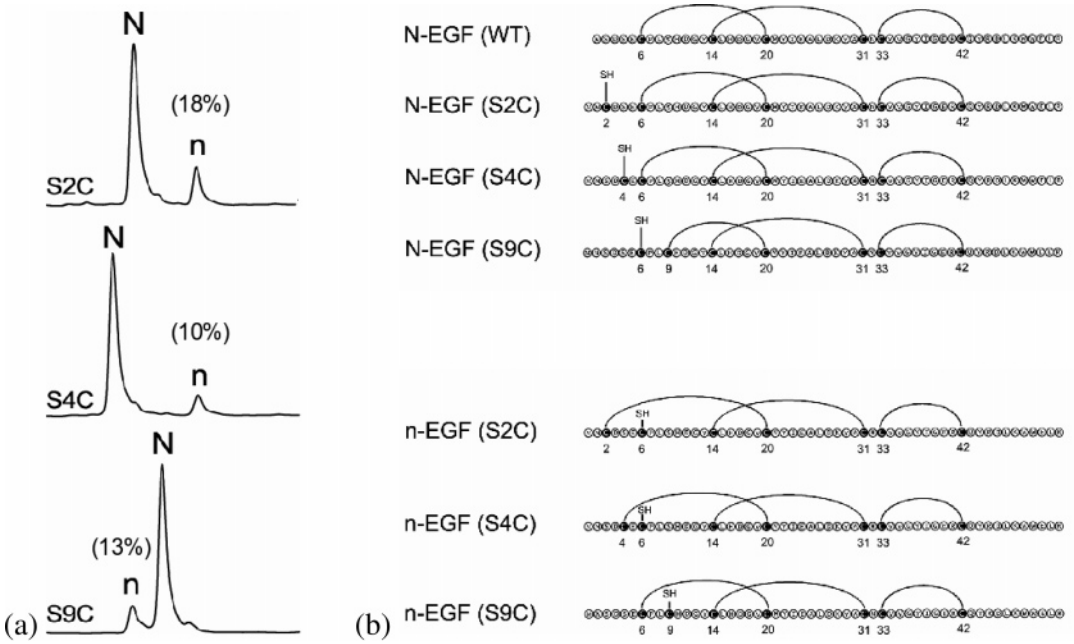

FIGURE 2: Disulfide scrambling of EGF molecules. (a) Probability analysis of purified EGF (mutants) by RP-HPLC; experimental details are in ref 14. N stands for the thermodynamically favorable isomer, and n stands for the less favorable isomer. (b) One-dimensional disulfide structures of the wild-type EGF and its mutants, which were derived from the LC-MS analysis of thermolysin-digested peptides. N and n have the same meanings as those in a.

various biological systems indicate that the ACM technique allows for much more extensive sampling in conformational space than the standard simulations to provide functional insights into allosteric mechanisms and the conformational variability of macromolecules (*19*, *21*, *22*).

A brief outline of this article is as follows. In Materials and Methods, details of the standard MD and ACM simulations of the EGF molecules as well as other computational details are described. In Results and Discussion, we give the simulation results of both standard and ACM simulations, and a comparison with the experimental data are also given. Also, the relationships between conformational stability, dynamics, and functional activity are discussed. Finally, we provide concluding remarks on the performance of the advanced ACM sampling technique to reveal large-scale protein dynamics.

MATERIALS AND METHODS

Standard MD and ACM Simulations. All of the standard MD simulations were performed with a parallel implementation of the GROMACS package (version 3.1.4) (*23*, *24*), using a GROMOS-87 (*25*) based force field with a united-atom model (*26*). Because no structure of the free human EGF was available, chain C (one of the two EGF molecules) of the crystal structure of the EGF−EGFR complex (pdb entry 1IVO) was used for the starting structure (Figure 1a), which contains only the coordinates of residues 5−51 (*4*). The protein was placed in a rectangular box such that the minimal distance between the solute and the box boundary was 1.2 nm. The box was then filled with SPC water molecules from an equilibrated cubic box containing 216 water molecules (*27*). The system, protein and water, was initially energy-minimized using the steepest descent method, until the maximum force on the atoms was less than 1000 kJ $mol^{-1}nm^{-1}$. Four Na^+ ions were added to compensate the net negative charges on the protein by replacing water molecules with the most favorable electrostatic potential. The system (protein, ions, and water molecules) was energy minimized again using the conjugate gradient algorithm with a force tolerance of 300 kJ $mol^{-1}nm^{-1}$, and a steepest descent step was done every 100 steps to make the minimization more efficient (*26*). A 100 ps simulation with position restraints was performed at 298 K. The system was simulated further for 8 ns to relax the structure because the initial EGF structure was taken from the EGF−EGFR complex, but we only simulated EGF itself here. Starting with the final structure of the preliminary simulation, we set up four systems: (1) wild-type EGF (N-WT) (Figure 1b); (2) the 2-disulfide intermediate EGF-II (Figure 1c); (3) we forced Cys6 and Cys20 to form a disulfide bond (n-S9C) after mutating Ser9 into Cys9 using the PROSSC procedure in GROMOS96 (*28*) (Figure 1d); and (4) we enforced a disulfide bond between Cys9 and Cys20 (N-S9C) using PSFGEN in the NAMD package, then minimized it by the NAMD2 program (*29*) with the CHARMM22 force field (*30*). This minimized structure was relaxed by a 6 ns MD simulation using GROMACS (*23*, *24*) with the united-atom force field (*25*, *26*) (Figure 1e). For these four EGF systems, we performed both standard MD and ACM simulations.

The Verlet integration scheme (leapfrog) (*31*) with an isothermal−isobaric simulation algorithm (*32*), and a 2 fs time step was used. The three groups (protein, ions, and the solvent) were coupled separately to a temperature bath of 298 K, with a relaxation time of 0.1 ps. The pressure was adjusted to 1 bar with a relaxation time of 1.0 ps, and the compressibility was 4.5×10^{-5} bar^{-1}. Covalent bonds in the protein were constrained using the LINCS algorithm (*33*). A twin-range cutoff was used for the van der Waals interactions: interactions within 0.9 nm were updated every step, whereas those within 1.4 nm were updated every 5 steps, together with the pair list. The long-range electrostatic

