# Unit-Vector RMS (URMS) as a Tool to Analyze Molecular Dynamics Trajectories

**Klara Kedem, Paul Chew and Ron Elber***

Department of Computer Science

Cornell University

4130 Upson Hall, Ithaca NY 14853

* corresponding author

phone 607-255-7416

fax 607-255-4428

e-mail ron@cs.cornell.edu

short title: Analysing trajectories with URMS

key words: root mean square distance, structure comparison, folding, C peptide, order parameter

## Abstract

The Unit-vector RMS (URMS) is a new technique to compare protein chains and to detect similarities of chain segments. It has a number of unique features that include exceptionally weak dependence on the length of the chain and efficient detection of substructure similarities. Two molecular dynamics simulations of proteins in the neighborhood of their native states are used to test the performance of the URMS. The first simulation is of a solvated myoglobin and the second is of the protein MHC. In accord with previous studies the secondary structure elements (helices or sheets) are found to be relatively, moving rigidly among flexible loops. In addition to these tests, folding trajectories of C peptides are analyzed, revealing a folding nucleus of seven amino acids.

# 1. Introduction

Molecular Dynamics (MD) simulations provide useful data on protein function by following the time evolution of protein structures [1]. The computed atomically detailed information can explain mechanisms of reaction and activation. However, the excessive data is difficult to analyze, necessitating the use of a wide range of tools to find the most interesting, unexpected events.

It is therefore not surprising that many different techniques have been used to study the protein shapes that are created by MD simulations. Examples of analysis tools vary from direct visualization of the dynamics, to coordinate RMS, contact matrices, and studies of a specific coordinate as a function of time (e.g. a torsion angle or a distance of prime interest) [2]. The above approaches, successful and extremely helpful in numerous cases, are far from optimal in the detection of *substructures* in the masses of protein data. It is the detection of substructures which is the focus of the present paper.

Classification and identification of substructures during dynamics of proteins is useful in a number of cases. Proteins are never uniformly rigid; rigid substructures are moving in a “fluid” of flexible loops and side-chains [3]. The structural fluctuations in the neighborhood of the native state have been shown to be relevant to ligand diffusion through a protein matrix, to active site adjustments, and to other functionally important motions. These fluctuations keep the global topology unchanged and cause relatively small and rigid shifts of secondary structure elements. If the rigid and

semirigid domains of a protein can be characterized, such identification can potentially lead to the design of MD algorithms with a reduced number of degrees of freedom and thus, to simpler representation of biomolecular dynamics.

In this paper we demonstrate how the *Unit-vector Root Mean Square distance (URMS)* provides new perspectives on folding processes and structural similarity problems. We use the URMS to (1) explore protein motions in the neighborhood of the native state and (2) identify significant substructures as they occur in a peptide folding-trajectory. We chose the topic of protein motions in the neighborhood of the native state because this topic has been studied extensively in the past using some of the various analysis tools mentioned above. As we argue below and as has been argued in a previous publication [9], the URMS provides a more natural measure for the problem at hand.

For a folding trajectory, we start with the dynamics of an unfolded chain. The set of structures consists of a "sea" of conformations that change rapidly. In the unfolded state it is unlikely that a persistent secondary substructure will be formed and observed. Nevertheless when the peptide starts to fold some secondary structure is initiated and is detectable. In the C peptide—the molecule that we study—the secondary structure is an $\alpha$-helix. We wish to address mechanistic questions such as: Where does the helical substructure start to form? Is it at the beginning of the chain or in the center? What is the necessary length for such a helix to be reasonably stable?

This last question can be recast as a search for a *post-critical nucleation* [4]. Initial small structures that are formed during the folding process are in general unstable. There are many events of rapid dissociation and reformation as a function of time in a folding trajectory. A critical size of a structural element is reached after which the growth continues rapidly with a strong sense of direction. The nucleus that leads to this kind of rapid "downhill" structure formation is called a *post-critical nucleation.* How large is the substructure at the critical size? Folding trajectories of C peptide are examined with the URMS measure to answer this and other questions.

