Supporting Information

Cote et al. 10.1073/pnas.1207083109

SI Text

Comparison Between Calculations Performed Over 400 ns (Five MD Runs) and Over 80 ns (for each MD Run). The FEPs (Fig. 1 and Fig. S2), the correlation coefficients R and $R_n(M)$ (Fig. 3 and Fig. S9), and the dPCA analysis were performed by using a long MD run of 400 ns built by joining the five MD runs of 80 ns described in ref. 1 together. The FEPs computed for each run of 80 ns were similar to those computed over 400 ns (Fig. 1). For example, the FEPs of MD run 1 is shown in Fig. S4. The FEPs vary in general by several degrees, say approximately 10°, and the largest activation barriers may vary typically by approximately 1 kT between each of five MD runs of 80 ns (1). The correlation coefficients R between the side-chain and main-chain motions were also computed for each MD run and averaged: the results are similar to a calculation of R from the 400 ns MD run (compare Fig. 3*A* and Fig. S8). The functions $R_n(M)$ (Fig. S5) were also computed for the entire 400 ns MD run (Fig. S9). They are similar to the functions $R_n(M)$ computed for each MD run of 80 ns of duration (compare Fig. S5 and S9, for example).

Calculation of the Pearson Correlation Coefficient *R* for the Dihedral Angular Steps. The Pearson correlation coefficient R computed between two functions $x_i(t)$ and $y_i(t)$ is given by (2)

$$R = \frac{\sum_{t=t_0}^{t_{\max}}[\Delta x_i(t)\Delta y_i(t)]}{\sqrt{(\sum_{t'=t_0}^{t_{\max}}[\Delta x_i(t')]^2)(\sum_{t''=t_0}^{t_{\max}}[\Delta y_i(t'')]^2)}}, \qquad \textbf{[S1]}$$

where $\Delta x_i(t) = x_i(t) - \langle x_i\rangle$, $\Delta y_i(t) = y_i(t) - \langle y_i\rangle$ and $t_{\max}$-t_0 is the time interval considered.

For the calculation of the correlation coefficient between the steps, $R_n(M)$, we used $x_i(t) = \Delta\gamma_n(t;M)$ and $y_i(t) = \Delta\delta_n(t;M)$. The values of $x_i(t)$ and $y_i(t)$ were recorded every *ps* and the averages $\langle\rangle$ were computed over the time duration $t_{\max}$-t_0 of the joined five MD runs; i.e., 400 ns. For the calculation of R between the trajectories of the CGDA, we cannot use $x_i(t) = \gamma_n(t)$ and $y_i(t) = \delta_n(t)$; i.e., the CGDAs computed for each MD snapshot are defined in $[-\pi, \pi]$ (Fig. S1), and the average cannot be computed correctly (for example, the simple average angle between $-\pi$ and $+\pi$ would give an angle of zero degrees). Therefore, to define the time averages $\langle\rangle$ of the dihedral angles in Eq. **S1**, we built two discrete time series S_γ and S_δ from the successive dihedral angle displacements $\Delta\gamma$ and $\Delta\delta$, respectively, on the unit circle (Fig. S1). For example, $S_\gamma(0) \equiv \gamma(0)$, $S_\gamma(1) \equiv S_\gamma(0) + \Delta\gamma(1;1)$,..., $S_\gamma(m) \equiv S_\gamma(m-1) + \Delta\gamma(m;1)$,..., $S_\gamma(M) \equiv S_\gamma(M-1) + \Delta\gamma(M;1)$ (see ref. 3). Use of the variable $S_\gamma(t)$ and $S_\delta(t)$ ensure that a jump from $-\pi$ to π is of length zero and that free diffusion on a circle corresponds to an exponent $\alpha = 1$ (3). The calculation of R between the trajectories was performed with Eq. **S1** using $x_i = S_\gamma(t)$ and $y_i = S_\delta(t)$.