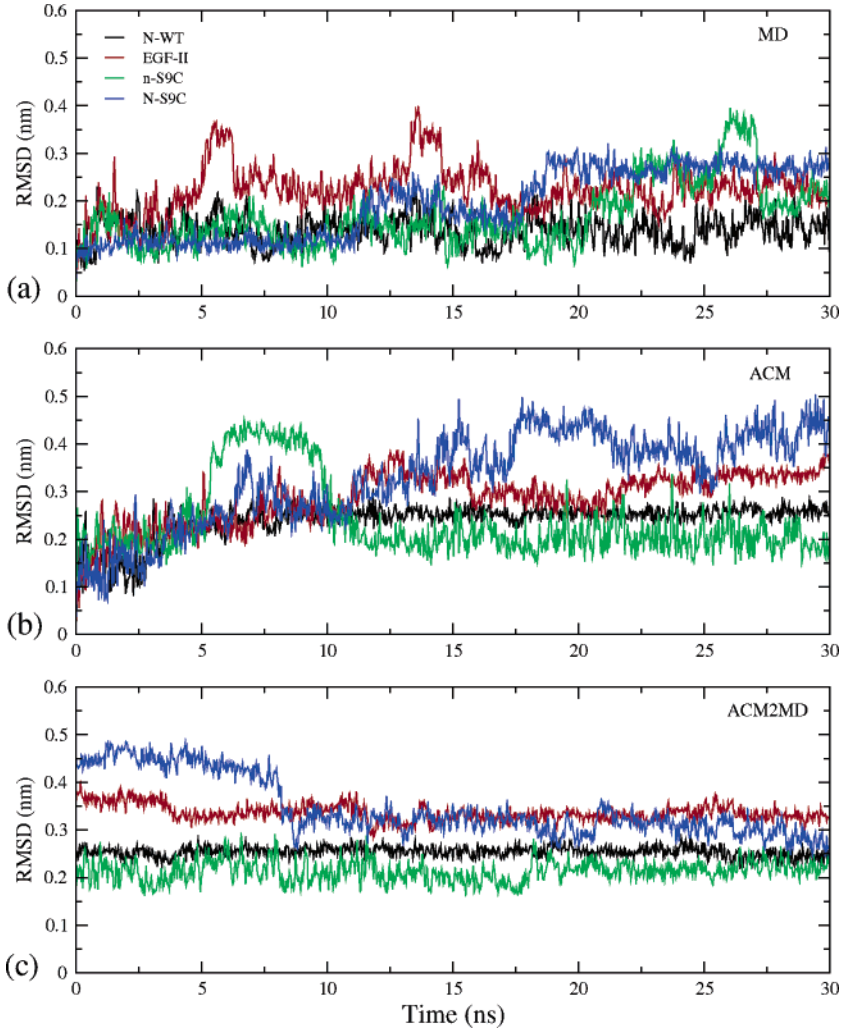

FIGURE 3: Root-mean-square deviations (rmsd's) of all of the simulations. Backbone atoms between residues 9−42 were used for calculations. (a) The rmsd with respect to the initial structure (Figure 1, b−e) in the 30 ns standard MD simulation of the four EGF molecules (S_X^{MD}). (b) The rmsd with respect to the initial structure in the 30 ns ACM simulation of the four EGF molecules (S_X^{ACM}). (c) The rmsd with respect to the initial structure in the 30 ns standard MD simulation following the ACM simulation of the four EGF molecules (S_X^{ACM2MD}). The starting structure of each ACM2MD simulation is the final structure of the corresponding ACM simulation.

interactions were treated by the PME algorithm (*34*), with a tolerance of 10^{-5} and an interpolation order of 4.

The ACM sampling technique was implemented in the GROMACS package (*23*, *24*). The ACM simulation parameters were the same as those in the standard MD simulations, except for the temperature coupling of the protein (*19*). The short- and long-range cutoff distances in ANM are 0.7 and 1.4 nm, respectively. The first three slowest collective modes were coupled to a higher temperature of 900 K, with a relaxation time of 0.006 ps. The other degrees of freedom in the protein were coupled to room temperature (298 K), with a relaxation time of 0.1 ps. The collective modes were recalculated every 50 time steps according to the current configuration of the protein.

Structural, Graphical, and Thermodynamic Properties. The subsequent trajectory analysis was performed using the tools in the GROMACS package (*23*, *24*). VMD (*35*) was used for trajectory visualization and graphical structure analysis. The calculation of free energy and entropy is challenging because they depend on the whole phase space of the system of interest (*36*). In this article, a semiempirical method called MM-GBSA (molecular mechanics generalized Born surface area) (*37*, *38*) was used to approximately estimate free energy values. An upper limit of the configurational entropy of each system was estimated by quasi-harmonic analysis of the trajectory introduced by Schlitter (*39*). Schäfer et al. recently re-examined this method and provided both analytical and numerical tests, which indicated that the method could give results with errors around 5% for condensed phase systems (*40*, *41*).

Essential Dynamics Analysis. Essential dynamics analysis can distinguish large-scale, global motions from small, localized ones (*42*). After eliminating the overall translational and rotational motions from a MD trajectory, we constructed a covariance matrix of atomic positional fluctuations $C_{ij} = \langle (x_i - \langle x_i \rangle)(x_j - \langle x_j \rangle) \rangle$, where x_i and x_j represent all of the Cartesian coordinates of the atoms we selected. The covariance matrix is diagonalized to produce eigenvalues, in the order of decreasing value, and corresponding eigenvectors. Only the motions along the first few eigenvectors (essential modes) describe significant motions in the protein that may be functionally related, and the motions along the other modes often correspond to small Gaussian-distributed random fluctuations (*43*, *44*).

Table 1: The rmsd Values of Various Simulations

	N-WT	EGF-II	n-S9C	N-S9C
MD[a]	0.14[b] (0.03[c])	0.22 (0.03)	0.24 (0.06)	0.27 (0.02)
ACM[d]	0.26 (0.01)	0.32 (0.02)	0.20 (0.03)	0.40 (0.04)
ACM2MD[e]	0.25 (0.01)	0.33 (0.01)	0.22 (0.01)	0.30 (0.02)

[a] Standard MD simulations. [b] Average values estimated from the last 10 ns in each simulation, in nm. [c] Standard deviations of the average. [d] ACM simulations. [e] Standard MD simulations after ACM sampling. The backbone atoms from residues 9−42 were used to calculate rmsd's.

RESULTS AND DISCUSSION

Conformational Stability of Wild-Type EGF. We performed a 30 ns standard MD simulation of the wild-type EGF (S^{MD}_{N-WT}) and calculated the root-mean-square deviations (rmsd's) from the initial structure (Figure 3a, black) to investigate the conformational stability. The structure of N-WT was fairly stable during the simulation, and the average rmsd of the last 10 ns was 0.14 nm (Table 1). However, the low rmsd of N-WT in the standard simulation suggests that the structure was trapped in a local minimum and was unable to explore other thermally accessible conformational states within the limited simulation time. To alleviate this problem, we carried out a 30 ns ACM simulation of N-WT (S^{ACM}_{N-WT}), starting with the same initial structure of S^{MD}_{N-WT}. In contrast to the standard MD, N-WT moved very fast (after 4 ns) to a new state in the ACM simulation (Figure 3b, black), and the system maintained the new conformation in the remaining simulation time with an average rmsd of 0.26 nm (Table 1). Considering that a number of collective modes were coupled to a high temperature at all times in S^{ACM}_{N-WT}, we cooled the system by another 30 ns standard MD simulation (S^{ACM2MD}_{N-WT}), continuing with the final structure of the ACM simulation. The protein continued to maintain the new conformation as indicated by the low fluctuation of rmsd values (0.25 ± 0.01 nm, Table 1). This stability suggests that the new state of N-WT, obtained after the ACM simulation, is energetically more favorable than the initial one.