In the following sections we (1) describe the protocols used to generate the Molecular Dynamics trajectories, (2) explain how the URMS measure is calculated, and (3) show examples of how the URMS can be used to analyze the MD simulation data. In particular, we use the URMS to visualize rigid substructures of myoglobin and MHC (Major Histocompatibility Complex) as they fluctuate about their native state and to determine a post-critical nucleation and other properties of folding trajectories of C peptide.

## 2. The Molecular Dynamics Protocols

For our myoglobin and MHC data, we used straightforward Molecular Dynamics simulations to collect data in the neighborhood of their native state. For the C peptide, we have employed a novel simulation methodology that generated molecular dynamics trajectories with a very large time step [5]. The new methodology enables us to collect significant statistics on folding trajectories of C peptide.

All the Molecular Dynamics computations described in the present manuscript were performed using the program MOIL [6]. The MOIL suite of programs uses the AMBER-OPLS force field [7] to perform energy minimization and dynamic simulations. It is available (free source code) for a variety of platforms. Three Molecular Dynamics studies were performed:

A. The first was a "high quality" computation of a molecular dynamics trajectory of myoglobin. A sperm whale myoglobin with an (unbound) CO molecule at the pocket was solvated in a box of 2320 water molecules. An $SO_4^{2-}$ ion, found in the crystallographic studies, was added to the system and two sodium ions were also included to ensure system neutrality. The dimensions of the box were $50 \times 45 \times 50 \overset{o}{\text{A}}$. The cutoff distance for van der Waals interactions was $9 \overset{o}{\text{A}}$ and the Particle Mesh Ewald was used to account for long-range electrostatic interactions. All the bonds were constrained to their equilibrium values using the SHAKE algorithm, and the constraint matrix was solved explicitly for the water molecules. The time step was 1 femtosecond for a total computation of 500 picoseconds of trajectory time. Structures were saved each 1000 steps (a picosecond) for further analysis. Myoglobin is a highly helical protein, and one of the goals of the simulation was to test the capacity of the URMS in identifying the relevant rigid substructures.

B. The second simulation was of the protein MHC. The structure includes a bound peptide (Gly-Ile-Leu-Gly-Phe-Val-Phe-Thr-Leu) with no water solvation. As in the previous simulation all the bonds were fixed at their equilibrium position, the time step was 1 femtosecond, and 500 structures were saved for analysis from the 500-picosecond simulation. The cutoff distances for non-bonded interactions were

$10\mathring{A}$ for van der Waals, and $14\mathring{A}$ for electrostatic interactions. MHC has both --$\beta$ sheets and $\alpha$ helices; both are expected to be relatively rigid (compared to loops). Our expectation was that the URMS should detect the above mentioned substructures.

C. The third simulation was of the helix formation in C peptide. The simulation is described in detail elsewhere [5]. In brief, C peptide is a fragment of the protein Ribonuclease A. It is known to have a significant tendency to form an $\alpha$-helix even with the absence of the rest of the protein [8]. The C peptide was placed in a box of size $54 \times 33 \times 28\mathring{A}$ with 1376 water molecules. The long dimension was design to accommodate the extended conformation of this peptide. In the stochastic path approach, which was used to compute the folding trajectory, the end points of the trajectory are fixed and a plausible trajectory between the initial and final state is determined. The trajectory is computed by global optimization techniques starting from an initial guess. The protocol is therefore very different from the usual computations of Molecular Dynamics trajectories that solve initial value differential equations. An advantage of the optimization protocol is the filtering of all the high frequency motions with a period shorter than the time step. The filtering makes it possible to use large time steps, larger by orders of magnitude compared to integration steps employed in straightforward Molecular Dynamics computations. In the simulations of C peptide folding, the time step was 500 picoseconds as compared to 0.001 picoseconds in ordinary Molecular Dynamics simulations. Of course the evaluation of a single conformation along the trajectory is considerably more expensive in the optimization protocol. To make the optimization cost effective, it is also necessary to use significantly

sparser time representation. Twelve trajectories of 17 nanoseconds each, leading from an extended chain to a folded conformation are analyzed.

