

Contents

1	Abstract	2
2	Introduction	2
2.1	Proteins and protein folding	2
2.1.1	Protein folding	4
2.1.2	Forces of protein folding	5
2.2	Simplified models	6
2.3	Monte Carlo methods	7
2.3.1	Dynamical interpretation	11
3	Aim of the present study	16
4	Methods	16
4.1	The models	17
4.2	Parameters and definitions	19
4.2.1	Discrimination measures	22
5	Results and discussion	22
5.1	Comparing polymer dynamics on sc lattice versus fcc lattice, allowing for one and two point moves	22
5.2	Investigating how the absence of two point moves affects the dynamics	23
5.3	Comparing the dynamics of the Metropolis algorithm versus that of the Glauber algorithm	27
5.4	Dead end attempts	28
5.5	Conclusions	29
6	Acknowledgments	30

1 Abstract

The enormous number of conformations available to proteins makes exhaustive sampling impossible, yet proteins find their unique native conformations in seconds (the Levinthal paradox). In trying to resolve this paradox earlier studies have used Monte Carlo simulations on cubic lattices to examine differences between slowly folding sequences and sequences that fold rapidly. It has been suggested that sequences with either a large energy gap or strong local interactions would fold fast. To examine the model dependence of these results we have performed a series of Monte Carlo simulations on different lattices using different transition probabilities and motions. The conditions studied show no significant differences. This supports the idea that conclusions drawn from simple lattice simulations may be valid for real proteins.

2 Introduction

The understanding of protein folding is one of the major challenges of biochemistry. The great complexity of the process makes computer simulations an important tool in the search for the principles that governs protein folding. Recently several Monte Carlo simulations on simple lattice models have contributed to a better understanding of the folding mechanism. Different studies have represented proteins in different ways and made different assumptions about their dynamics. There has only been few investigations of how this affects the simulations. This makes the comparison and interpretation of the results more complicated. The ambitious purpose of this report is to help putting these problems aside.

2.1 Proteins and protein folding

Proteins play a central role in many biological processes. They transmit chemical and physical signals between molecules, act as receptors on the cell surface, control the activity of DNA and of other proteins. They transport oxygen, lipids and metals in the blood and act as storage proteins. Some proteins control the flow of ions and other molecules across the cell membrane, others participate in the transfer of electrons in photo synthesis. Proteins also play a major protective role in the immune system and some proteins act as important structural and functional components in the cell. Of crucial importance to the functionality of most proteins is their three dimensional structure.

The monomeric building blocks of proteins are the 20 natural α -L-amino acids. For each of them there exists some information (base pair triplets) in the genetic material.

The amino acids all have the α -carbon atom connected to a carboxy-group and to an amino-group. In addition all amino acids except proline have the α -carbon atom connected to a hydrogen atom. The differences between the amino acids are due to the characteristic side chain R.

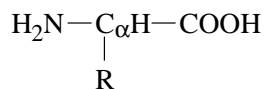


Figure 1: The common structure of amino acids.

In proteins the amino acids are condensed into long unbranched chains. By emitting water the α -carboxy group and the α -amino group of different amino acids form a peptide bond. The peptide bond shows resonance between two limiting structures.

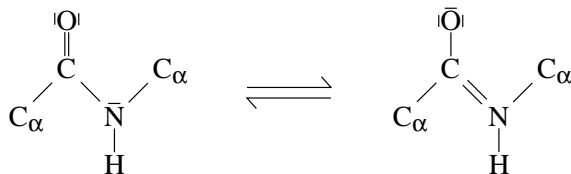


Figure 2: The two limiting structures of the peptide bond.

This results in a large dipole moment and in all six atoms of the peptide bond being forced into roughly the same plane surface. Therefore (except for small and fast fluctuations around the average structure) the only degrees of freedom available to the protein backbone are two rotational coordinates for each peptide bond. These are called the dihedral angles.

The chemical properties, the conformation and the function of the protein are all determined by the information that lies in the sequence of side chains. This sequence is called the primary structure of the protein.

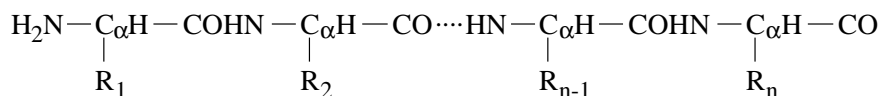


Figure 3: Schematic view of the structure of a protein. The sequence $\text{R}_1 \cdots \text{R}_n$ is called the primary structure of the protein.

Regularly repeated patterns of dihedral angles give rise to ordered structures called secondary structures. All proteins seem to contain larger or smaller parts with secondary structures. Many proteins are very compact and of almost spherical shape. These proteins are called globular. In such proteins parts with secondary structure are connected with irregular parts and packed into very compact bodies. The structure of these is referred to as tertiary structure. When several such units are aggregated, their structure is referred to as quaternary.

Under appropriate conditions a given amino acid sequence adopts a unique average structure. This structure is called the native state of the protein. Experimental studies have demonstrated that a protein does not have a static structure, but undergoes significant fluctuations relative to this average structure [Gurd & Rothgeb, 1979]. The native state is thus not a single point in the configuration space, but a subspace having local minima separated by barriers, see figure 4. The root mean square difference (rmsd) between the coordinates of these local minima are of order 0.1 to 1 Å [Elber & Karplus, 1987]. The dynamics and thermodynamics of this subspace is rather well studied through molecular dynamics simulations [Brooks et al., 1988]. The subspace that constitutes the native state is sampled in times 10^{-13} to 10^{-8} s at physiological temperatures.

The native state constitutes a very small subspace to the whole configuration space. The rest represents the denatured state. The rmsd of the coordinates in the denatured state can be tens of Å, and characteristic times for moves in the denatured state are in the range of ns to hours.

Preaveraging over the fast degrees of freedom leads to a coarser description of the protein on time and length scales relevant in the case of protein folding. In this coarse grained picture, moving in the configuration space means jumping between local energy minima of the protein configuration [Ansari et al., 1985].

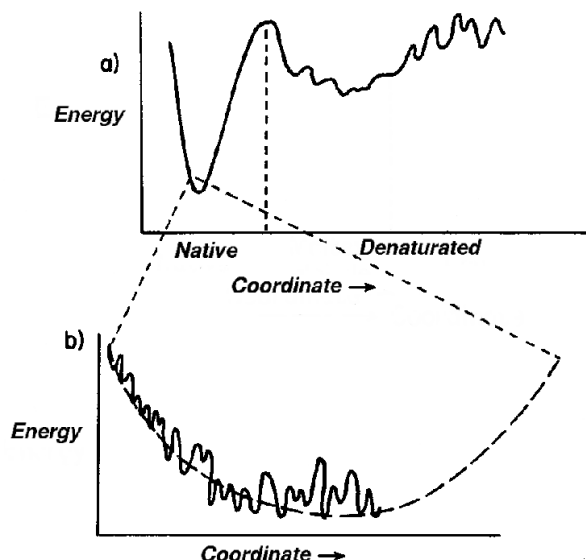


Figure 4: Schematic representation of the energy of a protein as a function of a configurational coordinate: a) complete space. b) enlarged view of the native state.

2.1.1 Protein folding

Several studies have shown that proteins fold into their unique compact native state when placed in physiological conditions. There is evidence that this process is thermodynamically reversible for many small single-domain globular proteins [Santoro & Bolen, 1988] and also for some multi-domain and coiled-coil proteins [Privalov, 1982]. This suggests that the native state coincides with the energetic ground state of the system.

In a well known argument C. Levinthal stated that this is not the case [Levinthal, 1969]. Levinthal reasoned that it is impossible for a protein to find the state with the lowest energy among the enormous number of possible conformations within a reasonable time. This is because the size of the conformational space scales exponentially with the number of amino acid residues. Thus if every amino acid has only two possible conformations, the number of possible conformations for a protein with 100 amino acids is $2^{100} \approx 10^{30}$. If it takes the protein 1 ps to explore each conformation, then the time required to explore all conformations is approximately 10^{10} years. Levinthal postulated that the native state must instead be determined by kinetic pathways.

In any case, even if folding is kinetically controlled and proteins do not fold to a structure that is a global minimum, they must still reliably fold to a unique structure. But the same argument as that used by Levinthal can be used to question how a protein could find any particular configuration. This is called Levinthal's paradox.

The weak point in Levinthal's argument is that it assumes all conformations to be equally likely along the path from the unfolded to the folded state. There are several suppositions as how proteins could solve Levinthal's paradox. Conformations with relatively low free energy that are smoothly connected to the native state could form funnels that guide the proteins towards the native states. The existence of folding funnels is the key concept that has been introduced to explain the rapid folding of natural proteins [Leopold et al., 1992]. Another possibility is the concept of a molten globula. An initial collapse of the protein chain into a compact globule would considerably reduce the number of conformations for the protein to explore. Exactly how real proteins do solve the Levinthal paradox is not known.

For small proteins there is evidence that the transition from unfolded to folded is a cooperative all-or-none process [Friere & Biltonen, 1978]. This means that the energy surface is relatively smooth and that there is one time constant that largely

dominates the relaxation behavior. For larger proteins this does not seem to be true [Privalov, 1982], instead they behave as if they have a very rough energy surface.

Rough energy landscapes occur in problems with many competing interactions, which can appear in proteins in several ways. For example bringing two favorable interacting residues of a protein into contact might require simultaneously bringing two unfavorable residues close. This is often referred to as energetic frustration.

Another cause for roughness could be geometric constraints. Because the polymer chain of the protein can not pass through itself, two conformations can have very similar shape and similar free energy but still be unable to reconfigure from one to the other. This is called excluded volume and can cause very large barriers between close local minima in the energy landscape.

Protein folding in general exhibits behaviors that are characteristic of both smooth and rough energy landscapes. A good description of a protein should therefore involve an energy surface that is neither very rough nor very smooth but something in between. Such an energy surface would have many closely related structures with similar free energy which are not dynamically connected within reasonable times. One should expect a protein with these properties to have great difficulties in finding reliably a unique native conformation [Bryngelson et al., 1995] within limited time. Thus a paradox similar to that of Levinthal arises because of a totally different reason.

Among all possible amino acid sequences, proteins are chosen by nature from the subgroup that reliably can find a unique configuration. It is likely that the requirements for the kinetic accessibility of the native state should be encoded in the primary sequence itself, and the question arises what it is that distinguishes these sequences from others.

Earlier studies have suggested that the sequences that reliably fold to their native conformation have large energy gaps between the native and the lowest non-native structure [Sali et al., 1994]. This result has been questioned by R. Unger and J. Moult who instead state that the folding sequences have strong possible interactions between residues close to each other in the sequence [Unger & Moult, 1996]. The statement by [Sali et al., 1994] that folding sequences have large energy gaps have also been called in question by D. K. Klimov and D. Thirumalai [Klimov & Thirumalai, 1996].

2.1.2 Forces of protein folding

Of crucial importance to the understanding of protein folding is the knowledge of the magnitude and nature of the participating forces. The following summary closely follows the review by Ken A. Dill [Dill, 1990].

The electrostatic forces present in protein folding can be divided into the classical electrostatic forces, ion pairing, hydrogen bonding and van der Waals interactions.

For charged proteins there arises a non specific repulsive electrostatic force. As the charge density is greater for the folded than for the unfolded protein, this force destabilizes highly charged proteins. Acids and bases therefore works in a destabilizing way on native proteins.

Ion pairs are sometimes formed in proteins. Although ion pairing can contribute to stability, ion pairing is so rare that it is clear that it can not be the dominant force of protein folding.

Hydrogen bonds and van der Waals interactions can occur between different parts of the peptide chain or between the peptide chain and the molecules of the solution. If the interactions within the protein are energetically favorable, this would support folding. The magnitudes of these forces are difficult to estimate, and there are no strong arguments why the hydrogen bonds between different parts of the protein backbone should have lower energy than those of the unfolded chain to water.