We calculated the average free energy from the final 10 ns time series of S^{MD}_{N-WT} and S^{ACM2MD}_{N-WT}. The free energy of the new state of N-WT is lower than that of the initial state (Table 2), which supports that the new state is more favorable. The major structural differences between the two states can be attributed to the three loops, especially the B-loop and *C*-terminal region (Figure 4). The initial structure of N-WT (Figure 1b) was taken from the EGF−EGFR complex (Figure 1a), where the major contacts are due to hydrophobic and hydrogen-bonding interactions (*4*). In this article, we only performed simulations of free EGF itself; therefore, it is not surprising that the EGF structure adjusts to the loss of its contacts with EGFR. For example, the distance between Ile23 in the B-loop and Leu47 in the *C*-terminal region, which are crucial for binding to EGFR (*4*), is 2.45 nm in the final structure of S^{MD}_{N-WT} (Figure 4a, red) but only 1.68 nm in the final structure of S^{ACM2MD}_{N-WT} (Figure 4b, red). In the complex, the B-loop and *C*-terminal region bind to different sites in EGFR (Figure 1a). When EGF is free, they close as indicated by the distance between Ile23 and Leu47, which buries a part of the structure. The average solvent-accessible surface area (SASA) of the last 10 ns simulation of S^{MD}_{N-WT} is 37.74 ± 0.67 nm^2, and the corresponding value of the new state is 36.34 ± 0.62 nm^2. Also, the closed new state exhibits more intramolecular hydrogen bonds (Figure 5b) than the initial state (Figure 5a). Therefore, the simulation results of wild-type EGF indicate that a new state of free EGF with lower free energy can be reached by the enhanced ACM sampling in 30 ns. However, this state cannot be reached by the standard MD simulation in the same time scale, which remains close to the initial structure because of its limited sampling efficiency.

Structural Dynamics of EGF-II Suggests an Entropy Barrier in the Folding Pathway. As mentioned in the Introduction, the disulfide scrambling experiments indicate that there is only a single stable 2-disulfide intermediate, EGF-II, along the folding pathway. The formation of the third native disulfide bond Cys6−Cys20 for EGF-II is very slow and does not occur directly in most conditions. The major pathway from EGF-II to the native structure leads through 3-disulfide scrambled EGF isomers, such as EGF-IIIA (Cys6−Cys42, Cys14−Cys33, and Cys20−Cys31) or EGF-IIIB (Cys6−Cys14, Cys20−Cys31, and Cys33−Cys42), which means the two native disulfide bonds of EGF-II will break and form again by disulfide reshuffling. These experimental results came as a surprise. One wonders why EGF would choose a more complicated pathway instead of a simpler one in its oxidative folding. We, therefore, planned to elucidate this paradox based on the conformational dynamics of EGF-II.

We carried out a 30 ns standard MD simulation of EGF-II (S^{MD}_{EGF-II}). Compared with S^{MD}_{N-WT}, the rmsd values in S^{MD}_{EGF-II} are larger (Figure 3a, red) with an average of 0.22 nm (Table 1), especially for the residues near the *N*-terminus (Figure 6a). Because there is no disulfide bond between Cys6 and Cys20 in EGF-II, the *N*-terminal A-loop is more disordered compared to that in N-WT (Figure 4a). However, we noticed that the global fold of EGF-II (Figure 6a) remained similar to that of N-WT (Figure 4a) with a rmsd of 0.20 nm. A solution structure of (Abu6, 20)mEGF, which lacks the Cys6−Cys20 disulfide bond, has been determined at 300 K and pH 2.8 by NMR spectroscopy (*10*). The loss of this disulfide bond did cause significant local structural

Table 2: Free Energies of EGF Molecules from Various Simulations

		N-WT	EGF-II	n-S9C	N-S9C
MD[a]	$\langle G \rangle$[b]	−1434.9[c] (12.0[d])	−1442.1 (13.0)	−1432.8 (13.3)	−1438.6 (12.3)
	TS[e]	<174.2[f]	<198.2	<197.3	<190.6
ACM2MD[g]	$\langle G \rangle$	−1451.8 (12.4)	−1455.4 (12.6)	−1437.7 (12.2)	−1438.2 (13.3)
	TS	<182.9	<192.2	<188.5	<203.0

[a] Standard MD simulations. [b] Free energy calculated by MM-GBSA (*37*, *38*). [c] Average values estimated from the last 10 ns in each simulation, in kcal/mol. [d] Standard deviations of the average. [e] Contribution from configurational entropy in kcal/mol. [f] Upper limit estimated by quasi-harmonic analysis (*39*) (backbone atoms were used to construct the covariance matrix). [g] Standard MD simulations after ACM sampling.

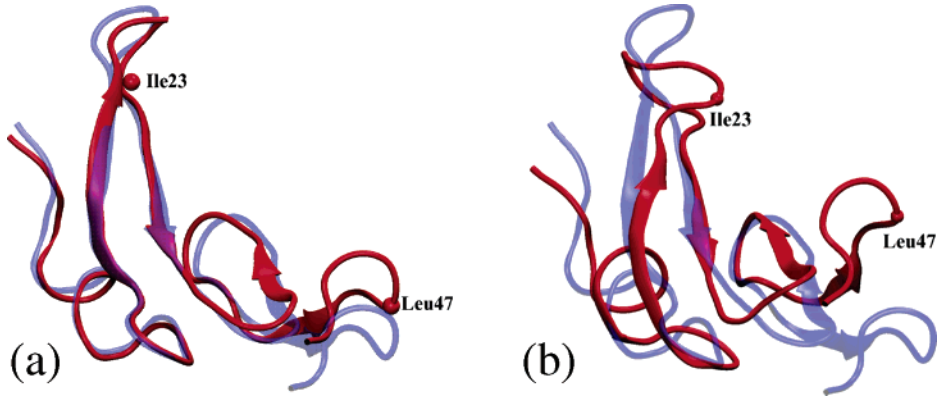

FIGURE 4: Comparison of the structures in the simulations of N-WT. (a) Superimposed structures between the initial structure of N-WT (Figure 1b) and the final one in S^{MD}_{N-WT}. (b) Initial structure of N-WT and the final one in S^{ACM2MD}_{N-WT}. Backbone atoms between residues 9−42 were used for the least-squares fit. The initial structure is colored light blue and the final structure solid red. The two residues, Ile23 and Leu47, are indicated by spheres.

