Supporting Information for

"New Insights into the Folding of a β-Sheet Miniprotein in a Reduced Space of Collective Hydrogen Bond Variables: Application to a Hydrodynamic Analysis of the Folding Flow"

by I. V. Kalgin, A. Caflisch, S. F. Chekmarev and M. Karplus

Collective variables

The algorithm to determine the collective variable that we used in the paper proceeds as follows:

1. Introduce a state vector $\mathbf{p}^m = (p_1^m, p_2^m, ..., p_D^m)$ which is conjugate to the vector of the hydrogen bond distances $\mathbf{h}^m = (h_1^m, h_2^m, ..., h_D^m)$, where D is the number of possible hydrogen bonds (dimension of conformation space), and m is the current number of the protein conformation among the total M conformations that are stored ($1 \leq m \leq M$). The bond distances are measured in angstroms.

2. Set the p_i^m to be equal to 1 if the hydrogen bond is formed and 0 otherwise. A bond was considered to be formed if $h_i^m \leq 3.6$Å and $\angle(\mathrm{N-H-O})_i^m \geq 130^\circ$ [1]. To avoid the noise introduced by the spatial fluctuations characteristic of the C- and N-terminal residues, the hydrogen bonds that these residues formed with the other residues were not included (see Fig. 1 in the main text). This reduces the total number of possible hydrogen bonds to $D = 272$.

3. Apply the standard PCA algorithm [2] to the data set $\mathbf{p}^m = (p_1^m, p_2^m, ..., p_D^m)$ ($1 \leq m \leq M$), i.e. calculate the means $\overline{p}_i = \mathrm{E}(p_i)$, the covariance matrix $\mathbf{C}(p_i, p_j) = \mathrm{E}[(p_i - \overline{p}_i)(p_j - \overline{p}_j)]$ and its eigenvalues λ and eigenvectors $\mathbf{x}$ that satisfy the equation $\mathbf{Cx} = \lambda\mathbf{x}$ (the symbol E denotes the expected value).

4. Choose the desired dimension of the reduced space (K) and direct the collective variables $g_1, g_2, ..., g_K$ along the eigenvectors $\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_K$ corresponding to the largest eigenvalues $\lambda_1, \lambda_2, ..., \lambda_K$. The protein conformation with bond distances $h_1, h_2, ..., h_D$ is determined by the values of the collective variables $g_j = \sum_{i=1}^{D} w_{ij} h_i$, where w_{ij} indicates the contribution of i bond into j variable.

5. Map the representative points of the protein in the original $\mathbf{h} = (h_1, h_2, ..., h_D)$ space onto the reduced $\mathbf{g} = (g_1, g_2, ..., g_\mathrm{K})$ space of collective variables.

As in the conventional PCA method, this algorithm transforms the original configuration space to a set of orthogonal collective variables (in our case, **h** to **g**) and is easy to implement. Since the collective variables are linear combinations of the original variables, they are measured in the same units as the latter, i.e. in angstroms. In what follows we refer to this algorithm to as the Hydrogen Bond PCA (HB PCA) method.

Principal Component Analysis (PCA), the atomic coordinate space

TABLE S1: Clusters of Protein Conformations

Cluster[a]	W_{clst}[b]	N_{str}[c]	Most populated structure[d]	W_{str}[e]	Cluster type[f]
1	24.29	467	-EEEETTEEEEEETTEEEE- -EEEETTEEEEEETTEEE--	33.38 31.77	Native
2	5.21	1613	-EEEETTEEEEEEEEEE-- -EEEETTEEEEEEEEEE-	6.37 5.58	Cs-or
3	11.73	1137	-EEE-SSS-EEEETTEEEE- -EEEETTEEEEEETTEEEE-	14.55 9.77	Native + Ns-or
4	4.39	3881	--HHHHHHHHHHT------ --SS--HHHHTTT-------	0.46 0.32	Helical
5	4.3	2991	-B-SSSSS-EEEETTEESB- -B-S-SSS-EEEETTEESB-	1.98 1.84	Ch-curl
6	2.59	1017	-B-SSSSS--EEETTEE-B- -B-S-SSS--EEETTEESB-	5.63 3.86	
7	2.91	760	-B-SSSS--EESSSSS-EE- -B-SSSSS-EEEETTTEEE-	5.16 4.95	
8	6.5	5502	-----SS--EEEETTEEEE- ----SSS--EEEETTEEEE-	0.57 0.31	Others
9	6.62	1992	--EEETTEEEEEETTEEEE- -EEEETTEEEEEETTEEEE-	4.25 4.07	
10	8.06	6247	-B--SSS---EEETTEE-B- -B--SSS---EEETTEESB-	0.98 0.6	
11	4.9	4132	-EEEETTEEEE---SS--B- -EE-SSS-EE---SS---B-	0.67 0.55	
12	3.98	3721	--SSSS---EEEETTEEEE- -EEEEETTEEEEESSS-EE-	0.18 0.18	
13	3.96	3015	-B--SSS---EEETTEE-B- ----SSS--EEEETTEEEE-	1.14 0.76	
14	4.4	3911	-B-SSSSS--EEETTEE-B- -B-SSSS----EETTEE-B-	0.7 0.43	
15	6.17	4086	-EEEETTEEEE-SSS----- --EEETTEEEEEEEEEEE-	0.71 0.71	

[a] Cluster number.

[b] Cluster weight equal to the number of the representative points in the cluster relative to the total number of the points (in %).

[c] The number of conformations that have different secondary structure strings.

[d] The secondary structure strings of the most populated conformations.

[e] Weight of the given conformation in the cluster (in %).

[f] Corresponds to the clusters of Fig. S1.