The hydrogen bonds and van der Waals interactions are probably not the dominant interactions of protein folding either. But a compact configuration without many

hydrogen bonds would indeed be energetically unfavorable. It is therefore believed that hydrogen bonds play an important role in the formation of secondary structures.

The conformational preferences of neighboring residues arising from the sum of all forces between connected residues are called intrinsic propensities. It has been suggested that these preferences account for the helical stability in globular proteins. But there is evidence that intrinsic propensities by themselves are insufficient and other forces must contribute to stabilize helices in globular proteins [Dill, 1990].

Hydrophobic interaction is used to describe the change in free energy upon non polar solvation processes in polar solutes.

There are strong reasons to believe that hydrophobicity is the dominant force of protein folding[Dill, 1990].

The hydrophobic interaction originates from both an entropic and an enthalpic contribution. The major part is rather due to loss of free energy when the polar solution molecules rearrange around non polar groups, than to favorable interaction between non polar groups themselves.

There is also an opposing entropic effect from the loss of conformational entropy when the protein is folded. This effect is almost equal to the hydrophobic in magnitude.

Stabilizing and destabilizing forces largely cancel out. The free energy difference between folded and unfolded state is of the order 5-20 kcal/mol for a typical protein [Privalov, 1979]. This is at room temperature less than $(1/10)k_B T$ (where k_B is Boltzmann's constant and T the temperature). Therefore, even weak interactions can contribute significantly to protein stability.

2.2 Simplified models

A detailed simulation of the physics of protein folding would require a detailed spatial description of the protein and a very detailed energy function. Furthermore one would have to use a time resolution smaller than the characteristic time of the fastest degrees of freedom in the system when simulating. The great spread of time and length scales on which protein folding occurs therefore makes it difficult to extend the time scale of the simulation up to the relaxation time of the slow degrees of freedom. The computational cost gets too large.

At the same time as the separated scales render detailed simulations impossible, they also enable a simplified description.

As a first simplification of the protein energy function we introduce the idea of a mean force potential. The precise forces acting on every atom during their fast fluctuations are not really relevant to protein folding. It is therefore possible to average these fluctuations and denote the potential of the resulting forces as a mean force potential. As the actual atomic forces are not known in detail we have to make some further simplifying assumptions about the mean force potential.

One approach that has been used is to use statistical information from experimentally determined protein structures about preferred distances between certain types of residues or atoms. Alternatively one could try to describe the physical forces actually present in proteins as detailed as possible using a molecular mechanics potential with an extra entropy term.

Very rough models have also been used, where one makes grave simplifications but still tries to capture the characteristic properties of proteins.

In the HP model the residues are classified as being either hydrophobic or polar, and the interaction between two residues depend solely on which classes they belong to. This model has been applied for instance by Dill and coworkers [Dill et al., 1995].

Another possibility is the route followed by Bryngelson and Wolynes [Bryngelson & Wolynes, 1987], Shacknovich, Sali, and Karplus [Sali et al., 1994], and others. They have used the random energy model (REM). This is based on the observation that the resulting forces

between residues in proteins are built up by many small competing contributions and therefore likely to be distributed approximately in a Gaussian way. In the REM the interactions between residues are drawn randomly from some distribution. By choosing a negative mean of the distribution, account can be taken of the preferred compactness of proteins.

The common approach to simplify protein structures is to represent the proteins as chains of a limited number of interacting centers per residue. To restrict the large conformational space every residue in the model is often allowed only to adopt some discrete conformational states. In some models one allows only some number of discrete pseudo dihedral angles between residues. This has been used by amongst others Rooman et al [Rooman et al., 1991]. In lattice models one allows the amino acid residues to occupy only discrete points in space with the allowed points forming a repeated pattern in space. Lattice studies of proteins have been made by for example Dill and coworkers [Dill et al., 1995], Shakhnovich and coworkers [Sali et al., 1994], and Klimov and Thirumalai [Klimov & Thirumalai, 1996].

When trying to predict protein structures, it is important that the geometric representation of the proteins is as simple as possible, but still able to depict the structural characteristics of protein conformations. The virtual bond and torsion angles of different geometric representations match the angles preferred by proteins more or less well, and, when restricted to these representations, the most native like structure of a modeled protein will deviate from the true native structure more or less much. Some off-lattice models using rather few discrete pseudo dihedral angles can be optimized to match secondary structures of real proteins very well. When the modeled proteins are restricted to move on lattices, it seems to hold approximately that deviation $\propto (\text{complexity})^{-1/2}$, where complexity means number of possible states for every residue [Park & Levitt, 1995]. To obtain a geometrical representation with a lattice model, that is as good as is possible for simple optimized off-lattice models, one has to make the lattice very complicated. But there are some features that still makes lattice models interesting.

In lattice models it is rather easy to implement local moves, that is moves were only a few residues of the chain change position. It is likely that such moves are important to the folding dynamics. The ability to index all positions of the residues can give considerable computational gain. To know when the ground state of the model system is reached in a simulation is in general a difficult problem because of the fast growing conformational space (, again Levinthal's paradox). On lattices it is possible to enumerate the complete set of conformations for short chains and thereby determine which of them is the ground state. This has also been accomplished in some off-lattice models, like in the model by G. M. Crippen and Y. Z. Ohkubo [Crippen & Ohkubo, 1998] where a small number of fixed secondary structure elements are combined.

The attempts made to study the folding dynamics of proteins have mainly been restricted to Monte Carlo simulations starting from randomly chosen initial conformations, to the enumeration of transition rates between classes of conformations which have the same number of contacts and are a given number of kinetic steps away from the native state [Chan & Dill, 1993], and to mode analyses of phenomenological master equations [Pitard & Orland, 1998].

2.3 Monte Carlo methods

The name Monte Carlo refers to a class of mathematical techniques. Common to them is that games of chance are used to study some phenomenon of interest.

Simple sampling A task where Monte Carlo methods are often applied, is the evaluation of multidimensional integrals.

Let $X = (x_1, x_2, \dots, x_d)$ be a point in a d -dimensional space, and $dX = dx_1 dx_2 \dots dx_d$ the volume element. We want to evaluate the integral over the domain Ω , $G = \int_{\Omega} f(X) dX$. The basic Monte Carlo method of calculating the integral is to draw a set of random values X_1, X_2, \dots, X_N uniform distributed (simple sampling) in Ω . The arithmetic mean $G_N = \frac{1}{N} \sum_{i=1}^N f(X_i)$ is then used as an estimator of G .

According to the law of large numbers, G_N will converge to the value of G with probability one. Furthermore the central limit theorem tells us that for sufficiently large N , G_N is a random variable distributed according to Gaussian distribution with mean G and variance

$$\frac{\langle f(X)^2 \rangle - \langle f(X) \rangle^2}{N} \quad (1)$$

The error δ of the integration will be of the order $\delta \propto N^{-1/2}$.

As the computational cost c for a Monte Carlo integration is proportional to N , the cost as function of a given error will be

$$c(\delta) \propto N \propto \delta^{-2} \quad (2)$$

The standard deterministic routine to solve one-dimensional integrals numerical is to chose the points X according to a regular grid. If the same technique is applied to evaluate multidimensional integrals points are chosen in the volume L^d on a grid with spacing h in each dimension. Here L denotes the length in each dimension d of the volume to integrate over. The computational cost c is proportional to the number of grid points

$$c \propto (L/h)^d \quad (3)$$

The order of the error δ of the calculation on the other hand is independent of the dimension of the integral but depends on the order n of the algorithm used.

$$\delta \propto (L/h)^{-n} \quad (4)$$

The computational cost as a function of a given error therefore grows like

$$c(\delta) \propto (L/h)^d \propto \delta^{-d/n} \quad (5)$$

For small dimensions d one can easily make the exponent d/n smaller than 2, but for every algorithm of order n the Monte Carlo algorithm will be faster in higher dimensions $d > 2n$. The order of typical algorithms are seldom better than $n = 4$. This is why Monte Carlo methods are often preferred for multidimensional integration. In practice the boundary lies around dimensions $d \simeq 4$.

How fast the Monte Carlo integration converges does not only depend on the dimension, but also on the function to integrate. If $f(X)$ varies strongly within Ω , some of $f(X_1), f(X_2), \dots, f(X_N)$ will contribute only little to the mean G_N , and will therefore be produced almost in vain.

A solution to this problem is given by importance sampling.

Importance sampling What is needed to solve the problem of slow convergence for some functions $f(X)$, is some technique to draw the values included in the average G_N not uniformly in Ω but rather according to some probability $p(X)$ which preferentially gives values X who contributes significantly to G_N .

The estimator for G is then given by

$$G_N = \frac{1}{N} \sum_{i=1}^N \frac{f(X_i)}{p(X_i)} \quad (6)$$

and for sufficiently large N we have

$$G = \int_{\Omega} f(X) dX = \int_{\Omega} \frac{f(X)}{p(X)} p(X) dX \approx G_N \pm \sqrt{\frac{\sigma^2}{N}} \quad (7)$$

where

$$\sigma^2 = \left\langle \frac{f(X)^2}{p(X)^2} \right\rangle - \left\langle \frac{f(X)}{p(X)} \right\rangle^2 \quad (8)$$

The idea of importance sampling is to choose $p(X)$ so that the quotient $f(X)/p(X)$ becomes constant. Exactly then σ^2 will disappear. The value of G_N is then independent of X , and it is sufficient with $N = 1$ to evaluate G_N exactly. Of course there is a snag in it. To make $f(X)/p(X)$ constant means choosing $p(X) = \frac{1}{G} f(X)$. That is: to draw X according to $p(X)$ requires the knowledge of the integral we want to calculate. Practically one chooses a distribution $p(X)$ that is easily generated and that follows the variation of $f(X)$ as close as possible, or one produces the X by a technique called dependent sampling.

Dependent sampling The idea of dependent sampling is to produce not a set of independent values X_1, X_2, \dots, X_N , but to use a state X_i and modify into a new state X_{i+1} in what is called a Monte Carlo step. The rule to generate next point X_{i+1} given the point X_i is described by the transition probability $W(X_i \rightarrow X_{i+1})$.

Consider an ensemble of such chains with different start points X_0 distributed according to some function $v_0(X)$. Let $v_j(X)$ be the distribution of the j -th points in each chain. The transition probability $W(X_i \rightarrow X_{i+1})$ is chosen to satisfy:

$$W(X \rightarrow X') \geq 0 \quad (9)$$

$$\sum_X W(X' \rightarrow X) = 1 \quad (10)$$

$$\sum_X W(X \rightarrow X') p(X) = p(X') \quad (11)$$

$$W^n(X \rightarrow X') > 0 \quad n \in \mathbb{N}, \quad \forall X, X' \quad (12)$$

The condition (11) means that $p(X)$ is an eigenvector with eigenvalue one to the matrix W . When all the conditions (9) to (12) are valid, the Frobenius theory of positive matrices assures that the norm of the other eigenvalues will be less than one. The distribution $v_i(X)$ will therefore converge to the distribution $p(X)$ in the sense that

$$\lim_{i \rightarrow \infty} \sum_X |v_i(X) - p(X)| = 0. \quad (13)$$

A sufficient condition for (11) is “detailed balance”:

$$W(X \rightarrow X') p(X) = W(X' \rightarrow X) p(X'). \quad (14)$$

In practice detailed balance plays an important role, as the proof of detailed balance often is simple for a suggested algorithm.

Ensemble averages The Monte Carlo method is often used in equilibrium thermodynamics to obtain approximate values for the partition function or for thermal averages for a finite system using a Markov chain of phase-space points instead of integrating over the whole phase space [Binder, 1987].

