J. Math. Biol. 44, 341–359 (2002)
Digital Object Identifier (DOI):
10.1007/s002850100133

Mathematical Biology

Mario Abundo[1,2,*] · Luigi Accardi[1,2] · Nicola Rosato[1,3,5] · Lorenzo Stella[4]

Analysing protein energy data by a stochastic model for cooperative interactions: comparison and characterization of cooperativity

Received: 2 October 2000 / Revised version: 4 April 2001 /
Published online: 14 March 2002 – © Springer-Verlag 2002

Abstract. In the frame of a Markov chain model for cooperative interactions in proteins, previously introduced by us, we deal here with estimation of unknown parameters from protein energy data. One of these parameters characterizes the cooperativity of a protein; we propose to measure it also by the so-called approximate entropy. By our computations the approximate entropy turns out to be a decreasing function of the cooperativity. We analyse both simulated data of the Markov chain, and protein energy data obtained by molecular dynamics simulation. Moreover, we compare two rubredoxin proteins at different temperatures, according to their degrees of cooperativity.

1. Introduction

Understanding the dynamic behaviour of protein molecules is one of the main goals of contemporary biophysics. Structural fluctuations of the peptidic chain are involved in many processes of great biological relevance, like protein folding, ligand binding and equilibrium dynamics [19]. The principal determinant of all these phenomena is the cooperativity of intramolecular interactions: mainly for entropic reasons, the formation of bonding interactions favours the formation of additional bonds. In the protein folding process, this behaviour results in a very fast transition from the unfolded to the native conformation, that in many cases are the only significantly populated states [11]. The cooperativity of the interactions is apparent also in equilibrium fluctuations of the protein conformation, that are usually dominated by large scale correlated motions [6]. These properties are shared by the great majority of proteins, irrespective of the variety of their tertiary structure; therefore, it should be possible to define a very general model of the dynamic behaviour of protein interactions that is independent from the specific protein or its structural

* *Corresponding author:* Dipartimento di Matematica, Università "Tor Vergata", via della Ricerca Scientifica, 00133 Roma, Italy. e-mail: abundo@mat.uniroma2.it

[1] Centro Vito Volterra, [2] Dipartimento di Matematica, [3] Department of Experimental Medicine and Biochemical Science, and INFM, [4] Dipartimento di Scienze e Tecnologie Chimiche, Università "Tor Vergata", Roma, Italy, [5] AFaR - Dipartimento di Neuroscienze, Ospedale Fatebenefratelli, Roma, Italy

Key words or phrases: Cooperative interactions – Markov chain – Parameter estimation – Simulation

features. Many simplified models of protein folding have been proposed, but they all maintain some description of the protein structure [7], [11], [13], [14]. In [1] we have introduced a phenomenological model for cooperative behaviour in proteins that is based only on the dynamics of intramolecular interactions, without any reference to the specific protein conformation. The basic model [1], that has been extended in [2] and [3], is described by a homogeneous Markov chain (MC), which depends only on two parameters: the average probability p to form an interaction (that is physically related to the mean free energy for the formation of a bond), and the maximum increase in this probability that can be caused by the previous formation of other bonds, Δp. This last parameter describes the coupling capacity and therefore it consitutes a measure of the degree of cooperativity. The main quality of this model is that it summarizes the cooperative behaviour of a given protein in this single parameter; Δp can be estimated from data regarding real proteins and it can be used as a simple way to compare molecules with different cooperativity.

In general, our model can describe both the transition of a protein between two different states (e.g. the folding/unfolding process, caused by a variation in temperature) and the conformational fluctuations at equilibrium: the two phenomena are associated to the non stationary and to the stationary behaviour of the MC, respectively. Previous papers were focused to the first application [1], [2], [3], while here we will use the model to study protein dynamics in the steady-state. Our first aim is to describe and compare several ways to estimate model parameters from the time evolution of an observable quantity (e.g. the system potential energy). After testing the various estimation methods on simulated data (Section 2), we will estimate Δp for two real proteins (under two different temperature conditions), as an example of practical application (Section 4); we will show that the degree of cooperativity can be related to the protein functional status.

Since the degree of cooperativity influences the time evolution of the protein system, in principle it should be correlated with other parameters that quantify the complexity of the protein dynamics. Recently, Pincus [16] has introduced the so-called approximate entropy (ApEn), that determines the degree of randomness in time series (or simply in sequences of numbers). Here we show that ApEn calculated on the time evolution of the interactions determined by our model is a decreasing function of the degree of cooperativity of a protein, and therefore it can also be used as a useful parameter to describe the cooperative behaviour of proteins (Section 3).

Now, we briefly recall from [1] the model. Let $(X_k)_{k\in\mathbf{N}}$ be a homogeneous Markov chain with state space $S = \{0, 1, \ldots, N\}$ and transition probabilities

$$p_{ij} \doteq Pr\{X_{k+1} = j | X_k = i\} = \binom{N}{j} p_i^j (1 - p_i)^{N-j}, \tag{1.1}$$

where $p_i \doteq p - \Delta p + 2\Delta p \frac{i}{N}$.

The state of the chain at time k represents the number of existing bonds (or interactions) at the discrete instant k, among amino acidic residues, during the protein folding. N represents the maximum number of pairings (bondings) which are allowed.

The parameters p and Δp are supposed to satisfy the obvious constraint

$$0 \leq p \pm \Delta p \leq 1 \tag{1.2}$$

in order that all probabilities (1.1) are indeed numbers between 0 and 1.

For $p = \Delta p = 1/2$ the model coincides with the well-known Fisher-Wright model in population genetics ([10]). When $p > \Delta p > 0$ and $p + \Delta p < 1$, the MC is irreducible because $p_{ij} > 0 \forall i, j \in S$ (see [1]); so there exists a unique stationary, invariant distribution $\{\pi_j\}$, $j \in S$ such that

$$\pi_j = \lim_{n\to\infty} p_{ij}^{(n)} \ \ \forall i, j \in S, \tag{1.3}$$

where $p_{ij}^{(n)}$ is the probability that the system goes from the state i to the state j in n steps. In [1] the qualitative behaviour of the stationary distribution was studied as a function of the parameters p and Δp; in fact the stationary probabilities were numerically found, i.e. the π_j were exactly computed by means of a computer, by using a software for finding eigenvalues of large-dimension matrices. Indeed, the probabilities π_j are the components of the left eigenvector relative to the eigenvalue 1, of the transition probability matrix given by (1.1) (see e.g. [1] or (2.4) of the next section).

