The evolution dynamics of model proteins

Guido Tiana
Department of Physics, University of Milano, via Celoria 16, 20133 Milano, Italy and INFN, Sez. di Milano, via Celoria 16, 20133 Milano, Italy

Nikolay V. Dokholyan
Department of Biochemistry and Biophysics, University of North Carolina at Chapel Hill, School of Medicine, Chapel Hill, North Carolina 27599

Ricardo A. Broglia
Department of Physics, University of Milano, via Celoria 16, 20133 Milano, Italy; INFN, Sez. di Milano, via Celoria 16, 20133 Milano, Italy; and The Niels Bohr Institute, Bledgamvej 17, 2100 Copenhagen, Denmark

Eugene I. Shakhnovich
Department of Chemistry and Biological Chemistry, Harvard University, Cambridge, Massachusetts 02138

(Received 26 February 2004; accepted 12 May 2004)

Explicit simulations of protein evolution, where protein chains are described at a molecular, although simplified, level provide important information to understand the similarities found to exist between known proteins. The results of such simulations suggest that a number of evolutionary-related quantities, such as the distribution of sequence similarity for structurally similar proteins, are controlled by evolutionary kinetics and do not reflect an equilibrium state. An important result for phylogeny is that a subset of the residues of each protein evolve on a much larger time scale than the other residues.
[DOI: 10.1063/1.1768513]

I. INTRODUCTION

The scope of evolutionary biology is quite ambitious: starting from the analysis of fossil records, to reconstruct the whole process which took place over billions of years.[1] New insight towards the fulfillment of this quest has come from the molecular description of the genetic information as well as of proteins (cf., e.g., Refs. 2, 3 and references therein). In particular, the major advantage in focusing on the molecular evolution of proteins is the soundness of the corresponding information. In fact, genomic and proteomic studies are providing large amount of detailed and consistent sets of data. The resulting protein sequences and structures can be compared in a qualitatively unambiguous fashion, and their degree of similarity assessed arguably more easily than in the case of complex organisms.

One of the main problems in studying the evolution of proteins is that their sequence, which is directly linked to the genotype, cannot be mapped in any simple fashion onto their structure, structure which determines the phenotype. This problem (that is the protein folding problem) has been studied in depth for the last few years, leading to important findings,[4–7] findings which can be used at profit to shed light over aspects of protein evolution. The use of simple models to investigate how protein sequences and structures change during evolution has been widely developed in Refs. 8, 9. The results of simulated evolution, analyzed at the light of the whole simulated evolutionary process, may provide new tools to study the history of real proteins and a benchmark to test the variety of hypotheses assumed in such studies. In keeping with these expectation we shall analyze by means of a lattice model of proteins the evolutionary dynamics leading to the creation of structurally similar proteins, as well as the associated sequence similarity. Moreover, special attention is paid to the role of different parts of the protein play in the evolution process.

II. THE MODEL

We represent a protein as a chain of beads on a cubic lattice.[4–7] The beads, representing the amino acids, can be of twenty types and interact through the potential

$$U(\{r_i\},\{\sigma_i\})=\sum_{i<j} B(\sigma_i,\sigma_j)\Delta(|r_i-r_j|), \tag{1}$$

where r_i and σ_i are, respectively, the position and type of the ith amino acid. The contact function $\Delta(|r_i-r_j|)$ takes the value 1 if the ith and jth amino acid are neighbors in the lattice but not consecutive along the chain, and zero otherwise, while two of them are not allowed to occupy the same site. The quantity $B(\sigma_i,\sigma_j)$ is the element of a 20×20 matrix, which sets the interaction energy between each pair of amino acid types. While the precise choice of the matrix is not critical,[10] we use, in what follows the matrix listed in Table VI of Ref. 11. The evolutionary dynamics is implemented by means of a selective pressure focused on the folding ability of proteins. Starting from a given ‘‘ancestor’’ protein, a point mutation is performed at random in its sequence. Three Monte Carlo simulations of the mutated sequence are carried out in conformational space, in order to assess

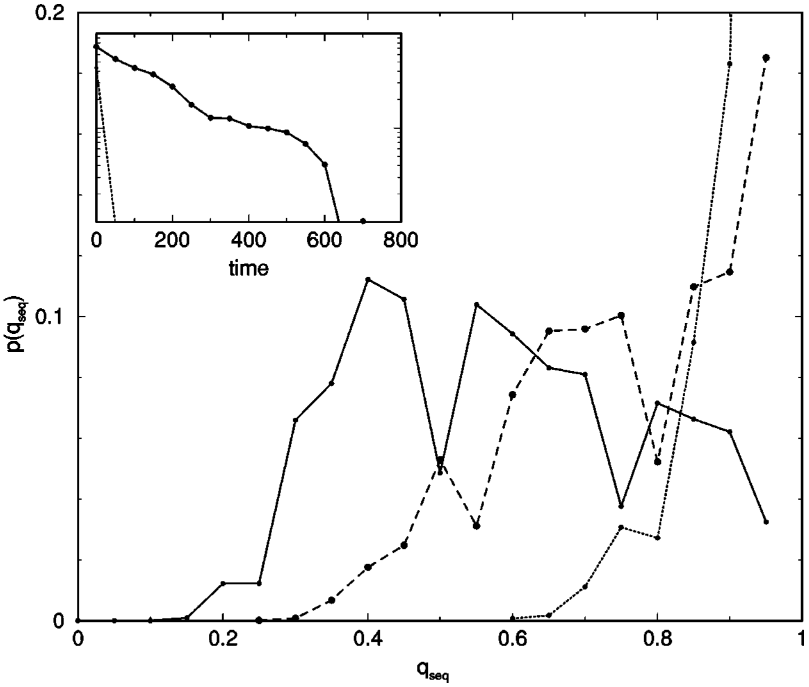

FIG. 1. The stationary distribution of $q_{\rm seq}$ for structurally similar conformations (i.e., $q_{\rm str}>0.7$) obtained comparing each pair of sequences generated by simulated evolutions without stability constraint (dotted curve), with a stability threshold $P_{\rm th}=0.05$ (dashed curve), and with a stability threshold $P_{\rm th}=0.15$ (solid curve). In the inset, the probability distribution for the time distances between similar conformations in linear-log scale.

whether the resulting sequence folds to a unique native conformation and whether this conformation is stable. More precisely, we require that the lowest energy conformations found in the three independent runs display a similarity parameter $q_{\rm str}$ larger than[12] 0.7. We also require the stability P_N of this conformation, defined as the relative amount of time that the protein spends in this unique conformation, to be larger than a threshold value $P_{\rm th}$. If the mutated sequence satisfies such a requirement, the mutation is accepted. This procedure is repeated for 5×10^4 evolutionary steps (ES). The set of (folding) sequences generated through this procedure are collected and their properties analyzed.