If we use importance sampling to estimate the thermal average of some observable $A(X)$ we get

$$\overline{A(X)} = \frac{\sum_{i=1}^N A(X_i) \exp(-\frac{H(X_i)}{kT}) / p(X_i)}{\sum_{i=1}^N \exp(-\frac{H(X_i)}{kT}) / p(X_i)} \quad (15)$$

where H is the Hamiltonian of the system and k the Boltzmann constant. By choosing $p(X_i) \propto \exp(-\frac{H(X_i)}{kT})$, this become

$$\overline{A(X)} = \frac{1}{sN} \sum_{i=1}^N A(X_i) \quad (16)$$

This result will be valid only if we can draw X_i independently from $p(X_i)$. In the case of dependent sampling, $X_i, X_{i+1}, X_{i+2}, \dots$ will be correlated, and obviously the result (16) will not be valid.

By introducing a time, the correlations between $X_i, X_{i+1}, X_{i+2}, \dots$ can be thought of as time correlations for a trajectory in phase space.

Define a time scale such that one transition $X_i \rightarrow X_{i+1}$ is performed in unit time and think of the configuration at step i in the Markov chain as the configuration appearing at time $t = i$.

As we have made the configuration $X_i = X(t = i)$ a function of time, we can think of the observable $A(X(t)) = A(t)$ as a function of system time itself. So we can equally well think of (16) as an estimate for the time average of the observable

$$\overline{A(X)} = \frac{1}{N} \int_{t=0}^{t=N} A(t) dt \quad (17)$$

Algorithms for dependent sampling. There are several well known algorithms for dependent sampling. One simple and general is the metropolis algorithm [Metropolis et al., 1953]. Here the transition probability $W(X \rightarrow X')$ is divided into two parts. Given X the new point X' is proposed with probability $T(X, X')$. The proposal is then accepted with probability $A(X, X')$.

$$W(X \rightarrow X') = A(X, X')T(X, X') + (1 - A(X, X'))\delta(X, X') \quad (18)$$

Of the acceptance probability is required that

$$A(X, X') = \min(1, \frac{p(X')}{p(X)}) \quad (19)$$

In equilibrium thermodynamics we want the distribution to converge to the Boltzmann distribution, and the acceptance probability becomes

$$A(X, X') = \min(1, \exp(-\frac{H(X') - H(X)}{k_B T})) \quad (20)$$

where $H(X)$ is the Hamiltonian of the system, k_B Boltzmann's constant and T the temperature.

Further we require from the proposal probability that

$$\sum_{X'} T(X, X') = 1 \quad (21)$$

$$T(X, X') = T(X', X) \quad (22)$$

Another algorithm common in equilibrium thermodynamics is the Glauber algorithm [Glauber, 1963]. This differs from the Metropolis algorithm only through the choice of the acceptance probability which for the Glauber algorithm is

$$A(X', X) = \frac{1}{2} \left(1 - \tanh\left(\frac{H(X') - H(X)}{2k_B T}\right) \right) \quad (23)$$

Detailed balance and the conditions (9),(10) are easily confirmed for the Metropolis and the Glauber algorithm. When designing the proposal probability $T(X, X')$ care is to be taken that also the ergodic condition (12) is satisfied. Then it is guaranteed that the probability distribution $v_i(X)$ will eventually converge to the distribution $p(X)$ as $i \rightarrow \infty$.

2.3.1 Dynamical interpretation

With the time scale introduced above, the dependent sampling Markov chain can be interpreted as a trajectory of the system in phase space. Instead of calculating time averages of the system when it has reached equilibrium, we are now interested in the average behavior of its dynamics. By producing several Markov chains by dependent sampling, we can simulate the dynamics of an ensemble of trajectories in phase space.

If $\nu(X, t) = \nu_i(X)$ is the probability of the system being in configuration X at time $t = i$, then

$$\nu(X, t+1) = \nu(X, t) - \sum_{X'} W(X \rightarrow X') \nu(X, t) + \sum_{X'} W(X' \rightarrow X) \nu(X', t) \quad (24)$$

and the average behavior of the system thus simulated, evolves according to the master equation

$$\frac{\Delta \nu(X, t)}{\Delta t} = - \sum_{X'} W(X \rightarrow X') \nu(X, t) + \sum_{X'} W(X' \rightarrow X) \nu(X', t) \quad (25)$$

A physical system need not necessarily obey a master equation as in (25). In fact there is in general little similarity between the artificial stochastic dynamics described by (25) and the actual dynamics of a system, although they lead to the same thermal equilibrium distribution.

For the kinetics of a physical process to be a Markov process, the transitions between sub-states of the system have to be rare. Rare in this case means that the event of transition from one state X to some other state X' is so infrequent that the system when entering a transition bears no previous transitions in memory. When the mean time between transitions is orders of magnitude larger than the time scale of atomic vibrations, then the phonons can be approximated as a heat bath which induces random transitions in the system.

In this study we assume the dynamics to show no memory effects. We call this the Markov hypothesis.

There is experimental support for the conception that protein folding may be a Markov process. For example Li and Scheraga in 1987 managed to fold [*Met*⁵] *enkephalin* using Monte Carlo simulation with transitions between local minimum conformations [Li & Scheraga, 1987]. For the folding of larger proteins there are on the other hand some reports of memory dependence [Privalov, 1979].

The master equation Given the Markov hypothesis we can justify a master equation description of the dynamics. But we are still not able to say anything more specific about the transition probabilities than that they must lead to the experimentally observed long time limits.

We will here outline the derivation of the master equation from the quantum mechanics that governs the microscopic dynamics with the ambition to make the underlying assumptions and approximations explicit.

The following account will closely parallel the the treatment in [van Kampen, 1992], [van Kampen, 1954], [Kreuzer, 1981], and [Pauli, 1928].

First we notice that all our measurements of some property of a protein will be afflicted with inaccuracy. We will therefore never be able to determine the microscopic state of the system with such precision that we can make a straight forward use of the deterministic equations of quantum mechanics. If the derivation of a master equation shall make any sense, the states X in (25) must therefore refer to macroscopic states.

Following van Kampens lead, we start with the construction of macroscopic states and macroscopic commuting operators of in general non commuting observables. As noted above, any measurement of some property A can only be made with a certain precision ΔA . We therefore divide the spectrum of accessible values of A into small cells A_1, A_2, \dots . A measurement only tells us within which cell A_n the value of A lies. This formation of cells can be made for any macroscopic property, but we will now focus on the energy of a system. The accuracy of an energy measurement is ultimately bounded according to Heisenberg's uncertainty principle

$$\Delta E \gg \frac{\hbar}{\tau} \quad (26)$$

where ΔE is the inaccuracy of the obtained value and τ is the duration of the measurement. As we can not distinguish between the micro states having energies in the range ΔE from the obtained value, we treat them as belonging to the same macro state. Instead of using the microscopic Hamiltonian operator $\hat{H} = \sum_i E_{n_i} |n_i\rangle \langle n_i|$, we construct a macroscopic operator $\{\hat{H}\} = \sum_n E_n |n\rangle \langle n|$, where the summation covers all energy cells E_n and where $|n\rangle \langle n| = \sum_{i=1}^{N_n} |n_i\rangle \langle n_i|$ with the sum including all N_n eigenstates of \hat{H} having energies in the range $E_n \pm \Delta E$.

Now we pay attention to some operator \hat{A} , which does not commute with \hat{H} . Heisenberg's uncertainty principle demands that

$$\delta E \delta A \geq \frac{1}{2} \left| \langle [\hat{H}, \hat{A}] \rangle \right| \quad (27)$$

where

$$(\delta E)^2 = \left| \langle \hat{H} - \langle \hat{H} \rangle \right|^2 \quad (28)$$

$$(\delta A)^2 = \left| \langle \hat{A} - \langle \hat{A} \rangle \right|^2 \quad (29)$$

and the brackets $\langle \rangle$ denotes the quantum mechanical expectation value: $\langle \hat{A} \rangle = \text{Tr}(\hat{A}\hat{\rho})$, where $\hat{\rho}$ is the density matrix. In the representation of the eigenstates of the Hamiltonian operator, the commuter $[\hat{H}, \hat{A}]$ has the matrix elements

$$[\hat{H}, \hat{A}]_{ij} = \langle n_i | [\hat{H}, \hat{A}] | n_j \rangle = (E_{n_i} - E_{n_j}) \langle n_i | \hat{A} | n_j \rangle = (E_{n_i} - E_{n_j}) A_{ij} \quad (30)$$

As far as orders of magnitude are concerned, one may therefore write

$$(E_{n_i} - E_{n_j}) A_{ij} \sim \delta E \delta A \quad (31)$$

We further observe that for a macroscopic measurement

$$\Delta E \Delta A \gg \delta E \delta A \quad (32)$$

and we can thus conclude

$$\Delta E \Delta A \gg (E_{n_i} - E_{n_j}) A_{ij} \quad (33)$$

Now we choose the two states so that $(E_{n_i} - E_{n_j}) \simeq \Delta E$ and obtain $\Delta A \gg A_{ij}$. This means that the matrix elements A_{ij} connecting two states having an energy difference of measurable size, are much smaller than the accuracy of measure ΔA and can therefore be neglected.

The matrix A_{ij} of an operator \hat{A} is thereby reduced to a strip of non zero elements along the main diagonal. Although most of the non zero matrix elements A_{ij} will have both states $|n_i\rangle$ and $|n_j\rangle$ in the same energy cell, there will be some non zero matrix elements A_{ij} connecting different energy cells. Our first crucial assumption is that these matrix elements can be neglected. We then have a matrix representation for \hat{A} divided into sub matrixes where every sub matrix corresponds to a particular energy cell.

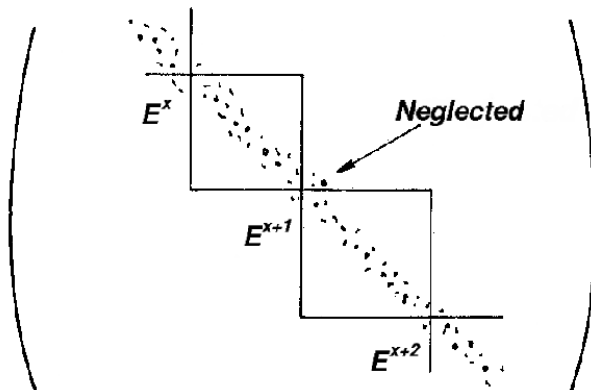


Figure 5: A slowly varying operator in the energy representation. The elements decrease with increasing distance from the main diagonal.

Now perform a unitary transform of the representation in each energy cell so that the new representation of the observable \hat{A} becomes diagonal in each energy cell. The diagonalize operator of energy cell E_n will be denoted \hat{A}_n and its eigenvalues $A_{n,s}$. We now split up this eigenvalue spectrum in cells of size ΔA , just in the same fashion as we did with the energy spectrum. Next we set all eigenvalues $A_{n,s}$ in the same cell equal to some average value.

The operator corresponding to the matrix obtained by this procedure will be denoted $\{\hat{A}\}$ and is the macroscopic operator of the observable A .

The procedure described above, can be repeated over and over again with other observables B, C, D, \dots ordered according to decreasing inaccuracy $\Delta B, \Delta C, \Delta D, \dots$ if only the condition corresponding to (32) holds true. But for each new observable included, the neglect of diagonal matrix elements connecting different cells, gets more and more questionable.

When all observable relevant in our macroscopic system have been included, we have a coarse grained description of the system. Each cell then corresponds to a macroscopic state, and all the relevant observables are simultaneously measurable. We label these cells with x .

To each cell there corresponds many microscopic states. For all the microscopic states $|x, i\rangle, i = 1, \dots, N_x$ constituting the cell x , a macroscopic observable $\{\hat{O}\}$ will

have the same eigenvalue $\{\widehat{O}\} |x, i\rangle = O_x |x, i\rangle$.