Calculation of the Pearson Correlation Coefficient *R* and Similarity Index *h* for the FEPs. The correlation coefficient R between the FEP was computed with Eq. **S1** using $x_i = \tilde{V}(\gamma) = -kT\ln[\tilde{P}(\gamma)]$ and $y_i = \tilde{V}(\delta) = -kT\ln[\tilde{P}(\delta)]$, where $\tilde{V}$ and $\tilde{P}$ are the FEPs and the PDFs after alignments of their deepest minimum and maximum, respectively (Fig. S2). The Boltzmann constant is k, and the temperature is T. The similarity index h between the FEPs was adapted from ref. 4 and defined by

$$h = \frac{2\sum_{\theta=-\pi}^{+\pi}[\tilde{P}(\gamma=\theta)\tilde{P}(\delta=\theta)]}{(\sum_{\gamma=-\pi}^{+\pi}[\tilde{P}(\gamma)]^2) + (\sum_{\delta=-\pi}^{+\pi}[\tilde{P}(\delta)]^2)}. \qquad \textbf{[S2]}$$

The Dihedral Principal Component Analysis. We performed a dPCA analysis of the CGDAs γ_n and δ_n over 400 ns (the five MD runs were joined as explained in point 1). In short, dPCA consists of diagonalizing the covariance matrix of the Cartesians components X and Y of the 2-D vectors $\mathbf{u}_n(t)$ representing a dihedral angle (Fig. S1), say $\mathbf{u}_n(t) = \{\cos[\gamma_n], \sin[\gamma_n]\}$, computed from an MD run (5). The eigenvalues λ of the covariance matrix are ordered by decreasing value: $\lambda_1 > \lambda_2 > \ldots\lambda_m$ ($m = N$-$2 = 43$ angles γ_n), and each collective mode k is characterized by λ_κ and by the corresponding eigenvector $\mathbf{e}_k = [\boldsymbol{e}_{k,2}, \boldsymbol{e}_{k,3}, \ldots\boldsymbol{e}_{k,m}]^t$ where $\boldsymbol{e}_{k,i} = [e_{k,i}(x), e_{k,i}(y)]$ is the amplitude of the displacement of the vector $\mathbf{u}_i$ in mode k. The contribution of a CGDA to a mode is quantified by the so-called influence $v_{k,i} \equiv [e_{k,i}(x)]^2 + [e_{k,i}(y)]^2$ (5). The sum of the eigenvalues is equal to the total mean-square-fluctuations of the vectors $\boldsymbol{u}_n$, i.e., $\sum_{i=1}^{m}\lambda_i = \sum_{n=2}^{N-2}\langle\mathbf{u}_n^2(t')\rangle_{t'} - \langle\mathbf{u}_n(t')\rangle_{t'}.\langle\mathbf{u}_n(t')\rangle_{t'}$ where the averages are computed over all times of the MD run. Therefore, λ_1 is the eigenvalue of the mode which contributes the most to the structural fluctuations of the CGDAs γ_n in the MD run. The largest values of the influence $v_{1,n}$ reveal the dihedral angles which contribute the most to the fluctuations in this mode. In normal mode analysis, the eigenvalues of the covariance matrix of the Cartesians displacements are proportional to the inverse of the square of the frequencies of the normal modes. This is the reason why large values of λ_n are named slow modes (6). The projection of the trajectory on the eigenvector $\mathbf{e}_k$ is named the principal component (PC_k). For example, $PC_1(t) = \sum_{n=2}^{N-2}[\mathbf{u}_n(t) - \langle\mathbf{u}_n\rangle].\boldsymbol{e}_{1,n}$. More details can be found in refs. 5, 7, and 8. Each PC_k is associated with an FEP V_k corresponding to the projection of the FEL of the protein along the collective coordinate PC_k; i.e., $V_k = -kT\ln[P(PC_k)]$, where $P(PC_k)$ is the PDF of the coordinate PC_k. We also computed the 2-D projections of the FEL, $V_{1,2} \equiv -kT\ln[P(PC_1, PC_2)]$. The 1-D $P(PC_k)$ and 2-D PDFs $P(PC_k, PC_l)$ were computed for each snapshot of the joined five MD runs (400 ns). We found that the cosine contents (9) of PC_1, PC_2 and PC_3 were 0.01, 0.002, and 0.01, respectively. This implies that the MD statistics is large enough to interpret the FEP V_k correctly (9). In addition, we found that the FEPs V_1 and V_2 were the only multiple-minima V_k FEPs. This justifies using only the two modes 1 and 2 as the minimal representation of the FEL of VA3 (10, 11).