Although an explicit theoretical formula for the stationary probabilities π_j cannot be found, various approximations of π_j, in the limit $N \to \infty$ can be obtained. One rather crude asymptotic approximation for N large and Δp far from $1/2$ is (see [1])

$$\pi_j \sim \binom{N}{j} \left(\frac{p - \Delta p}{1 - 2\Delta p}\right)^j \left(1 - \frac{p - \Delta p}{1 - 2\Delta p}\right)^{N-j} \tag{1.4}$$

for every fixed $j \in S$.

The right-hand side of (1.4) coincides with the probability function of a two-type mutation population model considered earlier by Feller ([9]), where mutations from type 1 to type 2 occur with some probability a and from type 2 to type 1 with some probability b. Setting $a \doteq p - \Delta p$ and $b \doteq 1 - p - \Delta p$, in Feller's notation (1.4) reduces to

$$\pi_j \sim \binom{N}{j} a^j b^{N-j} / (a + b)^N. \tag{1.4$'$}$$

The obvious constraints are $a + b > 0$ and $a, b \in [0, 1]$.

A more rigorous estimate of the stationary probability π_j for N large can be obtained by the continuous approximation of the MC via a diffusion process in $[0, 1]$, which has a stationary distribution with beta density (see [5]).

Notice that the convergence to the equilibrium (1.3) implies the validity of the ergodic theorem, i.e. $(X_0 + \ldots + X_M)/(M + 1)$, which represents the time average of the state of the system during the time interval $[0, M]$, converges almost surely as M tends to infinity to the first moment of the stationary distribution $m \doteq \sum_{j=0}^{N} j\pi_j$, that is

$$(X_0 + \ldots + X_M)/(M + 1) \to m(M \to \infty). \tag{1.5}$$

The argument presented in (2.5) shows that

$$m = N \frac{p - \Delta p}{1 - 2\Delta p}. \tag{1.6}$$

The value of the sample mean, obtained by computer simulation, agrees very well with the right-hand side of (1.6) (see [2]).

Alternatively, an estimate of the stationary probability π_j can be obtained by finding, for every $j \in S$, the sample frequency of the state j in a long enough trajectory. Of course, while for simulated data one can construct a statistic by generating several trajectories with the same initial value, for real data one has to estimate the probabilty π_j by using only one trajectory (even if a long one).

Several others methods are available to calculate numerically the stationary distribution, but we do not treat this topic, here.

In the following section we present four methods to estimate parameters, and we compare them for what concerns their efficiency when using simulated data of the MC. In Section 3, we define ApEn and we discuss some general analytical and numerical results about it. Section 4 is devoted to the elaboration and interpretation of the numerical results for two rubredoxin proteins at different temperatures. After processing the protein data by means of our model, we have obtained estimates of the parameters and approximate entropy; then we present a comparison of the proteins according to their degrees of cooperativity. These data, obtained by molecular dynamics simulation, are available from the authors on request.

2. Parameter estimation

If we put $\alpha = p - \Delta p$ and $\beta = 2\Delta p$, the transition probabilities (1.1) become

$$p_{ij} = p_{ij}(\alpha, \beta, N) = \binom{N}{j} \left(\alpha + \beta \frac{i}{N}\right)^j \left(1 - \alpha - \beta \frac{i}{N}\right)^{N-j} \tag{2.1}$$

with

$$0 \leq \alpha, \quad \alpha + \beta \leq 1. \tag{2.2}$$

Suppose a discrete trajectory $\{x_0, x_1, \ldots, x_M\}$ of the process $X_k,\ k = 0, \ldots, M$ is given such that M is the final time of observation. Our aim is to show how the unknown parameters α, β and N can be estimated by these data. We analyse four different methods to estimate the parameters, namely: *(i) the maximum likelihood method*, *(ii) the method of moments*, *(iii) the least square method*, *(iv) the diffusion approximation of the MC*. By using simulated data, we compare the four methods, in order to study the quality of the estimates and the numerical efficiency of each algorithm.

2.1. *Estimation of parameters by the maximum likelihood method*

For a given data set $\{x_0, x_1, \dots, x_M\}$ we consider the likelihood function, conditionally on the initial value being x_0 :

$$\mathcal{L}(\alpha, \beta, N) = \prod_{k=0}^{M-1} p_{x_k x_{k+1}}(\alpha, \beta, N), \tag{2.3}$$

where $p_{ij}(\alpha, \beta, N)$ is given by (2.1). By maximizing the logarithm of the likelihood function corresponding to the given data set, we can find the estimate $(\hat{\alpha}, \hat{\beta}, \hat{N})$ of (α, β, N).

In order to reduce the number of independent variables of the likelihood function (this is only for numerical convenience), we shall use the following argument, by combining maximum likelihood estimation with the methods of moments.

Disregarding the trivial cases in which the extreme states of the MC (0 and N) are absorbing, for $\alpha > 0$ and $\alpha + \beta < 1$, the MC turns out to be ergodic (see [1]); then there exist the stationary probabilities π_i, $i \in S$ and they are given by (1.3). Moreover

$$\pi_j = \sum_{i=0}^{N} p_{ij} \pi_i \ , \ j \in S. \tag{2.4}$$

The first moment of the stationary distribution satisfies

$$\begin{aligned} m \doteq \sum_j j \pi_j &= \sum_j j \sum_i p_{ij} \pi_i \\ &= \sum_i \pi_i \sum_j j p_{ij} =^* \sum_i \pi_i N(\alpha + \beta i/N) \\ &= N\alpha + \beta m \end{aligned} \tag{2.5}$$

where the equality (*) is based on the fact that the expectation of a random variable with binomial distribution $B(N, q)$ is Nq. Thus

$$m = N\alpha/(1-\beta) \ , \quad \beta \neq 1 \tag{2.6}$$

or in Feller' s notation (see (1.4')) $m = Na/(a+b)$.

If $\hat{m}$ is the sample mean of the data set, i.e.

$$\hat{m} = \frac{1}{M+1} \sum_{k=0}^{M} x_k, \tag{2.7}$$

then, due to the ergodicity, for large M the approximation

$$\hat{m} \approx m = N\alpha/(1-\beta) \tag{2.8}$$

holds, or equivalently

$$\alpha \approx (1-\beta)\hat{m}/N. \tag{2.9}$$

Thus for large M the number of independent arguments of the likelihood function $\mathcal{L}$ in (2.3) reduces to only two (they are β and N). This is very convenient when one searches for the maximum of $\mathcal{L}$ by a numerical algorithm, expecially in the present case where the likelihood function has many local maxima whose values are high and close to each other.

