Analysis of the distributed computing approach applied to the folding of a small β peptide

Emanuele Paci*†, Andrea Cavalli*, Michele Vendruscolo‡, and Amedeo Caflisch*

*Biochemisches Institut der Universität Zürich, Winterthurerstrasse 190, CH-8057 Zürich, Switzerland; and ‡Department of Chemistry, University of Cambridge, Lensfield Road, Cambridge CB2 1EW, United Kingdom

Edited by Alan Fersht, University of Cambridge, Cambridge, United Kingdom, and approved May 8, 2003 (received for review January 15, 2003)

In the recently proposed distributed computing approach to protein folding a very large number of short independent simulations is performed. Using this method, folding events on a time scale orders of magnitude shorter than the experimental one have been reported. However, it has also been observed that the folding process is not an elementary kinetic step and that the presence of initial lag phases can bias short simulations toward atypical pathways. We study here a 20-residue three-stranded antiparallel β-sheet peptide whose equilibrium properties can be characterized by atomistic molecular dynamics simulations. We found that the folding rate of this peptide is estimated correctly by the distributed computing approach when trajectories >≈1/100 of the equilibrium folding time are considered. We also found that the fastest folding events occur through high-energy pathways, which are unlikely under equilibrium conditions. These very fast folding pathways do not relax within the equilibrium denatured state that is stabilized by the transient presence of both native and non-native interactions, and they are characterized by the nearly simultaneous formation of the two β-hairpins and a very small number of non-native contacts.

Molecular dynamics simulations of the protein folding process represent a formidable challenge because the equations of motion of a system of thousands of atoms must be integrated over a time scale that ranges from microseconds to seconds. At present, on a single processor, the trajectory of a protein can be followed for a few nanoseconds (in water) or a few hundreds of nanoseconds (in implicit solvent). The simultaneous use of several thousand processors can in principle close the gap between observed folding times and simulated ones. There are two important aspects to parallel simulations. The first is the splitting of the system into smaller parts that can be processed separately. However, the interactions between the subsystems and the communication between the processors constitute a serious problem when thousands of processors are to be used simultaneously (see, e.g., www.research.ibm.com/bluegene). The second aspect is the splitting of the total simulation time into shorter times that can be processed independently. This is a difficult problem, ultimately because time is essentially serial.

Pande and coworkers (1, 2) have recently applied a new approach to protein folding, the "distributed computing." The method is based on the idea that overcoming an energy barrier is a stochastic process characterized by a distribution of crossing times. In a first-order process of time constant τ, the number, δN, of simulations that go to completion in a time $\delta t \ll \tau$ is $\delta N = N\delta t/\tau$, where N is the total number of simulations performed. For example, if $\tau = 10^{-5}$ s, then one in 1,000 simulations should be completed within a time $\delta t = 10^{-8}$ s (3).

A complication of the method is that escaping from a metastable minimum in a multidimensional energy landscape may involve several different transition states, corresponding to different transition times (1, 3). Moreover, the spontaneous search of the energy minimum may involve a succession of metastable minima (1, 3). In its simplest implementation, the distributed computing approach is capable of describing correctly the folding process when: (*i*) the entire folding process follows first-order kinetics, i.e., there are no additional lag phases, and (*ii*) the distribution of the folding times is such that the fastest folding times are shorter than few tens of nanoseconds, i.e., the time that can be simulated currently on a single processor.

In this study we determined the entire distribution of folding times for GS, a 20-residue designed peptide (4), ranging from extremely short ones (≤0.3 ns) to those required by the normal folding pathways (≈100 ns).

Methods

Model. The molecular dynamics simulations and part of the analysis of the trajectories were performed by using an all-atom model of the protein (5, 6). An implicit solvation model based on the solvent-accessible surface was used to describe the main effects of the aqueous solvent (7). The same force field and implicit solvent model have been used recently in molecular dynamics simulations of folding of structured peptides (α-helices and β-sheets) ranging in size from 15 to 31 residues (7–9) and small proteins of ≈60 residues (10, 11). Despite the absence of collisions with water molecules, in our simulations with an implicit solvent the separation of time scales is comparable with that observed experimentally. Helices fold in ≈1 ns (12), β-hairpins in ≈10 ns (12), and triple-stranded β-sheets in ≈100 ns (13), whereas the experimental values are ≈100 ns, (14), ≈1 μs (14), and ≈10 μs (4), respectively.