Relation Between the Pearson Correlation Coefficient and Anomalous Diffusion. In the section *Correlation between the CGDA Trajectories and Steps* of the main text, we stated that the Pearson correlation coefficient $R_n(M)$ was given by

$$R_n(M) = C_n(0;M)M^{-(\alpha+\alpha')/2}/\sqrt{4D_\alpha D_{\alpha'}}, \qquad \textbf{[S3]}$$

where the cross-correlation function $C_n(0;M)$ was defined by

$$C_n(0;M) = \langle\Delta\gamma_n(t';M)\Delta\delta_n(t';M)\rangle_{t'} \qquad \textbf{[S4]}$$

Eq. **S3** can be deduced immediately from the definition of R as follows. First, we found numerically that the displacements of each dihedral angle after M *ps*, averaged over the entire MD trajectory (all t') is null, i.e., $\langle\Delta\gamma_n(t';M)\rangle_{t'} = 0$ and

$\langle \Delta\delta_n(t';M)\rangle_{t'} = 0$, as the random walkers on the γ_n-circle and δ_n-circle (Fig. S1) are doing the same number of forward and backward steps on average. By using the definition of the correlation coefficient (ref. 2 and Eq. **S1**), we have

$$R_n(M) = \frac{\sum_{t=t_0}^{t_{\max}}[\Delta\gamma_n(t;M)\Delta\delta_n(t;M)]}{\sqrt{\left(\sum_{t'=t_0}^{t_{\max}}[\Delta\gamma_n(t';M)]^2\right)\left(\sum_{t''=t_0}^{t_{\max}}[\Delta\delta_n(t'';M)]^2\right)}}. \quad \textbf{[S5]}$$

By definition, the MSDs of the CG dihedral angles are exactly:

$$\mathrm{MSD}_\gamma(n;M) = \frac{1}{t_{\max}-t_0}\sum_{t'=t_0}^{t_{\max}}[\Delta\gamma_n(t';M)]^2, \quad \textbf{[S6]}$$

$$\mathrm{MSD}_\delta(n;M) = \frac{1}{t_{\max}-t_0}\sum_{t'=t_0}^{t_{\max}}[\Delta\delta_n(t';M)]^2. \quad \textbf{[S7]}$$

As we have proven that the RCFs of the CGDAs are stretched exponentials in ref. 1 and here, we know (1) that the MSDs are power-laws of time and are given (1) by:

$$\mathrm{MSD}_\gamma(n;M) = 2D_\alpha M^\alpha, \quad \textbf{[S8]}$$

$$\mathrm{MSD}_\delta(n;M) = 2D_{\alpha'} M^{\alpha'}, \quad \textbf{[S9]}$$

where (α, D_α) and (α', $D_{\alpha'}$) are the exponents and diffusion constants of the RCFs of the CGDAs γ_n and δ_n, respectively (Fig. 2 of the main text). On the other hand, by definition, the cross-correlation function between the steps $\Delta\gamma_n(t';M)$ and $\Delta\delta_n(t'+\tau;M)$ is

$$C_n(\tau;M) = \frac{\sum_{t'=t_0}^{t_{\max}}[\Delta\gamma_n(t';M)\Delta\delta_n(t'+\tau;M)]}{t_{\max}-t_0}. \quad \textbf{[S10]}$$

Inserting Eqs. **S8–S10** in Eq. **S5**, we deduce Eq. **S3**.

1. 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* 107:19844–19849.
2. Pearson K (1896) Mathematical contributions to the theory of evolution. III Regression, heredity and panmixia. *Phil Trans R Soc Lond A* 187:253–318.
3. Senet P, Maisuradze GG, Foulie C, Delarue P, Scheraga HA (2008) How main-chains of proteins explore the free-energy landscape in native states. *Proc Natl Acad Sci USA* 105:19708–19713.
4. Hodgkin EE, Richards WG (1987) Molecular similarity based on electrostatic potential and electric field. *Int J Quantum Chem* 14:105–110.
5. Altis A, Nguyen PH, Hegger R, Stock G (2007) Dihedral angle principal component analysis of molecular dynamics simulation. *J Chem Phys* 126:244111.
6. Nicolay S, Sanejouand YH (2006) Functional modes of proteins are among the most robust. *Phys Rev Lett* 96:078104.
7. Maisuradze GG, Leitner DM (2007) Free energy landscape of a biomolecule in dihedral principal component space: Sampling convergence and correspondence between structures and minima. *Proteins* 67:569–578.
8. Maisuradze GG, Liwo A, Scheraga HA (2009) Principal component analysis for protein folding dynamics. *J Mol Biol* 385:312–329.
9. Hess B (2002) Convergence of sampling in protein simulations. *Phys Rev E* 65:031910.
10. Hegger R, Altis A, Nguyen PH, Stock G (2007) How complex is the dynamics of peptide folding? *Phys Rev Lett* 98:028102.
11. Maisuradze GG, Liwo A, Scheraga HA (2009) How adequate are one- and two-dimensional free energy landscapes for protein folding dynamics. *Phys Rev Lett* 102:238102.