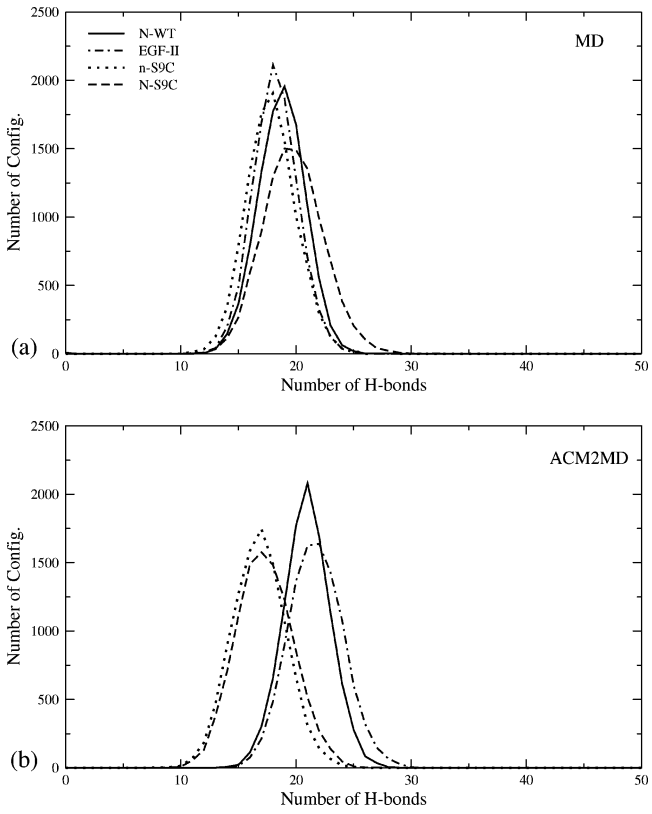

FIGURE 5: Distributions of the number of hydrogen bonds in the simulations of N-WT, EGF-II, and the two S9C isomers. (a) The last 10 ns (10,000 configurations) in each simulation of S^{MD}_{X} were used. (b) The last 10 ns (10,000 configurations) in each simulation of S^{ACM2MD}_{X} were used. Only the hydrogen bonds within the loops and between them were counted. The criteria for a hydrogen bond are a maximal distance of 0.25 nm between hydrogen−acceptor and a maximal 60° for the angle donor−hydrogen−acceptor.

changes near the *N*-terminus but likewise did not affect the global fold of mEGF. Our standard MD simulation results of EGF-II show agreement with this experimental phenomenon.

Because ACM offered us better sampling for the wild-type EGF, we also performed the same protocol on EGF-II. The final average rmsd in the 30 ns standard MD simulation (S^{ACM2MD}_{EGF-II}) following the ACM sampling (S^{ACM}_{EGF-II}) is 0.33 nm (Table 1), which indicates that the global fold of EGF-II is distorted but still basically remains a native fold. The number of intramolecular hydrogen bonds in EGF-II is comparable with that in N-WT (Figure 5b), which supports the experimental data that EGF-II is a quite stable 2-disulfide intermediate (*13*). However, the local structure of terminal residues (both *N*- and *C*-termini) changes much more significantly by enhanced sampling (Figure 6b) than by the standard MD (Figure 6a). According to the MM-GBSA calculations based on the standard MD simulations after ACM sampling, there is only a marginal free energy difference between N-WT and EGF-II except that EGF-II has a greater configurational entropy than N-WT (Table 2, results of ACM2MD). Our results suggest that there is an entropy barrier between EGF-II and N-WT, which may explain why EGF-II usually converts to the native structure slowly. The formation of the two native disulfide bonds, Cys14−Cys31 and Cys33−Cys42, can be easily accomplished because there is no significant entropy loss. There is an antiparallel β-sheet in both the B-loop and the C-loop providing rigidity and thereby maintaining low entropy. However, the *N*-terminal A-loop is largely devoid of secondary structure and highly flexible (Figure 6). Forming the third disulfide bond Cys6−Cys20 in one step would rigidify the flexible loop and significantly decrease the configurational entropy. How can EGF-II cross the entropy barrier and convert to N-WT? In general, the free energy of protein folding involves a cancellation between a decrease in enthalpy and a decrease in entropy (*45*). Our interpretation is that the 3-disulfide scrambled isomers, EGF-IIIA or EGF-IIIB, are more suitable intermediates because they can capture the unstructured *N*-terminal loop as it explores non-native conformations, which would decrease the enthalpy and lower the entropy only partially relative to those of EGF-II. In this hypothetical scenario, the EGF-III intermediates can then better reach the native EGF by disulfide reshuffling, which is highly supported by the fact that conversion of EGF-II to N-WT could be greatly accelerated in the presence of thiol catalysts that catalyze disulfide reshuffling (*13*).

Configurational Entropy of the Two EGF(S9C) Isomers. To investigate the stability of the one mutant that experimentally favors a non-native disulfide bridge, we performed 30 ns standard MD simulations for the two EGF(S9C) isomers, n-S9C and N-S9C (S^{MD}_{n-S9C} and S^{MD}_{N-S9C}, respec-

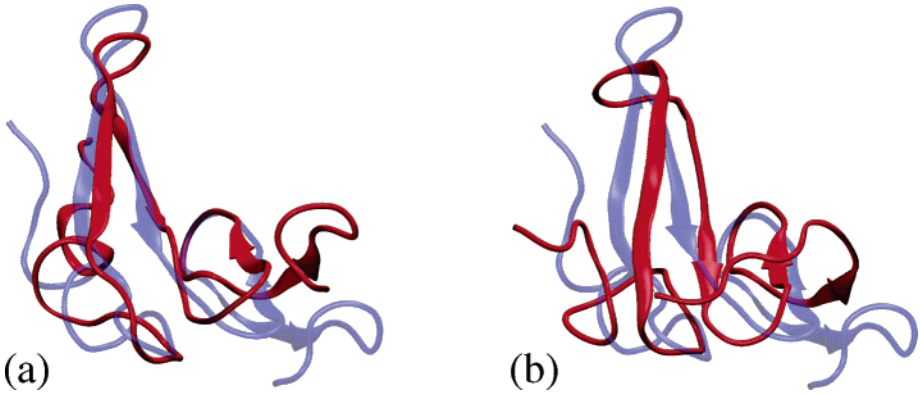

FIGURE 6: Comparison of the structures in the simulations of folding intermediate EGF-II. (a) Initial structure of EGF-II (Figure 1c) and the final one in S^{MD}_{EGF-II}. (b) Initial structure of EGF-II and the final one in S^{ACM2MD}_{EGF-II}. Backbone atoms between residues 9−42 were used for the least-squares fit. The initial structure is colored light blue and the final structure solid red.

tively). This was followed again in each case by a 30 ns ACM simulation (SC^{ACM}_{n-9} and S^{ACM}_{N-S9C}) and finally by 30 ns standard MD simulation (S^{ACM2MD}_{n-S9C} and S^{ACM2MD}_{N-S9C}).