Definition of Progress Variables. For the GS peptide a list of $Q = 26$ native contacts was determined by observing contacts present at least 50% of the time in a 200-ns simulation at 300 K (8). The C_α rms deviation between the average structure over the 200-ns simulation at 300 K and the average NMR conformation is 1.9 Å (1.7 Å if one neglects the first and last residues). Average interproton distance violations were derived from the 200-ns run as $d_{viol} = \langle r(t)^{-6}\rangle^{-1/6} - r_{exp}$, where $r(t)$ is the interproton distance at simulation time t, r_{exp} is the nuclear Overhauser effect (NOE) upper distance limit (4), and $\langle\ \rangle$ represents a time average. During the 200-ns simulation at 300 K, 23 of the 26 experimental NOEs (4) are satisfied ($d_{viol} < 0$ Å for 16 distances and $d_{viol} < 1$ Å for 7 distances). The three violations involve the distances between $C_{\varepsilon 3}$H Trp-2–$C_{\beta'}$H Asn-13, C_εH Tyr-11–$C_\gamma H_3$ Ile-18, and C_αH Thr-1–$C_\delta H_3$ Ile-3 with d_{viol} values of 3.9, 2.7, and 3.3 Å, respectively. The four NOEs with strong and medium-strong intensity concerning C_αH–C_αH distances are satisfied with $d_{viol} < 0$ Å (Trp-2–Tyr-11, Gln-4–Lys-9, Trp-10–Tyr-19, and Gln-12–Lys-17). Only type II′ turns were observed at 300 K.

The native contacts include 10 backbone hydrogen bonds (five in each β-hairpin) and 16 contacts between side chains. The number of native contacts between strands 1–2 and 2–3 are, respectively, $Q_{12} = 11$ and $Q_{23} = 11$ (see table 2 of ref. 8). A hydrogen bond is considered formed if the distance between the hydrogen atom of a NH group and a main-chain oxygen is <2.6 Å. A side-chain contact is defined as a pair of side chains whose

This paper was submitted directly (Track II) to the PNAS office.

†To whom correspondence should be addressed. E-mail: paci@bioc.unizh.ch.

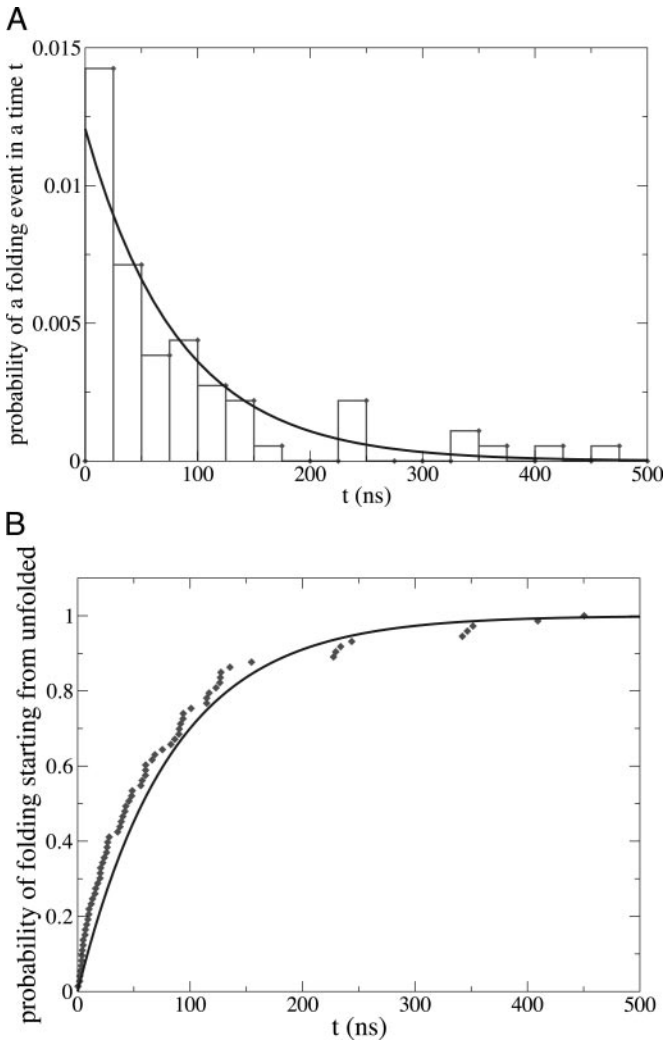

Fig. 1. (*A*) Probability that a folding event occurs in a time *t*; we also show the Poisson distribution $P(t) = k \exp(-kt)$, with $k = 1/83$ ns^{-1}. (*B*) Probability of folding in a time *t* starting from the unfolded state; the continuous line represents $P(t) = 1 - \exp(-kt)$, with $k = 1/83$ ns^{-1}.

center of geometry is closer than 6.5 Å. A folding event is defined when $Q \geq 23$; an unfolding event is defined when $Q \leq 3$.