The necessity of a simplified model arises due to the high computational cost of testing at each step the folding ability of the mutated protein. It is also for this reason that the simulations are performed at a temperature ($T=0.28$, in the units of the interaction matrix and setting Boltzmann constant equal to 1) which causes the folding to be particularly fast, although the stability of the native state is only marginal. The acceptance rate of point mutations decreases as a function of the stability constraint $P_{\rm th}$. It takes the value 0.32 for simulations performed setting $P_{\rm th}=0$ (no stability constraint) and gets down to 0.11 setting $P_{\rm th}=0.15$. This will be the maximum constraint used for the simulations, again for computational cost reasons.

The model of evolutionary dynamics described above accounts only for the folding properties of proteins, and not for other selective features, such as their function and interaction properties. On the other hand, the ability to fold and to be stable are necessary conditions a protein must satisfy, and consequently provide a basic framework for studying protein evolution.

III. PROPERTIES OF STRUCTURALLY SIMILAR PROTEINS

In order to analyze the results of the simulations, we make use of two parameters $q_{\rm seq}$ and $q_{\rm str}$, which measure the similarity between pairs of sequences and between pairs of native conformations, respectively. They are defined as

$$q_{\rm seq}(\{\sigma_i\},\{\sigma'_i\})=\frac{1}{N}\sum_i \delta(\sigma_i,\sigma'_i), \tag{2}$$

where the function δ assumes the value 1 if the ith site of the protein is occupied by the same kind of amino acid in both the sequences $\{\sigma_i\}$ and $\{\sigma'_i\}$, and zero otherwise, while N is the length of the protein. To simplify the alignment of the sequences, we will consider all proteins of the same length, although a generalization is straightforward and can be treated with standard alignment tools. We also define

$$q_{\rm str}(\{r_i\},\{r'_i\})=\frac{1}{n_c}\sum_{i<j}\Delta(|r_i-r_j|)\Delta(|r'_i-r'_j|), \tag{3}$$

where Δ is the contact function defined above and $n_c=\max[\Sigma_{i<j}\Delta(|r_i-r_j|),\Sigma_{i<j}\Delta(|r'_i-r'_j|)]$ is the numbers of native contacts which, between the two proteins, is largest.

Simulations of evolution starting from a folding sequence (that of Ref. 10) are performed for 5×10^4 ES, using different stability constraints $P_{\rm th}$. The distributions of sequence similarity $q_{\rm seq}$ calculated comparing each pair of structurally similar proteins ($q_{\rm str}>0.7$) generated in 5×10^4 ES at each value of $P_{\rm th}$ are presented in Fig. 1. These distributions are stationary (as shown in the following section) and represent the sequence similarity of the whole ensemble of structurally similar model proteins selected by evolution, that is, current-day proteins (within the framework of the model) and all their ancestors. The distribution associated with the simulation without stability constraint (dotted line) is sharply peaked around $q_{\rm seq}=1$, indicating that in this case proteins displaying similar structure are only those with very similar sequence. The dotted curve in the inset to Fig. 1, showing the corresponding distribution of time distance between structurally similar proteins, also indicate that these are very close in evolutionary time (the distribution drops to zero in 50 steps).

As the stability constraint is made finite, the peak in the distribution of q_{seq} shifts towards lower values. At $P_{\text{th}}=0.15$ it displays a peak around $q_{\text{seq}}=0.4$, associated with proteins displaying low sequence similarity but high structural similarity, and a long tail towards higher degrees q_{seq}. The latter result (i.e., that at $P_{\text{th}}=0.15$) is qualitatively in agreement with the fact, well known to biologists, that, as a rule, there is a large number of dissimilar sequences displaying the same native conformation. In making this comparison, one has to be aware of the fact that, for such short model proteins as the 36mers employed here, the limit of dissimilar sequences is already reached at $q_{\text{seq}}=0.4$ (see below and Ref. 13), due to the low number of degrees of freedom in sequence space, while in the case of longer proteins one usually speaks of dissimilar sequences when q_{seq} ranges between 0.2 and 0.3. From this point of view, the absence of pairs of highly dissimilar sequence ($q_{\text{seq}}<0.2$) in Fig. 1 is likely to be connected with the short length of the chain (note, however, that simulations with longer chains are so far out of question, due to their high computational cost).

The overall picture which emerges from the model results is that pairs of structurally similar model proteins most likely display quite dissimilar sequences ($q_{\text{seq}}=0.4$), less likely display higher sequence identity (up to $q_{\text{seq}}=1$), and seldom display sequence identity below $q_{\text{seq}}=0.4$. Note that this result does not imply that designing a pair of model proteins with the same fold and $q_{\text{seq}}=0.4$ is easier than designing proteins with larger or smaller values of q_{seq}, but only implies that structurally similar proteins *which belong to an evolutionary trajectory* display preferably $q_{\text{seq}}=0.4$. This is because, as will be discussed in more detail later, evolution arises from the competition between two effects, due to the fact that accepted mutations change both the sequence (thus lowering q_{seq}) and the structure (lowering q_{str}) of proteins.[14] Consequently, although in principle it is possible to design pairs of structurally similar proteins displaying any value of q_{seq}, it is not granted that this pair belongs to a stationary solution of evolutionary dynamics.

A straightforward result of the simulations performed at $P_{\text{th}}=0.15$ is that, while dissimilar sequences can give rise either to dissimilar or to similar conformations, similar sequences produce similar structures. In Fig. 2 is displayed the distribution of structural similarity q_{str} for pairs of sequences displaying $q_{\text{seq}}>0.6$, showing that the probability that such pairs correspond to structurally dissimilar proteins is low. This result can be compared with the findings by Sander and Schneider, who have shown that two proteins of length ~30 and sequence similarity larger than 0.5 are necessarily structurally similar[13] (this threshold is length dependent and decreases to 0.2 for alignments larger than 100 residues).

IV. EVOLUTIONARY DYNAMICS

An interesting fact associated with model evolution simulations is that it is possible to study directly the dependence of sequence and structural properties on evolutionary time. In the following we will study a simple model of evolution, where the protein at evolutionary time t is generated by a point mutation from the protein at evolutionary time $t-1$. This is meant to describe a single branch of the evolutionary tree. A more realistic model which allows for branching consists in choosing as parent to the protein at time t any of the proteins already generated. Since this model makes the study of dynamics more complicated, we will now concentrate on the in-line model and leave the branching model to a later work on the description of the statistical properties of evolutionary-generated networks.