## 3. The URMS Measure

The structural similarity measure we apply here is the *Unit-vector RMS distance (URMS)* (also called the *orientation RMS*). This measure was originally suggested and discussed in Chew et al. [9] where the URMS was used to find large common 3D substructures in pairs of proteins.

**Definition of the URMS:** Consider a protein described by its sequence of α-carbons. Rather than encoding the coordinates of these atoms, we use an orientation-based representation of the protein: for each successive pair of α-carbons along the backbone chain, we record the unit vector in the direction from α-carbon $i$ to α-carbon $i + 1$. Note that because the separation of these successive atoms is essentially a fixed distance ($3.84\,\text{Å}$), this representation captures the structure of the backbone. By chaining the unit vectors end to end, we obtain the standard model of a protein as a sequence of α-carbons in space. Alternatively, we can place all of the unit vectors at the origin; and the backbone is thus mapped into vectors in the unit sphere. For the URMS distance between two proteins *A* and *B*, we first compute their representations *U* and *V* as unit vectors mapped to the unit sphere. We then determine the rotation (rotation only, no translation) that minimizes the sum of the squared distances between corresponding unit vectors. The square root of resulting minimum sum is defined as the URMS distance between the original proteins *A* and *B*.

Note how this differs from the standard RMS distance between *A* and *B*. For the standard RMS distance, we first align the centers of mass of *A* and *B* and then compute the rotation that minimizes the sum of the squared distances between corresponding points of *A´* and *B* (*A´* is the translated and rotated version of *A*). The RMS distance is then the square root of the resulting minimum sum.

A related representation of the protein chain was proposed by Rackovsky and Goldstein [10]. However, the analysis of structure similarity they propose is purely local and is based on discrete differential geometry in terms of the virtual bond vectors.

The URMS distance offers a number of advantages over the standard RMS distance. Advantages include:

- The URMS is insensitive to outliers. Consider imaginary proteins *A* and *B*, which are identical except for a straight "pole" sticking outward. In *B*, the angle of this pole is a few degrees different from the angle of the pole in *A*. Depending on the length of the pole, the minimum RMS distance between *A* and *B* can be arbitrarily large, while the URMS distance between *A* and *B* is quite small, regardless of the pole's length. For the URMS the maximum squared distance between any two-unit vectors is 4, thus no small portion of a protein can have a large effect on the URMS distance. As a result, the URMS exhibits a natural resilience to the presence of outliers.

- The URMS weighs all portions of the protein equally. The standard minimum RMS distance has the property that points far from the center of mass are, in effect, weighted more heavily than points near the center of mass. Consequently, a common core shared by two structures is difficult to detect using the minimum RMS distance. This is problematic in the case of proteins, where some crucial similarities occur near the center of mass.

- The minimum URMS has an upper bound (independent of *n*, the number of α-carbons) and does not grow significantly as the length of the protein. For the minimum RMS distance, there is no upper bound on the distance between two proteins. It has been observed [11] that the minimum RMS grows as the length of the protein increases. Thus there is no absolute size of RMS distance for which we consider proteins to be similar. There have been a number of methods to compensate for this observation [11,12]. Analysis (see [9]) and experiment (see Figures 4 and 5) indicate that the value of the URMS distance for proteins does not depend strongly on the length of the compared chain. The analysis in [9] shows that the expected minimum URMS distance between two random sets of *n* unit vectors is $\sqrt{2-\frac{2.84}{\sqrt{n}}}$. Thus, for randomly generated chains of length above 10 residues, the expected URMS is between 1.1 and 1.4. Of course protein shapes are far from random, so the URMS values that we get are even smaller than those predicted for random sequences and the dependence of the URMS on *n* is significantly weaker than for the RMS.