Remark. The argument used in (2.5) yelds also the variance $s^2 \doteq \sum_{i \in S}(i-m)^2 \pi_i$ of the stationary distribution. Indeed, it holds

$$s^2 = \frac{abN}{(a+b)^2(1-(1-a-b)^2(1-1/N))} = \frac{\alpha(1-\alpha-\beta)N}{(1-\beta)^2(1-\beta^2+\beta^2/N)}. \tag{2.10}$$

Formula (2.10) will be proved in Section 2.2.

Note that s^2 is larger than or equal to, and even not asymptotically equal to the variance of a binomial random variable with parameters N and $\alpha/(1-\beta) = a/(a+\beta)$. This is an additional indication that the approximation (1.4) is not very satisfying. (1.4) holds in the sense of probabilities, but functionals like the variance are not respected. However, although better approximations can be numerically found (see the discussion at the end of Section 1), (1.4) furnishes a convenient analytical expression to approximate the stationary probability π_j, $j \in S$. Another estimation of the stationary probability π_j is given by the normal approximation

$$\pi_j \approx \frac{1}{s\sqrt{2\pi}} e^{-(j-m)^2/2s^2}$$

which is simple to calculate, and in many cases even better than (1.4).

2.2. *Estimation of parameters by the method of moments*

The method of moments consists in comparing for a given trajectory $\{x_0, \dots, x_M\}$ the sample moments

$$\hat{m_h} \doteq \frac{1}{M+1} \sum_{k=0}^{M} x_k^h, \quad h = 1, 2, \dots \tag{2.11}$$

with the moments m_h of the stationary distribution $\{\pi_i\}$ defined via

$$m_h \doteq \sum_{j=0}^{N} j^h \pi_j, \quad h = 1, 2, \dots \tag{2.12}$$

For any positive integer h explicit formulae for m_h can be obtained. For the first three moments one has

$$m_1 \doteq m = \alpha N/(1-\beta), \tag{2.13}$$

$$m_2 = (1-\beta^2+\beta^2/N)^{-1}[(N\alpha(1+\alpha(N-1)) + m_1\beta(1+2\alpha(N-1))] \tag{2.14}$$

and

$$\begin{aligned} m_3 = (1-\beta^3(N-1)(N-2)/N^2)^{-1}\{&N\alpha[\alpha^2(N-1)(N-2) \\ &+3\alpha(N-1)+1]+m_1\beta[3\alpha^2(N-1)(N-2)+6\alpha(N-1)+1] \\ &+3\beta^2 m_2(1-1/N)(\alpha(N-2)+1)\}. \end{aligned} \tag{2.15}$$

Since the variance of the stationary distribution satisfies $s^2 = m_2 - m_1^2$, (2.10) follows from(2.13) and (2.14).

As (2.13) is already derived in (2.5), we briefly report how (2.14) and (2.15) are obtained.

For what concerns m_2, we have:

$$m_2 = \sum_j j^2\pi_j = \sum_j j^2 \sum_i p_{ij}\pi_i = \sum_i \pi_i \sum_j j^2 p_{ij}.$$

As the second moment of a binomial random variable with parameters N and q is $Nq(1+q(N-1))$, the expression above (with $q = \alpha + \beta i/N$) becomes

$$\begin{aligned} m_2 &= N\sum_i \pi_i(\alpha+\beta i/N + (\alpha+\beta i/N)^2(N-1)) \\ &= N\alpha + \beta m_1 + N(N-1)\alpha^2 + 2\alpha\beta(N-1)m_1 + (1-1/N)\beta^2\sum_i i^2\pi_i. \end{aligned}$$

Thus:

$$m_2(1-(1-1/N)\beta^2) = N\alpha + \beta m_1 + N(N-1)\alpha^2 + 2\alpha\beta(N-1)m_1$$

from which (2.14) follows.

For m_3, we have

$$m_3 = \sum_j j^3\pi_j = \sum_j j^3 \sum_i p_{ij}\pi_i = \sum_i \pi_i \sum_j j^3 p_{ij}.$$

Recalling that the third moment of $B(N,q)$ is $Nq[1+3q(N-1)+q^2(N-1)(N-2)]$, it follows that

$$\begin{aligned} m_3 &= \sum_i \pi_i[N(N-1)(N-2)(\alpha+\beta i/N)^3 + 3N(N-1)(\alpha+\beta i/N)^2 \\ &\quad +N(\alpha+\beta i/N)] \\ &= N(N-1)(N-2)(\alpha^3 + 3\alpha^2\beta m_1/N + 3\alpha\beta^2 m_2/N^2 + \beta^3 m_3/N^3) \\ &\quad +3N(N-1)(\alpha^2 + 2\alpha\beta m_1/N + \beta^2 m_2/N^2) + N(\alpha + \beta m_1/N) \end{aligned}$$

and (2.15) follows immediately.

To find the estimates of the unknown parameters, one could solve the algebraic system obtained from (2.13), (2.14), (2.15), by replacing the quantities m_h with the sample values $\hat{m}_h$, $h = 1, 2, 3$. However, doing so, the solution of the system may not satisfy the natural constraints of the three parameters. Thus, the estimates

$\hat{N}$, $\hat{\alpha}$, $\hat{\beta}$ of N, α, β are obtained by finding the values of the arguments at which the function

$$Q(N,\alpha,\beta) \doteq \sum_{h=1}^{3}(m_h - \hat{m_h})^2 \tag{2.16}$$

takes its minimum under the constraints $N \in \mathbf{N}$, $\alpha, \beta \geq 0$, $\alpha + \beta \leq 1$. From $\hat{N}$, $\hat{\alpha}$, $\hat{\beta}$, the estimates of the original parameters of biological interest, N, p, Δp, are easily recovered.

2.3. Estimation of parameters by least squares method

This is a variant of the previous method. Here we use the estimate (1.4) of the stationary probability for large N, i.e.

$$\begin{aligned}\pi_j \sim \hat{\pi_j} &\doteq \binom{N}{j}(m/N)^j(1-m/N)^{N-j}\\ &= \binom{N}{j}\left(\frac{\alpha}{1-\beta}\right)^j\left(1-\frac{\alpha}{1-\beta}\right)^{N-j}, \quad \beta \neq 1.\end{aligned} \tag{2.17}$$

Indeed, the further β is away from 1, the better the agreement between the true value of π_j and that of $\hat{\pi}_j$. By using (2.17), the estimate of the stationary probability π_j becomes a function of N and m/N. Now, let $\tilde{\pi}_j$ be the sample frequency of the state j (that is the number of $j's$ in the sequence $\{x_0, x_1, \ldots, x_M\}$, divided by $M+1$).