Our next concern are the dynamic properties of the coarse grained system. We assume that our system is exposed to some perturbation, such as the coupling to a heat bath. The microscopic Hamiltonian of our system plus perturbation is $\widehat{H}_{tot} = \widehat{H} + \lambda \widehat{\Phi}(t)$, where λ is the coupling strength and $\widehat{\Phi}(t)$ is the operator of the perturbation.

Consider an arbitrary microscopic state $|\psi(t)\rangle$ and expand it in the eigenstates of the unperturbed Hamiltonian:

$$|\psi(t)\rangle = \sum_{x_i} C_{x_i}(t) |x_i\rangle = \sum_x \sum_{i=1}^{N_x} C_{x,i}(t) e^{-iE_x t/\hbar} |x, i\rangle \quad (34)$$

The expansion coefficients $C_{x,i}(t)$ must fulfill the set of differential equations [Kreuzer, 1981]

$$i\hbar \frac{\partial}{\partial t} C_{x,i}(t) = \sum_{x'_i} \langle x, i | \widehat{\Phi}(t) | x', i' \rangle \exp\left[\frac{i(E_x - E_{x'})t}{\hbar}\right] C_{x',i'}(t) \quad (35)$$

This system of equations was solved in first order perturbation theory by Pauli [Pauli, 1928], who found that for times $t \gg \frac{\hbar}{|E_x - E_{x'}|}$

$$C_{x,i}(t) = \sum_{x', i'} \langle x, i | \widehat{\Phi}(t) | x', i' \rangle C_{x',i'}(t=0) \quad (36)$$

where

$$\langle x, i | \widehat{\Phi}(t) | x', i' \rangle = \sum_{x_i} \langle x, i | x_i \rangle e^{-iE_{x_i} t/\hbar} \langle x_i | x', i' \rangle \quad (37)$$

The expectation value of an operator $\{\widehat{A}\}$ is then given by

$$\langle \psi(t) | \{\widehat{A}\} | \psi(t) \rangle = \sum_x A_x \sum_i |C_{x,i}(t)|^2 = \sum_x A_x P_x(t) \quad (38)$$

where $P_x(t) = \sum_{i=1}^{N_x} |C_{x,i}(t)|^2$ is the probability of finding the system in cell x. Inserting (38) gives

$$P_x(t) = \sum_{x', i', x'', i''} \left(\sum_i \langle x, i | \widehat{\Phi}(t) | x', i' \rangle \langle x, i | \widehat{\Phi}(t) | x'', i'' \rangle \right) C_{x',i'}(t=0) C_{x'',i''}^*(t=0) \quad (39)$$

This expression for the probabilities $P_x(t)$ is still not what we want. To calculate $P_x(t)$ according to (39) all $C_{x,i}(t=0)$ have to be known. That is we would need exact knowledge of the initial microscopic state.

To come further we have to make some new assumptions. First we observe that in the first sum of (39), the diagonal terms with $i'=i''$ and $x'=x''$ are real and non negative, whereas all off diagonal elements are complex. We now assume that the off diagonal elements are so spread out in the complex plane, that they will soon cancel out. The only significant contribution that remains, is the one from the diagonal terms in (39).

$$P_x(t) = \sum_{x', i'} \left(\sum_i \left| \langle x, i | \widehat{\Phi}(t) | x', i' \rangle \right|^2 \right) |C_{x',i'}(t=0)|^2 \quad (40)$$

Next we make the assumption that the probability density inside cell x'_i is uniformly spread over the whole cell. We can thus make the substitution

$$|C_{x',i'}(t=0)|^2 \rightarrow \frac{\sum_{i'} |C_{x',i'}(t=0)|}{N_{x'}} = \frac{P_x(t=0)}{N_{x'}} \quad (41)$$

in equation (40), and we get

$$P_x(t) = \sum_{x', i'} \left(\sum_i \left| \langle x, i | \widehat{\Phi}(t) | x', i' \rangle \right|^2 \right) \frac{P_x(t=0)}{N_{x'}} = \sum_{x'} T_{x, x'}(t) P_{x'}(t=0) \quad (42)$$

where

$$T_{x, x'}(t) = \frac{1}{N_{x'}} \sum_{i, i'} \left| \langle x, i | \widehat{\Phi}(t) | x', i' \rangle \right|^2 \quad (43)$$

can be interpreted as the probability of the system to be in state x at time t , given that the system was in state x' at time 0.

The assumption that the off diagonal elements in (39) cancel in times of order τ , and that the substitution (41) is justifiable for the distribution at $t=0$, led to

$$P_x(\tau) = \sum_{x'} T_{x, x'}(\tau) P_{x'}(t=0) \quad (44)$$

If the same assumptions are justifiable also at time τ , then we can repeat the same arguments and we get

$$P_{x''}(2\tau) = \sum_x T_{x'', x}(\tau) P_x(t=\tau) \quad (45)$$

and if we insert (44) for $P_x(\tau)$ we get

$$P_{x''}(2\tau) = \sum_{x, x'} T_{x'', x}(\tau) T_{x, x'}(\tau) P_{x'}(t=0) = \sum_{x'} [\mathbf{T}^2(\tau)]_{x'', x'} P_{x'}(t=0) \quad (46)$$

where $\mathbf{T}^2(\tau)$ is the second power of the matrix $\mathbf{T}(t)$ in the sense of matrix multiplication.

In the same manner as above, if we can assume the probability densities to be uniformly distributed within each phase cell at all times $t = m\tau$, we end up with

$$P_x(m\tau) = \sum_{x'} [\mathbf{T}^m(t)]_{x, x'} P_x(t=0) \quad (47)$$

This is the master equation in a slightly different form than (25), but we get the equation (25) if we take $\Delta t = \tau$ and define the transition probabilities $W(x' \rightarrow x)$ by demanding

$$T_{x, x'} = \delta_{x, x'} (1 - \tau \sum_{x''} W(x' \rightarrow x'')) + \tau W(x' \rightarrow x) \quad (48)$$

At the times $t = m\tau$ repeating the assumption of the probability densities being uniformly distributed within each phase cell, is equivalent with assuming the validity of the Chapman-Kolmogorov equation

$$T_{x, x'}(t_1 + t_2) = \sum_{x''} T_{x, x''}(t_2) T_{x'', x'}(t_1) \quad (49)$$

The physical interpretation of (49) is that if one at $t=0$ takes an ensemble with constant density in the phase space cell x' and zero density outside, and measures the fraction of the ensemble that is found in some state x at time $t = t_1 + t_2$. Then, (49) states, one gets the same result as if one had interrupted the evolution of the system at some intermediate time $t = t_1$ and redistributed the density within each phase cell, so that it became uniform distributed in each cell, and then let the system continue to evolve until the time $t = t_1 + t_2$. The intermediate redistribution within each phase cell should not affect the final fractions.

The Chapman-Kolmogorov equation is equivalent with the master equation (25), so we have in some sense assumed precisely what we wanted to derive. Hopefully the above treatment have made the assumption of the Markov hypothesis more transparent.

We can further conclude from the above treatment, that quantum mechanics can not give us any precise information about the transition probabilities.

We do know that the probability density should in the long time limit approach the Boltzmann distribution, and after some time the local accessible degrees of freedom should change in accordance with the Boltzmann weights of the corresponding conformations.

At the beginning of real protein folding, the transition probabilities may not obey the Boltzmann distribution. But if we assume that folding is a Markov process, the initial non equilibrium character of the transition probabilities should be forgotten after a long enough time since protein folding occurs on a time scale much longer than that needed for local equilibrium.

But there is still some ambiguity in the choice of transition probabilities even when the detailed balance condition is enforced.

It has been indicated that the random energy model gives different relaxation asymptotic depending on the details of the probability of transition between different states, even though they obey detailed balance [Koper & Hilhorst, 1987].

3 Aim of the present study

Many different approaches have been followed to gain insight to protein folding. The results have often been difficult to interpret, and there is an ongoing discussion about the physical correctness of the models used, and their relevance for proteins.

Our primary purpose is not to answer questions about the nature of protein folding, but to examine the validity of some models used for protein folding investigations. We will devote ourself to lattice Monte Carlo simulations.

In lattice models one makes many artificial simplifications and it is not obvious what are artifacts of the models and what are common features of the models and the folding of real proteins. We try to investigate how consistent the results from lattice Monte Carlo simulations are when we switch the model.

In all kinetic Monte Carlo simulations we assume the dynamics to be Markovian. We make no investigation of this underlying hypothesis, but the "Markov assumption" is not sufficient to specify the dynamics of the modeled proteins. We also need to know the ease of transitions between connected states, the transition rates. However due to the microscopic origin of the connectedness and the transition rules, it becomes difficult to account for these parameters correctly. We therefore have used different assumptions about the transition rates and examined how the choice affects the results.

There are many questions about protein folding to be answered and we leave several important aspects out of this investigation. In particular we say very little about the energy functions used. The energy function itself constitutes an important and difficult problem well worthy investigating, but we think the way to go is one step at a time.

4 Methods

To easier the comparison with earlier studies, we have in large parts used the same terminology and methods as [Sali et al., 1994].

4.1 The models

The models we have compared are all lattice models. We have represented proteins as chains of interaction centers with each amino acid residue represented by one bead. We number the beads along the chain from 0 to $N-1$, and denote by \mathbf{r}_n the position of the n th bead. The beads are restricted to move on a three dimensional lattice with no more than one bead occupying the same lattice point. The conformation of a protein is specified if we know either the set $(\mathbf{r}_0, \mathbf{r}_1, \dots, \mathbf{r}_{N-1})$ of bead positions or the set $(\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_{N-1})$ of bond vectors, where $\mathbf{a}_n = \mathbf{r}_n - \mathbf{r}_{n-1}$ is the n th bond vector.

The lattices we have investigated are the simple cubic (sc) lattice and the face centered cubic (fcc) lattice. On the sc lattice the bond vectors can take values that are any permutation of $(0, 0, \pm 1)$. Thus on the sc lattice every lattice point has six nearest neighbors. On the fcc lattice every lattice point has twelve nearest neighbors and the bond vectors can be all permutations of $(0, \pm 1, \pm 1)$.

On both the sc and the fcc lattice we have investigated the effect of allowing the chains to do different types of local moves, distinguishing between single bead and double bead moves. We define a single bead move for bead n to be the most general move which change the position of this bead but leave the positions the other beads unchanged. Inside the sc lattice this only means a corner flip, whereas there are three different types of single point moves inside the chain on the fcc lattice. On the sc lattice we define two point moves to be all possible moves inside the chain where two neighboring beads gets new positions and the coordinates of all other beads stay unchanged. Thus a two point move on the sc lattice is equivalent with a 90 or 180 degree rotation of a crankshaft. The beads n to $n+3$ are said to form a crankshaft if the distance between bead n and bead $n+3$ is 1.

On the fcc lattice there are several more possibilities to change the coordinates of only two beads. We could define two point moves just as we did for the sc lattice, but then we would include moves where the chain could pass through itself, see figure 6. We excluded all two point moves where this is geometrical possible and ended up with two cases when we allowed the beads $n+1$ and $n+2$ to make a two point move. These are the cases when the beads n and $n+3$ are separated the distance $\sqrt{2}$ or $\sqrt{4}$.

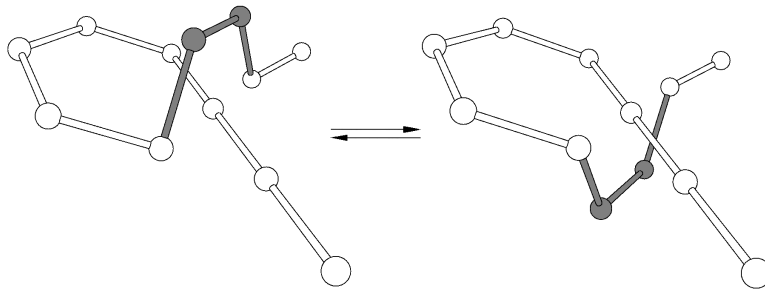


Figure 6: Example of a two point move where the chain passes through itself.