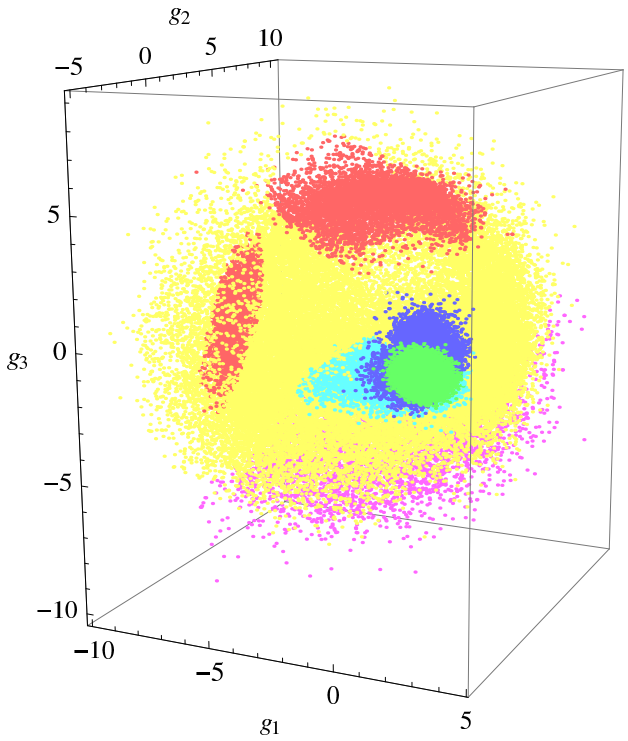

FIG. S1: Distribution of the representative points of the protein in the $\mathbf{g} = (g_1, g_2, g_3)$ space. The green, cyan, blue, magenta, red and yellow points correspond to Native, Cs-or, Native+Ns-or, Helical, Ch-curl and others conformations respectively. The representative points of the protein in the $\mathbf{g} = (g_1, g_2, g_3)$ space are clustered using MCLUST method [3].

Comments: The Ns-or cluster is not separated from the native cluster.

Principal Component Analysis (PCA), the hydrogen bond distance space

TABLE S2: Clusters of Protein Conformations (Notations are as in Table S1)

Cluster[a]	W_{clst}[b]	N_{str}[c]	Most populated structure[d]	W_{str}[e]	Cluster type[f]
1	14.72	1670	-EEEETTEEEEEETTEEE--	36.34	Native
			-EEEETTEEEEEETTEEEE-	25.49	
2	5.37	2572	-EEEETTEEEEEETTEEEE-	19.11	
			--EEETTEEEEEETTEEEE-	10.85	
3	5.81	2918	-EEEETTEEEEEETTEEE--	36.1	
			-EEEETTEEEEEETTEE---	8.47	
4	2.75	3696	-EEEETTEEEEEEEEEE--	9.57	Cs-or
			-EEEETTEEEEEEEEEEE-	9.42	
5	2.58	1272	-EEE-SSS-EEEETTEEEE-	19.47	Ns-or
			-EEEESSSEEEEETTEEEE-	19.16	
6	11.73	2593	-EEEETTEEEEEETTEEEE-	35.03	Ns-or+Native
			-EEE-SSS-EEEETTEEEE-	10.02	
7	6.61	35199	--HHHHHHHHHHT------	0.52	Helical 1
			---HHHHHHHHHT------	0.27	
8	4.36	33520	----SS---BTTTTBSS---	0.21	Helical 2
			-----SHHHHHHBTTTB---	0.16	
9	7.34	8920	-B-SSSSS--EEETTEE-B-	2.6	Ch-curl
			-B-SSSSS-EEEETTTEEE-	1.98	
10	9.91	34301	-----SS--EEEETTEEEE-	0.85	Others
			-B--SSS---EEETTEE-B-	0.73	
11	4.92	8266	-EEEETTEEEEEETTEEEE-	4.7	
			---EETTEEEEEETTEEEE-	3.05	
12	11.6	68766	--EEETTEEEEETTTEEEE-	0.34	
			--EEETTEEEEETTTEES--	0.16	
13	4.7	13997	-EEEETTEEEEEETTEE---	2.42	
			-EEEETTEEEEEETTEEE--	1.76	
14	2.56	18376	--EEETTEEE--BTTTB---	0.5	
			---EETTEE---BTTTB---	0.46	
15	5.04	35813	-----SS-SS-EETTEE---	0.18	
			-EEETTTEEEETTTEEEE--	0.16	

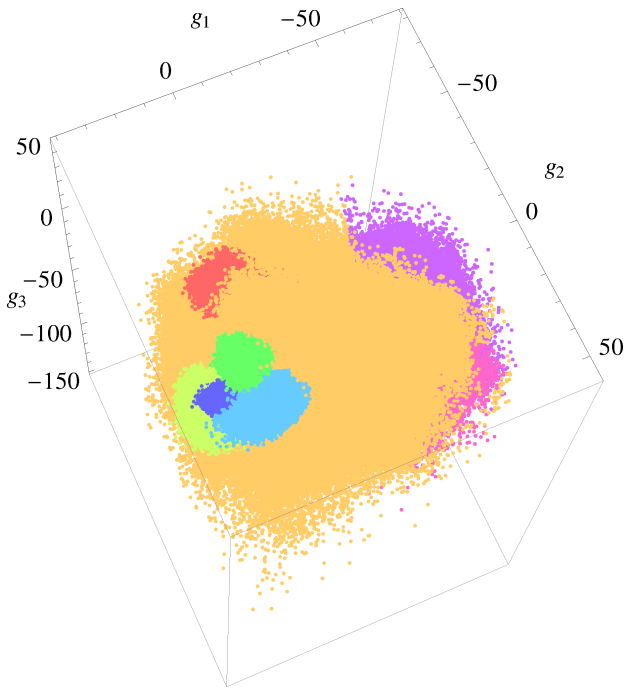

FIG. S2: Distribution of the representative points of the protein in the $\mathbf{g} = (g_1, g_2, g_3)$ space. The green-yellow, green, light blue, blue, purple, magenta, red and yellow points correspond to Native, Cs-or, Ns-or, Ns-or+Native, Helical 1, Helical 2, Ch-curl, and others conformations respectively. The representative points of the protein in the $\mathbf{g} = (g_1, g_2, g_3)$ space are clustered using MCLUST method [3].