The least squares method consists in finding the values of the arguments at which the minimum (with the obvious constraints) is obtained for the function

$$\tilde{Q}(m/N, N) = \sum_{j=0}^{N}(\hat{\pi}_j - \tilde{\pi}_j)^2$$

Note that this method allows only to find an estimate of N and $m/N = \alpha/(1-\beta)$ and not of the parameters α and β independently. It can be useful when one is interested to estimate the ratio $\alpha/(1-\beta)$.

2.4. Estimation of parameters by using the diffusion approximation of the MC

We make use of the following approximation result for $N \to \infty$ (see [3], [4], [5]).

Theorem 2.1. *Let α and β depend on N such that the limits*

$$\lambda = \lim_{N\to\infty} N \cdot \alpha(N) \text{ and } \mu = \lim_{N\to\infty} N \cdot (\beta(N) - 1) \tag{2.18}$$

exist. Assume further that X_0/N converges in distribution to some constant y_0. Then, as $N \to \infty$, the normalized process $(\frac{1}{N}X_{[Nt]})_t$ with values in $K_N = \{\frac{i}{N}, i = 0, 1, \ldots, N\}$ converges weakly to Y_t in the Skorohod space $\mathcal{D}_{[0,1]}([0,\infty))$, (see e.g.

[8]) where Y_t denotes the diffusion process with values in $[0, 1]$*, which is the strong solution of the stochastic differential equation (SDE):*

$$dY_t = (\lambda + \mu Y_t)dt + \sqrt{Y_t(1 - Y_t)}dB_t, \quad Y_0 = y_0 \tag{2.19}$$

where B_t denotes standard Brownian motion.

This means that, if, for large N

$$\alpha \sim \lambda/N, \quad \beta \sim \mu/N + 1, \tag{2.20}$$

then the normalized process approximately satisfies (2.19).

Diffusion equations such as (2.19) arise e.g. from Fisher&Wright-like models in population genetics ([10], [17], [18]) and from stochastic models for neural activity ([15]). By discretization of (2.19) one obtains $x_0 = y_0$ and

$$x_{n+1} = x_n + (\lambda + \mu x_n)h + \sqrt{x_n(1 - x_n)}\Delta B_n \tag{2.21}$$

where x_n, $n = 0, 1, \ldots$ denotes the process $(Y_t)_t$ evaluated at the time $t_n = nh$, $h \sim 1/N$, and $\Delta B_n = B_{t_{n+1}} - B_{t_n}$ is the increment of a standard Brownian motion. The relations (2.21) mean that the random variable X_{n+1} conditionally to $(X_n = x_n)$ is distributed according to a Gaussian with expextation $x_n + (\lambda + \mu x_n)h$ and variance $x_n(1 - x_n)h$. Then, given the sequence of data $(x_n)_{n=0,1,\ldots M}$, we obtain the likelihood function, conditionally on the initial value being x_0:

$$L(\lambda, \mu) = \prod_{n=0}^{M} \frac{1}{\sqrt{2\pi h x_n(1 - x_n)}} \cdot \exp\{-[x_{n+1} - x_n - (\lambda + \mu x_n)h]^2/2hx_n(1 - x_n)\} \tag{2.22}$$

The maximum likelihood estimates $\hat{\lambda}$, $\hat{\mu}$ of the diffusion parameters λ, μ are obtained by setting to zero the partial derivatives of the log-likelihood function with respect to its arguments. In this way, we obtain

$$\hat{\lambda} = \frac{1}{hA}\left(D + \frac{B(CD - EA)}{A^2 - CB}\right) \text{ and } \hat{\mu} = \frac{1}{h}\frac{EA - CD}{A^2 - CB}, \tag{2.23}$$

where

$$A = \sum_n \frac{1}{1 - x_n}, \quad B = \sum_n \frac{x_n}{1 - x_n}, \quad C = \sum_n \frac{1}{x_n(1 - x_n)},$$
$$D = \sum_n \frac{x_{n+1} - x_n}{1 - x_n}, \quad E = \sum_n \frac{x_{n+1} - x_n}{x_n(1 - x_n)}.$$

Finally, from $\hat{\lambda}$, $\hat{\mu}$, the estimates $\hat{\alpha}$, $\hat{\beta}$ of the parameters α and β are easily recovered, in the approximation *N large,* by using the relations (2.20).

We emphasize that the above procedure does not allow to estimate N, since it is taken ∞ in the diffusion limit.

2.5. *Comparison of the four methods*

In this section, we focus on the original parameters of biological interest p and Δp instead of α and β, although the former parameters can be simply obtained from the latter ones. We have performed a number of simulation runs with given input values of p, Δp, N, each of them consisting of 20000 data points, and we have applied independently the four methods mentioned above. The simulation of the MC trajectories and the estimation of parameters have been obtained running FORTRAN computer programs specifically written to this end. Their execution requires from few minutes of CPU time for simulation, up to some tens of minutes for parameter estimation, by using an ALPHA Server 800 computer.

Then, by the first and second method, we have recovered the estimates of all three parameters; by the third method we have found the estimates of N and $(p - \Delta p)/(1 - 2\Delta p)$ (i.e. $\alpha/(1 - \beta)$); finally the fourth method has allowed us to estimate p and Δp. Notice that, although the estimates obtained by the diffusion approximation do not provide the value of N, the fourth method is very efficient to refine any estimate of $(\hat{p}, \hat{\Delta} p, \hat{N})$ obtained by the first or the second one (i.e. when $\hat{N}$ has been already found). In fact, while the first three methods need a numerical algorithm for function maximization (or minimization), and therefore their execution is time consuming, the estimate $(\hat{p}, \hat{\Delta} p)$ (that is obtained by finding the point $(\hat{\lambda}, \hat{\mu})$ at which the likelihood function (2.22) takes its maximum) can be analytically found with the fourth method, by explicit calculation.

Although the diffusion limit holds in the approximation $\alpha \approx 0$, $\beta \approx 1$, i.e. $p \approx \Delta p$, $\Delta p \approx 1/2$ (see (2.19)), the fourth method provides good estimates of p and Δp, also for p and Δp far from $1/2$, in the case when N is a large integer which is known in advance. Roughly speaking, the estimates of these parameters, obtained by applying brutally the method, are good enough also in the cases when we are not entitled to use the diffusion approximation.

In Table 1, we report, for a set of simulation runs, the input values of the parameters and their estimates recovered with the four methods, for comparison.