The total number of contacts (shown in Fig. 6), including non-native ones, is obtained by counting all hydrogen bonds and side-chain contacts between all pairs of residues at least three positions apart in the sequence; we also add contacts between side chains of the pairs of residues 8–10, 16–18, and 18–20, because they are considered native contacts.

The simulations described in this article, for a total of 33.4 μs (12.6 μs for the equilibrium folding simulations and 20.8 μs for the 14,300 short simulations), required a total of 500 days on an Athlon MP 2100+ processor and were performed on a Beowulf cluster.

Results

Equilibrium Simulations. The equilibrium behavior of the GS peptide at 330 K has been studied by performing four simulations for a duration ranging from 2.7 to 4.4 μs, started from the folded structure. Over the total simulation time of 12.6 μs we observe 72 folding events and 73 unfolding events. The average time to go from a completely unfolded state to a completely folded one, i.e., the folding time, is $\tau_{eq} = 83$ ns. Folding times vary between 0.34 and 450 ns. The histogram of the folding times τ_f (Fig. 1*A*) can be approximately fitted by a Poisson distribution. The limited number of available folding events, 72, however, does not allow us to identify unambiguously possible deviations from a Poisson distribution. The estimated probability of folding in a time *t* is shown in Fig. 1*B*.

The free energy surface of GS can be calculated from the equilibrium simulations as a function of two progress variables (Q_{12} and Q_{23}, see *Methods*), as shown in Fig. 2. This plot shows an equilibrium property and does not give direct information on the kinetics of the folding events. For this reason we analyzed in detail the last 1 ns before reaching the folded state in the folding events observed in the equilibrium simulations. This time span represents the postcritical phase in the folding process, where the system has already crossed (or is about to cross) the rate-limiting barrier and folding proceeds fast. For the last 1 ns of six randomly chosen equilibrium folding events, we show in Fig. 3*A* the total number of contacts, the number of native contacts Q, the number Q_{12} of native contacts between strands 1 and 2 (hairpin 12), and the number Q_{23} of native contacts between strands 2 and 3 (hairpin 23); see *Methods* for the exact definitions. One important finding is that the total number of contacts (native and non-native) can be much larger than the number of native contacts. Therefore, several non-native interactions are always present until the very last phase of the folding process where the total number of contacts and the number of native contacts are highly correlated. Even in the postcritical phase of equilibrium folding events, several non-native contacts are present. In four of six cases the hairpin 23 forms first, in one case the hairpin 12 forms first, and in one case the order of hairpins formation is ambiguous.

The equilibrium folding mechanism also has been investigated by considering all 72 folding events we recorded and calculating the histograms of Q_{12} and Q_{23} at three stages of the postcritical phase of equilibrium folding events (see Fig. 4*A*). These three stages are defined as the last time before a folding event the total number of native contacts Q is equal to 10, 12, or 14, respectively. Immediately before folding it is more likely that strands 2 and 3 form more contacts than strands 1 and 2. The sequence of events along the most probable pathway can be identified from the histogram: it involves the formation of hairpin 23 first, followed

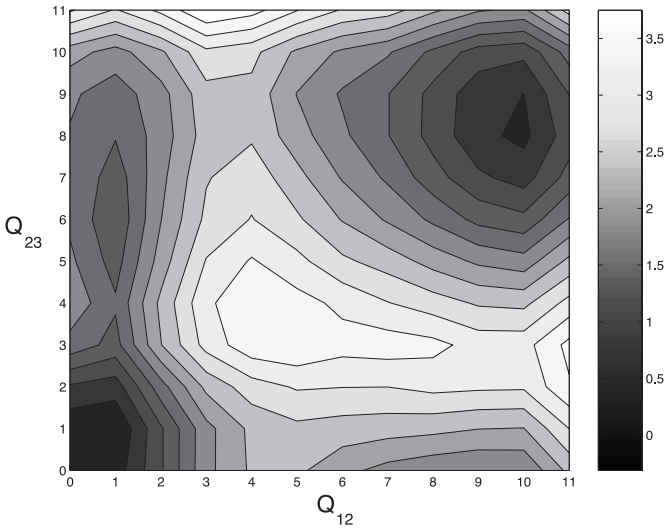

Fig. 2. Free energy as a function of the number of native contacts between strands 1 and 2 (Q_{12}) and strands 2 and 3 (Q_{23}) in units of k_BT. The free energy surface was computed from 12.6-μs equilibrium simulations using a snapshot every 20 ps.