- Another advantage of the URMS is its computational efficiency. For the applications presented here, this advantage is not really significant since both

methods run in time linear in the input size. For other applications, such as searching for common substructure in two proteins, the URMS can be more efficient than the standard RMS distance because proteins can be compared using the URMS in time $O(n \log n)$ by using the FFT (Fast Fourier Transform), while the standard RMS requires $O(n^2)$ time (see [13]).

## 4. Applying the URMS

We present four methods using the URMS to visualize and analyze MD simulation data:

1. We use the URMS to visualize the structural fluctuations of proteins in the neighborhood of their native state. This technique uses an $\alpha$-helix and a $\beta$-strand (a relatively straight, single-strand portion of a $\beta$-sheet) as probes to determine how close portions of a protein are to the shapes of these secondary structures.
2. We analyze the folding of the C peptide by determining the average URMS distance (averaged over each position along the peptide) between the peptide and an idealized $\alpha$-helix.
3. The URMS is used to compare the C peptide as it folds with its native state. The technique used here shows how close each portion of the peptide is to its final native fold.
4. This final visualization method presents a compact summary of the data developed in method 3. For each peptide conformation a single number (order parameter) measures its similarity to the native structure. This is used to present a graph of the peptide's similarity to its native state as a function of folding time-steps.

### 4.1 Visualization of Thermal Fluctuations near the Native State

For this visualization method we provide a visual display of the contents of a matrix that we call M. The rows of M correspond to the residue numbers of the protein and the columns of M correspond to time steps (e.g., a protein consisting of 150 residues simulated for 250 time steps would correspond to a 149×250 matrix). Note that the matrix rows actually correspond to the unit vectors along the α-carbon backbone, thus there is one less matrix-row than the number of residues.

For this example, we use a small probe that is moved over the whole length of the protein chain comparing shape (using the URMS) at each position. Here are the details of how the matrix M is computed.

Algorithm M:

1. For each time step $t$=1,...,$T$, let $P$ be the current protein structure. Perform the following two steps.
2. For each position $k$=1,...,$n$-$w$+1 along the length of $P$ (where $n$ is the length of $P$ and $w$ is the length of the probe), compute the URMS distance between the probe sequence and the subsequence of $P$ from $P_k$ to $P_{k+w-1}$. Assign this value to L($k$,$t$) (L is a temporary matrix.)
3. Assign M($i$,$t$) = min{L($k$,$t$): $k$=$i$-$w$+1,...,$i$} for each position $i$ along the protein $P$.

The intuition behind the last step is to ensure that each residue is assigned a value that indicates the best of all the comparisons in which this residue participates. The

minimization done here is just one of several ways that this intuitive idea of "best" might be implemented. This choice appears to work well in practice.

We analyze the dynamics of the protein MHC. To maintain clarity we show only the analysis of residues 1 to 275. Similar results are obtained for residues 276 to 375. We use two probes to examine the set of structures generated by the Molecular Dynamics protocols. One probe is an idealized α-helix of ten amino acids (Figure 1). The second probe is a β-strand of the same length (see Figure 2). Both figures use colors to indicate the relative values contained in the matrix M: dark blue for URMS=0 (similar shape) and red for URMS=1.3 (dissimilar shape). The figures suggest that substructures similar to either the α-helix or the β-strand remain similar throughout the simulation. This supports the notion of rigid secondary structure elements among flexible loops [3].

Similar color-coded matrices illustrate the comparison of myoglobin with the idealized α-helix of length 10 residues (See Figure 3). MHC is an α/β protein while myoglobin is a helical protein. We were interested in comparing the dynamics of proteins with significantly different topologies. Moreover, the two computations were performed under different simulation conditions: the MHC computations were performed without solvation shell, while the myoglobin study was of "high quality" with significant solvation and summation of long range forces. The results of the MHC simulation are qualitatively similar to the simulation of myoglobin, supporting the idea of rigid secondary structure elements moving in otherwise plastic protein.