The estimates of the parameters obtained by the first and second method are both excellent; the estimates found by the third method are good in the cases when $p - \Delta p$ is large enough and Δp is small, and rather bad otherwise. This is due to the fact that (2.17) is only an approximation of π_j (see [1]).

3. The approximate entropy

In [16] Pincus introduced approximate entropy (ApEn) to quantify the concept of changing complexity. Usually, the parameters utilized to measure chaos associated to a given set of data are e.g. Hausdorff and correlation dimension, K-S entropy, and the Lyapunov spectrum (see [16] for a discussion). While for computing one of those parameters, the amount of data typically required to achieve convergence is impractically large, estimation of ApEn(m, r) (see below for the definition) can be achieved with relatively few points. In fact, as shown in [16], with only 1000 points, and $m = 2$, ApEn(m, r) is able to distinguish a wide variety of system behaviour. Indeed, it can potentially separate deterministic systems from stochastic ones, periodic from chaotic systems.

Table 1. Estimates of parameters obtained by the four methods of section 2, for simulated data of the MC with transition probabilities (1.1). For six simulation runs, the input values of the parameters and their estimates are reported, for comparison; $(p_i, \Delta p_i, N_i)$ denotes the estimate of $(p, \Delta p, N)$ obtained by the method i $(i = 1, 2, 3, 4)$. The four columns after the first one contain the input values of the parameters, and the value of $\theta := (p - \Delta p)/(1 - 2\Delta p) = \alpha/1 - \beta$, to make more convenient the comparison with column 12, containing the estimate θ_3 of the ratio obtained by the third method. Each run consists of 20000 steps of simulation.

Run	p	Δp	θ	N	p_1	Δp_1	N_1	p_2	Δp_2	N_2	θ_3	N_3	p_4	Δp_4
1	0.90	0.05	0.944	1000	0.8957	0.050	1005	0.8965	0.0505	1004	0.9463	999	0.8937	0.052
2	0.89	0.10	0.987	1000	0.8894	0.0995	1000	0.8893	0.0995	1002	0.9470	1041	0.8907	0.0992
3	0.30	0.05	0.277	800	0.2924	0.050	825	0.2935	0.0505	820	0.2800	769	0.2910	0.0472
4	0.30	0.10	0.250	1000	0.2928	0.0995	1034	0.2935	0.0995	1030	0.2499	1000	0.2908	0.0952
5	0.55	0.40	0.750	1000	0.5460	0.390	1026	0.550	0.395	1025	0.9000	804	0.556	0.399
6	0.70	0.10	0.750	1000	0.6850	0.099	1025	0.6851	0.099	1023	0.9000	801	0.687	0.0945

Now, we will recall from [16] the definition of ApEn. Let us suppose we are given a time-series of data $\{x_1, x_2, \ldots, x_M\}$ equally spaced in time. Fix a positive integer m and let r be a positive number. Then, let us form a sequence of vectors $\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_{M-m+1}\}$ in $\mathbf{R}^m$ defined by

$$\mathbf{v}_i = (x_i, x_{i+1}, \ldots, x_{i+m-1})^T \tag{3.1}$$

Next, define for each i, $1 \le i \le M - m + 1$,

$$C_i(m, r) = \frac{(\text{number of } j \text{ such that } d(\mathbf{v}_i, \mathbf{v}_j) \le r)}{M - m + 1}, \tag{3.2}$$

where the distance $d(\cdot, \cdot)$ between two vectors is defined by

$$d(\mathbf{v}_i, \mathbf{v}_j) = \max_{k=1,\ldots,m} |x_{i+k-1} - x_{j+k-1}|. \tag{3.3}$$

The $C_i(m, r)$ values measure within a tolerance r the frequency of patterns similar to a given pattern of window lenght m. Now define

$$\Phi(m, r) = \frac{\sum_{i=1}^{M-m+1} \log C_i(m, r)}{M - m + 1} \tag{3.4}$$

and

$$ApEn\ (m, r) = \lim_{M \to \infty} (\Phi(m, r) - \Phi(m + 1, r)). \tag{3.5}$$

Given M data points, the formula (3.5) can be implemented by defining the statistics

$$ApEn(m, r, M) = \Phi(m, r) - \Phi(m + 1, r). \tag{3.6}$$

Heuristically ApEn measures the logarithmic likelihood that runs of patterns that are close for m observations, remain close on the next incremental comparison. A greater likelihood of remaining close (i.e. regularity) produces smaller ApEn values, and viceversa.

On the basis of the analysis of simulated data, Pincus showed that for $m = 2$ and $M = 1000$, choices of r ranging from 0.1 to 0.2 times the standard deviation (SD) of the x_i data produce reasonable statistical validity of ApEn(m, r, M).

The following analytical result shows that for a Markov chain, ApEn coincides with the Kolmogorov-Sinai entropy (see e.g. [20]).

Theorem 3.1 ([16]). *Let X_k be a homogeneous, stationary MC with discrete state space $S = \{x_1, x_2, \ldots\}$ and transition probabilities*

$$p_{ij} = P(X_{k+1} = x_j | X_k = x_i), \quad i, j \in S.$$

Let $\{\pi_i\}$ be the vector of the stationary probabilities, such that

$$\pi_i = \lim_{n \to \infty} p_{ij}^{(n)},$$

where $p_{ij}^{(n)}$ denotes the $n-$step transition probability.
Then, if $r < \min\{|x - y|,\ x \neq y,\ x, y \in S\}$, for any m, a.s. it holds:

$$ApEn\ (m, r) = -\sum_i \sum_j \pi_i p_{ij} \log p_{ij}. \tag{3.7}$$

After recalling the definition of ApEn, we go to show some results about the approximate entropy of data relative to simulated trajectories of the MC with transition probabilities (1.1). Since the applicability of formula (3.7) is impractical when the number of states N is large, due to the heavy computation required to obtain the stationary probabilities π_i, it is more convenient to calculate ApEn numerically by (3.6) for sufficiently large M.