We also examined the effect of changing the transition probabilities of the Monte Carlo simulations. We compared the Metropolis algorithm and the Glauber algorithm, which are both functions of the difference in energy between the initial conformation and end conformation of the transition. It has been suggested that one should expect different characteristics of the kinetics when using different transition probabilities [Pitard & Orland, 1998].

Both the Metropolis and the Glauber algorithm involves the selection of a proposed transition. The proposal is then accepted with some probability. In the Metropolis algorithm every transition that lowers the energy is accepted with probability one,

whereas in the Glauber algorithm the transition gets more probable the more the transition lowers the energy. Furthermore the acceptance probability of the proposal is a smooth function for the Glauber algorithm but not for the Metropolis algorithm.

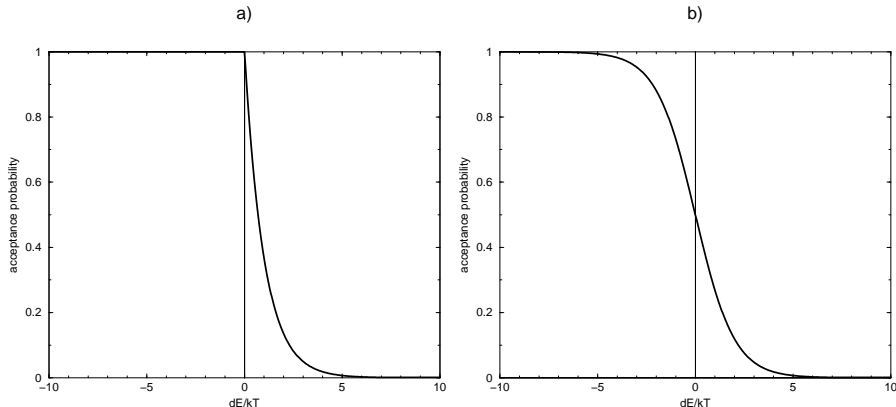


Figure 7: The acceptance probability as function of the energy difference for a) the Metropolis algorithm and b) the Glauber algorithm.

In the simulations we have looked for correlations between kinetic properties of the modeled proteins and properties of their energy landscapes. The energy landscapes are defined for every amino acid sequence through the set $\{B_{i,j}\}$ of possible interaction energy strengths for all pairs of beads i and j . We will call one such set a sequence.

The Hamiltonian of the model proteins is set to be

$$H = \sum_{i,j} B_{i,j} \Delta_{i,j} \quad (50)$$

where $\Delta_{i,j}$ is equal to 1 if the beads i and j are geometrical nearest neighbors on the lattice but are not neighbors along the chain, otherwise $\Delta_{i,j}$ is 0. We will call beads for which $\Delta_{i,j} = 1$ to be in contact.

The interaction strengths $B_{i,j}$ are chosen as Gaussian numbers which are centered around a negative mean to provide an overall attraction, emulating the hydrophobic effect of real proteins. This model was introduced on sc lattices by Shakhnovich and Gutin [Shakhnovich & Gutin, 1990 (b)].

On the sc lattice we choose the polymer length to be $N=27$, and on the fcc lattice $N=19$. Thus total number of self-avoiding conformations on the sc lattice are approximately $(6-1)^{(27-1)} \approx 10^{18}$, and on the fcc lattice approximately $(12-1)^{(19-1)} \approx 10^{19}$. This means that in both cases we are unable to investigate every conformation to identify the global minimum. Levinthal's paradox is present.

To recognize when the global minimum is reached in our simulations, we need to know the conformations with the minimum energies. This can be achieved by letting the average of the contact energies $B_{i,j}$ be sufficiently negative. Then the global minimum is in general found among the conformations which have the largest possible number of contacts [Sali et al., 1994]. We define these conformations as maximal compact. These are few enough to be exhaustively searched.

The value required of the average $B_{i,j}$ to ensure that the global energy minimum with a high probability is found among the maximal compact configurations, might be unrealistic low for proteins [Sali et al., 1994].

On the sc lattice with $N=27$ the maximal compact conformations are the self-avoiding walks on a $3 \times 3 \times 3$ cube, and have 28 contacts. There are 103,346 such walks

unrelated by symmetry [Shakhnovich & Gutin, 1990 (a)].

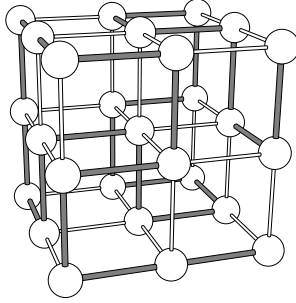


Figure 8: Example of maximal compact 27-bead self-avoiding path on the sc lattice.

On the fcc lattice the maximum number of contacts for a chains of length $N=19$ is 42. This is obtainable on two geometrically different sub lattices and in total 32,971,430 different maximal compact conformations unrelated by symmetry are possible.

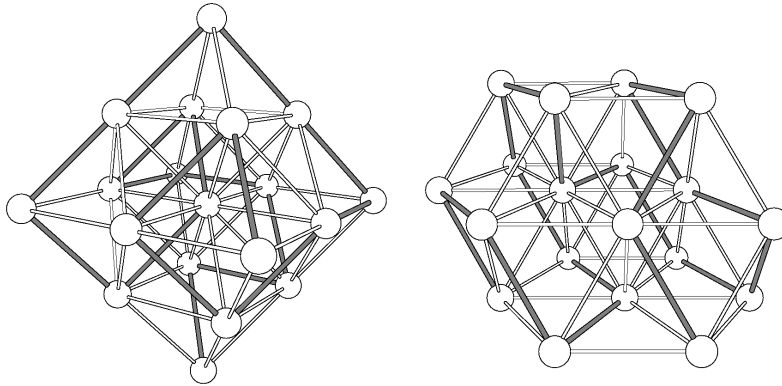


Figure 9: Examples of self-avoiding walks on the two maximal compact 19 bead sub lattices of the fcc lattice.

The scheme we used to generate all possible maximal compact configurations was to first select a symmetric unique starting path on one of the sub lattices where the maximal number of contacts can be obtained. In the cases considered by us, this means one of the lattices shown in figure 8 and figure 9. We then let the starting chain grow by choosing the next step from the current available nearest neighbors on the lattice. Finally we repeated the second step until all lattice points of the sub lattice were occupied once, or until a dead end was reached. The generation was stopped when all possible combinations of steps had been considered. This procedure was repeated for all symmetric unique starting paths possible.

4.2 Parameters and definitions

We will denote the conformation which corresponds to the global energy minima of a sequence as its native state. In our simulations we investigate which sequences find their native state within a reasonable time and which do not. The fraction of simulations where a sequence found its native state will be called the foldicity, and we define folding sequences as those with a foldicity 0.4 or larger.

The results obtained in the Monte Carlo simulations depends strongly on the temperatures used. A typical dependence of foldicity on temperature is shown in figure 10. The lower the temperature gets, the easier the chain can get trapped in

local minima, and thus hindered in finding the native state. But when the temperature is higher, the chain will spend less time in its native state at equilibrium. Therefore there is a trade off between at which conditions a chain is likely to be thermodynamic stable in its native state and at which conditions it is able to fold to its native state. We found some sequences that did not seem to fold at any temperature, but these were so rare that we could draw no conclusions about them.

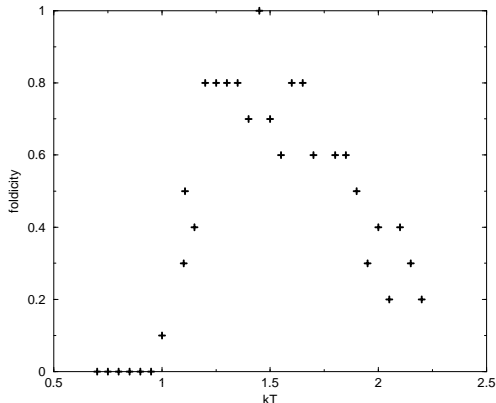


Figure 10: The foldicity at different temperatures for one sequence on the fcc lattice. The qualitative behavior is typical and common to most sequences on both the fcc and the sc lattice.

[Sali et al., 1994] argued that for a sequence to be defined as a folding sequence, it should not only be able to fold to its native state under some conditions, but it must be able to do so under the specific conditions when the native conformation is thermodynamic stable. They therefore selected the simulation temperature unique for each sequence as the temperature T for which the expression

$$X(T) = 1 - \sum_i \left(\frac{e^{-H_i/kT}}{\sum_j e^{-H_j/kT}} \right)^2 \quad (51)$$

has the value of 0.8. The summations were done for all maximal compact conformations. This choice of temperature was motivated as the highest temperature for which the ground state still thermodynamic dominates the distribution of states reasonable much. $X(T)$ is an order parameter that is approximately 1 if many states have comparable Boltzmann weights, and approaches 0 when one state largely dominates the Boltzmann distribution [Sali et al., 1994],[Shakhnovich & Gutin, 1990 (b)].

The above described choice of temperature has been criticized by among others [Unger & Moulton, 1996]. Sali et al. came to the conclusion that strongly folding sequences are those which have a large gap between the lowest and second lowest energies of the sequence. Ungler and Moulton showed that this energy gap is strongly correlated with the temperature fulfilling the criteria $X(T)=0.8$. They argued that the relation between foldicity and energy gap found by Sali et al depends rather on the choice of simulation temperatures than on the energy gaps themselves.

We have performed simulations at a broad range of temperatures. As the simulation results depends strong on the temperatures used and we wanted to compare the results between themselves, we needed some unambiguous way to choose the simulation temperatures. We have therefore specially investigated the temperatures for which $X(T)=0.8$. Although this choice is somewhat arbitrary, we believe our comparisons between the different models to be correct as $X(T)$ does not depend on the dynamics of the underlying model.

Also the tradeoff between high foldicity and a low value of $X(T)$ look quite alike when compared for two strong folding sequences on the sc and fcc lattice. This is shown in figure 11.

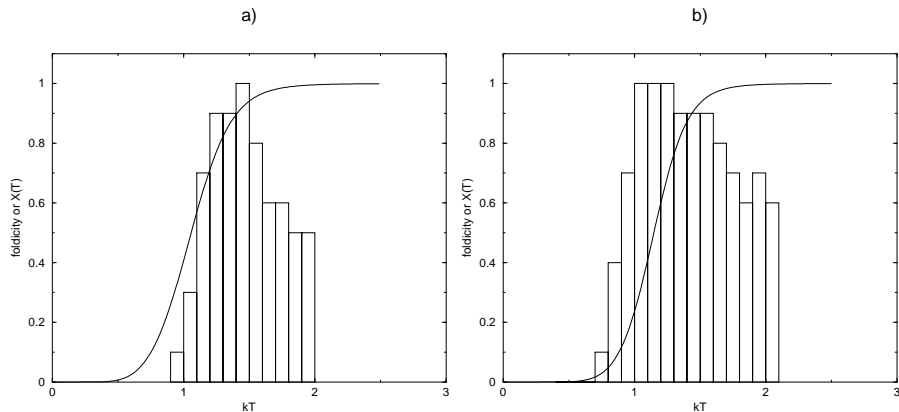


Figure 11: The histograms show the foldicity at various temperatures, and the curve show $X(T=0.8)$ on the same scale. a) sc lattice. b) fcc lattice

In figure 12 the frequencies of temperatures for which $X(T)=0.8$ is shown for 200 sets $\{B_{i,j}\}$ for chains of length $N=27$ on the sc lattice and for chains of length $N=19$ on the fcc lattice.