### 4.2 Determining the Length of a Post-Critical Folding Nucleus

An unfolded protein chain exists in a multitude of structures, flipping rapidly among alternative conformations. Occasionally, during the folding process, the protein chain adopts a fraction of native fold. If the fraction of the correct fold is above critical size the rest of the folding process proceeds rapidly.

We want to determine the critical length for the formation of a helix in C peptide. To do this, we analyze folding trajectories of C peptide using α-helix probes of varying length. The idea is that if the probe is of the same length as of the critical nucleus, a very abrupt trajectory is anticipated. Once the critical size is reached, the process is more or less downhill and structural similarity increases in a "free fall". On the other hand, if the probe is too small a more gradual increase in the content of structural similarity will be observed. The sharp detection of a new fast time scale as a function of the probe size indicates the length of the post-critical nucleus.

In Figure 4 we present a summary of 12 different trajectories. The average URMS, computed by summing the rows of the matrix M and then dividing by the number of rows, is plotted as a function of time, with the different trajectories ordered sequentially on one graph. Several plots show how probes of different sizes (4 to 8 residues long, or 3-7 unit vectors) search for helical segments in the structures of the trajectory. Large helical content is indicated by small URMS value. As expected all trajectories approach a minimal value at the end in which the C peptide is forced (by the simulation set-up) to form a helix.

A more interesting feature is revealed when examining the URMS changes as a function of the probe size. When the probe is short (e.g., 4 amino acids) the approach to the "native" helical state is gradual and occurs at all times. This is inconsistent with our view of the post-critical nucleation in which rapid folding followed its formation. Instead helices of 4 amino acids form and reform many times at different locations along the peptide until the collections of helical residues added up to the complete helical segment.

As the size of the probe increases, a different view of the dynamics, more in line with the formation of a post-critical nucleation, is seen. For probe lengths of 6 and 7, we see an "incubation" period in which the URMS remains roughly a constant; the system is unable to generate a helix of that length during this initial period. However, once such a segment is created the formation of the rest of the helix quickly follows. As a result we identify the critical size as seven amino acids.

An alternative interpretation (given in reference 5) is that the final state of the helix is less stable. However the computations of the progress of the chain toward the helix conformation (measured by short segments) suggest that there is a constant drift toward the helix. The drift would not be directed to the helix if the helix were not stable.

Another point of interest is related to the coordinate RMS (the more common measure). As noted earlier, the coordinate RMS depends strongly on the length of the structures compared. Difficulties are thus expected when comparing the different RMS values as a function of the probe size. In Figure 5 we repeat the search for a

post-critical nucleation using the standard RMS distance. The plots show greater variability in absolute RMS value as well as a less-certain picture of the folding dynamics as a function of the probe size.

### 4.3 Finding Structural Patterns in a Folding Trajectory

Here we are searching for more detailed information about a folding trajectory. Besides the critical size of the post-critical nucleus, we want to identify its position along the chain and to understand in general the process by which the helix is formed.

Traditionally, a folding trajectory is visualized with a 3-dimensional animation showing the protein as it folds. Such a representation is very helpful and we use it as well. However, projections of three-dimensional objects onto two-dimensional plots (and computer screens) can result in loss of information; interesting events might be hidden or the viewing angle may be such that events are overlooked. It is therefore useful to have an alternate visualization of relevant trajectory information for more convenient analysis.

One approach taken for such visualization is the use of the distance or the contact matrix [2]. However, as we believe is demonstrated below, the *G* matrix we introduce here is easier to understand.

Here, a single protein conformation *P* is compared with its own native fold. The previously discussed M matrix encapsulated the whole simulation in one rectangular matrix. Here we compute a matrix G for each simulation time step *t*. The rows of G,

just as for matrix M, represent residue positions along the protein backbone. The columns of G correspond to probe lengths, short probes on the left, longer probes to the right.