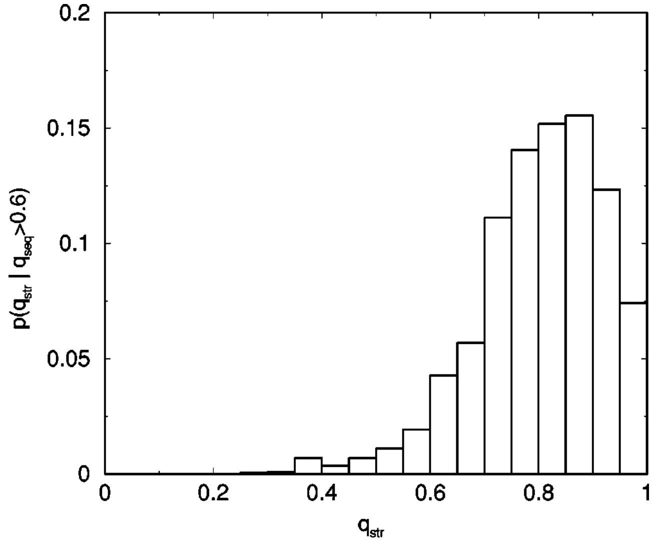

FIG. 2. The distribution of structural similarity q_{str} for proteins generated at $P_{\text{th}}=0.15$ and displaying $q_{\text{seq}}>0.6$.

Useful quantities for the study of the dynamics are the sequence similarity $\langle q_{\text{seq}}(\Delta t)\rangle$ and the structural similarity $\langle q_{\text{str}}(\Delta t)\rangle$ between pairs of proteins separated by an evolutionary time Δt, averaged over all proteins produced in the simulation. These two quantities are displayed in Fig. 3 as a function of the evolutionary time, and for different values of the stability threshold P_{th}. From this figure it can be seen that all the curves decrease monotonically. In the case of $P_{\text{th}}=0$ [Fig. 3(a)] and after 10^3 ES $\langle q_{\text{seq}}(\Delta t)\rangle$ reaches the value ≈0.05 corresponding to pairs of completely uncorrelated sequences, while $\langle q_{\text{str}}(\Delta t)\rangle$ falls to ≈0.10, a situation which describes the case of two structures having little in common except the fact of both being compact.[15] This implies that, after a number of evolutionary steps, proteins lose sequence and structure similarity.

The persistence in evolutionary time of sequence and structural similarity depends on the stability requirement P_{th}. At zero or low values of P_{th}, structural similarity decays faster than sequence similarity, and consequently structurally similar proteins are only those displaying high sequence similarity. Increasing the threshold P_{th}, one observes a crossover [at $P_{\text{th}}\approx 0.08$, cf. Figs. 3(b) and 3(c)] where structural similarity starts decaying slower than sequence similarity. This produces structurally similar proteins with low sequence similarity.

The structural and sequence similarities are well fitted by the sum of two exponentials, in the form

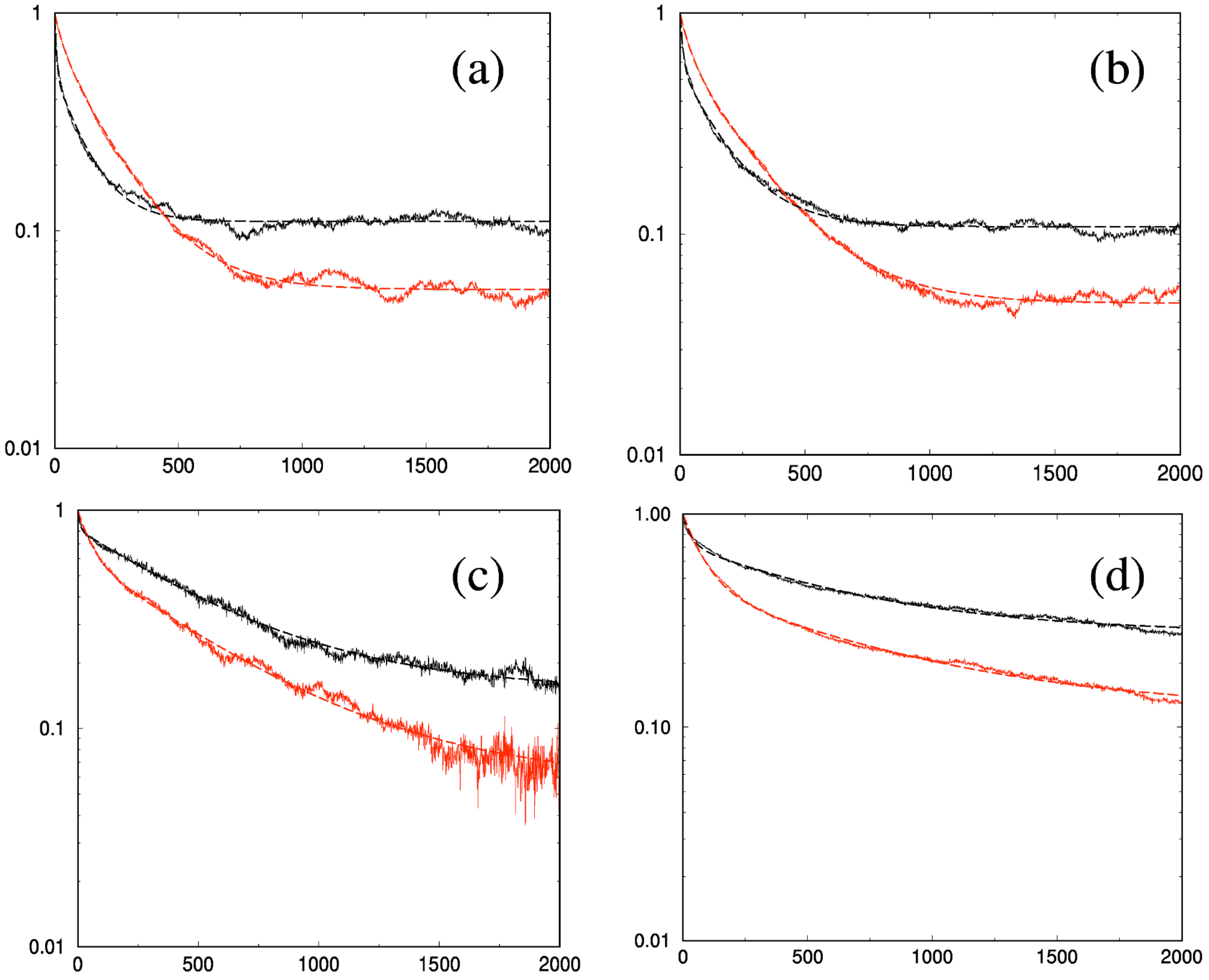

FIG. 3. The dependence of the average quantity $\langle q_{\rm str}(t-t')\rangle$ (black curve) and $\langle q_{\rm seq}(t-t')\rangle$ (gray curve) associated with simulations (a) without stability constraint, (b) with $P_{\rm th}=0.01$, (c) with $P_{\rm th}=0.10$, and (d) with $P_{\rm th}=0.15$. The plots are in linear-log scale. Dashed curves indicate the fit made with a double exponential function (see text).