We have performed various simulation runs of the MC, each consisting of 1000 data points. In the first group we have taken $N = 300$, $p = 0.5$ and we have let Δp vary from 0 to 0.5. For every input value of Δp, we have calculated

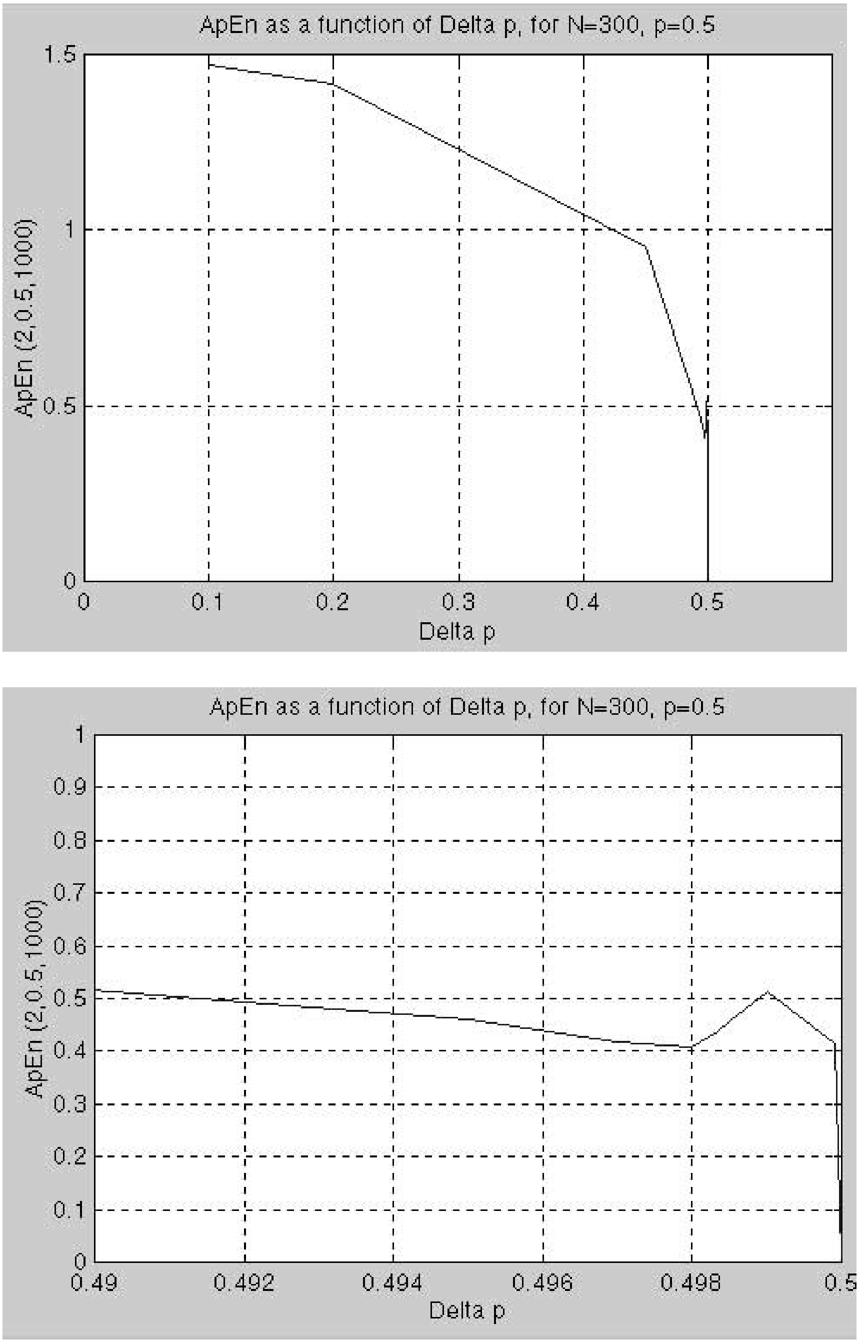

Fig. 1. (a) Plot of *ApEn* as a function of Δp for data relative to Table 2. (b) Zoom of Fig. 1a, for Δp near 0.5.

ApEn$(2, 0.5, 1000)$, finding that its shape as a function of Δp is substantially decreasing, except for the presence of a local maximum, and some minor oscillations, near the critical value Δp_0 in correspondence of which (see [1]), the stationary distribution of the system is almost uniform. The value Δp_0 was numerically detected in [1], showing that for $p = 0.5$ and $\Delta p = \Delta p_0$, the system becomes very erratic and its trajectory very complex, giving the maximum value of the fractal dimension (of covering) (see [1]). Maintaining fixed $p = 0.5$, and letting vary Δp, an increase of N in the simulation runs results in a progressive disappearing of the oscillation behaviour near the corresponding critical value Δp_0 mentioned above, since, as shown in [5], $\Delta p_0 \sim \frac{1}{2}(1 - \frac{1}{N})$, so the larger N, the more Δp_0 is shifted at right towards the value $1/2$. In Tables 2 and 3 we summarize the numerical results obtained for two groups of simulation runs; those relative to the first group are also reported graphically in Fig. 1 . For data of Table 2, we have put $r = 0.5$, since the SD is of order 5, in this case. For Table 3, we have taken $r = 0.1$ times the SD of data. Undoubtedly, ApEn$(2, r, 1000)$ appears to be a substantially decreasing function of the degree of cooperativity Δp.

4. Numerical results for protein energy data

In this section, we deal with estimation of parameters, calculation of ApEn and comparing cooperativity, for protein data obtained by molecular dynamics simulations and referring to a couple of rubredoxin proteins. The first protein comes from a bacterium living at normal temperature ($35\,^\circ C$), the other one comes from an organism that lives at about $100\,^\circ C$([12]). It remains yet unknown the reason why the two proteins present a different temperature stability even though their structures are extremely similar (the structure of the first protein breaks if it is carried to $100\,^\circ C$). For this, the comparison between the two proteins is particularly interesting. The rubredoxin living at $35\,^\circ C$ is indicated by the code 1iro, that resisting at $100\,^\circ C$ by 1caa.

Table 2. Approximate entropy for simulated data of the MC with transition probabilities (2.1). Here $N = 300$, $p = 0.5$, $r = 0.5 (SD \cong 5)$ and $m = 2$; Δp varies from 0.1 to 0.5.

Δp	$ApEn(m, r, 1000)$
0.1	1.470
0.2	1.417
0.4	0.950
0.49	0.786
0.495	0.462
0.497	0.418
0.498	0.409
$0.49834 = \Delta p_0$	0.436
0.4985	0.457
0.499	0.470
0.4995	0.51
0.4999	0.415
$0.5 - 0$	0.040

Table 3. Approximate entropy for other simulation runs of the MC. Here, $N = 150000$, r is taken one tenth of the standard deviation (SD).

p	Δp	SD	r	$ApEn(2, r, 1000)$
0.72	0.26	104	10.4	1.465
0.63	0.36	72	7.2	1.29
0.649	0.345	72	7.2	1.3263
0.666	0.329	63.5	6.3	1.30

For each protein we have analysed data referring to two different temperatures: at 35 °C (308K) and 100 °C (373K) and concerning the total electrostatic energy of proteins. The data refer to trajectory segments relative to the time interval from 500 to 2000 ps, with step 0.05 (data relative to the first 500 ps have been disregarded, since at the beginning the situation is far from the equilibrium), and consist of 30000 points. These are contained into 4 data files which are available from the authors, on request.