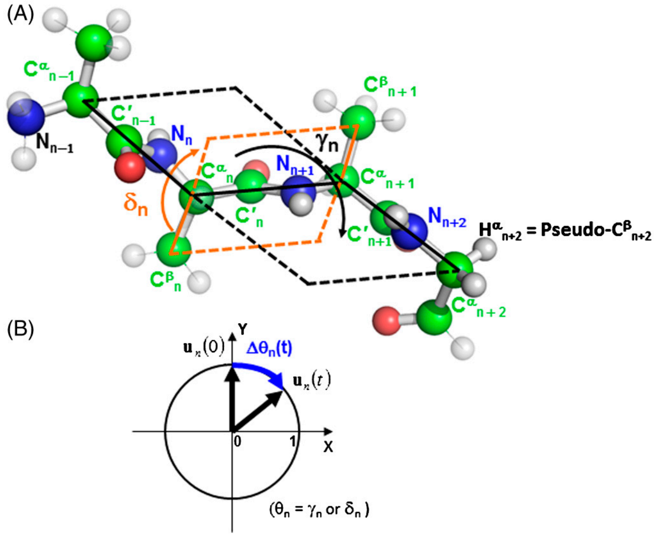


Fig. S1. (*A*) Definitions of the coarse-grained dihedral angle (CGDA) γ_n constructed from four consecutive C$^\alpha$ atoms and of the CGDA δ_n built from two consecutive C$^\alpha$ atoms and their respective C$^\beta$ atoms (for Gly residues, a pseudo-C$^\beta$ atom is defined as the position of its side chain H atom). The colors are as follows: C (green), N (blue), O (red) and H (white). (*B*) Definition of the 2-D unit vector $\mathbf{u}_n$ for a dihedral angle θ_n and of the angular displacement $\Delta\theta_n$ on the unit circle.

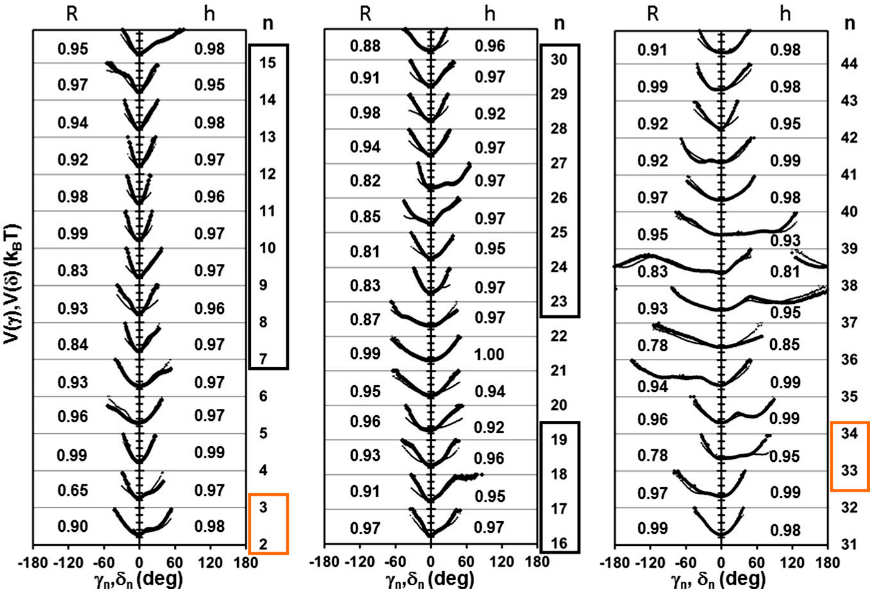


Fig. S2. Effective FEP $V(\gamma_n)$ (thick black lines) and $V(\delta_n)$ (thin black lines) computed from MD along the primary sequence (γ_n, and δ_n, $n = 2$ to 44). The FEPs were first computed from the concatenation of the five MD runs (Fig. 1 of the main text) and were aligned next on their deepest minimum. Residues n located in a helix are in a black box and residues located in a β-sheet are in an orange box. For each residue n, the number in the inset is the value of the correlation coefficient R and of the similarity index h computed between these aligned FEPs and $V(\gamma_n)$ and $V(\delta_n)$.