$$\langle q_{\rm str}(\Delta t)\rangle = a\,\exp(-\Delta t/\tau_a)+b\,\exp(-\Delta t/\tau_b)-a-b+1, \tag{4}$$
$$\langle q_{\rm seq}(\Delta t)\rangle = c\,\exp(-\Delta t/\tau_c)+d\,\exp(-\Delta t/\tau_d)-c-d+1,$$

and the parameters obtained by from fit are listed in Table I. This is consistent with the results of Ref. 16, which highlight two different time scales in the evolution of real proteins (cf. also following section). If we subscribe to this interpretation, we can conclude that structural changes take place at two different time scales, that is τ_a and τ_b. Part of the protein, involving between one fourth and one half of the contacts (cf. 2nd column in Table I), changes in few evolutionary steps (cf. 3rd column in Table I). The rest of the protein changes in times which are consistently longer and increase as the stability constraint is increased (cf. 5th column in Table I). This suggests that only a part of the contacts is involved in the overall stability of the protein, while the rest is essentially free to rearrange.

The dynamics found for $\langle q_{\rm seq}(\Delta t)\rangle$ and $\langle q_{\rm str}(\Delta t)\rangle$ can help in understanding the distribution of $q_{\rm seq}$ for structurally

TABLE I. The parameters associated with the double exponential fits of $\langle q_{\rm str}(\Delta t)\rangle$ and $\langle q_{\rm str}(\Delta t)\rangle$ (see text). The parameters a, b, c, d indicate the ratio of structure and sequence which evolve on the time scale τ_a, τ_b, τ_c, τ_d, respectively.

	$\langle q_{\rm str}(\Delta t)\rangle$				$\langle q_{\rm seq}(\Delta t)\rangle$			
stab	a	τ_a	b	τ_b	c	τ_c	d	τ_d
0	0.490±0.002	7.8±0.1	0.402±0.002	117.9±0.8	0.295±0.003	36.3±0.6	0.653±0.003	195.5±0.8
0.01	0.432±0.001	11.2±0.2	0.454±0.002	163±1	0.332±0.003	38.7±0.7	0.617±0.003	235±1
0.05	0.242±0.001	9.6±0.2	0.655±0.001	385.1±0.9	0.44±0.002	67.1±0.5	0.509±0.001	572±2
0.10	0.189±0.001	9.3±0.5	0.66±0.001	513±4	0.40±0.003	59.6±0.8	0.559±0.002	551±5
0.15	0.283±0.001	24±1	0.436±0.001	639±3	0.532±0.003	80±1	0.343±0.002	713±4

similar proteins (Fig. 1). In fact, it is possible to write the distribution $p(q_{\rm seq})$ associated with an evolutionary simulation using its definition

$$p(q_{\rm seq}) = K\int_0^\infty \delta[q_{\rm seq}(\{\sigma_i(t)\},\sigma_i(t'))\} - q_{\rm seq}]\,\theta[q_{\rm str}(\{r_i(t)\},r_i(t')\}) - 0.7]\,dt\,dt',$$

where $\{\sigma_i(t)\}$ is the sequence of the protein at evolutionary time t and K is a normalization constant. The step function θ constraints the integral to pairs of sequences whose structural similarity $q_{\rm str}$ is above 0.7. One can now calculate $p(q_{\rm seq})$ in a mean field approximation, that is, identifying the variables $q_{\rm str}$ and $q_{\rm seq}$ with their averages $\langle q_{\rm str}(\Delta t)\rangle$ and $\langle q_{\rm seq}(\Delta t)\rangle$, disregarding the fluctuations that these variables display. This approximation presents two advantages: first, it allows to use the general expressions found in Eq. (4), expression which depends on a small number of parameters, instead of the exact functions of $q_{\rm str}$ and $q_{\rm seq}$, which are defined pointlike and do not have an analytical expression. Note that the soundness of such mean field approximation is supported by the low value of the autocorrelation function $\langle q^2_{\rm str}(\Delta t)\rangle - \langle q_{\rm str}(\Delta t)\rangle^2 \approx 10^{-8}$ as obtained from the simulations. Moreover, since $\langle q_{\rm str}(\Delta t)\rangle$ is a decreasing function of Δt, one can estimate that structurally similar proteins are those which are separated by a time interval Δt smaller than $\Delta t_{\rm str}$ such that $\langle q_{\rm str}(\Delta t_{\rm str})\rangle = 0.7$.

A further simplification consists in approximating $\langle q_{\rm seq}(\Delta t)\rangle$ by the single exponential of longest characteristic time τ_d, assuming that the other characteristic time τ_c is so short that it has already reached its asymptotical value at the time scale of interest (cf. Table I). Consequently, one obtains

$$p(q_{\rm seq}) = K\int_0^{\Delta t_{\rm str}} \delta(\langle q_{\rm seq}(\Delta t)\rangle - q_{\rm seq})\,d(\Delta t) \approx \begin{cases} \dfrac{k}{q_{\rm seq}+c+d-1} & \text{if } q_{\rm seq} > \exp[-\Delta t_{\rm str}/\tau_d], \\ 0 \quad \text{otherwise.} \end{cases} \tag{5}$$

Further approximating $\langle q_{\rm str}(\Delta t)\rangle$ as a single exponential, one obtains $\Delta t_{\rm str} = -\tau_b \ln[(1-a-b)/0.7]$. Consequently the interval where $p(q_{\rm seq})$ is larger than zero is

$$q_{\rm seq} > \left(\frac{1-a-b}{0.7}\right)^{\tau_b/\tau_d}.$$

This analytical distribution $p(q_{\rm seq})$ displays a long tail towards $q_{\rm seq}=1$ and a sharp cutoff towards low values of $q_{\rm seq}$. Using the parameters listed in Table I one obtains, in the case of $p_{\rm th}=0.15$, a cutoff at $q_{\rm seq}=0.45$, consistent with the distribution obtained from the simulations and displayed in Fig. 1 (solid line). The main difference between the analytical and the computational distributions is that in the latter the cutoff is blurred due to the fluctuations that we have neglected in the mean-field treatment.