Comments: The Cs-or and Ns-or clusters are not symmetric with respect to the Native cluster along g_2. A cluster similar to the intermediate Ns-or+Native cluster, i.e. a Cs-or+Native cluster, is not observed.

Local Linear Embedding (LLE) [4], the atomic coordinate space

TABLE S3: Clusters of Protein Conformations (Notations are as in Table S1)

Cluster[a]	W_{clst}[b]	N_{str}[c]	Most populated structure[d]	W_{str}[e]	Cluster type[f]
1	9.44	854	-EEEETTEEEEEETTEEE--	37.89	Native
			-EEEETTEEEEEETTEEEE-	15.18	
2	17.91	615	-EEEETTEEEEEETTEEEE-	39.27	
			-EEEETTEEEEEETTEEE--	24.37	
3	3.08	1205	-EEEETTEEEEEEEEEE--	8.71	Cs-or
			-EEEETTEEEEEEEEEEE-	8.54	
4	2.37	580	-EEEESSSEEEEETTEEEE-	19.25	Ns-or
			-EEE-SSS-EEEETTEEEE-	17.1	
5	7.06	3599	-EEE-SSS-EEEETTEEEE-	9.85	Ns-or + Others
			--EEETTEEEEEETTEEEE-	1.56	
6	3.46	1013	-B-SSSS--EEEETTTEEE-	4.48	Ch-curl
			-B-SSSS--EESSSSS-EE-	4.45	
7	2.96	1210	-B-SSSSS--EEETTEE-B-	6.02	
			-B-S-SSS--EEETTEESB-	3.45	
8	3.02	1872	-EEE-SSS-EEEETTEEEE-	2.62	
			-B--SSS---EEETTEE-B-	2.35	
9	6.59	5503	-EEE-SSS-EEEETTEEEE-	0.52	Others
			-EEETTTEEEETTTEEEE--	0.3	
10	7.49	6027	-EEE-SSS-EEEETTEEEE-	1.72	
			-EEEETTEEEEEETTEE---	0.61	
11	6.93	2358	-EEEETTEEEEEETTEEE--	6.48	
			-EEEETTEEEEEETTEEEE-	5.98	
12	4.72	1596	-EEEETTEEEEEETTEEEE-	12.94	
			--EEETTEEEEEETTEEEE-	9.14	
13	3.59	3023	-EEEETTEEEEEETTEE---	0.97	
			-EEEETTEEEEEETTEEE--	0.95	
14	0.98	879	--EEEEETTEEEETTEEEE-	0.92	
			--EEEE-SSEEEETTEEEE-	0.72	
15	8.53	6357	-EEEETTEEEEEETTEEEE-	0.83	
			-EEE-SSS-EEEETTEEEE-	0.42	
16	4.72	3728	-EE-SSS-EE---SS---B-	1.17	
			--EEETTEEEEEETTEEEE-	1.02	
17	7.16	6262	--SSSS---EEEETTEEEE-	0.36	
			--SSSSS--EEEETTEEEE-	0.31	

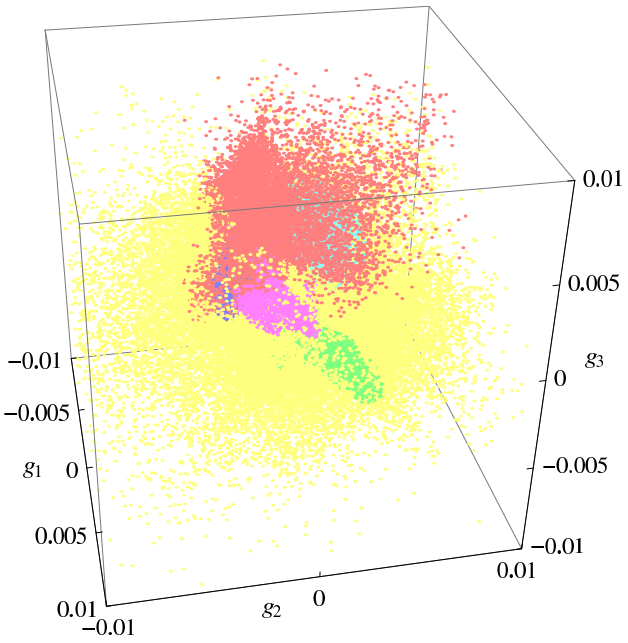

FIG. S3: Distribution of the representative points of the protein in the $\mathbf{g} = (g_1, g_2, g_3)$ space. The green, cyan, blue, magenta, red and yellow points correspond to Native, Cs-or, Ns-or, Ns-or+Native, Ch-curl and others conformations respectively. The representative points of the protein in the $\mathbf{g} = (g_1, g_2, g_3)$ space are clustered using MCLUST method [3].

Comments: All clusters are more or less separated, but the picture is vague.

Full Correlation Analysis (FCA) [5, 6], the atomic coordinate space

TABLE S4: Clusters of Protein Conformations (Notations are as in Table S1)