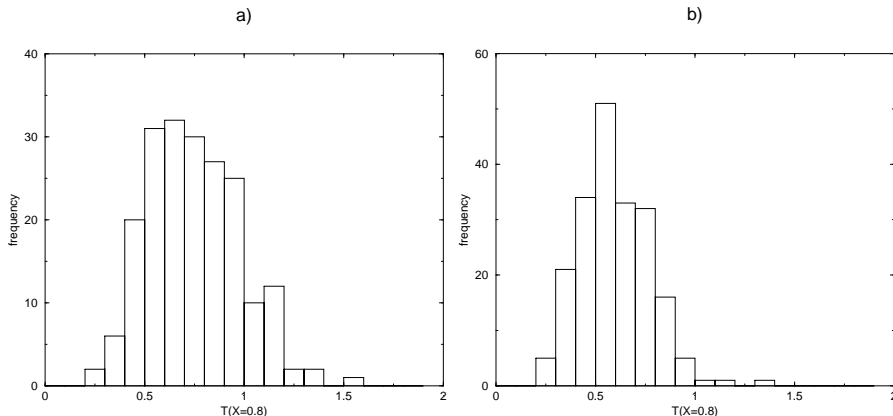


Figure 12: The frequency of temperatures which fulfilled the criteria $X(T)=0.8$ for 200 random sequences: a) on the sc lattice. b) on the fcc lattice.

In the simulations where both one and two point moves were allowed, we tried to construct two point moves four times more often than one point moves. This proportion was used by [Sali et al., 1994], and motivated by the fact that two point moves are likely to result in double occupancy. Proposing a single point move only every fifth time results in a ratio of one point versus two point moves tested for the acceptance criterion, of approximately 5. The fraction of proposals accepted are about 8% and 0.5% for one point moves respectively two point moves. We found approximately the same relationship on the fcc lattice as well, and therefore used the same proportions as Sali et al.

The simulations were interrupted when the native state was found or when the sequence had failed to find the native state within a reasonable time. With reasonable time is meant approximately three times the average folding time for the folding sequences. In the Metropolis and Glauber simulations with both one and two point

moves this means that we interrupted the simulations if they had not led to the native state of the simulated sequence after 50,000,000 Monte Carlo steps. The dynamics when only one point moves were allowed, was considerable slower. Therefore these simulations were allowed to continue for 500,000,000 Monte Carlo steps.

4.2.1 Discrimination measures

As quantitative measures for how well we can discriminate between folding and non folding sequences, we will use specificity, sensitivity and and Mathews' correlation coefficient.

Mathews' correlation coefficient is defined as

$$C_{Mathews} = \frac{(P^t N^t) - (P^f N^f)}{\sqrt{(N^t + N^f)(N^t + P^f)(P^t + N^f)(P^t + P^f)}}$$

where P^t is the number true positives (folding sequences that were correctly identified by the discriminator to be folding), P^f is the number of false positives (the number of non folding sequences that were falsely identified as folding), N^t denotes the number of true negatives (the number of non folding sequences that correctly classified of the discriminator as non folding), and N^f is the number of false negatives (the number of folding sequences that were falsely classified as non folding).

$C_{Mathews}$ equals 1 if the discriminator can separate the folding and non folding sequences perfectly. The worse the classification gets, the more $C_{Mathews}$ decreases. $C_{Mathews}$ becomes zero for a total random guess and -1 for the worst possible prediction [Mathews, 1975].

Specificity and sensitivity are defined as

$$sens = \frac{P^t}{P^t + N^f}$$

and

$$spec = \frac{P^t}{P^t + P^f}$$

Sensitivity thus gives the fraction of the folding sequences that were classified as folding by the discriminator. The specificity is the fraction of the sequences predicted to be folding, that actually were folding.

For a perfect separation all three measures equals one.

5 Results and discussion

5.1 Comparing polymer dynamics on sc lattice versus fcc lattice, allowing for one and two point moves

In our first study, we tried to reconstruct the investigation made by [Sali et al., 1994]. On the sc lattice and on the fcc lattice we repeated the simulations performed by Sali et al on the sc lattice. The simulations on both lattice types showed the same tendency as Sali et al had observed, but slightly less pronounced because we did not find as many strong folding sequences as Sali et al. did. Among the 200 sequences they investigated, they found 30 sequences with a foldicity 0.4 or larger, whereas we only found 9 folding sequences on the sc lattice and 16 on the fcc lattice (see figure 13).

Sali et al claimed, that a sequence will be folding if and only if the gap between the two lowest energies of the sequence is large [Sali et al., 1994]. Indeed in our simulations we can also discriminate between most of the folding and non folding

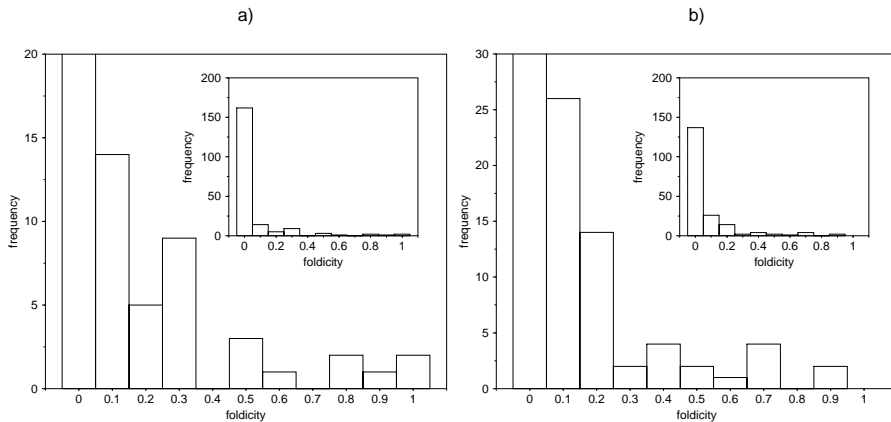


Figure 13: The total fold frequency for the 200 investigated sequences. a) On the sc lattice. b) On the fcc lattice.

Table 1: Discrimination performances when separating the folding and non folding sequences with an energy gap cut-off and with a simulation temperature cut-off.

	$C_{Mathews}$	sensitivity	specificity
sc lattice, energy gap separation	0.69	1	0.5
fcc lattice, energy gap separation	0.80	0.92	0.73
sc, simulation temperature separation	0.74	0.67	0.86
fcc, simulation temperature separation	0.95	0.92	1

sequences by using an energy gap as cut-off and even better by using a simulation temperature cut-off (see table 1 and figure 14).

Although the foldicity increases with energy gap and with temperature, this dependence is very weak. Figure 14 shows only non folding and folding sequences. The total dependence of foldicity on energy gap and temperature is shown in figure 15 and 16.

But, just as Sali et al, we found no dependence of the foldicity on the number of local interactions found in the native state, neither could we find any significant difference between the sc and the fcc lattice. The frequency of contacts for a given separation along the chain of the contact pair is shown in figure 17.

As reviewed above, earlier studies have suggested that the foldicity of a hetero polymer might depend on its ability to form local interactions. This we have not investigated.

We could not detect any significant difference in the dynamic simulations on the two lattices. We take this as a first indication of that the exact geometric representation of a protein is not crucial in Metropolis Monte Carlo simulations.

As we will report below, this view is also supported by the fact that we found no qualitative difference between the sc and the fcc lattice when using other models.

5.2 Investigating how the absence of two point moves affects the dynamics

One of the difficulties that one encounters with when trying to perform Monte Carlo simulations of proteins using off-lattice models, is how to define local moves within the chain. It is therefore of interest to investigate the difference between simulations using different types of local moves.

Because of shortage of time, we did not investigate these models using as many

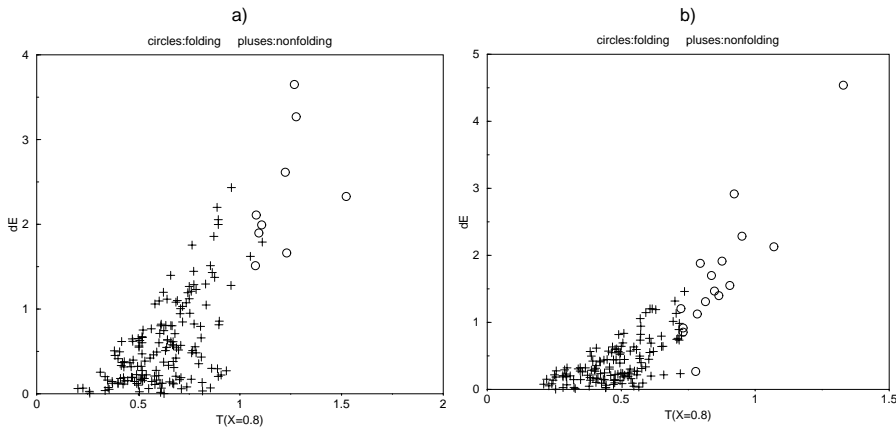


Figure 14: Simulation temperature and energy gap for folding (circles) and non folding (pluses) sequences. Simulations performed on the sc lattice (a) and on the fcc lattice (b).

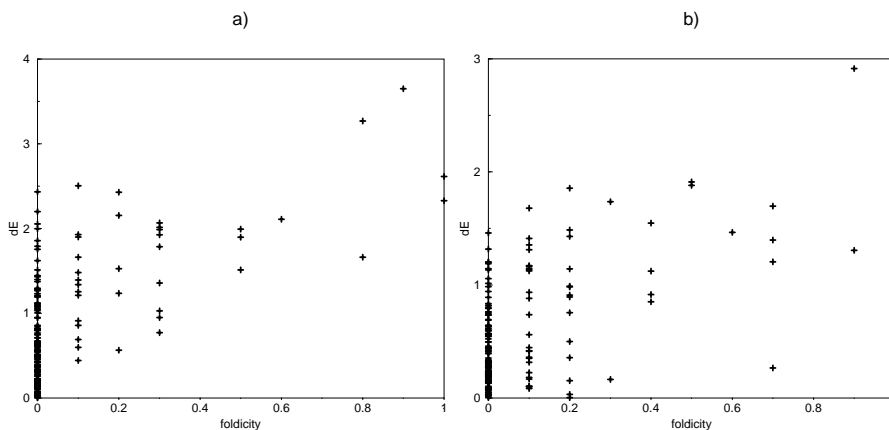


Figure 15: Foldicity as function of gap size between the two lowest energies of the sequences. a) Simulated on the sc lattice. b) Simulated on the fcc lattice.

sequences as we did in the first study. The significance of the above results, was very much limited by that we found so few sequences with high foldicity. We therefore performed these simulations, and the next ones we will report of, on a set of only 20 sequences, where we have included the ten sequences on the sc respectively on the the fcc lattice that had highest foldicity in the first study. The rest of the sequences we choose by random among the weak and non folding sequences.

When sequences folded with only one point moves, they needed on average ten times longer before they found their native state than the sequences in the above described simulations. To make sure we did not just observe a rescaling of the time, we allowed all simulations to last ten times longer than when also using two point moves.

To certify that extending the simulations ten times was sufficient, on both the sc lattice and the fcc lattice the simulations of two sequences that folded good when two point moves were allowed, but not when restricted to one point moves, were continued another 500,000,000 Monte Carlo steps. This did not change their foldicity to a significant extent.

In 1975 H. J. Hilhorst and J. M. Deutch performed Monte Carlo simulations of polymer chain relaxation on the sc lattice [Hilhorst & Deutch, 1975]. They found that

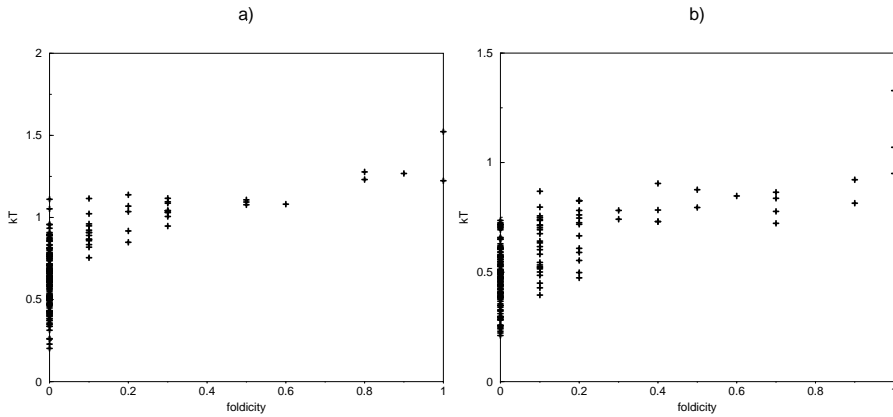


Figure 16: Foldicity as function of simulation temperature on a) the sc lattice, and b) on the fcc lattice.