The n-S9C system shows a drastic instability after 20 ns standard MD (Figure 3a, green). The average rmsd in the last 10 ns of S^{MD}_{n-S9C} is 0.24 nm with a large variability of 0.06 nm (Table 1), which is caused by the single-site mutation. Obviously, n-S9C has not yet stabilized. In S^{ACM}_{n-S9C}, the structure also exhibits a large conformational change after 5 ns, and the rmsd is over 0.40 nm (Figure 3b, green). However, after only 5 more nano seconds the rmsd settles at a relatively low 0.20 nm (Table 1). In the standard MD simulation following ACM (S^{ACM2MD}_{n-S9C}), the rmsd of n-S9C remains near 0.22 nm with little variability of 0.01 nm (Table 1). This result indicates that the conformational space of n-S9C is sampled more extensively by the ACM simulation compared to the standard MD simulation. The structure is sufficiently relaxed by ACM and then reaches its equilibrium state in the second standard MD simulation (Figure 7b).

The rmsd of S^{MD}_{N-S9C} reaches a high plateau of 0.27 nm in the last 10 ns (Table 1). The initial model of N-S9C, based on the wild-type structure, is not stable and, therefore, quickly moves to another state (Figure 3a, blue). In the ACM simulation of N-S9C (S^{MD}_{N-S9C}), the structure escapes further from the initial model, with an average rmsd of 0.40 nm (Table 1) because of enhanced sampling. However, the rmsd decreases again to 0.30 nm (Table 1) in the standard MD simulation following ACM (Figure 3c, blue), which indicates that the ACM sampling helped the system cross energy barriers to reach a new stable state (Figure 7d).

Again, the simulation results after ACM sampling provide a good explanation of the known experimental data. Because of the strain of entrapment, we observed fewer intramolecular hydrogen bonds in n-S9C and N-S9C compared to those in N-WT (Figure 5b). The hydrogen bonds are less distinguishable in the standard MD simulations (Figure 5a). Also, the MM-GBSA free energies of both trapped states n-S9C and N-S9C are significantly higher than that of the unscrambled N-WT (Table 2, results of ACM2MD), but there is no such difference from the calculations of the standard MD simulations (Table 2, results of MD). The free energies of n-S9C and N-S9C are approximately at the same level except that N-S9C allows a greater configurational entropy than n-S9C (Table 2, results of ACM2MD) because of the reduced steric constraints on the *N*-terminus in the case of the Cys9−Cys20 bond (Figure 7d). According to experimental disulfide scrambling, N-S9C is slightly more stable than n-S9C by a small free energy difference of 1.1 kcal/mol (*14*). Our simulations with enhanced sampling agree with the experiments in that N-S9C is slightly more favorable than n-S9C mainly because of the larger configurational entropy of N-S9C. This also explains why in the mutant S2C the native disulfide bond Cys6−Cys20 is encountered more often than Cys2−Cys20 (and likewise in S4C when compared to Cys4−Cys20) (*14*). In all cysteine mutations including the S9C case studied here, structures with the least restrained *N*-terminus (i.e., highest numbered binding partner for Cys20) are favored by entropy.

Collective Motions. From the simulation results in Figure 6, we see that EGF-II basically retains the global fold of the native EGF (rmsd from N-WT (Figure 4b): 0.27 nm), whereas N-S9C is distorted because of the non-native disulfide bond (rmsd from N-WT: 0.39 nm). We investigated whether these two molecules exhibit comparable functional activity to the wild-type system. A reduced binding affinity with EGFR was observed for the synthetic analogue (Abu6, 20)mEGF due to the absence of the native Cys6−Cys20 bond (*10*). The authors ascribed the activity loss to greater configurational entropy in the unbound structure, reducing the free energy of binding to the receptor. Also, structural changes in the *N*-terminus of (Abu6, 20)mEGF were expected to affect functionally important residues in the A-loop, further interfering with receptor binding.

The proposed greater entropy of (Abu6, 20)mEGF is consistent with our simulations of human EGF-II and N-S9C, which exhibit greater configurational entropies than N-WT because both lack the native Cys6-Cys20 bond. The *N*-terminal A-loop fluctuates dramatically in EGF-II because of the lack of the Cys6−Cys20 bond (Figure 6b), and the *N*-terminal region in N-S9C changes significantly because of the non-native disulfide bond Cys9−Cys20 (Figure 7d). The analogy with (Abu6, 20)mEGF suggests EGF-II and N-S9C may lose at lease part of their biological activities.

It has been widely accepted that protein dynamics is critical to function. We have investigated the collective motions in the different EGF molecules (N-WT, EGF-II, n-S9C, and N-S9C) from their MD trajectories by essential dynamics