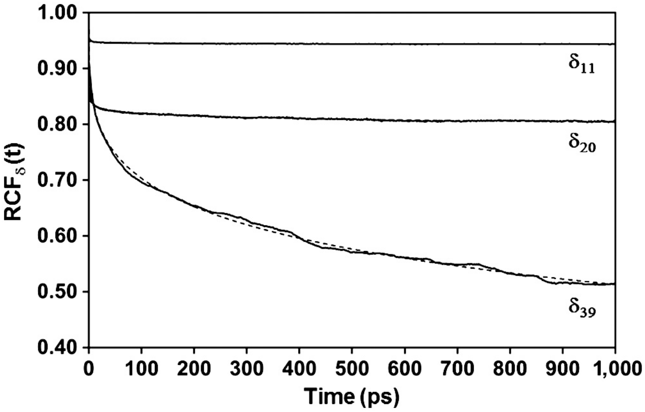


Fig. S3. Typical results for the RCFs T_2 of the coarse-grained (CG) dihedral angles δ_n up to 1 *ns* (full lines) computed by using Eq. **1** of the main text from MD of the protein VA3. Results are shown for MD run 1. Results are presented for the three types of RCFs corresponding to three types of 1-D free-energy profiles (FEP) (Fig. 1 of the main text and Fig. S4): a harmonic FEP (δ_{11}), a wide single-minimum FEP (δ_{20}) and a multiple-minima FEP (δ_{39}). The RCFs computed from MD were fitted by stretched exponentials, $\exp(-4D_\alpha t^\alpha)$, up to 1 ns. The fits (dashed lines) and the RCFs computed from MD are hardly distinguishable.

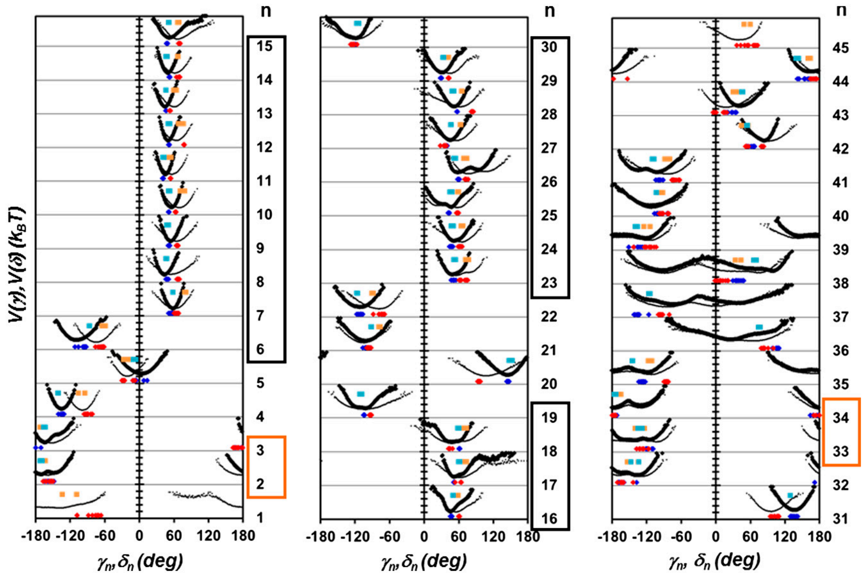


Fig. S4. Effective FEP $V(\gamma_n)$ (thick black lines) and $V(\delta_n)$ (thin black lines) computed from MD run 1 along the primary sequence (γ_n, $n = 2$ to 44 and δ_n, $n = 1$ to 45) compared to the NMR-derived structural data [blue diamonds (γ_n) and red diamonds (δ_n)] calculated from the different models of VA3 (PDB ID: 1ED0), and to x-ray data (PDB ID: 1OKH) [blue square (γ_n) and orange square (δ_n)]. Residues n located in a helix are in a black box and residues located in a β-sheet are in an orange box.