Algorithm G:

Let $P$ represent the current protein conformation and let $S$ represent the native state of the protein (both are of length $n$).

1. For each probe size $w = 1,\ldots,n$ perform the following steps.
2. For each position $k = 1,\ldots,n\text{-}w\text{+}1$ along the length of $P$, compute the URMS distance between the subsequence of $P$ from $P_k$ to $P_{k+w-1}$ and the corresponding subsequence of $S$ ($S_k$ to $S_{k+w-1}$). Assign this value to F($k$,$w$) (F is a temporary matrix.)
3. Assign G($i$,$w$) = min{F($k$,$w$): $k=i\text{-}w\text{+}1,\ldots,i$} for each position $i$ along the protein $P$.

Just as for matrix M, the intuition behind the last step is to ensure that each residue is assigned a value that indicates the best of all the comparisons in which this residue participates. This step also has the effect of maintaining a clear correspondence between a residue position and a row of the G matrix. The minimization done here is just one of several ways that this intuitive idea of "best" might be implemented; this choice appears to work well in practice.

Intuitively, the entries in row $i$ indicate how close the residues near residue $i$ are to their native fold. The different entries along the row correspond to larger and larger matching portions of the protein. Thus a low URMS value (dark blue in the figures)

on the left means that a small portion of the protein conforms to the native fold while a low URMS value on the right means that a large portion of the protein conforms to its native fold.

In Figures 6-8 we display both the G matrix and a sketch of the protein backbone (blue line) superposed on the native fold (red line) for several time steps of the simulation. The sequence of plots at different times provides a useful summary of the trajectory. We show just the data for time steps 17, 25, and 26 of the MD simulation. In these figures, dark blue in the matrix *G* corresponds to URMS=0 (similar shape) and red corresponds to URMS=1.3 (dissimilar shape). Note that the entire matrix G is not shown; some columns on both the left and the right are excluded simply because they don't convey significant information (for example, windows of size 1 or 2 residues always indicate that an exact shape-match exists).

As the fold progresses in time, the protein becomes more similar to the native fold, and the map of *G* becomes more dominated by blue (see structure 26). We find that the plot of *G* conveys more information than the plot of the protein (Figures 6-8). Besides the effective projections of the whole structure to two dimensions, the blue coloring of the structure draws immediate attention to the areas most similar to the native fold. For example, comparing structure 17 with the native fold (Figure 6), we find residues 2-6 closer to the native fold, as compared to the rest of the chain. The center of the chain deviates most from the native conformation.

### 4.4 An Order Parameter: A Compact Measure of Proximity to Native State

It is useful to have a single number (an order parameter) that measures the proximity of the current structure to the native fold. We define two such order parameters here; both based on the matrix G used above. The first, denoted by AR(*t*), measures a relative area: the area (the number of entries) of G for which the URMS value is less than the 70% of the URMS for a random structure. This is normalized by dividing by the total area of G. The second parameter, denoted by AV(*t*), is based on the average value of entries in G.

Previous methods have included the use of the traditional RMS, the function Q [14], or the contact correlation functions [15]. The traditional RMS is not really satisfactory since conformations with a fraction of the native fold do not have significantly lower RMS values than that achieved for random structures. This is problematic since the initial structure formation is of course important for the understanding of folding dynamics.

Q [14], a popular and useful function for measuring closeness to a native conformation, is defined as the fraction of the native contacts that are present in the current conformation. Changes in Q emphasize long range contacts (along the chain). An advantage of our new measures is that similarity is examined on all scales—local and extended—on equal footing. Moreover, to compute Q it is necessary to select a cutoff distance to define what is meant by a "contact". This cutoff distance is somewhat arbitrary and may influence the result. No such loosely defined parameter is required by AV(*t*). Thus, AV(*t*) does not require any cutoff parameter and in that sense it is less arbitrary than Q. Note though that AR(*t*) does have a cutoff parameter: we use a URMS value based on 70% of the expected value for a random structure.