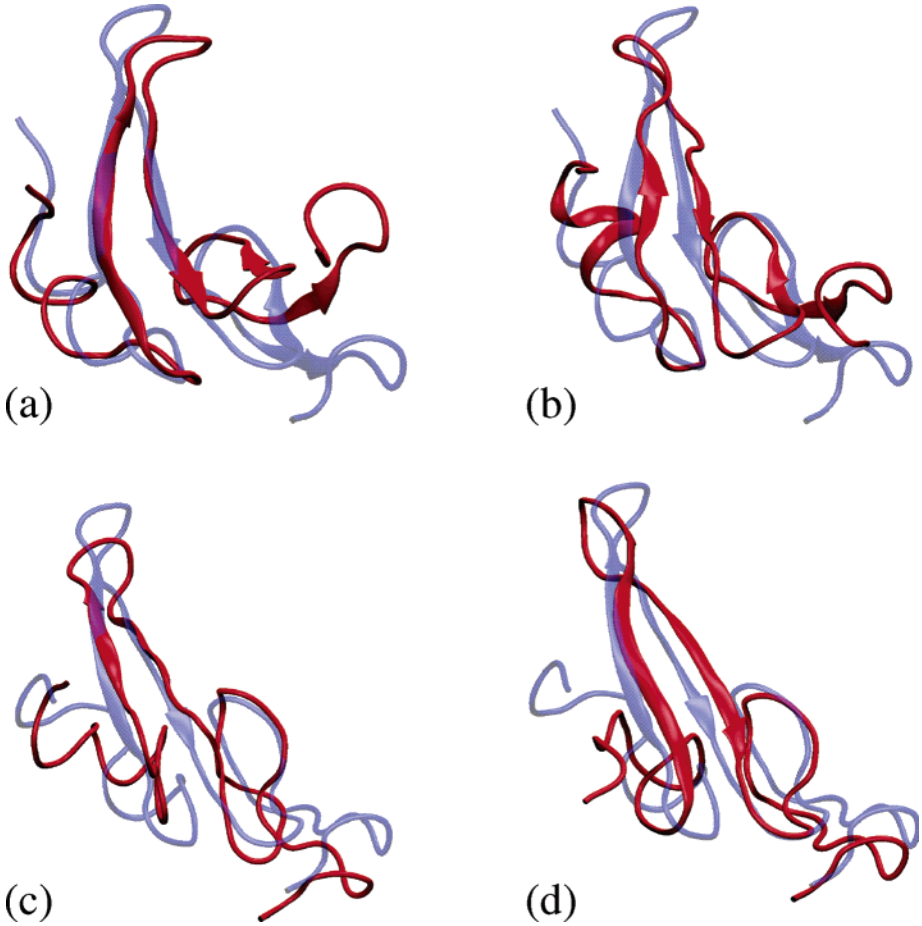

FIGURE 7: Comparison of the structures in the simulations of S9C isomers. (a) Initial structure of n-S9C (Figure 1d) and the final one in S^{MD}_{n-S9C}. (b) Initial structure of n-S9C and the final one in S^{ACM2MD}_{n-S9C}. (c) Initial structure of N-S9C (Figure 1e) and the final one in S^{MD}_{N-S9C}. (d) Initial structure of N-S9C and the final one in S^{ACM2MD}_{N-S9C}. In all cases, the backbone atoms between residues 9−42 were used for the least-squares fit. The initial structure is colored light blue and the final structure solid red.

analysis (EDA). Because the standard MD simulations after ACM (S^{ACM2MD}_X) show better agreement with the experimental data than standard MD without ACM (S^{MD}_X), we used the trajectories after ACM to perform EDA. Configurations in the last 10 ns of each trajectory were considered, and only C_α atoms were used to construct the covariance matrix. In Figure 8, we show the first essential mode of each simulation. In the case of N-WT, we can see clearly that the collective motions of the three loops are well correlated across the entire structure (Figure 8a). The three prominent loops of the native EGF structure are globally coupled because of the three disulfide bonds and other native interactions. This global consistency of the protein dynamics may be very important to its function because each loop binds to one site in EGFR (Figure 1a), which would enable EGF to provide an allosteric coupling between the binding sites in the receptor.

The long-range consistency of motion is reduced in two simulated cases (EGF-II and n-S9C). Here, the motion among the loops is more localized and not globally correlated (Figure 8b and c) even in the first (most global) essential mode. The absence of the disulfide bond Cys6−Cys20 in EGF-II, although not affecting the overall fold, limits the allosteric coupling between the nearby A loop with the more distant B- and C-loops. In n-S9C, the mutation (S9C) also perturbs the coupling among the three loops, although the three native disulfide bonds are still there. There are considerable perturbations in the *N*-terminus of n-S9C because of the mutation (Figure 7b). In the case of N-S9C, we can still see the global coupling between the loops, except within the C-loop (Figure 8d). However, the essential dynamics in N-S9C looks quite different from that in N-WT (Figure 8a) because of the differences in overall fold, which may also affect the binding affinity. Although we were mainly interested here in the presence or absence of correlated motion, the details of the modes are also of interest and could be investigated in future work.

CONCLUSIONS

In this article, we have presented a systematic study of the wild-type human EGF (N-WT), its 2-disulfide intermediate EGF-II, and one of its mutants S9C, which includes two isomers (n-S9C and N-S9C), in order to explain the experimental data of disulfide scrambling (*13*, *14*). Besides the standard MD simulations, we also carried out an ACM simulation for each system. In all cases, the results from the ACM simulations show better agreement with the experiments. The initial open structure of N-WT is from the bound state of EGF with EGFR. A more closed structure was obtained by the ACM simulation (Figure 4b) with lower free energy, whereas with 30 ns standard MD, the EGF structure

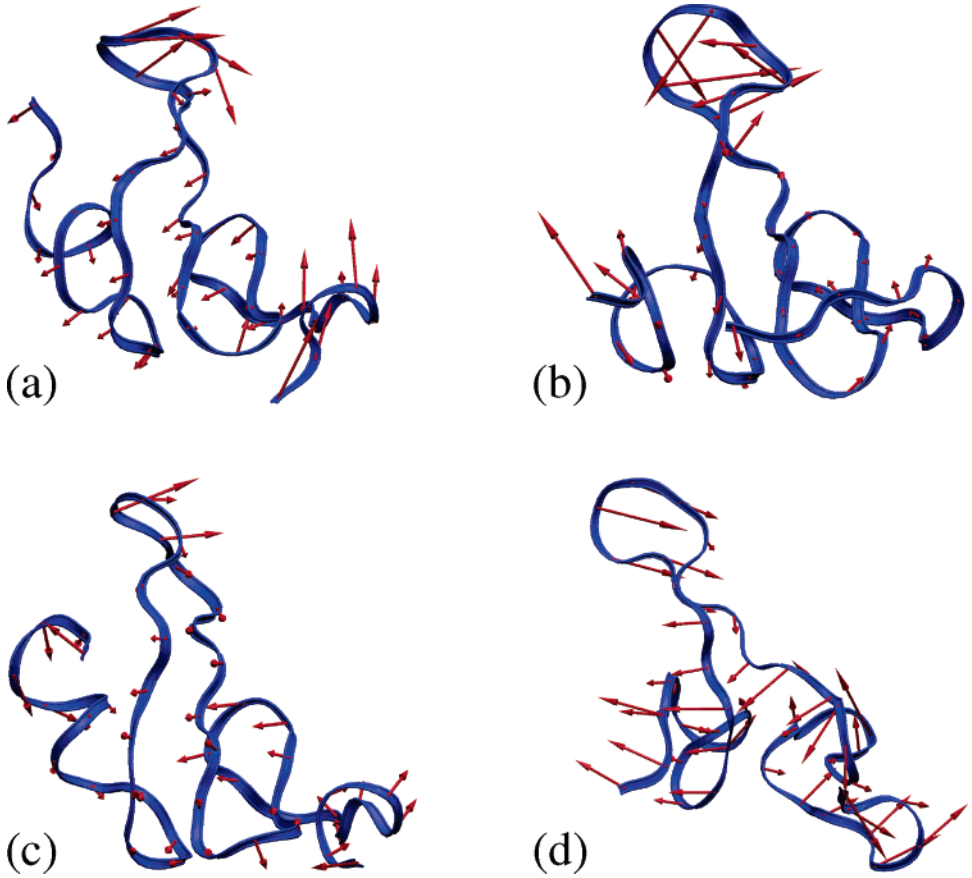

FIGURE 8: Essential dynamics analysis (EDA) of the EGF simulations. (a) N-WT, (b) EGF-II, (c) n-S9C, and (d) N-S9C. For each simulation of S_X^{ACM2MD}, C_α atoms were used to perform EDA from the last 10 ns. The ribbons structure in blue is the configuration in the trajectory (20−30 ns) that has the most negative or positive projection on the first essential mode, and the red arrows indicate the motions of the C_α atoms along this mode.

was trapped in the initial state (Figure 4a) because of limited sampling.