Cluster[a]	W_{clst}[b]	N_{str}[c]	Most populated structure[d]	W_{str}[e]	Cluster type[f]
1	31.02	1584	-EEEETTEEEEEETTEEEE- -EEEETTEEEEEETTEEE--	25.91 23.42	Native
2	3.76	1525	-EEEETTEEEEEETTEEEE- -EEEETTEEEEEETTEE---	11.88 8.48	
3	5.66	1420	-EEEETTEEEEEETTEEE-- -EEEETTEEEEEETTEE---	16.45 5.14	
4	8.14	2713	-EEEETTEEEEEETTEEEE- --EEETTEEEEEETTEEEE-	12.16 5.24	
5	2.45	605	-EEEETTEEEEEEEEEE-- -EEEETTEEEEEEEEEEE-	12.14 11.24	Cs-or
6	2.21	1675	-EEEETTEEEE--SS----- -EEEETTEEEEEEEEE---	1.22 1.22	Ns-or
7	2.23	2009	--HHHHHHHHHHT------ --TTTTHHHHHHS------	0.54 0.31	Helical
8	3.82	3011	-B--SSS---EEETTEE-B- -B--SSS---EEETTEESB-	2.67 1.39	Ch-curl
9	2.85	1068	-B-SSSSS--EEETTEE-B- -B-SSSS---EEETTEE-B-	5.47 3.71	
10	3.52	2357	-B-SSSSS-EEEETTEESB- -B-S-SSS-EEEETTEESB-	2.47 2.24	
11	2.89	765	-B-SSSS--EESSSSS-EE- -B-SSSS--EEEETTTEEE-	5.3 5.02	
12	3.41	2824	-EEETTEEEETTEEEE-- -EEEETTEEEEEEEEEE--	0.7 0.56	Others
13	5.	3640	--SSSSS--EEEETTEEEE- -EE-SSS-EE---SS---B-	0.92 0.8	
14	3.53	3090	-----SS--EEEETTEEEE- -B---SSS-----SSS--B-	0.34 0.25	
15	6.52	5303	-EEEETTEEEEEETTEE--- --EEETTEEEEEETTEEEE-	0.69 0.43	
16	6.43	5444	-B-SSSSS--EEETTEE-B- -EEEETTEEEE---SS--B-	0.53 0.51	
17	6.56	5952	-B--SSS---EEETTEE-B- ----SSS-----SSS-----	0.24 0.24	

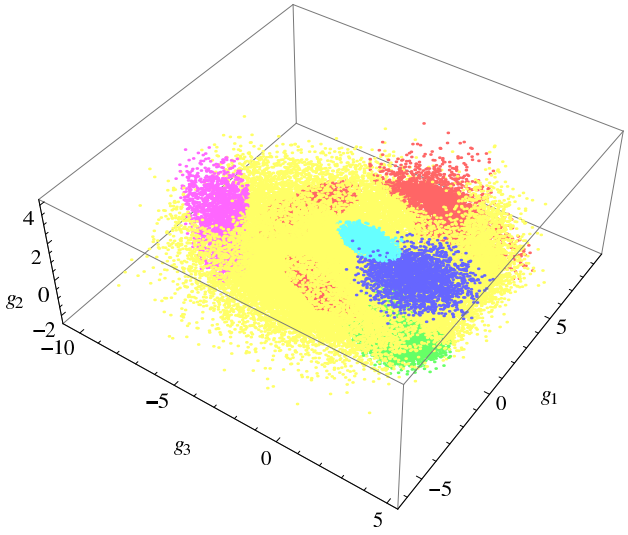

FIG. S4: Distribution of the representative points of the protein in the $\mathbf{g} = (g_1, g_2, g_3)$ space. The green, cyan, blue, magenta, red and yellow points correspond to Native, Cs-or, Ns-or, Helical, Ch-curl and others conformations respectively. The representative points of the protein in the $\mathbf{g} = (g_1, g_2, g_3)$ space are clustered using MCLUST method [3].

Comments: Separation into clusters is reasonable, but the Ns-or cluster (blue) is positioned between the Cs-or (cyan) and Native (green) clusters.

Manifold Sculpting (MS) [7], the hydrogen bond distance space

TABLE S5: Clusters of Protein Conformations (Notations are as in Table S1)

Cluster[a]	W_{clst}[b]	N_{str}[c]	Most populated structure[d]	W_{str}[e]	Cluster type[f]
1	6.07	425	-EEEETTEEEEEETTEEEE-	48.53	Native
			--EEETTEEEEEETTEEEE-	6.31	
2	12.8	820	-EEEETTEEEEEETTEEE--	35.25	
			-EEEETTEEEEEETTEEEE-	15.77	
3	10.08	572	-EEEETTEEEEEETTEEE--	35.28	
			-EEEETTEEEEEETTEEEE-	21.59	
4	3.4	239	-EEEETTEEEEEETTEEEE-	53.52	
			-EEEETTEEEEEETTEEE--	5.39	
5	2.58	577	-EEEETTEEEEEEEEEE--	12.28	Cs-or
			-EEEETTEEEEEEEEEE-	10.3	
6	3.25	320	-EEE-SSS-EEEETTEEEE-	43.28	Ns-or
			-EEEESSSEEEEETTEEEE-	5.48	
7	2.66	262	-EEEESSSEEEEETTEEEE-	19.04	
			-EEE-SSS-EEEETTEEEE-	18.59	
8	6.47	5050	--HHHHHHHHHHT------	0.53	Helical
			---HHHHHHHHHS------	0.34	
9	0.93	748	-B-SSSSS-EEETTTTEEE-	3.35	Ch-curl
			-B-SSSS--EEETTTTEEE-	2.7	
10	1.03	276	-B-SSSSS-EEEETTEESB-	8.49	
			-B-S-SSS-EEEETTEESB-	7.9	
11	1.88	429	-B-SSSS--EEEETTTEEE-	8.4	
			-B-SSSS--EESSSSS-EE-	8.4	
12	0.94	170	-B-SSSS--EE-SSS--EE-	13.89	
			-B-SSSSS-EE-SSS--EE-	13.68	
13	2.84	605	-B-SSSSS--EEETTEE-B-	7.44	
			-B-SSSS---EEETTEE-B-	4.48	
14	1.92	1377	--EEETTEEEEETTTEEEE-	1.46	Others
			-EEEETTEEEE-SSS-----	1.3	
15	5.56	2160	-EEEETTEEEEEETTEEEE-	4.36	
			--EEETTEEEEEETTEEEE-	3.11	
16	2.32	1432	-EE--SSSEEEEETTEEEE-	1.98	
			-----SSS--EEETTEEE--	1.38	
17	4.22	1339	-EEEETTEEEEEETTEEEE-	11.33	
			-EEEETTEEEEEETTEE---	7.73	
18	9.27	8548	----SS---BTTTTBSS---	0.15	
			----SSS-----SSS-----	0.11	
19	6.04	4144	-----SS--EEEETTEEEE-	1.01	
			----SSS--EEEETTEEEE-	0.71	
20	6.99	5782	-----SS--S-EETTEE---	0.3	
			-----SSS--EEETTEEEE-	0.29	
21	5.3	4346	-EEEETTEEEE---SS--B-	0.57	
			---EETTEE---SSS-----	0.51	
22	3.48	3127	-B---SSS-----SSS--B-	0.58	
			-B---SSS----BTTTB-B-	0.37	