We believe that the cutoff we have chosen is a useful one, but other values might also be used.

In Figures 9-11 we plot Q($t$), AR($t$), and AV($t$) for a folding trajectory of C peptide. It is intriguing that the URMS related measures—AR($t$) and AV($t$)—show distinct features that are not observed in Q($t$) (Figure 9). Both AV($t$) and AR($t$) show a sharp drop in value (increase in similarity) near structure number 26 (Figure 8), while Q($t$) rates structures 25 and 26 (Figures 7 and 8) as about equally similar to the native fold. Observe that structure 25 occupies about the same space as the native conformation, but the local chain structure and the hydrogen bonding are clearly wrong. This kind of difference from the native fold is hard to catch using Q($t$) because it is a long-range measure, emphasizing long range contacts. In contrast the two new measures are sensitive to local rearrangements in addition to long-range contacts; thus, they are able to identify the sharp locking-into-place of the secondary structure element.

## 5. Conclusions

We have shown that the new URMS measure for structural similarity can be used effectively in many analyses of protein dynamics simulations. The simulations require comparison of structural segments, of complete protein chains, and the simultaneous exploration of changes in time and space. The URMS has a number of unique advantages that include the weak dependence on the chain length, efficient computations, and compact presentation of massive data in the form of the M and G matrices.

The examples provided in this manuscript are from various areas of simulations and protein dynamics. We considered the thermal fluctuations at the neighborhood of the native state, an examination that is performed traditionally with the coordinate RMS measure of similarity. This measure has a strong dependence on the length of the chain, which makes it difficult to compare the fluctuations of proteins of different sizes. The URMS has weaker dependence on chain length and suggests a more meaningful comparison. In particular it was helpful in detecting/suggesting post-critical nucleation site. The nucleation was not observed using the coordinate RMS probably due (again) to the size dependence of this approach. Moreover, we have shown that the new measures capture local structural rearrangements during the folding process, which are difficult to detect with the widely employed order parameter Q.

## References

[1] J. Andrew McCammon and S.C Harvey, "Dynamics of Proteins and Nucleic Acids", Cambridge University Press, Cambridge, 1987

[2] T.F. Havel, I.D. Kuntz and G.M. Crippen, Bull. Math. Biol. 45,665(1983); W. A. Kabsch, "A solution for the best rotation to relate two sets of vectors", Acta Crystallogr. 1976;A32:922-923

[3] D. Rojewska and R. Elber, "Molecular Dynamics Study of Secondary Structure Motions in Proteins: Application to Myohemerythrin", Proteins, 1990;7:265-279

[4] Shakhnovich EI, "Folding nucleus: specific or multiple? Insights from lattice models and experiments", Folding and Design 1998;3:R108-11; D. Thirumalai, and

K.D. Klimov, “Fishing for folding nuclei in lattice models and proteins”, Folding and Design 1998;3:R112-8

[5] R. Elber, J. Meller and R. Olender, “A stochastic path approach to compute atomically detailed trajectories: Application to the folding of C peptide”, J. Phys. Chem., B, 1999;103:899-911

[6] R. Elber, A. Roitberg, C. Simmerling, R. Goldstein, H. Li, G. Verkhivker, C. Keasar, J. Zhang and A. Ulitsky, “MOIL: A program for simulations of macromolecules”, Computer Physics Communications, 1995;91:159-189

[7] S.J. Weiner, P.A. Kollman, D.A. Case, U.C. Singh, C. ghio, G. Alagons, S. Profeta Jr., and P. Weiner, A new force field for molecular mechanical simulation of nucleic acids and proteins”. J. Amer. Chem. Soc. 1984;106:765-784 ; W.L. Jorgensen and J. Tirado-Rives, “The OPLS [optimized potentials for liquid simulations] potential functions for proteins, energy minimization for crystals of cyclic peptides and crambin”, J. Amer. Chem. Soc. 1988;110:1657-1666