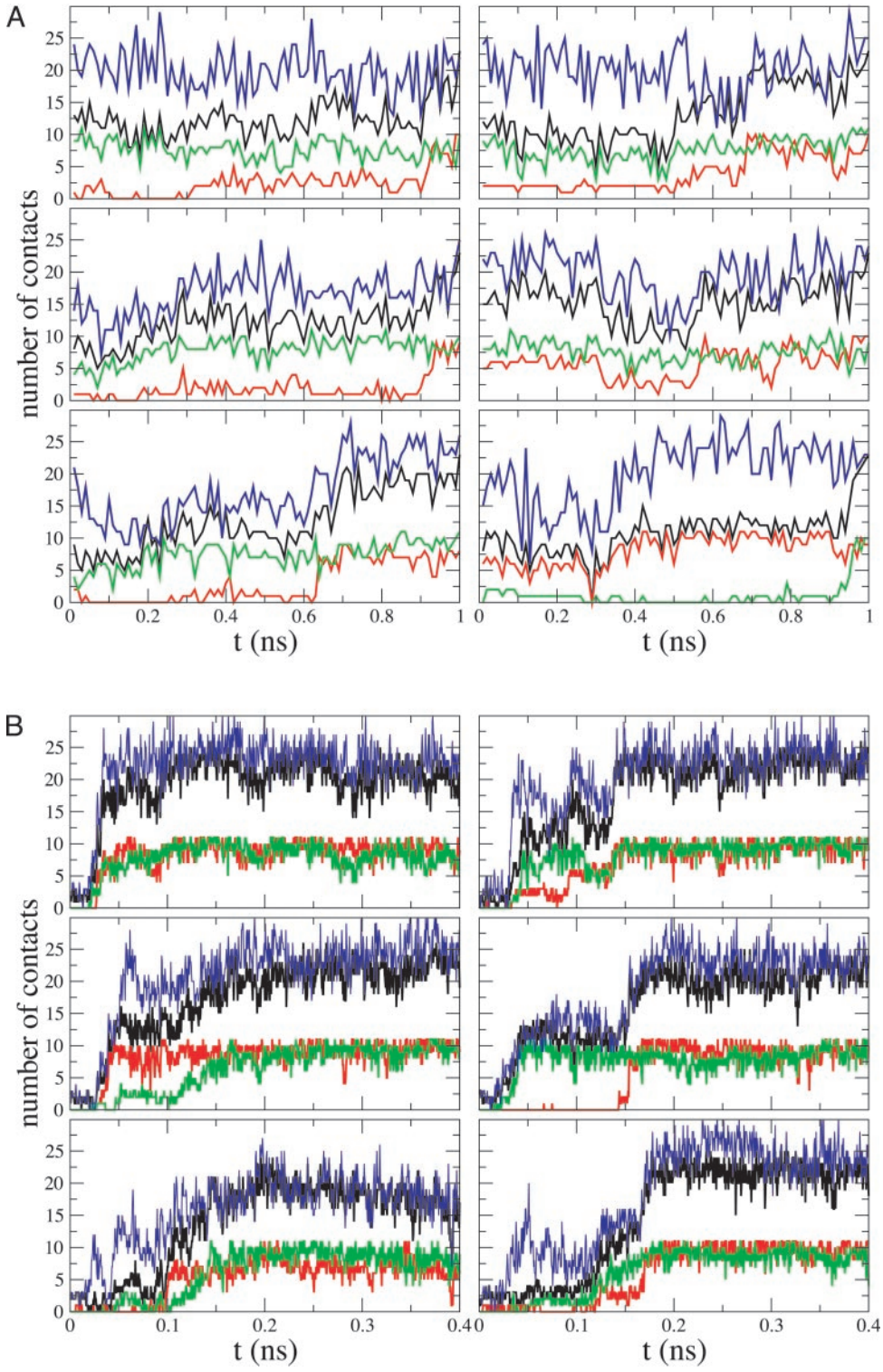

Fig. 3. Total number of contacts (blue line), number of native contacts Q (black line), contacts between strands 1 and 2 (red line), Q_{12}, and strands 2 and 3, Q_{23} (green line). (*A*) Results for the last 1 ns before folding for six randomly chosen folding events at equilibrium. (*B*) Results for the six fastest folding events in the 1-ns simulations.

by the formation of hairpin 12. The simultaneous formation of the two hairpins is unlikely.