The filenames are self-explicative, but for the sake of brevity, we recodify them, by referring to the temperature measured in centigrade degrees, as shown below:

1caa308KEEstot	c35
1caa373KEEstot	c100
1iro308KEEstot	g35
1iro308KEEstot	g100

To obtain an estimate of the average number of bonds in the considered protein, an average energy associated to every bond was needed. We have choosen the value $\Delta E_{es} = -1$ kcal/mole, that is an estimate of the average of the different electrostatic interactions among aminoacidics residues in a protein. In any case, the actual number of bonds should not be critical as long as it is large enough. Using the above value we have obtained the number of chemical bonds as a function of time, with time-step of 0.05 ps, in order to apply our model. Then, we have calculated for modified data of each of the files c35, c100, g35, g100, the minimum value, the maximum value (N), the mean value and the standard deviation (SD). These are reported in Table 4.

Table 4. Statistics of the first 1000 data points relative to the files c35, c100, g35 and g100.

Data file	Min	Max	Mean value	SD
c35	108230	110390	109270	317.9
c100	137600	141370	139636.8	601.38
g35	146340	149680	148047.3	487.38
g100	131030	134720	133051.4	571.79

Since the g35 data have resulted to have the greater value of $N(\sim 150000)$, we have rescaled the data relative to every file (except g35) so that the N-values were the same for all files, this way allowing a comparison between the protein data.

By means of the methods described in Section 2, we have estimated the parameters p and Δp by using all the 30000 data points, then we have recalculated the SD of rescaled data and, taking $r = 0.1 \times SD$, we have calculated ApEn(2, r, 1000), for each set of data. The results are summarized in Table 5 and reported graphically in Fig. 2 and Fig. 3.

The cooperativity is usually related to the biological activity of proteins. In this light, every physical modification of the environment, leading to a loss of biological activity, is related to a loss of cooperativity. The protein 1iro (g) is active at ambient temperature ($\sim 35\,^\circ C$) and then is thermically stressed by raising the temperature to $100\,^\circ C$. Hence, the cooperativity is reduced by this thermal shock. But also a

Table 5. Estimates of p and Δp and approximate entropy (calculated with 1000 points) for data relative to the files c35, c100, g35 and g100. Here, SD = $10\,r$.

Data file	r	p	Δp	$ApEn(2, r, 1000)$
c35	43	0.725	0.267	1.286
c100	63	0.631	0.364	1.136
g35	49	0.649	0.345	1.153
g100	63	0.666	0.329	1.203

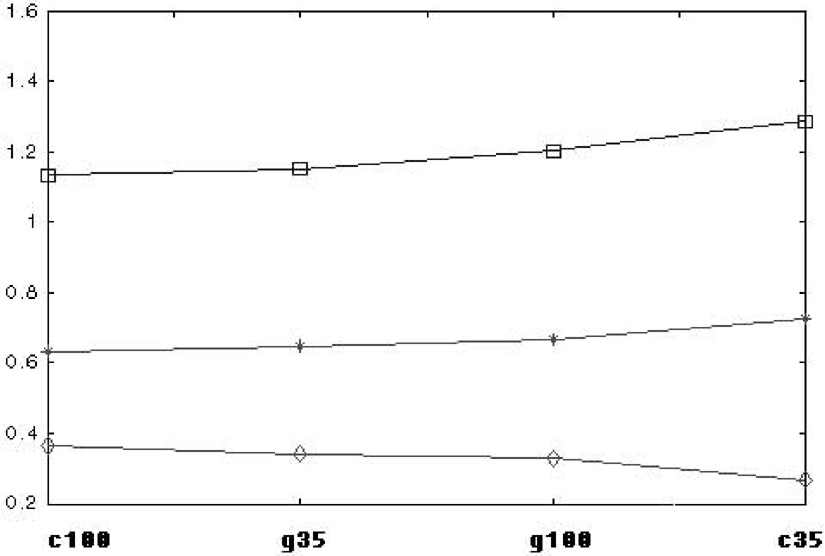

Fig. 2. Graphical representation of the behaviour of $ApEn(\Box)$, $p(*)$, $\Delta p(\diamond)$, as a function of data-files c35, c100, g35, g100 (see Table 5).

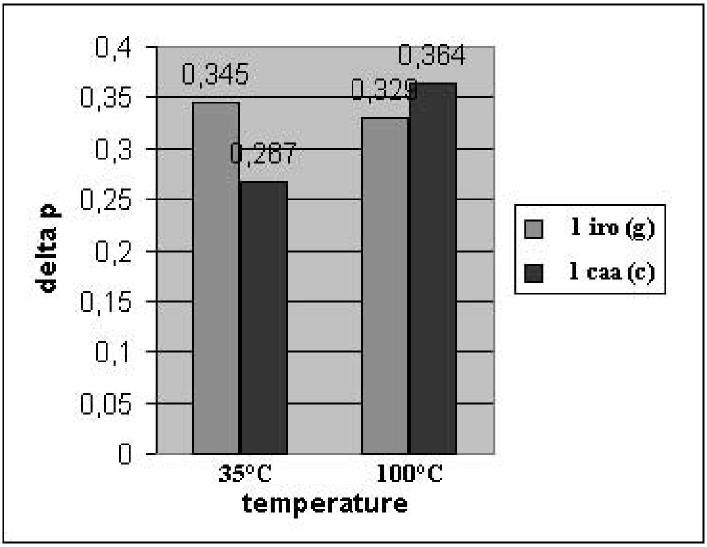

Fig. 3. Histogram representation of cooperativity Δp for 1iro (g) and 1caa (c) proteins, at 35 °C and 100 °C.