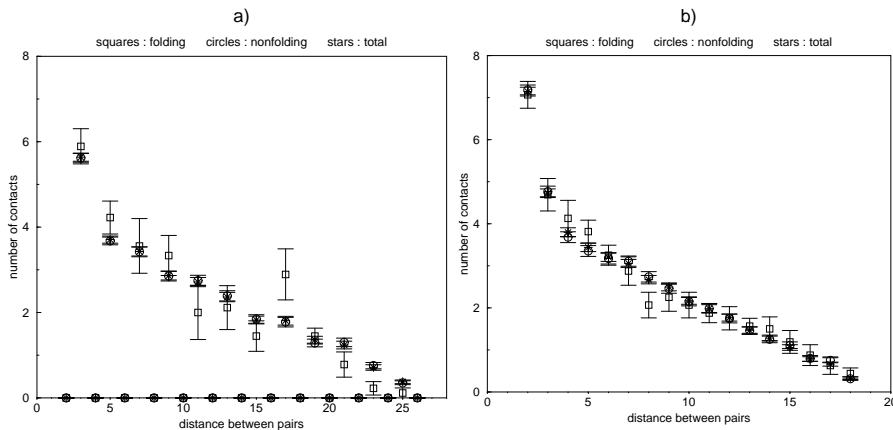


Figure 17: The figure shows the number of contacts in the native conformations as function of the distances along the chain between the residues of the contact pair. The error bars show the mean deviation from the average for folding sequences (squares), non folding sequences (circles) and totally for all simulated sequences. a) On the sc lattice. b) On the fcc lattice.

only allowing one point moves on the sc lattice restricted the model in an unrealistic way, thereby slowing down the dynamics considerably. Therefore Sali et al included two point moves in their simulations [Sali et al., 1994].

If a subchain constitutes a geometrical maximum in one of the bond directions on the sc lattice, and only one point moves are allowed, then the geometrical maximum can only be removed if a series of one point moves diffuses all the way from the subchain to the end of the chain. Geometrical extrema cannot disappear or be created except at the chain ends. The length of a subchain that constitutes a local extrema can grow or decrease, and the extrema can move within the chain. But a maximum and a minimum along the same direction in the lattice can not pass each other, but will stay disjoint and maintain their order along the chain.

These arguments for why a model with only one point moves should be unrealistic, does not apply to the fcc lattice. Nevertheless, we could observe no significant qualitative difference between the two lattice types.

First we can note that sequences that were strong folders in the previously reported simulation, not always had high foldicity when only one point moves were allowed.

The foldicity of the 20 simulated sequences in the two different models are shown in figure 18.

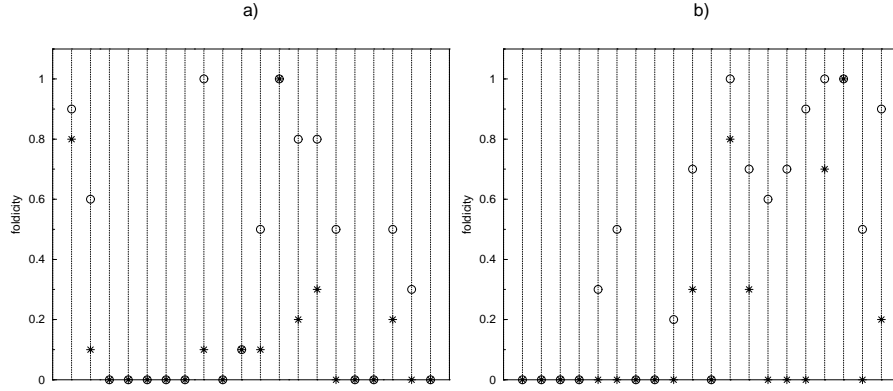


Figure 18: The plots show the foldicity when one and two point moves are used in the simulation (circles) and when only one point moves are used (stars). Every vertical line represents one sequence. a) On the sc lattice. b) On the fcc lattice.

We see that the sequences in general folded worse when no two point moves were present, but the same dependence of foldicity on temperature and energy gap can be seen as in the above study. As can be seen in figure 19, one can separate perfectly between the folding and non folding sequences with both an energy gap cut-off and with a simulation temperature cut-off. The foldicity as function of energy gap and as

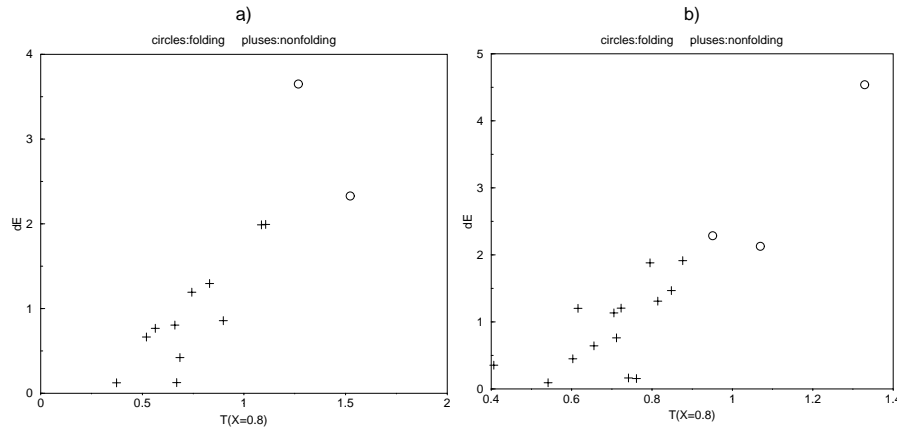


Figure 19: Simulation temperature and energy gap for folding (circles) and non folding (pluses) sequences. Simulations performed on the sc lattice (a) and on the fcc lattice (b).

function of simulation temperature is shown in figures 20 and 21.

Despite the fact that we get the same qualitative dependence of foldicity on temperature and energy gap for both the simulations with and without two point moves, we conclude that several sequences who folded in the first case, had a low foldicity in the second. We interpret this as an argument for that the dynamics of lattice simulations is sensitive to which local moves are allowed.

The fact that the behavior is similar on the fcc and the sc lattice can be seen as another indication of that the underlying geometric representation does not affect the dynamics noticeable.

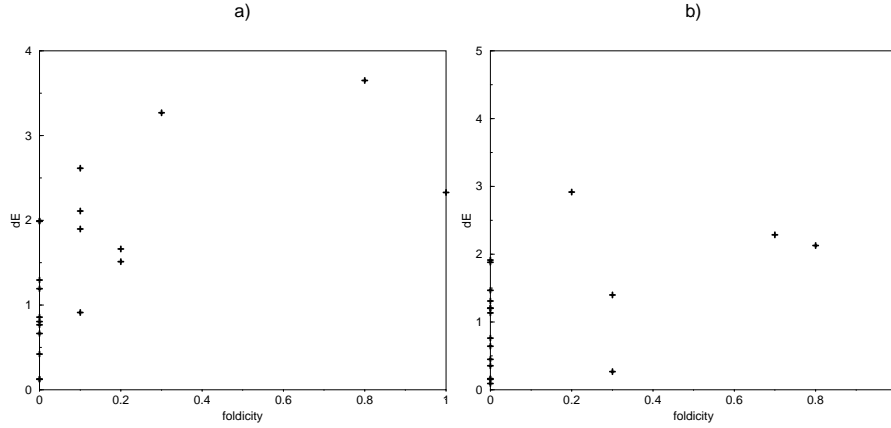


Figure 20: Foldicity as function of gap size between the two lowest energies of the sequences. a) Simulated on the sc lattice. b) Simulated on the fcc lattice.

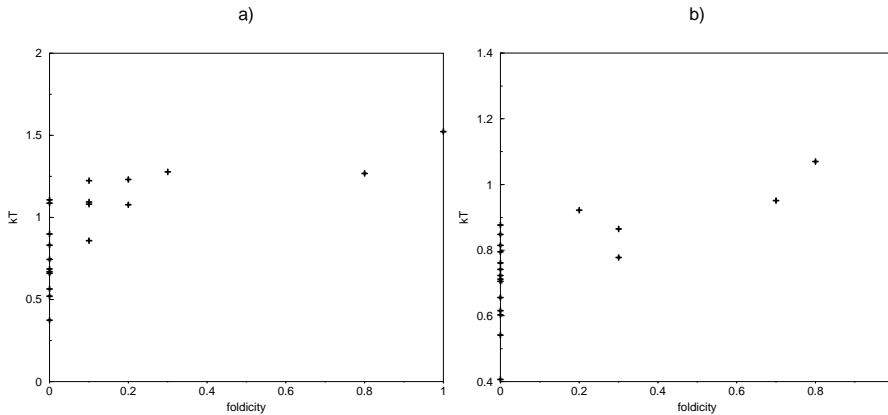


Figure 21: Foldicity as function of simulation temperature on a) the sc lattice, and b) on the fcc lattice.

5.3 Comparing the dynamics of the Metropolis algorithm versus that of the Glauber algorithm

Our third set of simulations we devoted to the transition probabilities of the Monte Carlo algorithm. On each lattice type, we used the same set of 20 sequences as mentioned above. We performed simulations with the Glauber algorithm [Glauber, 1963] and compared the results with the Metropolis simulations. The foldicity of the 20 sequences is shown in figure 22.

The plots of the results corresponding to those shown for the Metropolis simulation above, are shown in figures 23, 24 and 25. We observe no difference in the qualitative behavior, neither between the Glauber simulations and the Metropolis simulations, nor between the results of the Glauber simulations on the sc lattice and the fcc lattice. We can separate perfectly between folding and non folding sequences on both lattice types by using a simulation temperature cut-off. On the sc lattice the separation is also perfect when we use an energy gap cut-off, whereas we get $C_{Mathews}=0.88$, sensitivity=0.89 and specificity=1 on the fcc lattice.

We take the likeness of the results for the two simulation algorithms as yet another indication of that the dynamics of the simulations is not very sensitive to slight variations of the model.

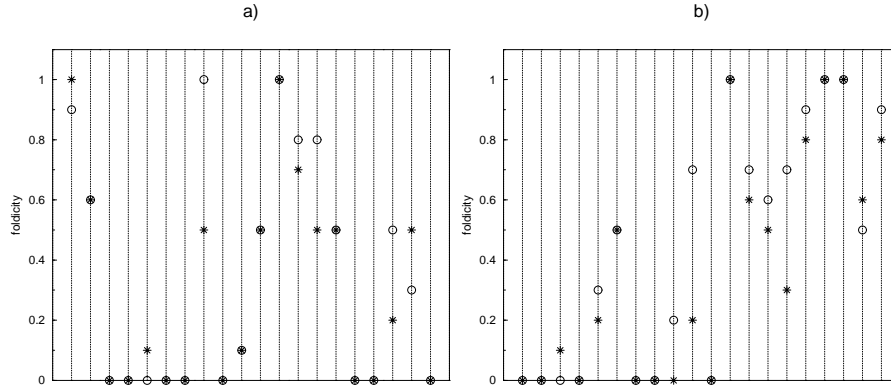


Figure 22: The plots show the foldicity when the dynamics is simulated using the Metropolis algorithm (circles) and when the Glauber algorithm were used (stars). Every vertical line represents one sequence. a) On the sc lattice. b) On the fcc lattice.