Short Simulations. We performed 14,300 simulations of 1 ns, starting from an extended conformation equilibrated at 330 K. Of these simulations, 1,300 were continued up to 6 ns. The trajectories differ by their initial velocities, which are sampled from a Maxwell distribution corresponding to a temperature of 330 K. For a single step reaction, the number of simulations that should fold in a time δt is

$$\delta N = N(1 - e^{-k\delta t}), \qquad [1]$$

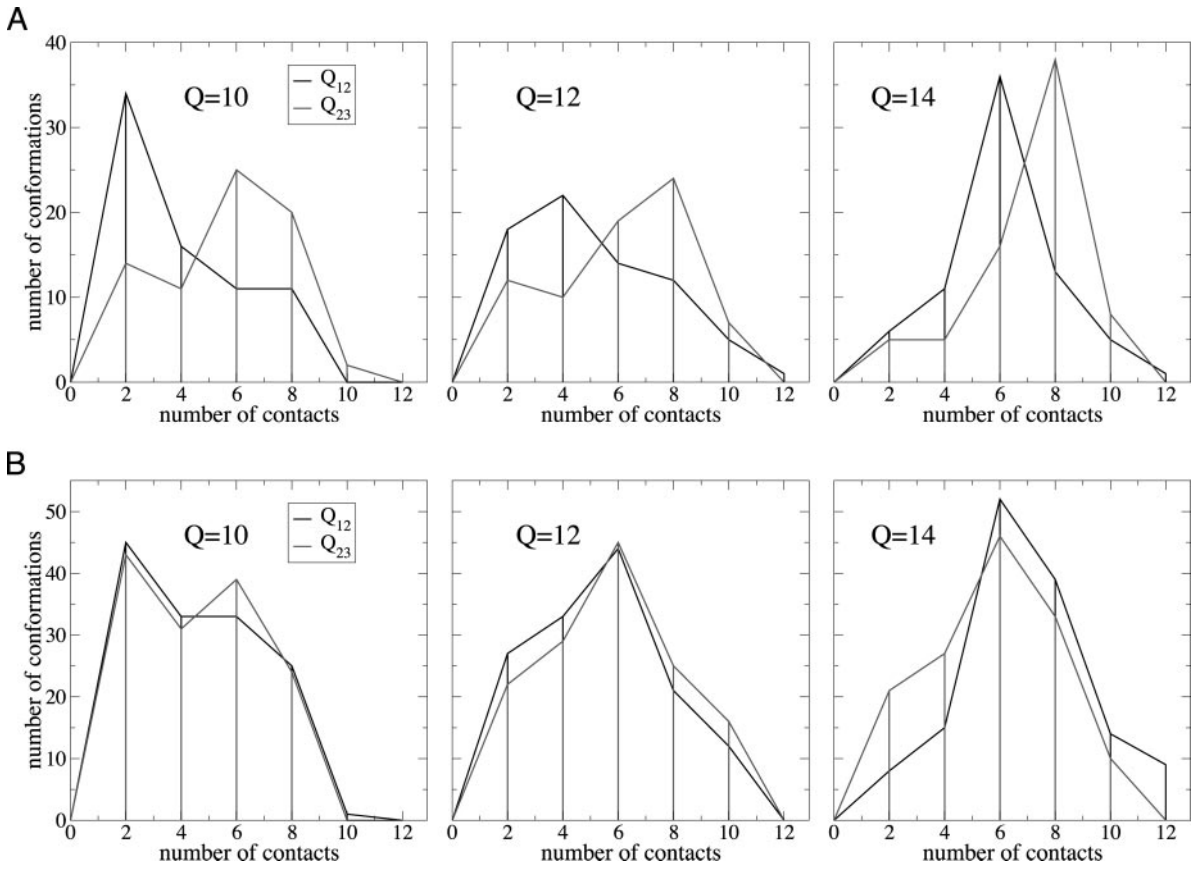

Fig. 4. Histogram of Q_{12} and Q_{23} in the postcritical phase of the folding process, i.e., when Q is, for the last time before the folding event, 10, 12, or 14. (*A*) Results for the 72 folding events observed in the equilibrium simulations. (*B*) Results for the 137 folding events observed in the 14,300 simulations of 1 ns.

where N is the total number of simulations performed. This formula can be solved for the rate k ($= 1/\tau$, where τ is the folding time)

$$k = -\frac{1}{\delta t}\ln\left(1 - \frac{\delta N}{N}\right). \quad [2]$$

The folding time estimated in this way is longer than that (τ_{eq}) estimated from the equilibrium simulations (see Fig. 5) when trajectories less than ≈1 ns are considered. Instead, for trajectories >1 ns the folding time estimated is very close to τ_{eq}.