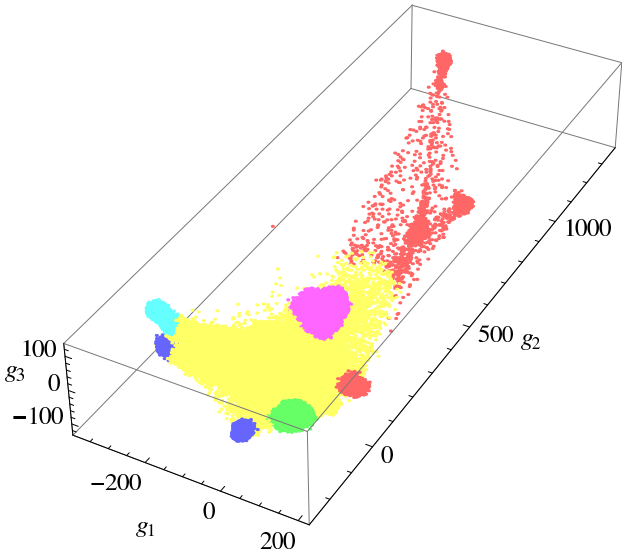

FIG. S5: Distribution of the representative points of the protein in the $\mathbf{g} = (g_1, g_2, g_3)$ space. The green, cyan, blue, magenta, red and yellow points correspond to Native, Cs-or, Ns-or, Helical, Ch-curl and others conformations respectively. The representative points of the protein in the $\mathbf{g} = (g_1, g_2, g_3)$ space are clustered using MCLUST method [3].

Comments: Clustering is good but the Ns-or and Ch-curl subclusters are very scattered.

TABLE S6: Comparison of the Weights (in %) of Consolidated Clusters

Cluster	KGA[a]	pfoldf[b]	REMD[c]	CTMD[d]	PCA[e]	PCA[f]	LLE	FCA	MS
Native	36.4	35.0	37.8	37.1	24.3	37.63	27.4	36.7-48.6[h]	33.1
Cs-or	3.6	2.6	5.3	5.3	5.21	2.75	3.1	2.5	2.6
Ns-or	7.4	6.2	7.3	6.3	11.7	2.58	2.4 - 9.5[g]	2.2	5.9
Helical	11.6	11.2	-	-	4.4	10.97	-	2.2	6.5
Ch-curl	6.0	4.9	3.9	1.8	9.8	7.34	9.4	13.1	7.6

[a] Kinetic grouping analysis (KGA), Refs. [8, 9].

[b] p_{fold} analysis based on an equilibrium kinetic network (pfoldf), Ref. [9].

[c] Replica exchange molecular dynamics (REMD), Ref. [10].

[d] Constant temperature molecular dynamics (CTMD), Ref. [10].

[e] PCA in coordinate space.

[f] PCA in hydrogen bonds space.

[g] Ns-or and Ns-or+Others.

[h] Clusters "1+3" and "1+2+3+4".

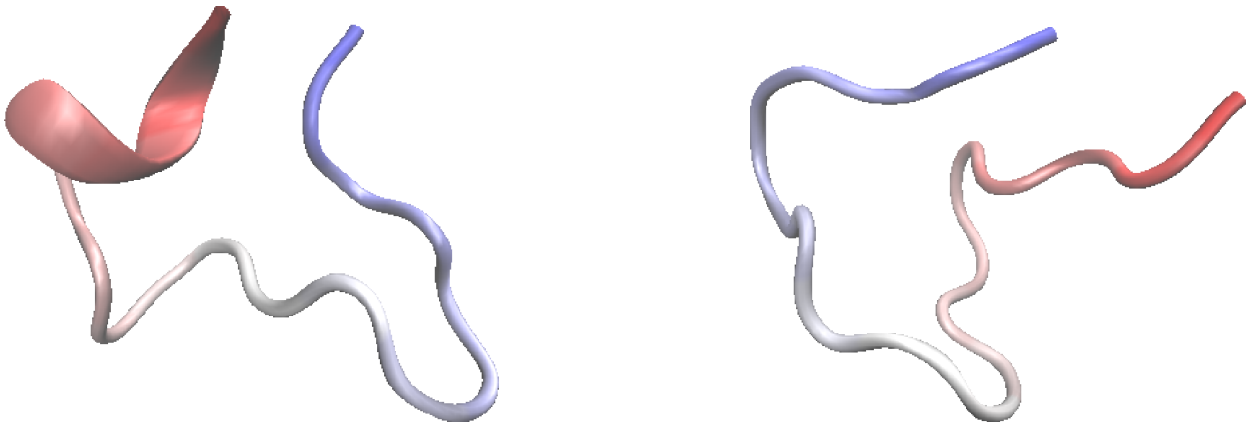

FIG. S6: Examples of unstructured conformers

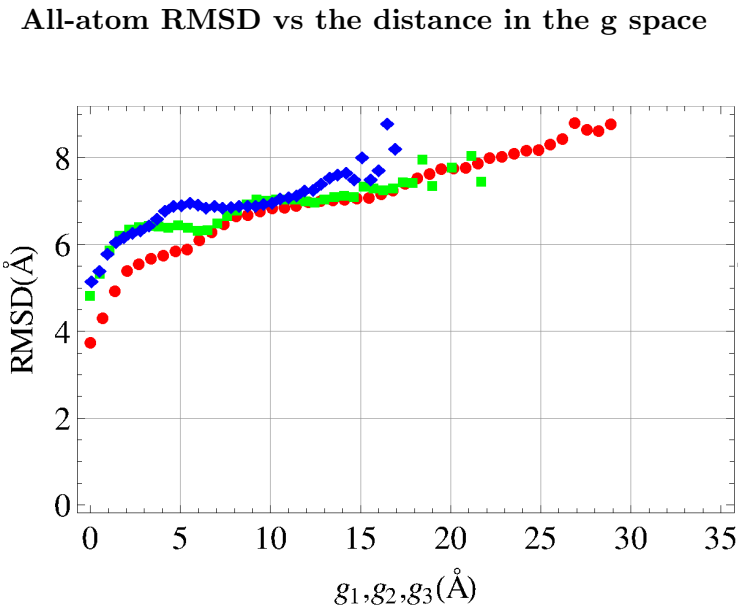

FIG. S7: Average all-atom RMSD between two conformations which correspond to different points (1 and 2) in the **g** space as a function of the distance $g_i = |g_i^2 - g_i^1|$, where $i = 1, 2, 3$, and the upper indices denote the points. The averaging is taken over all points in the $g_i = g_i^1$ and $g_i = g_i^2$ planes. The red, green and blue curves are for g_1, g_2 and g_3, respectively.