The distribution of $q_{\rm seq}$ for structurally similar natural proteins has been analyzed by Rost in the case of four entire genomes,[17] the corresponding results being displayed in Fig. 4. These distributions are more strongly peaked than that predicted by Eq. (5). The difference between the present theory and experimental data could be associated with the assumption done of an in-line evolution, not allowing for branching. In other words, at each evolutionary step we retain only the last protein, neglecting its past history and picturing evolution as a one-dimensional process. A more realistic model would take into account the fact that several different protein can fixate during evolution and simultaneously produce offsprings. If this is the case, the most appropriate topology for evolving protein is not a chain but a branching tree.

Since we are describing the set of evolving proteins as a dynamical process, it is necessary to compare all pairs of proteins, and not just the end-points of the tree. In order to make this calculation easier, avoiding border effects, we choose a Bethe lattice topology with coordination number k (i.e., a lattice where each site is connected to k neighbors and there are no loops). In such a topology, each protein is connected through Δt evolutionary steps to other $k^{\Delta t}$ proteins, picturing in this way the bulk of a branching tree. Repeating the same calculation as for Eq. (5) one obtains

$$p(q_{\rm seq}) = K\int_0^{\Delta t_{\rm str}} k^{\Delta t/T}\,\delta(\langle q_{\rm seq}(\Delta t)\rangle - q_{\rm seq})\,d(\Delta t),$$

where T is the unit used to measure time distances on the lattice. The solution of this equation decays as

$$(q_{\rm seq}+c+d-1)^{-\tau_d/T\ln k-1}, \tag{6}$$

providing a distribution which is more peaked at small values of $q_{\rm seq}$ than the distribution associated with Eq. (5). The experimental data are well described by Eq. (6) provided that k is of the order of 10 (cf. Fig. 4, where the values of k are found by mean of a least square optimization).

The genetic data, at the light of the present model, testify for a tree topology associated with a structure evolution very rich in branches. Of course the value of k we obtain has to be regarded as a mean value, and little can be said about its fluctuations just from the plots displayed in Fig. 4.

Another difference between the present model and genetic data is that, while in the former all pairs of sequences generated in the evolutionary process are compared, in the latter one considers only the currently living offsprings. In other words, the present model does not implement an extinction mechanism which eliminates individuals as a consequence of some kind of natural selection. The reason for this choice is that such a mechanism would require a characterization of model proteins which goes beyond the mere folding properties and selection rules which hardly would be under control, while a quality of the present model is to be simple. Thus, the approximation we have employed is that any protein has the same probability of surviving in evolution, provided that it folds to a stable conformation. What we have shown is that it is possible to explain the shape of experimental curves on the base of folding properties only, providing a good starting point to build more complicated theories.

Note also that the rather high values of k found for genetic data could again reflect the heterogeneity of fitness of

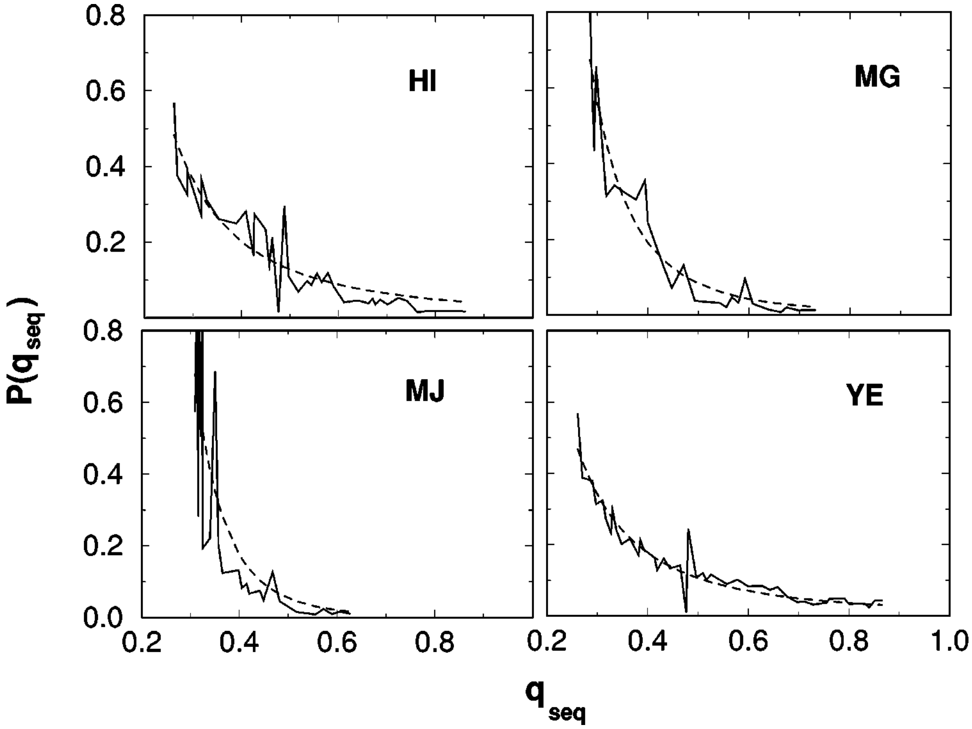

FIG. 4. The distribution of sequence similarity $p(q_{\mathrm{seq}})$ obtained by Rost Ref. 6 from the analysis of structurally similar proteins of four entire genomes (Haemophilus Influenzae, Mycoplasma Genitalis, Methanococcus Jannaschii, Yeast). The dashed curve indicates the prediction of Eq. (6), using the values $k=2.7$, $k=14$, $k=66$, $k=3.6$, respectively.

different proteins associated with features which are not considered in the present model. In other words, evolution could have selected correlated proteins because of their ability to carry out some biological task, because of their reluctance to aggregate, etc., giving rise, in such a way, to large values of k.