We analyzed in detail the behavior of the number of contacts in the fastest folding events in Fig. 3*B* as a function of time for the six trajectories that fold in <0.2 ns. These results show that the total number of contacts is essentially identical to the number of native contacts, except in a few cases in which the initial collapse generates unfavorable non-native interactions that disappear quickly. Therefore, chain collapse proceeds by formation of native contacts only.

There is an almost monotonic increase in Q, Q_{12}, and Q_{23}, as opposed to the equilibrium behavior where the system dwells in a series of metastable states and unfolds completely several times before folding (8). The very fast folding pathways can even cross the highest free energy region in the energy surface (see Fig. 2) and are thus unlikely under equilibrium conditions. Thus, for fast folding events in Fig. 3*B* there is no clear pathway preference. The two most common equilibrium pathways, involving the sequential formation of the two hairpins, are sporadically observed in the very short simulations; most frequently the two hairpins are formed almost simultaneously. This is also confirmed by the statistical analysis of all of the fast folding events. In Fig. 4*B* the position of the peak is identical for Q_{12} and Q_{23}, showing that the most likely folding pathway involves the formation of an equal number of contacts between strands 1 and 2 and strands 2 and 3.

To further clarify the way in which the very fast folding pathways are special, we considered the effective energy, defined as the intramolecular energy plus the solvation free energy, as a

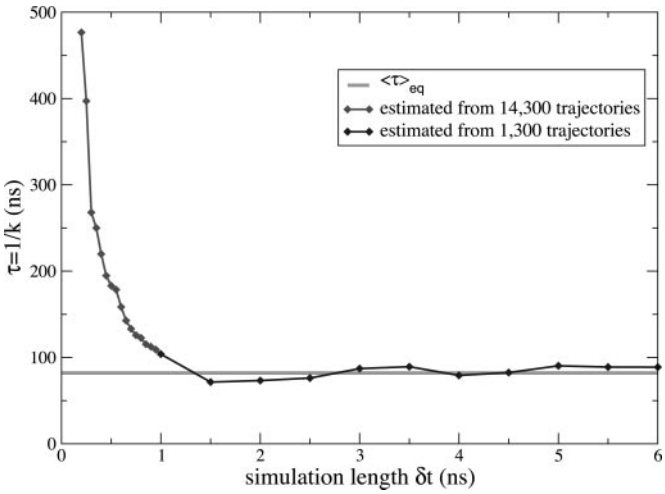

Fig. 5. Folding time (inverse folding rate) estimated from the fraction of folding events within a time δt: $k = -1/\delta t \ln(1 - \delta N/N_0)$. From trajectories >1 ns a folding time close to the equilibrium one is estimated, but shorter trajectories lead to an overestimation of the folding time.