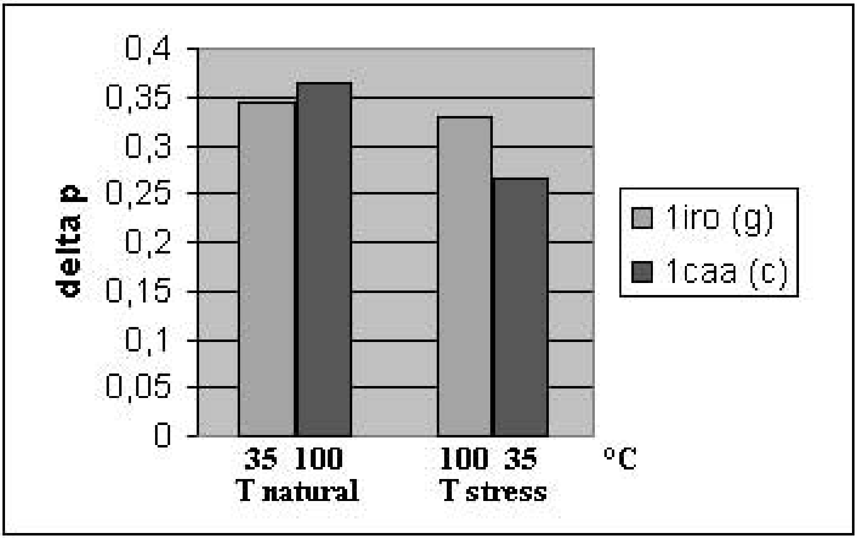

Fig. 4. Comparison of cooperativity for 1iro (g) and 1caa (c) proteins, at the natural and stressed temperature.

strong reduction of the temperature from the usual value, where the protein is active, that is about 100 $°C$ for 1caa (c), to a low 35 $°C$, can lead to a loss of activity and then to a reduction of cooperativity (see Fig. 4).

In physical terms this reduction of cooperativity can be interpreted in the first case (g) as a larger amplitude motion of atoms, that weakens long range correlations. In the latter case (c), the cold environment tends to reduce the atomic motion amplitude "freezing" the fluctuations in local and uncorrelated vibrations.

5. Concluding remarks

In this paper, analysing data relative to the time evolution of protein energy data, by a Markov chain model for cooperative interactions previously introduced by us, we have been able to characterize cooperativity in proteins. Indeed, we have estimated the degree of cooperativity of a protein, Δp, from data, so rendering possible a comparison between different proteins.

Another parameter we have used to characterize protein behaviour is the approximate entropy (ApEn); it turned out that ApEn is related to the cooperativity, in fact it appears to be a decreasing function of Δp. Thus, ApEn is a measure of cooperativity, alternative to Δp, and it is very effective, since a good estimate of ApEn can be easily obtained with relatively few data points and a small numerical effort.

We applied our procedures to both simulated data of the Markov chain, and energy data which have been obtained by molecular dynamics simulation.

Acknowledgements. We would like to express particular thanks to the referees for useful comments and suggestions which have improved the presentation of the whole paper. This work was supported by the CNR grant N. 98.00989.CT01 and the CNR target project MADESS II.

References

1. Abundo, M., Accardi, L., Rosato, N.: *A Markovian model for cooperative interactions in proteins*. Math. Models and Meth. in Appl. Sci. **5** (6), 835–863 (1995)
2. Abundo, M., Accardi, L., Rosato, N., Mei, G., Finazzi Agrò, A.: *A stochastic model for the sigmoidal behaviour of cooperative biological systems*. Biophysical Chemistry **58**, 313–323 (1996)
3. Abundo, M., Accardi, L., Rosato, N., Stella, L.: *A stochastic model for the cooperative relaxation of proteins, based on a hierarchy of bonds between aminoacidic residues*. Math. Models and Meth. in Appl. Sci., **8** (2), 327–358 (1998)
4. Abundo, M., Baldi, P., Caramellino, L.: *A diffusion approximation which models hierarchic interactions in cooperative biological systems*. Open Sys. & Information Dyn. **5**, 1–23 (1998)
5. Abundo, M., Caramellino, L.: *Some remarks about a Markov chain which models cooperative biological systems*. Open Sys. & Information Dyn. **3** (3), 325–343 (1995)
6. Amadei, A., Linseen, A.B.M., Berendsen, H.J.C.: *Essential dynamics of proteins*. Proteins: Struct. Funct. Genet. **17**, 412–425 (1993)
7. Berriz, G.F., Gutin, A.M., Shakhnovich, E.I.: *Cooperativity and stability in a Langevin model of proteinlike folding*. J. Chem. Phys. **26**, 271–287 (1996)
8. Ethier, S.N., Kurtz, T.G.: *Markov processes. Characterization and convergence*. Wiley, New York, 1986

9. Feller, W.: *Diffusion processes in genetics*. Proc. Second Berkeley Symp. Math. Statistics and Probab., 227–246 (1950). University of California Press, Berkeley and Los Angeles, 1951
10. Fisher, R.A., Wright, S., Malecot, See G.: *Sur un probleme de probabilites en chaine que pose la genetique*. Comptes rendus de l'Academie des Sciences, **219**, 379–381 (1944)
11. Freire, E., Haynie, D.T., Xie, D.: *Molecular basis of cooperativity in protein folding. IV. CORE: a general cooperative folding model*. Proteins: Struct. Funct. Genet. **17**, 111–123 (1993)
12. Hernandez, G., Jenney, F.E., Adams, M.W.W., LeMaster, D.M.: *Millisecond time scale conformational flexibility in a hyperthermophile protein at ambient temperature*. Proc. Natl. Acad. Sci. USA, **97**, 3166–3170 (2000)
13. Klimov, D.K., Thirumalai, D.: *Cooperativity in protein folding: from lattice models with sidechains to real proteins*. Folding & Design **3**, 127–139 (1998)
14. Kolinski, A., Skolnick, J.: *Discretized model of proteins.1. Montecarlo study of cooperativity in hompolypeptides*. J. Chem. Phys. **97**, 9412–9426 (1992)
15. Lanska, V., Lansky, P., Smiths, C.E.: *Synaptic trasmission in a diffusion model for neural activity*. J. Theor. Biol. **166**, 393–406 (1994)
16. Pincus, S.M.: *Approximate entropy as a measure of system complexity*. Proc. Natl. Acad. Sci. USA, **88**, 2297–2301 (1991)
17. Shiga, T.: *Diffusion processes in population genetics*. J. Math. Kyoto Univ., **21** (1), 133–151 (1981)
18. Shimakura, N.: *Formulas for diffusion approximations of some gene frequency models*. J. Math. Kyoto Univ., **21** (1), 19–45 (1981)
19. Sneddon, S.F., Brooks III, C.L.: *Protein motions: structural and functional aspects*. In: Diamons, R. et al. (eds): "Molecular Structures in Biology". Oxford University Press, (1993)
20. Walters, P.: *An introduction to Ergodic Theory*. New York: Springer-Verlag, 1982