V. EVOLUTION OF PROTEIN CORE

In the following it will be shown that stable proteins display a kernel of strongly interacting sites, localized in each protein conformation, which have a persistence time larger than that of both sequence and structure similarity. In order to locate those sites which mostly attract their neighbors, and consequently are likely to make major contribution to the stability of the protein, we define for each site i a parameter

$$h(i)=\frac{\overline{\epsilon_i}-\overline{\epsilon}_{\mathrm{min}}}{\overline{\epsilon}_{\mathrm{max}}-\overline{\epsilon}_{\mathrm{min}}}, \tag{7}$$

where $\overline{\epsilon_i}\equiv\overline{\Sigma_j B(\sigma_i,\sigma_j)\Delta(|r_i-r_j|)}$ is the total interaction of an amino acid with its neighbors, $\overline{\epsilon}_{\mathrm{min}}$ and $\overline{\epsilon}_{\mathrm{max}}$ are the minimum and the maximum of this quantity in a given protein. The distribution of $h(i)$ for all sites of all proteins in the simulated evolution displays a bimodal shape whose minimum is 0.5, indicating a well-defined cluster of stabilizing sites. Due to their energetic properties, the sites displaying $h(i)>0.5$ are likely to be key elements in determining the stability of the protein. In comparing the results of the present model with experimental findings, one could make the ansatz that these sites are the "core" of the protein, that is, the subset of the residues that maintains the backbone of the structure at temperatures close to the folding transition temperature.[18,19] In the case of small globular proteins the core is often, but not always, identical to the folding nucleus. This is the case for our starting protein.[20] The ansatz is supported by physical intuition and by detailed analysis of few cases,[21,22] out of course the rigorous proof can only come from the analysis of all the proteins which build out an evolutionary trajectory, which is computationally too demanding. Note also that, although the position of the core is necessarily correlated with the number of contacts of its residues, using a complicated interaction matrix as that of Ref. 11 there is not a trivial partition of the core between bulk and surface sites.[23]

The similarity in the energetic pattern defining the core of a pair of proteins α and β can be defined as

$$q_{\mathrm{core}}^{\alpha\beta}\equiv\frac{1}{N}\sum_i \theta(h_i^{\alpha}-0.5)\,\theta(h_i^{\beta}-0.5), \tag{8}$$

where θ is Heavside's step function and N is the length of the protein. In Fig. 5 is displayed the decay of $\langle q_{\mathrm{core}}(\Delta t)\rangle$ obtained from the simulation at $p_{\mathrm{th}}=0.15$, similarly to what done for the sequence and for the similarity in the pattern of contacts. A good fit for this function is a double exponential

$$\langle q_{\mathrm{core}}(\Delta t)\rangle=0.02\exp\left(-\frac{\Delta t}{56}\right)+0.15\exp\left(-\frac{\Delta t}{2297}\right)+0.05. \tag{9}$$

The second term of this function indicates that, in average, 15% of the residues of a protein build a strong energetic pattern which stays in place for more than 2000 ES, approxi-

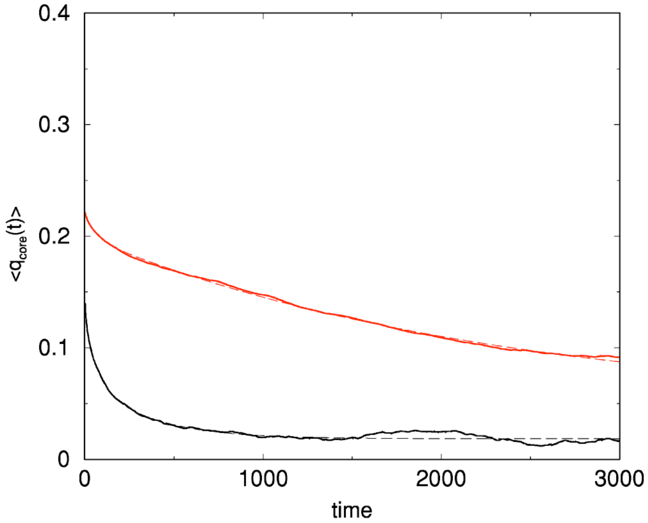

FIG. 5. The decay of the similarity $\langle q_{\mathrm{core}}(\Delta t)\rangle$ in the pattern of strongly interacting sites (solid curve) and the associated double-exponential fit (dashed curve). The black curve refers to the simulation without stability constraint, while the grey curve to that with constraint equal to 0.15.

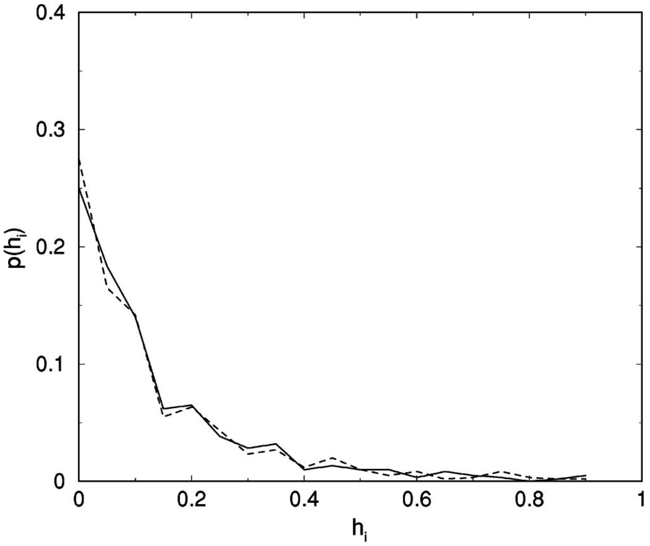

FIG. 6. The distribution of the parameter $h(i)$ for the sites where mutations were tried and accepted. The solid curve is associated with the value of h in the parent conformation and the dashed curve in the offspring conformation.

mately three times the persistence time of the whole sequence and of the structure of the protein. The result contained in Eq. (9) thus indicates that the position of the core survives not only to the variation in protein sequence, but also to the structure, so that different folds can conserve a similar position of the core (but not necessarily the same types of amino acids). The first term of Eq. (9) describes a fast process which takes place at the same time scale of the variation of the typical amino acid (cf. Table I). It involves a tiny amount of residues and is likely to be associated with the approximation done in locating the core by mean of Eq. (7).

The distribution of values of $h(i)$ associated with the sites where mutations are accepted is displayed in Fig. 6. Accepted mutations are mostly localized in sites not belonging to the core. Only 3% of accepted mutations are localized in the core and they are usually mutations which substitute an amino acid with another displaying similar energetic properties (e.g., E with D, K with R). It is difficult to assess any correlation between the value of $h(i)$ in the parent and in the offspring conformation (i.e., before and after the mutation).

We have repeated this calculation with a random interaction matrix instead of that of Table VI of Ref. 11, obtaining similar results, but with a negligible amount of accepted mutations in the core, as expected since in this case there are not amino acids displaying similar energetic properties. These results testify to the fact that our findings are independent on the particular realization of the interaction matrix.