Projections of the eigenvectors onto the hydrogen bond space

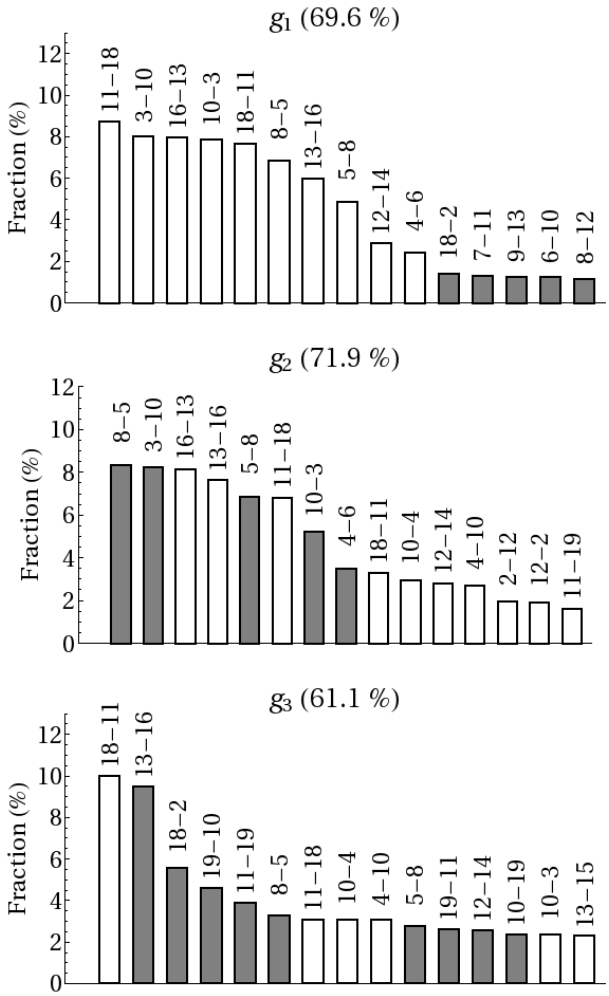

FIG. S8: Fractions of the hydrogen bonds which have a major contribution to the collective variables g_1, g_2 and g_3. The figures at the top of each bar denote the bond; the first figure is the number of the residue with the oxygen atom and the second figure is that with the nitrogen atom. The empty and solid bars are for the bond contributions to the negative and positive directions of the collective variable, respectively. The numbers in percentage at the top of each panel are the total contribution of the given bonds to the collective variable.

Contributions of hydrogen bonds to collective variables

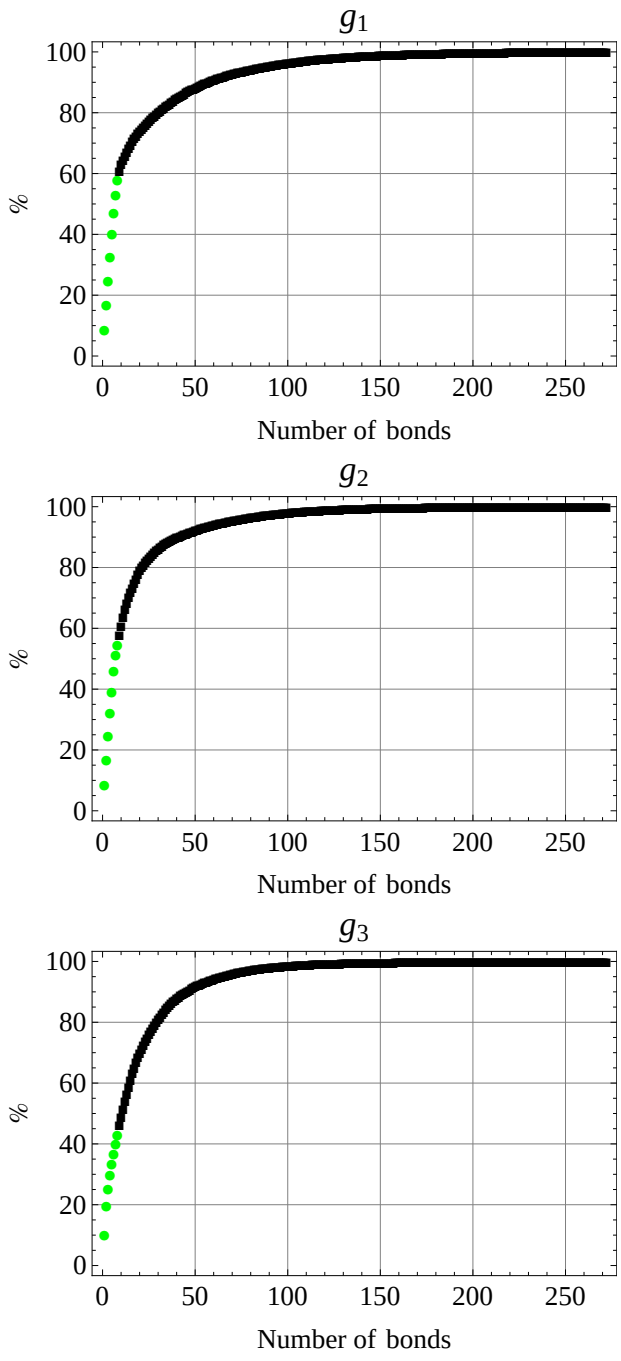

FIG. S9: Cumulative contributions of the hydrogen bonds to the collective variables g_1, g_2 and g_3. The green-colored points corresponds to the first eight bonds depicted in Fig. 6 of the main text.