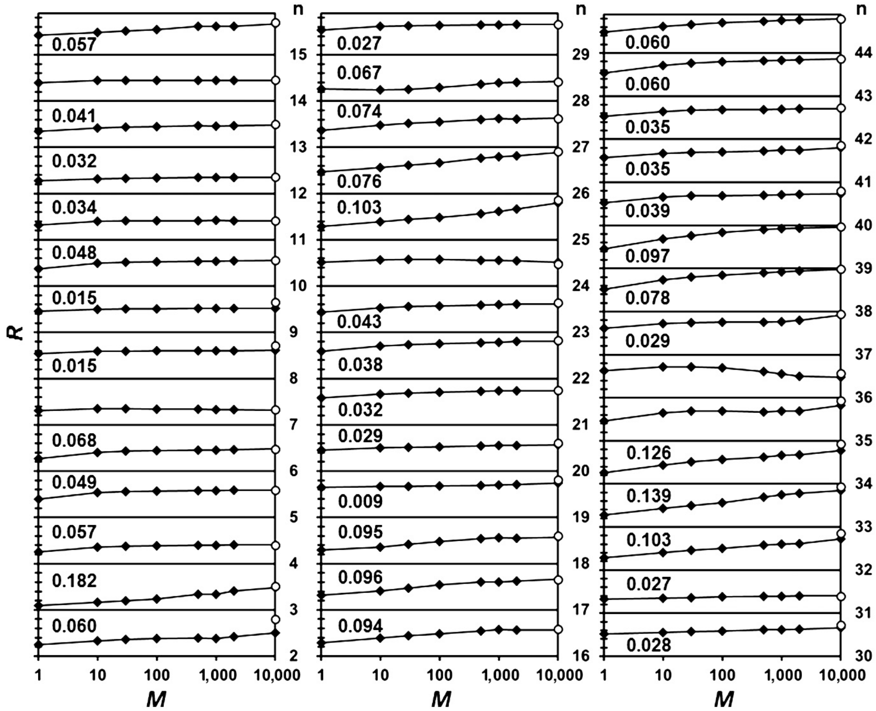


Fig. S5. Comparison of the evolution of the correlation coefficient $R_n(M)$ (lines and filled diamonds) between the displacements $\Delta\gamma_n(t;M)$ and $\Delta\delta_n(t;M)$, as function of the logarithm of the number M of steps, along the amino acid sequence of VA3. Results are shown for MD run 1. Each curve $R_n(M)$ for $n = 2$ to 44, is plotted between 0 and 1 with a tick mark every 0.2. The correlation coefficient $R_n(M)$ was computed for M equals to 1, 10, 30, 100, 500, 1,000, 2,000, and 10,000. The values of R computed between the time series $\gamma_n(t)$ and $\delta_n(t)$ extracted from MD run 1 are shown for comparison (circles) at $M = 10{,}000$ and are generally hardly discernable from the values of $R_n(10{,}000)$. The exponent β extracted from a fit of $R_n(M)$ to a power-law [$R_n(M) \sim M^{\beta}$] up to 1 ns is given as the number in the inset for each value of n for which the fit was a good approximation.