[8] J.J. Osterhout, R. L. Baldwin, E.J. York, J.M. Stewart, J.H. Dyson, P.E. Wright, “Proton NMR studies of the solution conformation of an analog of the C-peptide of Ribonuclease A”, Biochemistry, 1989;2:7059-7064

[9] L. P. Chew, D.P. Huttenlocher, K. Kedem and J. Kleinberg, “Fast Detection of Common Geometric Substructure in Proteins”, Proceedings of ACM RECOMB International Conference on Computational Molecular Biology, 1999, to appear. Also appears as Tech. Rept. No. 98-1705, Computer Science Department, Cornell University.

[10] S. Rackovsky, H.A. Scheraga, ”Differential geometry and polymer conformations 1. Comparison of protein conformations”, Macromolecules 1978;11:1168-1174 ; S. Rackovsky, H.A. Scheraga, “Differential geometry and

polymer conformations 2. Development of a conformational distance function", Macromolecules, 1980;12:1440-1453 ; Rackovsky S, Goldstein DA, "Protein comparison and classification: a differential geometric approach.", Proc Natl Acad Sci USA 1988;85:777-81

[11] B.A. Reva, A.V. Finkelstein, and J. Skolnick, "What is the probability of a chance prediction of a protein structure with an rmsd of 6 A?", Folding and Design, 1998;3:141-7

[12] V.N Maiorov VN, and G.M. Crippen, "Size-independent comparison of protein three-dimensional structures", Proteins 1995;22:273-283

[13] X. Pennec and N. Ayache, "A geometric algorithm to find small but highly similar 3D substructures in proteins", Bioinformatics, 1998;14:516-522.

[14] J.N. Onuchic, Z. Luthey-Schulten, P.G. Wolynes, "Theory of protein folding: the energy landscape perspective", Annu. Rev. Phys. Chem., 1997;48:545-600

[15] Z. Guo, D. Thirumalai, J.D. Honeycutt, "Folding kinetics of proteins: a model study", J. Chem. Phys., 1992;97:525-35

## Figure legends

1. The matrix M for the simulation of MHC probed by a helix of ten amino acids. Only residues 1 to 275 of the MHC protein are considered. Note the dark bands that correspond to the $\beta$ sheet domain or to loops that are significantly different from the helix. White domains that remain so throughout the simulation indicate the helices.
2. The same as Figure 1, this time the probe is of a $\beta$ strand of the same length.
3. Analysis for the myoglobin simulation. Myoglobin is a helical protein and therefore we show only the helical probe of length 10 as in Figure 1.

4. A summary of 12 different folding trajectories of C peptide. The trajectories are ordered sequentially on the same axes. The average URMS (computed by summing rows of the M matrix and dividing by the number of elements) is plotted as a function of time. The several plots are for probes of different sizes – 4 to 8 residues long or 3 to 7 unit vectors.
5. The same as Figure 4, this time for the traditional RMS. Note that the RMS value depends on the length of the probe; as a result it is difficult to detect the formation of the folding nucleus.
6. A display of the matrix G and a sketch of the alpha-carbon backbone showing an early state (MD time step 17) on the folding trajectory of C peptide. Note that residues 2-6 (in blue) are similar to the native fold while the rest of the chain is not. The center of the chain deviates the most from the native fold.
7. A display of the matrix G and a sketch of the alpha-carbon backbone of C peptide along the folding trajectory (MD time step 25).
8. Same as Figures 6-7 but for MD time step 26. The structure is near the native state; as a result the map of G is dominated by blue.
9. Application of the order parameter Q(t) to identify intermediates along the folding pathway of C peptide. Q(t) is the fraction of native contacts that are found in the present structure.
10. Application of the order parameter AR(t) to investigate the folding of C peptide. Note the similarity to the native structure which is observed near structure 26 that is not detected by Q(t)
11. The same as Figure 10, this time for the order parameter AV(t).