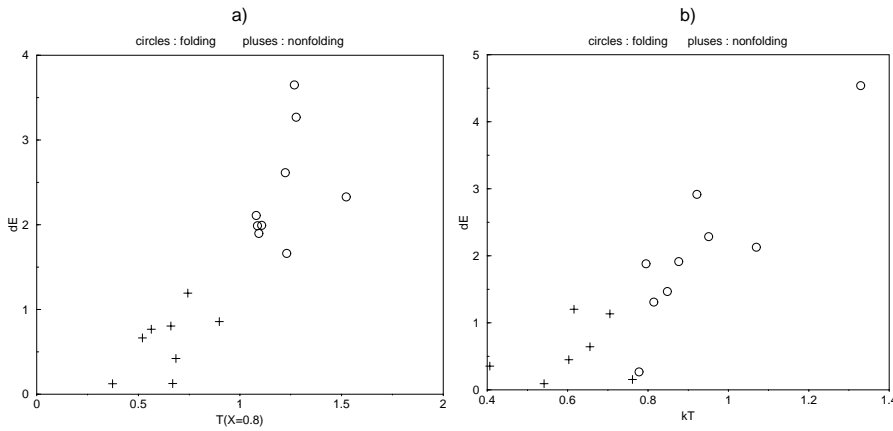


Figure 23: Simulation temperature and energy gap for folding (circles) and non folding (pluses) sequences. Simulations performed on the sc lattice (a) and on the fcc lattice (b).

5.4 Dead end attempts

One could easily imagine other transition probabilities than those used by the Metropolis and Glauber algorithms, that still would fulfill detailed balance. We have tried to perform Monte Carlo simulations with an algorithm that differs from the Metropolis and Glauber algorithms in that the acceptance probability depends mainly of $H(X)$, the energy of the initial state of the transition:

$$A(X \rightarrow X') \propto e^{H(X)/k_B T} \quad (52)$$

It has been stated that such an transition probability should give rise to a different dynamics than seen in our simulations [Pitard & Orland, 1998]. A problem when simulating with this transition probabilities, is that the exponent $H(X)/k_B T$ spans over two orders of magnitude as the conformation X varies from a random coil to a more compact state. We thus get very slow dynamics for a broad range of conformations.

Further we have tried a complete different approach, and by Monte Carlo simulations attempted to estimate the ease with which sequences can reconfigure between different maximal compact configurations. We started the simulations in the maximal compact configuration with the second lowest energy and investigated which

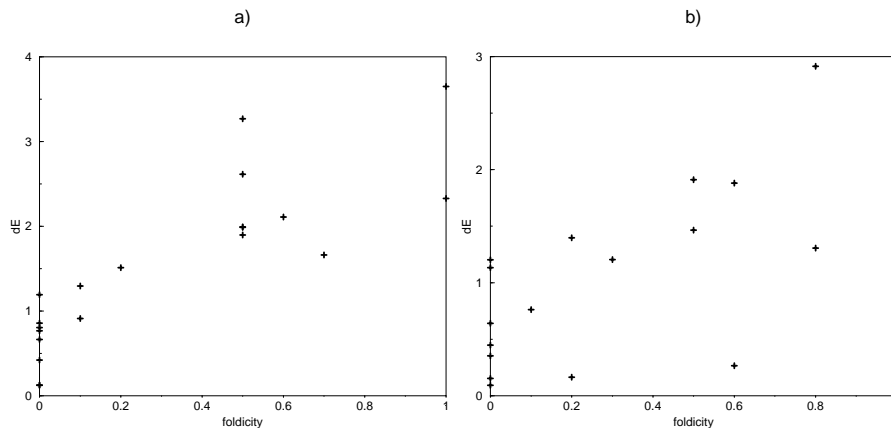


Figure 24: Foldicity as function of gap size between the two lowest energies of the sequences. a) Simulated on the sc lattice. b) Simulated on the fcc lattice.

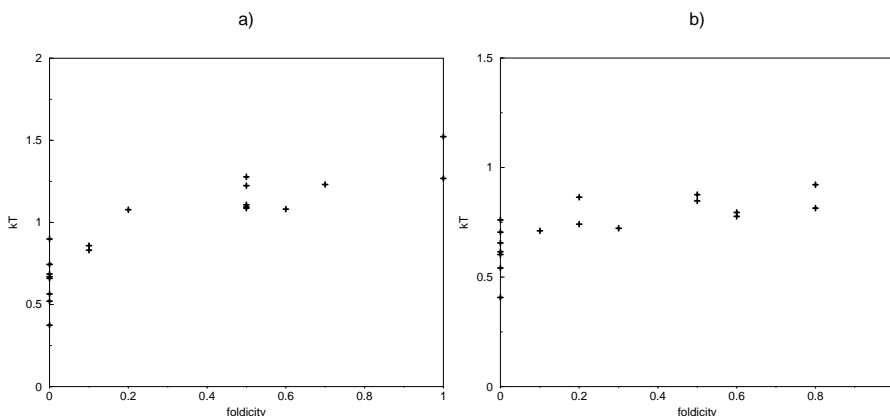


Figure 25: Foldicity as function of simulation temperature on a) the sc lattice, and b) on the fcc lattice.

sequences that found the lowest energy conformation. This simulation used a broad range of temperatures in the hope that we could find some common property among the sequences which could reconfigure without passing through a totally unfolded conformation. No such property was found.

5.5 Conclusions

We conclude from our investigations, that the dynamic of lattice simulations of proteins is not very sensitive to the details of the lattice model being used. However, which local moves that are allowed in a simulation seem to have a qualitatively affect on the results.

Exactly which set of local moves that would provide a model realistic dynamics, is a matter of further investigation. Also the influence of choosing other algorithms for the Monte Carlo simulations, is a topic that could turn out to be worth further studies.

In all, we see little qualitative variation in the results from our different simulations. We take this as a support for the idea that conclusions drawn from simple lattice simulations may be valid for real proteins.

6 Acknowledgments

I would like to thank Clas Blomberg, Olle Edholm and Lars Sandberg at the department of Theoretical Physics at The Royal Institute of Technology, for stimulating and clarifying discussions during my work, and Hans Öhman for proofreading this report. I also in particular wish to thank my supervisor Arne Elofsson at the department of Biochemistry at Stockholms University for discussions, ideas, for proofreading, and for being continuously helpful, Erik Lindahl at the department of Theoretical Physics at The Royal Institute of Technology, who has implemented a great part of the simulation programs I have used, and assisted with ideas and discussions, and Erik Wallin at the department of Biochemistry at Stockholms University for valuable help with computer related problems.

References

- [Ansari et al., 1985] A. Ansari, J Berendzen, S.F Bowne, H. Frauenfelder, I. E. T. Iben, T. B. Sauke, E. Shyamsunder, and R. D. Young: Proc. Natl. Acad. Sci. USA, 1985, 82, 5000-5004
- [Binder, 1987] K. Binder: Applications of the Monte Carlo Method in Statistical Physics (Topics in Current Physics, Vol 36), 1987, Springer Verlag
- [Brooks et al., 1988] C. L. III Brooks, M. Karplus, and B. M. Pettitt: Proteins: A Theoretical Perspective of Dynamics, Structure and Thermodynamics, Adv. Chem. Phys. LXXI, 1988, New York, John Wiley & Sons
- [Chan & Dill, 1993] H. S. Chan and K. A. Dill: J. Chem. Phys., 1993, 99, 2116
- [Crippen & Ohkubo, 1998] G. C. Crippen and Y. Z. Ohkubo: Proteins, 1998, 32, 425-473
- [Bryngelson & Wolynes, 1987] A. D. Bryngelson and P. G. Wolynes: Proc. Natl. Acad. Sci. U.S.A, 1987, 84, 7524-7528
- [Bryngelson et al., 1995] J. D. Bryngelson, J. N. Onuchic, N. D. Socci, and P. G. Wolynes: Proteins, 1995, 21, 167-195
- [Dill, 1990] K. A. Dill: Biochemistry, 1990, 29, 31: 7133
- [Dill et al., 1995] K. A. Dill, S. Bromberg, K. Yue, K. M. Fiebig, D. P. Yee, et al.: Protein Sci., 1995, 4, 561-602
- [Elber & Karplus, 1987] R. Elber and M. Karplus: Science, 1987, 235, 318-321
- [Friere & Biltonen, 1978] E. Friere and R. L. Biltonen: Biopolymers, 1978, 17, 463
- [Glauber, 1963] R.J. Glauber: Journal of Mathematical Physics, 1963, 4: 294
- [Gurd & Rothgeb, 1979] F. R. N. Gurd and T. M. Rothgeb: Adv. Protien Chem., 1979, 33, 73-165

- [Hilhorst & Deutch, 1975] H. J. Hilhorst and J. M. Deutch: *J. Chem. Phys.*, 1975, 63, 12, 5153-5161
- [van Kampen, 1954] N. G. van Kampen: *Physica*, 1954, XX, 603-622
- [van Kampen, 1992] N. G. van Kampen: *Stochastic processes in Physics and chemistry*, 1992, Amsterdam, Elsevier Science Publisher B.V
- [Klimov & Thirumalai, 1996] D. Klimov and D. Thirumalai: *Physicar Review Letters*, 1996, 76, 21, 4070-4073
- [Koper & Hilhorst, 1987] G. J. M. Koper and H. J. Hilhors: *Europhys. Lett.*, 1987, 3, 1213
- [Kreuzer, 1981] H. J. Kreuzer: *Nonequilibrium Thermodynamics and its Statistical Foundations*, 1981, Oxford, Clarendon Press
- [Leopold et al., 1992] P. E. Leopold, M. Montal, and J. N. Onuchic: *Proc. Natl. Acad. Sci. U.S.A.*, 1992, 89, 8721
- [Levinthal, 1969] C. Levinthal: *Mossbauer Spectroscopy in Biological Systems*, P. De-Brunner, J. Tsibris, E. Munck (eds). Urbana, IL: University Of Illinois Press, 1969, p. 22-24
- [Li & Scheraga, 1987] Z. Li and H. A. Scheraga: *Proc Natl Acad Sci USA*, 1987, 84, 6611-6615
- [Mathews, 1975] B. Mathews: *Biochim. Biophys. Acta*, 1975, 405, 442-451
- [Metropolis et al., 1953] Metropolis et al: *J. Chem. Phys.*, 1953, 21, 1087
- [Park & Levitt, 1995] B. H. Park and M. Levitt: *J. Mol. Biol.*, 1995, 249, 493-507
- [Pauli, 1928] W. Pauli: *Probleme der modernen Physik*, 1928, Verlag, Leipzig, p. 30-45
- [Pitard & Orland, 1998] E. Pitard, H Orland: pre print, 1998, cond-mat/9811252
- [Privalov, 1979] P. L. Privalov: *Adv. Protein Chem.*, 1979, 33, 167-241
- [Privalov, 1982] P. L. Privalov: *Adv. Protein Chem.*, 1982, 35, 1
- [Rooman et al., 1991] M. J. Rooman, J. A Kocher, and S. J. Wodak: *J. Mol. Biol.*, 1991, 221, 961-979
- [Sali et al., 1994] A. Sali, E. Shakhnovich and M. Karplus: *J. Mol. Biol.*, 1994, 235, 1614-1636
- [Santoro & Bolen, 1988] Santoro and Bolen: *Biochenistry*, 1988, 27, 8063
- [Shakhnovich & Gutin, 1990 (a)] E. I. Shakhnovic and A. M. Gutin: *J. Chem. Phys.*, 1990, 93, 5967-5071

- [Shakhnovich & Gutin, 1990 (b)] E. I. Shakhnovic and A. M. Gutin: *Nature*, 1990, 346, 773-775
- [Unger & Moulton, 1996] R. Unger and J. Moulton: *J. Mol. Biol.*, 1996, 259, 988-994