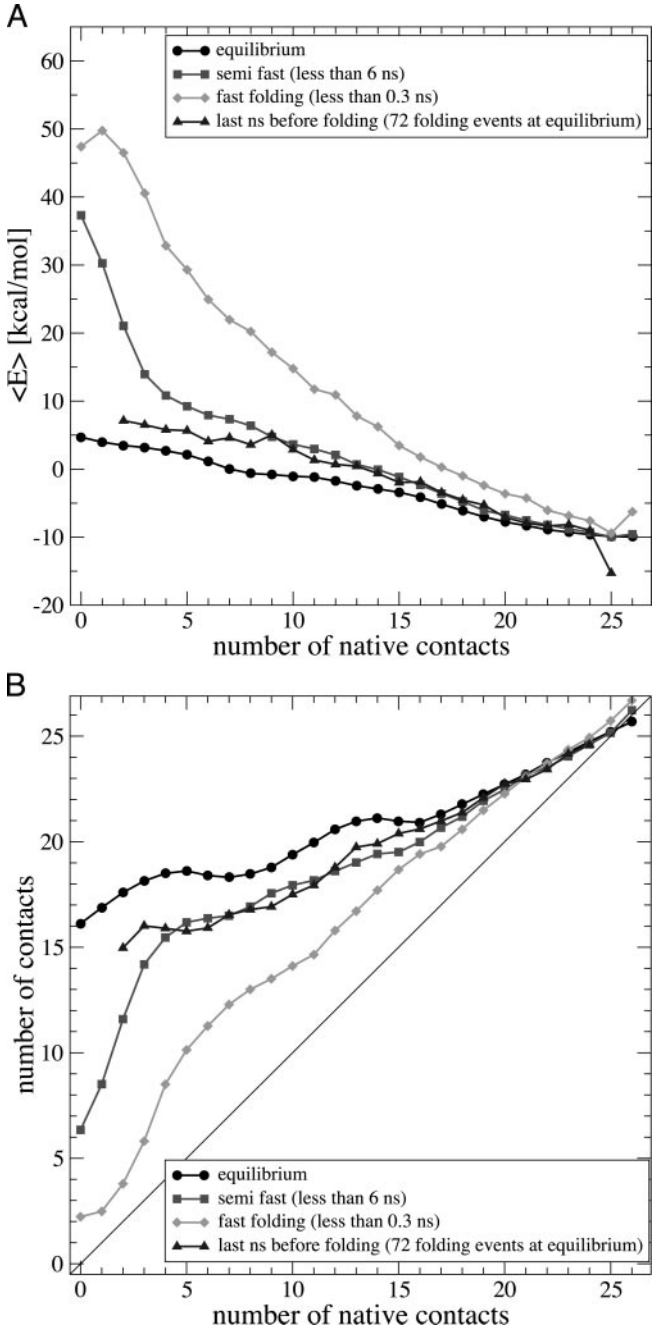

Fig. 6. Total effective energy, including the solvation free energy (*A*) and total number of contacts (*B*) as a function of the number of native contacts *Q*. Four different results are shown: the average over the folding events in the 1-ns simulations (◆), the average over the folding events in the 6-ns simulations (■), the average over the 12.6-μs equilibrium trajectory (●), and the average over the last 1 ns before a folding event in the equilibrium trajectories (▲).

function of the number of native contacts Q. The average effective energy over the fastest folding trajectories is shown in Fig. 6*A*. The average is taken over the 14 trajectories that folded in <0.3 ns, selected among 14,300 trajectories of 1 ns. For comparison, we also plot other effective energy averages: over the 12.6-μs equilibrium trajectory, over the last 1 ns before the 72 folding events, and over the 85 trajectories (of 1,300) where folding occurs within 6 ns. Fig. 6*A* shows clearly that in the fastest folding events the denatured state (corresponding to $Q \leq 10$) does not equilibrate and the energy is significantly higher than in equilibrium simulations. The respective effective energies are comparable when $Q \approx 13$–15. Effective energies are identical (within 0.3 SDs) when the peptide is folded ($Q \geq 23$). The average taken over the 6-ns simulations that reached the folded state shows a difference, especially in the region of low native contacts, with the equilibrium average. The reason for this difference is obtained from the observation of individual folding trajectories. The collapse and equilibration of a compact denatured state in most cases occurs rapidly, on the nanosecond time scale for the system considered here. Averaging only on trajectories that fold faster than 1 ns or even 6 ns, however, overestimates the weight of the fast folding events described above, which are unlikely at equilibrium. Except for low values of Q, the average over the 6-ns folding events is identical to the average result from the postcritical fraction of the equilibrium simulations. Taken together these results suggest that the folding behavior becomes closer to the dominant one at equilibrium for trajectories of increasing length.

The results shown in Fig. 6*B* add interesting information on the peculiarity of fast folding trajectories: the number of "misfolded" contacts, i.e., contacts not present in the native conformation, is very small at each stage of the folding process. Therefore, in the very fast folding events, only native contacts form, a behavior expected from a "Gō model" (15, 16), where only native interactions are favorable and stabilizing. In contrast, under equilibrium conditions, non-native contacts are present and stabilize the denatured state, as observed for denatured and transition states of larger proteins (17).

Conclusions

We have applied the distributed computing approach to study the folding process of a short β peptide. From long (12.6 μs) equilibrium trajectories, where several (72 total) folding-unfolding events were observed, we computed an average folding time of 83 ns. The estimate of the folding rate obtained from short simulations using the distributed computing approach is in good agreement with the equilibrium result if the short simulations are >1 ns (i.e., ≈ 100 times shorter than the equilibrium average folding time). If only very short trajectories are considered (100–500 times shorter than the equilibrium folding time), the estimated folding time turns out to be longer than the equilibrium one.

It has been already observed (18) that the distributed computing can provide an estimate of the folding time only within an order of magnitude, and our results are consistent with that conclusion. In the case of the GS peptide the error in the estimation of the folding rate obtained using only very short simulations originates from the peculiar behavior of the fast folding events that consist of atypical sequences of conformational transitions, not representative of the major folding pathways, such as the formation of only native contacts during the collapse and the nearly simultaneous formation of both hairpins.