In order to test the role of the core in the evolution process, we have repeated the simulation of evolution allowing mutations only in sites with $h(i)>0.5$. The result is that the protein fold remains fixed and equal to the initial one ($\langle q_{\mathrm{str}}(\Delta t)\rangle=1$). On the other hand, if we allow mutations only on those sites which display $h(i)<0.5$ we obtain a $\langle q_{\mathrm{str}}(\Delta t)\rangle$ very similar to that of the unconstrained evolution ($\tau_b=571$).

Note that stable proteins never display large structural changes associated with single mutations, almost all steps of the $P_{\mathrm{th}}=0.15$ simulation displaying $q_{\mathrm{str}}>0.7$. Only in four cases (out of 5×10^4 ES) the structure of the protein changed by a consistent amount ($q_{\mathrm{str}}=0.5$). This fact is mainly associated with the stability constraint. In fact, decreasing to zero the value of p_{th}, the distribution of q_{str} associated with point mutations moves from a peak sharply centred at $q_{\mathrm{str}}=1$ to a broad distribution which reaches $q_{\mathrm{str}}=0.2$.

These data suggest a picture for the evolution where the position of the core is conserved for a longer period [e.g., 2297 ES in Eq. (9)] than the sequence and even the conformation, in such a way that proteins displaying different folds can have the core in the same position. The position of the core could thus be used to assess the evolutionary relationship between proteins whose sequence and structural similarity is low. Snapshots of the evolution of the core are displayed in Fig. 7, the core being the cluster of residues colored in dark grey. Moreover, in the present model evolution seems not to proceed by mutating residues belonging to the core, which anyway would destabilize to a large extent the protein. Even mutations of wild type residues with residues which are chemically similar seem to be tolerated, but do not drive evolution, as testified by the fact that preventing substitutions in the core does not affect evolution in any way. On the contrary, sequence mutations in sites outside the core produce small structural changes. The accumulation of such changes allows the protein to change fold. The position and the kind of amino acids which build out the core changes as consequence of the evolution of the surrounding residues, some of them being stabilized and absorbed into the core, while some residues of the core being downgraded by unfavorable interaction with the surrounding amino acids. In other words, the only mechanism which can drive structural evolution within the framework of the present model is that the shell of residues surrounding the core change it, by inclusion and erosion. This effect is displayed in Fig. 8, which

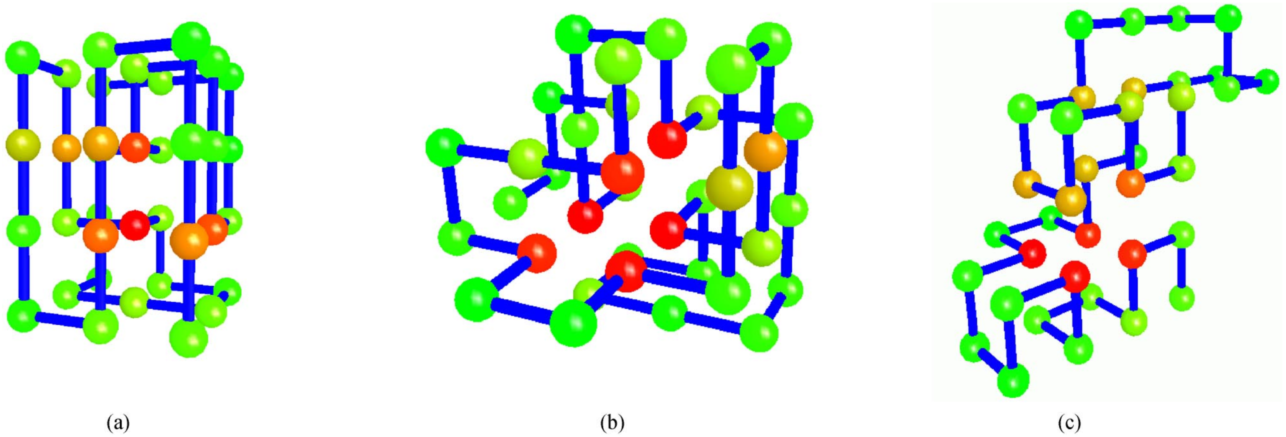

FIG. 7. Three snapshots of the evolutionary dynamics at $P_{\mathrm{th}}=0.15$. (a) The initial structure, where the most interacting residues are 2, 6, 7, 16, 27, and 31 (darker beads). (b) Structure after 1000 ES, displaying $q_{\mathrm{seq}}=0.12$ and $q_{\mathrm{str}}=0.14$ to the initial structure, but the core is the same (the only difference is the relative stabilization of residue 9 and the destabilization of residue 7). (c) The structure after 4525 ES, displaying a different core (residues 3, 8, 13, 16, 31, 34).

shows how the energy parameter per site $h(i)$ changes with evolutionary time. The darker columns indicate the core sites, where the stabilization energy is concentrated and which are quite persistent. Sometimes a darker column is interrupted and a new one starts, indicating that the core is moving [cf., e.g., the switch between sites 2 and 3 after ~3800 ES, corresponding to the switch between structure (b) and (c) in Fig. 7]. Comparing the results displayed in Fig. 8 with the associated native conformations, it emerges that the core remains a compact cluster of spatially localized residues (although not localized along the sequence) throughout the whole evolutionary dynamics. Another mechanism which could change residues in the core, but is not allowed in the present model, is by multiple mutations.[18] Correlated mutations could move some of the most stabilizing residues *in toto*, making the protein effectively cross the energy barrier in sequence space which separates two stable proteins. One can make a rough estimate of the time scale over which this mechanism can move the core: consider that it is composed of 1/4 of the amino acids of the proteins, that one can change it with four coordinated mutations and that the probability that each mutation is productive is 1/10. The resulting time scale would be $40^4 \approx 10^6$ ES, to be compared with the time scale 10^3 ES associated with the inclusion-erosion mechanism. That of multiple mutations would be definitely a rarer strategy, but not negligible over evolutionary times, proteins adopt to move the core, and consequently need a more careful investigation.

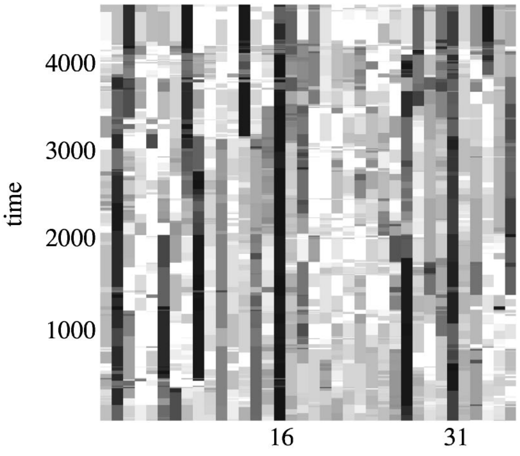

FIG. 8. The evolution of the parameter $h(i)$ which defines the position of the core for the simulation of evolution performed at $p_{\mathrm{th}}=0.15$.