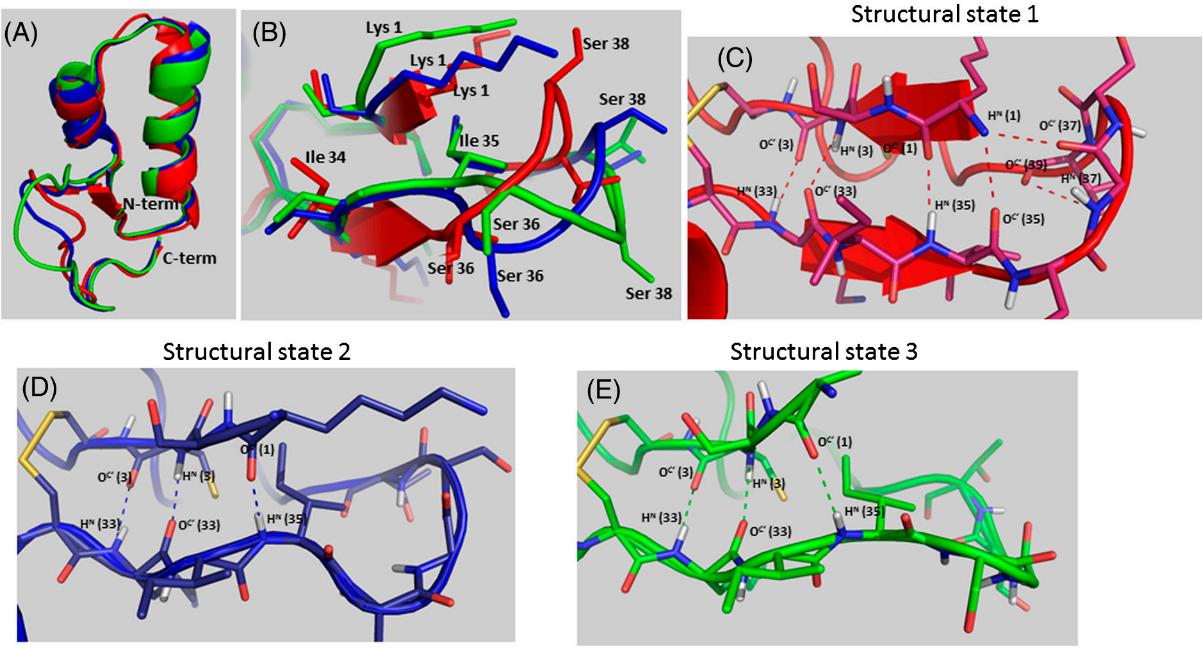


Fig. S6. Comparison between the three structures representative of structural states 1 (red structure), 2 (blue structure), and 3 (green structure) defined in the free-energy surface (FES) built on *PC1* and *PC2* (inset of Fig. 4 of the main text). The structures were extracted from the MD trajectory, and each structure represents a snapshot corresponding to the most probable molecular structure in basins 1, 2, and 3 of the FEL (*PC1, PC2*). (*A*) The three representative structures aligned on their backbone. (*B*) Same as *A* and zoomed in on the C-terminal loop of the protein involved in the collective modes discussed in the main text. Hydrogen bonds in structural state 1, 2, and 3 are shown in *C*, *D*, and *E*, respectively.

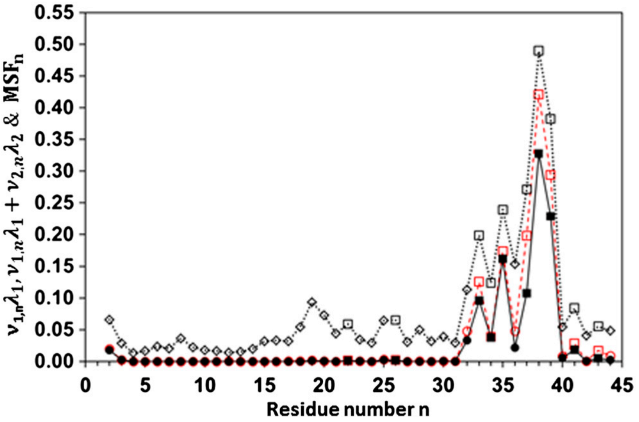


Fig. S7. Contribution of mode 1 to the MSF ($\lambda_1\nu_{1,n}$, filled symbols and full lines), contribution of modes 1 and 2 to the MSF ($\lambda_1\nu_{1,n} + \lambda_2\nu_{2,n}$, red empty symbols and red dashed lines), and the whole MSF (empty symbols and dotted lines) along the amino acid sequence of VA3. Dihedral angles with multiple-minima potentials (Fig. 1 of the main text) are shown by squares symbols. All calculations were performed from the five concatenated MD runs (over 400 ns) by applying dPCA to the vectors $\boldsymbol{u}_n(t) = \{\cos[\delta_n(t)], \sin[\delta_n(t)]\}$.

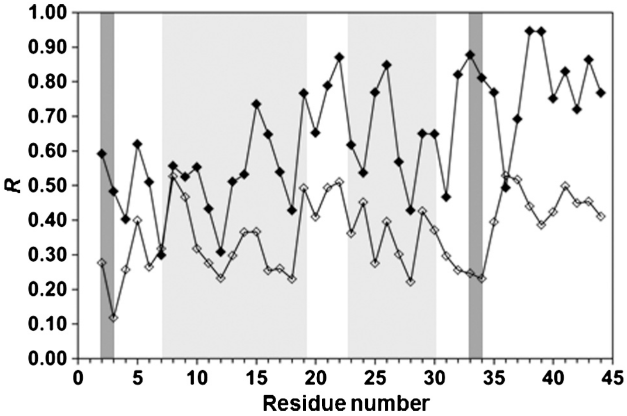


Fig. S8. Comparison of the correlation coefficient R between the trajectories $\gamma_n(t)$ and $\delta_n(t)$ (filled symbols) and between the steps $\Delta\gamma_n(t)$ and $\Delta\delta_n(t)$ (empty symbols) along the amino acid sequence of VA3 computed from each of the five MD runs (80-ns duration each). The average of the values of R of each of the five MD run is shown for each value of n. The α-helices and β-sheets are indicated by light gray and dark gray stripes, respectively.