The denatured state ensemble comprises structures that cover a broad energetic spectrum (see figure 1 of ref. 3), from low-energy conformations mainly stabilized by non-native interactions, to high-energy conformations, such as extended or partially extended ones with few contacts, either native or non-native. In the folding events that occur very quickly the system does not have time to relax within the denatured state. Therefore these events are not necessarily representative of the major folding pathways (3). In the case of the GS peptide, we found that in the fast folding events, folding occurs directly from the high-energy initial conformation, before the system has time to relax within the denatured state. Folding from these high-energy states involves a very rapid decrease in energy (see Fig. 6*A*) and proceeds through conformations of the transition state

ensemble, which are of high energy and not normally visited at equilibrium.

The use of massively parallel simulations has the potential to provide the large ensemble of trajectories between the unfolded and the folded state that is necessary to understand protein folding as a stochastic process. Folding rates can be estimated correctly, by considering trajectories not <2 orders of magnitude (i.e., of ≈ 1 ns) than the average folding time (83 ns). The pathways observed, however, might not be representative of the equilibrium pathways. As one considers trajectories of increasing length, the behavior becomes closer to the dominant one at equilibrium. To extract a correct description of folding mechanism using the distributed computing approach, we suggest to analyze simulations of increasing length, as a lack of convergence may indicate the presence of atypical folding pathways.

We thank Urs Haberthür and Francesco Rao for maintaining the Beowulf cluster where the simulations have been performed. We thank Armin Widmer for the program WITNOTP, which was used for visual analysis of the trajectories. M.V. is supported by a Royal Society University Research Fellowship. This work was supported in part by the Swiss National Competence Center in Structural Biology and the Swiss National Science Foundation (Grant 31-64968.01 to A. Caflisch).

1. Shirts, M. R. & Pande, V. S. (2001) *Phys. Rev. Lett.* **86,** 4983–4987.
2. Snow, C. D., Nguyen, H., Pande, V. S. & Gruebele, M. (2002) *Nature* **420,** 102–106.
3. Fersht, A. R. (2002) *Proc. Natl. Acad. Sci. USA* **99,** 14122–14125.
4. de Alba, E., Santoro, J., Rico, M. & Jimenez, M. A. (1999) *Protein Sci.* **8,** 854–865.
5. Brooks, B. R., Bruccoleri, R. E., Olafson, B. D., States, D. J., Swaminathan, S. & Karplus, M. (1983) *J. Comp. Chem.* **4,** 187–217.
6. Neria, E., Fischer, S. & Karplus, M. (1996) *J. Chem. Phys.* **105,** 1902–1921.
7. Ferrara, P., Apostolakis, J. & Caflisch, A. (2002) *Proteins* **46,** 24–33.
8. Ferrara, P. & Caflisch, A. (2000) *Proc. Natl. Acad. Sci. USA* **97,** 10780–10785.
9. Hiltpold, A., Ferrara, P., Gsponer, J. & Caflisch, A. (2000) *J. Phys. Chem. B* **104,** 10080–10086.
10. Gsponer, J. & Caflisch, A. (2001) *J. Mol. Biol.* **309,** 285–298.
11. Gsponer, J. & Caflisch, A. (2002) *Proc. Natl. Acad. Sci. USA* **99,** 6719–6724.
12. Ferrara, P., Apostolakis, J. & Caflisch, A. (2000) *J. Phys. Chem. B* **104,** 5000–5010.
13. Cavalli, A., Ferrara, P. & Caflisch, A. (2002) *Proteins* **47,** 305–314.
14. Eaton, W. A., Muñoz, V., Hagen, S. J., Jas, G. S., Lapidus, L. J., Henry, E. R. & Hofrichter, J. (2000) *Annu. Rev. Biophys. Biomol. Struct.* **29,** 327–359.
15. Zhou, Y., Vitkup, D. & Karplus, M. (1999) *J. Mol. Biol.* **285,** 1371–1375.
16. Shimada, J., Kussell, E. L. & Shakhnovich, E. I. (2001) *J. Mol. Biol.* **308,** 79–95.
17. Paci, E., Vendruscolo, M. & Karplus, M. (2002) *Biophys. J.* **83,** 3032–3038.
18. Zagrovic, B., Snow, C. D., Khaliq, S., Shirts, M. R. & Pande, V. S. (2002) *J. Mol. Biol.* **323,** 153–164.