VI. DISCUSSION AND CONCLUSIONS

We have developed a model for protein evolution which describes explicitly, although in a simplified way, the molecular structure of proteins. The model is based on a selective pressure focused on the folding properties of the evolving proteins. In spite of its simplicity, it reproduces a number of features observed in the analysis of real proteins, such as the distribution of sequence similarity for pairs of structurally similar chains, the presence of many structurally similar proteins displaying low sequence similarity, and, as will be described in a another paper, the evolution of protein designability[24] and the topology of the structural network.

From the analysis of the simulated evolutions it appears that the macroscopic features associated with the system (e.g., distribution of sequence similarity, etc.) do not reflect thermodynamics equilibrium in the space of proteins, but only a stationary state of evolution. In other words, they do not depend on time because of a dynamic balance between different factors, not because the system has sampled a consistent part of protein space. Using physical language, probability does not depend on time, but the probability flux is not zero. The first consequence of this fact is that all the time-independent quantities associated with evolution depend on the time constants (i.e., τ_a, τ_b, etc.) of the system and, possibly, on the initial conditions (i.e., the progenitor protein), something that would not happen if equilibrium were reached. This conclusion disagrees with that of Rost in Ref. 17 and with the hypothesis that many more protein sequences than folds are known because different folds display at equilibrium different weights due to different designability.[25] Designability is indeed important for evolution, but due to evolutionary dynamics bias.[22]

Furthermore, we observe that simulated evolution produces only divergent pathways. The average structural similarity $\langle q_{\mathrm{str}}(\Delta t)\rangle$ is a decreasing function of time and fluctuations are very small. Thus, *in average*, proteins becomes more structural dissimilar as time goes by and, within the assumptions of the model, never converge again. Although it is not possible to exclude that in billions of years some fluctuation has created pairs of structurally similar proteins beyond their structural correlation time (τ_b), this cannot be the typical, average mechanism. The present simulations are then not compatible with convergent evolution: structurally similar proteins are always homologous. It is certainly possible that convergent evolution is a consequence of evolutionary features which are not implemented here, such as functional requirements. Nonetheless, an important conclusion of this simplified model is that it is not necessary to introduce convergent evolution to account for the experimental data.

The presence of a highly conserved core of strongly interacting contacts seems to be an important feature of model evolution. This in an important result, because from the point of view of the Hamiltonian of the model, no site of the protein is advantaged with respect to the others, and consequently the core we observe is an emergent property of the system. Its position has a persistence in time which is longer than those of sequence and structure, indicating that studying its evolution (locating it by mean of static energy calculations performed, for example, with classical energy functions) could be a precious tool to compare proteins which evolved beyond their sequence correlation time. These results agree and complete those described in Ref. 19 about the conservation in protein families of residues important for stability and kinetics.

Since the core is responsible for the stability of the protein, its evolution is a key issue. Essentially, it cannot evolve on its own, since mutations in core sites cause destabilization of the whole protein. It can evolve because the surrounding residues can join the core or turn a core site into a normal site. In this perspective, both changes in the core and in the whole structure of the protein are collections of small steps, the system never displaying large conformational changes.

[1] J. W. Schopf, Science **260**, 640 (1993).
[2] M. V. Volkenstein, *Physical Approach to Biological Evolution* (Springer, Berlin, 1994).
[3] L. Cavalli-Sforza, *The History and Geography of Human Genes* (Princeton University Press, Princeton, NJ, 1994).
[4] E. Shakhnovich and A. M. Gutin, J. Chem. Phys. **93**, 5967 (1989).
[5] K. F. Lau and K. Dill, Macromolecules **22**, 3986 (1989).
[6] G. Tiana, R. A. Broglia, H. E. Roman, E. Vigezzi, and E. I. Shakhnovich, J. Chem. Phys. **108**, 757 (1998).
[7] R. A. Broglia and G. Tiana, Proteins **45**, 421 (2001).
[8] D. M. Taverna and R. A. Goldstein, Biopolymers **53**, 1 (2000).
[9] S. Govindarajan and R. A. Goldstein, Proteins **29**, 461 (1997).
[10] E. I. Shakhnovich and A. M. Gutin, Biophys. Chem. **34**, 187 (1989).
[11] S. Miyazawa and R. Jernigan, Macromolecules **18**, 534 (1985).
[12] The value 0.7 corresponds to the average value of q_{str} associated with the minimum of the distribution $p(q_{\mathrm{str}})$ which separates the nativelike to the unfolded states in simulations of 36mer folding sequences at the folding temperature.
[13] C. Sander and R. Schneider, Proteins **9**, 56 (1991).
[14] A first suggestion that this is the case comes from the distribution of time distances associated with pairs of structurally similar proteins, which display an exponential shape, indicating a Poisson process with characteristic time of 266 ES (cf. inset of Fig. 1).
[15] From Monte Carlo simulations at infinite temperature it is found that two uncorrelated conformations of 36mers display a value of $q_{\mathrm{str}}=0.03$. If we impose to the conformations to be compact, through the simulation of homopolymers with contact interactions of energy $-kT$, we find $q_{\mathrm{str}}=0.10$.
[16] N. V. Dokholyan and E. I. Shakhnovich, J. Mol. Biol. **312**, 289 (2001).
[17] B. Rost, Folding Des. **2**, S19 (1997).
[18] N. V. Dokholyan, S. V. Buldyrev, H. E. Stanley, and E. I. Shakhnovich, Folding Des. **3**, 577 (1998).
[19] L. A. Mirny and E. I. Shakhnovich, J. Mol. Biol. **291**, 177 (1999).
[20] V. I. Abkevich, A. M. Gutin, and E. I. Shakhnovich, Biochemistry **33**, 10026 (1994).
[21] G. Tiana and R. A. Broglia, J. Chem. Phys. **114**, 2503 (2001).
[22] R. A. Broglia and G. Tiana, J. Chem. Phys. **114**, 7267 (2001).
[23] M. Skorobogatiy and G. Tiana, Phys. Rev. E **58**, 3572 (1998).
[24] G. Tiana, B. Shakhnovich, N. Dokholyan, and E. I. Shakhnovich, Proc. Natl. Acad. Sci. U.S.A. **101**, 284 (2004).
[25] A. V. Finkelstein, A. M. Gutin, and A. Ya. Badretdinov, FEBS Lett. **325**, 23 (1993).