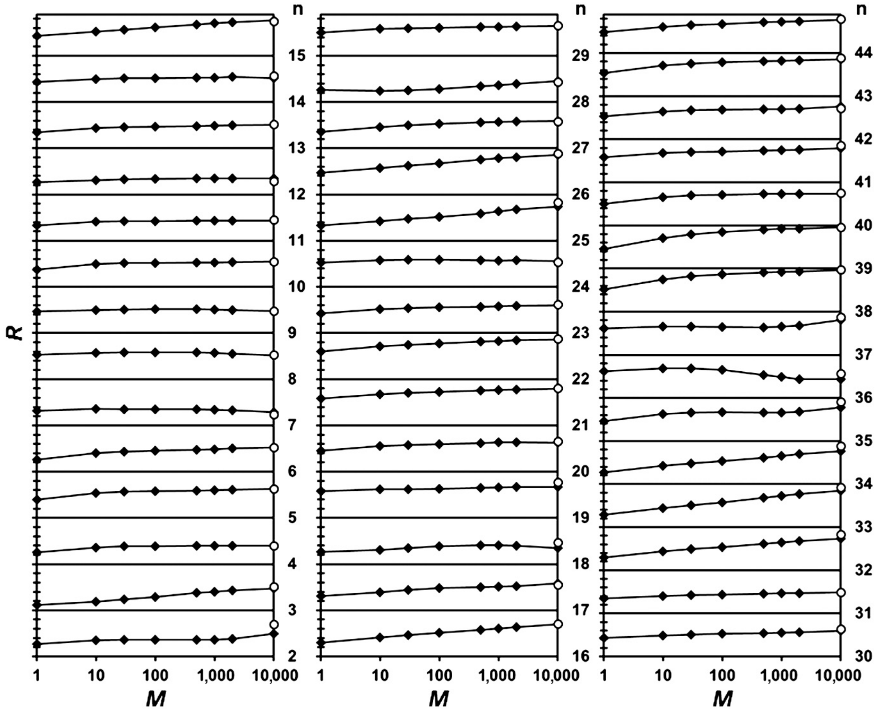
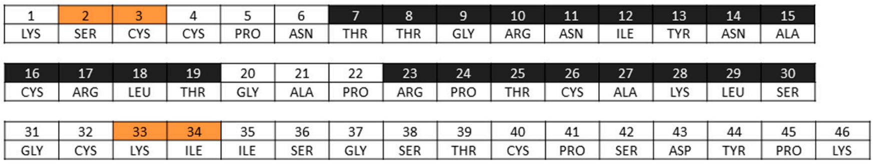


Fig. S9. Comparison of the evolution of the correlation coefficient $R_n(M)$ (lines and filled diamonds) between the displacements $\Delta\gamma_n(t; M)$ and $\Delta\delta_n(t; M)$ as function of the logarithm of the number M of steps, along the amino acid sequence of VA3. Results were computed from the five concatenated MD runs (400 ns). Each curve $R_n(M)$ for $n = 2$ to 44, is plotted between 0 and 1 with a tick mark every 0.2. The correlation coefficient $R_n(M)$ was computed for M equals to 1, 10, 30, 100, 500, 1,000, 2,000, and 10,000. The values of R computed between the time series $\gamma_n(t)$ and $\delta_n(t)$ extracted from the concatenated MD trajectory of 400 ns duration are shown for comparison (circles) at $M = 10{,}000$ and are generally hardly discernable from the values of $R_n(10{,}000)$.



1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
LYS	SER	CYS	CYS	PRO	ASN	THR	THR	GLY	ARG	ASN	ILE	TYR	ASN	ALA

16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
CYS	ARG	LEU	THR	GLY	ALA	PRO	ARG	PRO	THR	CYS	ALA	LYS	LEU	SER

31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46
GLY	CYS	LYS	ILE	ILE	SER	GLY	SER	THR	CYS	PRO	SER	ASP	TYR	PRO	LYS

Fig. S10. Amino acid sequence of VA3. Residues n located in a helix are in a filled black box and residues n located in a β-sheet are in a filled orange box.