Free energy profiles: Clustering over g_1 and the $\mathbf{g} = (g_1, g_2, g_3)$ space

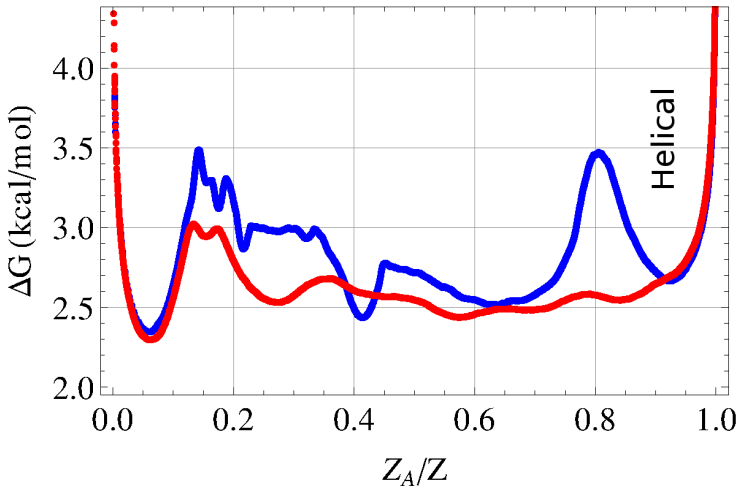

FIG. S10: Free energy profiles. The red curve corresponds to the clustering over the g_1 collective variable (Fig. 7 of the main text), and the blue curve to the clustering in the $\mathbf{g} = (g_1, g_2, g_3)$ space.

Comparison to the Zheng et al. study [12]

In their study of folding of beta3s miniprotein, Zheng et al. [12] used the LSDMap technique by Rohrdanz et al. [13] to reduce the conformation space of the protein. They have found that the first principal coordinate plays the role of the reaction coordinate for the folding process and the others, which correspond to smaller eigenvalues, discriminate between the clusters of representative conformations of the protein (basins on the FES). The LSDMap is a nonlinear dimensionality reduction technique, which treats the whole ensemble of configurations sampled from an MD simulation as a noisy data set that resides on a low-dimensional underlaying manifold. Calculating the radius in configuration space around every point within which the manifold of the representative points can be approximated by a hyperplane tangent to the manifold, a Gaussian kernel determining the affinity of the points is introduced. This kernel is considered as a transition probability matrix for a Markov process to explore the configuration space of the system (equivalently, the process is described by the Fokker-Plank equation). The eigenfunctions of the kernel matrix present collective coordinates corresponding to the (diffusive) motion of the system on different time scales. Specifically, the zeroth eigenfunction corresponds to the Boltzmann distribution, the first eigenfunction to the collective motion with the slowest time scale, the second eigenfunction to the second slowest collective motion, etc., so that the variables of Ref. [12] allow discrimination of the protein dynamics on different time scales. However, a consequence of this useful feature of the variables of Ref. [12] is that the spatial distributions constructed with use of these variables (FESs, spatial kinetic networks, etc.) will be "biased", because different variables correspond to different time scales. In contrast, the variables we use allow the construction of "unbiased" spatial distributions. Different time scales can be distinguished by comparison of the cross-sections of the tubes that connect the clusters, i.e. the larger the cross-section, the faster the motion.

Rates of transitions between the clusters of conformations vs the distances in the atomic coordinate space

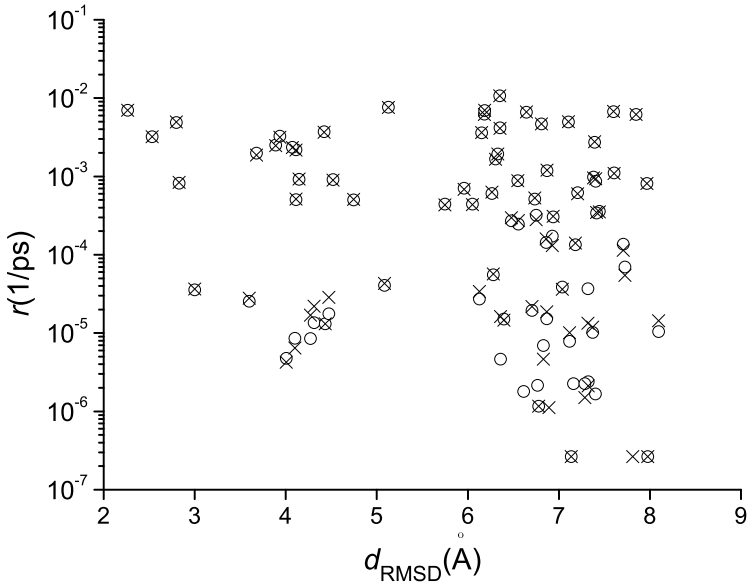

FIG. S11: Rates of transitions between the clusters of conformations vs the distances between the centers of the clusters in the atomic coordinate space. Crosses and circles are for the transitions from smaller and larger populated clusters, respectively.