EGF-II is the predominant 2-disulfide intermediate during oxidative folding of EGF, but according to the ACM simulation results, it would have to overcome too large of a configurational entropy loss to directly convert to the native structure.

The simulation data of the two EGF(S9C) isomers support the experimental results that a non-native disulfide bond Cys9−Cys20 in N-S9C is slightly more favorable than the native Cys6−Cys20 in n-S9C because N-S9C affords a greater configurational entropy than n-S9C.

The results from EDA may suggest an activity loss in EGF-II isomers because of the reduced correlation among the three receptor binding loops. The loss of coupling was also observed in the mutant n-S9C. Interestingly, the folds of n-S9C and EGF-II remained very close to that of N-WT, similar to what was experimentally observed for (Abu6, 20)-mEGF, which exhibited loss of activity. This suggests a possible mechanism of functional control through the modulation of coupling strength among the receptor binding sites. In the N-S9C case, a more conventional allosteric model emerged in which the non-native disulfide bond deformed the fold and generally increased disorder in the structure, while still retaining strong coupling among two loop regions A and B.

These results are quite promising and suggest potential applications of the ACM sampling method in simulated mutagenesis and simulated disulfide scrambling. Mutagenesis is widely used in the study of protein function. In the experiments of disulfide scrambling, the single site mutation S9C changes the pattern of disulfide bonds, misfolds the *N*-terminus, and eventually affects the conformational stability and biological activity of EGF. Using the enhanced sampling technique, we were able to study the conformational and dynamical differences between the wild-type EGF and its mutants much better than that by standard MD.

There are still improvements possible with respect to the efficiency and thermodynamic accuracy of the sampling. ACM is a nonequilibrium simulation technique, which raises questions on how to recover the Boltzmann distribution and how to calculate thermodynamics properties of the system. In the present study, we have combined ACM with subsequent (equilibrium) MD calculations to overcome this limitation and to compute free energies.

REFERENCES

1. Ullrich, A., Coussens, L., Hayflick, J. S., Dull, T. J., Gray, A., Tam, A. W., Lee, J., Yarden, Y., Libermann, T. A., Schlessinger, J., Downward, J., Mayes, E. L. V., Whittle, N., Waterfield, M. D., and Seeburg, P. H. (1984) Human epidermal growth factor receptor cDNA sequence and aberrant expression of the amplified gene in A431 epidermoid carcinoma cells, *Nature 309*, 418−425.
2. Brown, P. M., Debanne, M. T., Grothe, S., Bergsma, D., Caron, M., Kay, C., and O'Connor-McCourt, M. D. (1994) The extracellular domain of the epidermal growth factor receptor. Studies on the affinity and stoichiometry of binding, receptor dimerization and a binding-domain mutant, *Eur. J. Biochem. 225*, 223−233.