TABLE S7: Numbers of transitions between the clusters and residence times.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
1	164687	9333	16	12161	138	0	26291	1	1	0	0	128	1923	55	8	1	0
2	9473	21614	13	1108	10	0	1869	0	0	0	0	371	4487	77	7	11	0
3	18	8	20627	239	0	0	3	0	0	0	0	4993	20	0	165	128	0
4	12080	1172	236	14019	12	0	2141	0	0	0	0	715	933	7	61	2	0
5	135	5	0	11	26124	1307	467	0	0	0	0	0	501	1076	89	0	0
6	0	0	0	0	1309	21040	6	0	0	0	0	0	13	2128	52	0	0
7	26293	1930	4	2111	458	6	16273	0	0	0	0	34	2195	177	8	0	0
8	0	0	0	0	0	0	0	57715	8238	0	0	0	0	2	1085	29	9303
9	1	0	0	0	0	0	0	8299	38533	0	0	0	0	9	857	1	3678
10	0	0	0	0	0	0	0	0	0	31748	603	0	0	575	201	283	0
11	0	0	0	0	0	0	0	0	0	600	41472	1	0	31	1029	585	454
12	103	356	5003	722	1	0	33	3	0	0	1	33735	735	43	863	3944	221
13	1886	4506	19	944	508	8	2221	0	0	0	0	684	17723	3231	292	168	0
14	57	102	0	2	1075	2125	162	3	7	574	33	55	3194	70172	5578	671	22
15	9	6	167	58	80	62	22	1080	875	198	1017	873	301	5550	63472	4131	9202
16	1	9	115	3	0	0	1	26	0	290	605	3922	165	683	4081	20111	3571
17	0	0	1	0	0	0	0	9245	3724	0	441	252	0	16	9255	3517	70631

The headers are the cluster numbers according to Table I of the main text. The off-diagonal elements are the numbers of transitions between the clusters, and the diagonal elements are the numbers of transitions within the clusters (the residence times) registered in the course of the 20μs trajectory.

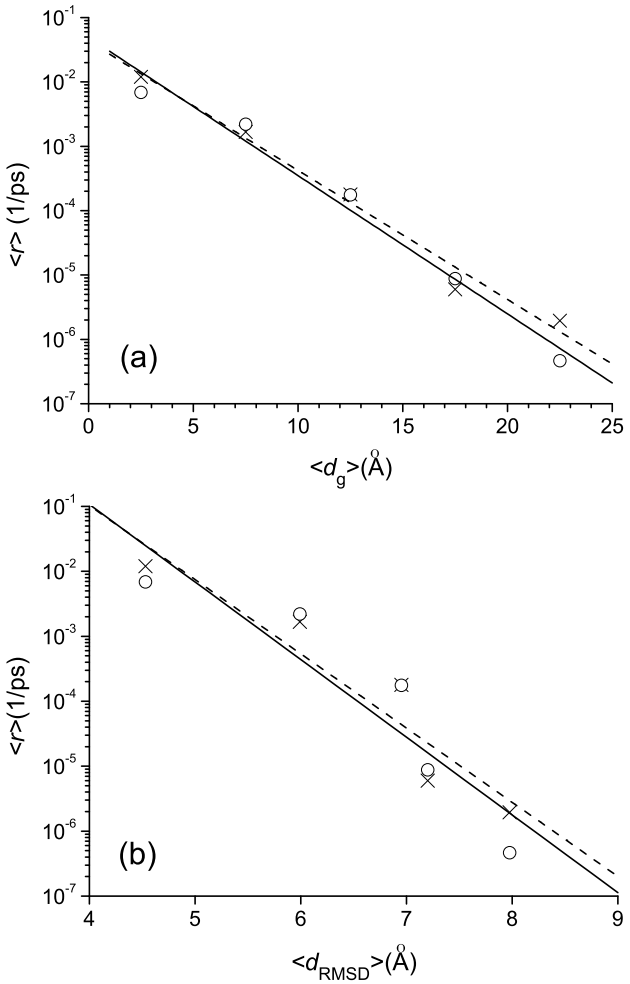

FIG. S12: Average rates of transitions between the clusters of conformations vs the average distances between the centers of the clusters in (**a**) **g** space and (**b**) atomic coordinate space. Crosses and circles are for the transitions from smaller and larger populated clusters, respectively. In panel (**a**), the dashed line corresponds to the best fit for the crosses [$\langle r\rangle \sim \exp(-0.46\langle d_{\rm g}\rangle)$], and the solid line to that for the circles [$\langle r\rangle \sim \exp(-0.49\langle d_{\rm g}\rangle)$]. Similar in panel (**b**), where the dashed line is the best fit for the crosses [$\langle r\rangle \sim \exp(-2.63\langle d_{\rm RMSD}\rangle)$], and the solid line is the best fit for the circles [$\langle r\rangle \sim \exp(-2.75\langle d_{\rm RMSD}\rangle)$].

[1] Seeber, M.; Cecchini, M.; Rao, F.; Settanni, G.; Caflisch, A. *Bioinformatics* **2007**, *23*, 2625-2627.

[2] Jolliffe, I. T. *Principal Component Analysis*; 2nd ed.; Springer: New York, 2002.

[3] Fraley, C.; Raftery, A. E. *J. Am. Stat. Assoc.*, **2002**, *97*, 611-631.

[4] Roweis, S. T.; Saul, L.,K. *Scince*, **2000**, *290*, 2323-2326.

[5] Lange, O.; Grubmuller, H. *Proteins* **2006** *62*, 1053-1061.

[6] Lange, O.; Grubmuller, H. *Proteins* **2008** *70*, 1294-1312.

[7] Gashler, M.; Ventura, D.; Martinez, T. In *Advances in Neural Information Processing Systems 20*; Platt, J. C., Koller, D., Singer, Y., Roweis, S., Eds.; MIT Press: Cambridge, MA, 2008, pp. 513-520.

[8] Muff, S.; Caflisch, A. *Proteins: Struct., Funct., Bioinform.* **2008**, *70*, 1185-1195.

[9] Krivov, S. V.; Muff, S.; Caflisch, A.; Karplus, M. *J. Phys. Chem. B* **2008**, *112*, 8701-8714.

[10] Muff, S.; Caflisch, A. *J. Phys. Chem. B* **2009**, *113*, 3218-3226.

[11] Muff, S.; Caflisch, A. *J. Chem. Phys.* **2009**, *130*, 125104.

[12] Zheng, W.; Qi, B.; Rohrdanz, M. A.; Caflisch, A; Dinner, A. R.; Clementi, C. *J. Phys. Chem. B* **2011**, *115*, 13065-13074.

[13] Rohrdanz, M. A.; Zheng, W.; Maggioni, M.; Clementi, C. *J. Chem. Phys.* **2011**, *134*, 124116.