3. Lemmon, M. A., Bu, Z., Ladbury, J. E., Zhou, M., Pinchasi, D., Lax, I., Engelman, D. M., and Schlessinger, J. (1997) Two EGF molecules contribute additively to stabilization of the EGFR dimmer, *EMBO J. 16*, 281−294.
4. Ogiso, H., Ishitani, R., Nureki, O., Fukai, S., Yamanaka, M., Kim, J. H., Saito, K., Sakamoto, A., Inoue, M., Shirouzu, M., and Yokoyama, S. (2002) Crystal structure of the complex of human epidermal growth factor and receptor extracellular domains, *Cell 110*, 775−787.
5. Carpenter, G., and Cohen, S. (1990) Epidermal growth factor, *J. Biol. Chem. 265*, 7709− 7712.
6. Ullrich, A., and Schlessinger, J. (1990) Signal transduction by receptors with tyrosine kinase activity, *Cell 61*, 203−212.
7. Lu, H. S., Chai, J. J., Li, M., Huang, B. R., He, C. H., and Bi, R. C. (2001) Crystal structure of human epidermal growth factor and its dimerization, *J. Biol. Chem. 276*, 34913−34917.
8. Ferguson, K. M., Berger, M. B., Mendrola, J. M., Cho, H. S., Leahy, D. J., and Lemmon, M. A. (2003) EGF activates its receptor by removing interactions that auto-inhibit ectodomain dimerization, *Mol. Cell 11*, 507−517.
9. Montelione, G. T., Wuthrich, K., Burgess, A. W., Nice, E. C., Wagner, G., Gibson, K. D., and Scheraga, H. A. (1992) Solution structure of murine epidermal growth factor determined by NMR spectroscopy and refined by energy minimization with restraints, *Biochemistry 31*, 236−249.
10. Barnham, K. J., Torres, A. M., Alewood, D., Alewood, P. F., Domagala, T., Nice, E. C., and Norton, R. S. (1998) Role of the 6−20 disulfide bridge in the structure and activity of epidermal growth factor, *Protein Sci. 7*, 1738−1749.
11. Chamberlin, S. G., Brennan, L., Puddicombe, S. M., Davies, D. E., and Turner, D. L. (2001) Solution structure of the mEGF/TGFα_{44-50} chimeric growth factor, *Eur. J. Biochem. 268*, 6247−6255.
12. Chang, J. Y., Schindler, P., Ramseier, U., and Lai, P. H. (1995) The disulfide folding pathway of human epidermal growth factor, *J. Biol. Chem. 270*, 9207−9216.
13. Chang, J. Y., Li, L., and Lai, P. H. (2001) A major kenetic trap for the oxidative folding of human epidermal growth factor, *J. Biol. Chem. 276*, 4845−4852.
14. Lu, B. Y., Jiang, C., and Chang, J. Y. (2005) Isomers of epidermal growth factor with Ser → Cys mutation at the N-terminal sequence: Isomerization, stability, unfolding, refolding and structure, *Biochemistry 44*, 15032−15041.
15. McCammon, J. A., Gelin, B. R., and Karplus, M. (1977) Dynamics of folded proteins, *Nature 267*, 585−590.
16. Karplus, M., and McCammon, J. A. (2002) Molecular dynamics simulations of biomolecules, *Nat. Struct. Biol. 9*, 646−652.
17. Adcock, S. A., and McCammon, J. A. (2006) Molecular dynamics: survey of methods for simulating the activity of proteins, *Chem. Rev. 106*, 1589−1615.
18. Clarage, J. B., Romo, T., Andrews, B. K., Pettitt, B. M., and Philips, G. N., Jr. (1995) A sampling problem in molecular dynamics simulations of macromolecules, *Proc. Natl. Acad. Sci. U.S.A. 92*, 3288−3292.
19. Zhang, Z., Shi, Y., and Liu, H. (2003) Molecular dynamics simulations of peptides and proteins with amplified collective motions, *Biophys. J. 84*, 3583−3593.
20. Atilgan, A. R., Durell, S. R., Jernigan, R. L., Demirel, M. C., Keskin, O., and Bahar, I. (2001) Anisotropy of fluctuation dynamics of proteins with an elastic network model, *Biophys. J. 80*, 505−515.
21. He, J., Zhang, Z., Shi, Y., and Liu, H. (2003) Efficiently explore the energy landscape of proteins in molecular dynamics simulations by amplifying collective motions, *J. Chem. Phys. 119*, 4005−4017.
22. Wriggers, W., Zhang, Z., Shah, M., and Sorensen, D. C. (2006) Simulating nanoscale functional motions of biomolecules, *Mol. Simul. 32*, 803−815.
23. Berendsen, H. J. C., van der Spoel, D., and van Drunen, R. (1995) Gromacs: A message-passing parallel molecular dynamics implementation, *Comput. Phys. Commun. 91*, 43−56.
24. Lindahl, E., Hess, B., and van der Spoel, D. (2001) Gromacs 3.0: A package for molecular simulation and trajectory analysis, *J. Mol. Model. 7*, 306−317.
25. van Gunsteren, W. F., and Berendsen, H. J. C. (1987) *GROMOS-87 Manual*, BIOMOS b. v. Nijenborgh 4, 9747 AG Groningen, The Netherlands.
26. van der Spoel, D., van Buuren, A. R., Apol, E., Meulenhoff, P. J., Tieleman, D. P., Sijbers, A. L. T. M., Hess, B., Feenstra, K. A., Lindahl, E., van Drunen, R., and Berendsen, H. J. C. (2001) *Gromacs User Manual*, Version 3.1, Nijenborgh 4, 9743 AG Groningen, The Netherlands. Internet: http://www.gromacs.org.
27. Berendsen, H. J. C., Postma, J. P. M., van Gunsteren, W. F., and Hermans, J. (1981) *Interaction Models for Water in Relation to Protein Hydration* (Pullman, B., Ed.) pp 331−342, Reidel, Dordrecht, Netherlands.
28. Scott, W. R. P., Hünenberger, P. H., Tironi, I. G., Mark, A. E., Billeter, S. R., Fennen, J., Torda, A. E., Huber, T., Krüger, P., and van Gunsteren, W. F. (1999) The GROMOS biomolecular simulation program package, *J. Phys. Chem. A 103*, 3596−3607.
29. Kalé, L., Skeel, R., Bhandarkar, M., Brunner, R., Gursoy, A., Krawetz, N., Phillips, J., Shinozaki, A., Varadarajan, K., and Schulten, K. (1999) NAMD2: Greater scalability for parallel molecular dynamics, *J. Comput. Phys. 151*, 283−312.
30. MacKerell, A. D., Jr., Bashford, D., Bellott, M., Dunbrack, R. L., Jr., Evanseck, J. D., Field, M. J., Fischer, S., Gao, J., Guo, H., Ha, S., Joseph-McCarthy, D., Kuchnir, L., Kuczera, K., Lau, F. T. K., Mattos, C., Michnick, S., Ngo, T., Nguyen, D. T., Prodhom, B., Reiher, W. E., III, Roux, B., Schlenkrich, M., Smith, J. C., Stote, R., Straub, J., Watanabe, M., Wiórkiewicz-Kuczera, J., Yin, D., and Karplus, M. (1998) All-atom empirical potential for molecular modeling and dynamics studies of proteins, *J. Phys. Chem. B 102*, 3586− 3616.
31. Verlet, L. (1967) Computer 'experiments' on classical fluids: I. Thermodynamical properties of Lennard-Jones molecules, *Phys. Rev. 159*, 98−103.
32. Berendsen, H. J. C., Postma, J. P. M., van Gunsteren, W. F., DiNola, A., and Haak, J. R. (1984) Molecular dynamics with coupling to an external bath, *J. Chem. Phys. 81*, 3684− 3690.
33. Hess, B., Bekker, H., Berendsen, H. J. C., and Fraaije, J. G. E. M. (1997) A linear constraint solver for molecular simulations, *J. Comput. Chem. 18*, 1463−1472.
34. Essman, U., Perela, L., Berkowitz, M. L., Darden, T., Lee, H., and Pedersen, L. G. (1995) A smooth particle mesh ewald method, *J. Chem. Phys. 103*, 8577−8592.
35. Humphrey, W. F., Dalke, A., and Schulten, K. (1996) VMD−Visual molecular dynamics, *J. Mol. Graphics 14*, 33−38.
36. Leach, A. R. (2001) *Molecular Modelling: Principles and Applications*, Pearson Education Limited, Essex, UK.
37. Kollman, P. A., Massova, I., Reyes, C., Kuhn, B., Huo, S., Chong, L., Lee, M., Lee, T., Duan, Y., Wang, W., Donini, O., Cieplak, P., Srinivasan, J., Case, D., and Cheatham, T. E., III (2000) Calculating structures and free energies of complex molecules: Combining molecular mechanics and continuum models, *Acc. Chem. Res. 33*, 889−897.
38. Zhu, J., Shi, Y., and Liu, H. (2002) Parametrization of a generalized Born/solvent-accessible surface area model and applications to the simulation of protein dynamics, *J. Phys. Chem. B 106*, 4844−4853.
39. Schlitter, J. (1993) Estimation of absolute and relative entropies of macromolecules using the covariance matrix, *Chem. Phys. Lett. 215*, 617−621.
40. Schäfer, H., Mark, A. E., and van Gunsteren, W. F. (2000) Absolute entropies from molecular dynamics simulation trajectories, *J. Chem. Phys. 113*, 7809−7817.
41. Schäfer, H., Smith, L. J., Mark, A. E., and van Gunsteren, W. F. (2002) Entropy calculations on the molten globule state of a protein: side-chain entropies of α-lactalbumin, *Proteins: Struct.. Funct., Genet. 46*, 215−224.
42. Amadei, A., Linnsen, A. B. M., and Berendsen, H. J. C. (1993) Essential dynamics of proteins, *Proteins: Struct.. Funct., Genet. 17*, 412−425.
43. Kitao, A., and Go, N. (1999) Investigating protein dynamics in collective coordinate space, *Curr. Opin. Struct. Biol. 9*, 164−169.
44. Berendsen, H. J. C., and Hayward, S. (2000) Collective protein dynamics in relation to function, *Curr. Opin. Struct. Biol. 10*, 165−169.
45. Makhatadze, G. I., and Privalov, P. L. (1995) Energetics of protein structure, *Adv. Protein Chem. 47*, 307−425.

BI0615083