NAIST-IS-DT0261028

Doctoral Dissertation

Two aspects of bioinformatics in different timescales: sampling problem on molecular dynamics simulation and perspective to the gliding mechanism of *Mycoplasma mobile*

Shoichi Metsugi

March 24, 2005

Department of Bioinformatics and Genomics
Graduate School of Information Science
Nara Institute of Science and Technology

A Doctoral Dissertation
submitted to the Graduate School of Information Science,
Nara Institute of Science and Technology
in partial fulfillment of the requirements for the degree of
Doctor of Science.

Thesis Committee:

Professor Shin Ishii	(Supervisor)
Professor Nobuhiro Go	(Co-supervisor)
Professor Toshio Hakoshima	(Member)

Two aspects of bioinformatics in different timescales: sampling problem on molecular dynamics simulation and perspective to the gliding mechanism of *Mycoplasma mobile**

Shoichi Metsugi

Abstract

Computational science has deeply penetrated into molecular biology for the last few decades and emerged as a new field called bioinformatics. Bioinformatics treats biological problems of extremely different timescales. One extreme is molecular dynamics (MD) simulation, which deals with the motion of protein molecules in the order of nanoseconds. Another extreme is the analysis of molecular evolution, which deals with the events that take place during the time scale of millions of years. Here, I would like to report my approach to two biological problems of two distinct time scales. One is the analysis of sampling problems in MD simulation of protein, and the second is the analysis of gliding mechanism of *Mycoplasma mobile*.

MD simulation has become one of general tools to estimate thermal property of proteins. A simulation of nanoseconds of protein motion can be carried out. To derive thermal properties of molecule from a MD trajectory, an assumption has to be made, that the simulation has been long enough to reconstruct the ensemble of the molecular motion. There is no simple criterion to make a judgment, if the simulation time was sufficiently long to estimate the physical property of the molecule. Many physical quantities are known to be expressed as a second moment of dynamical variables, which can directly deduce vibrational modes. Therefore, I have investigated the effect of vibrational modes on convergence of the second moment matrix component of atom positions. We found that the frequency modes higher than 68[cm-1] converges to the equilibrium value within 10% error value within 400 pico seconds. The difference between the eigenvalues of second moment matrix of all modes and those of only middle (10~68[cm-1]) and low (~10[cm-1]) modes is very small. These results mean that the

higher frequency components of second moment matrix converge very fast and hardly influence the convergence of middle and low frequency components. The middle frequency modes influence the convergence of lower modes but this influence is less significant, compared from the effect from low modes obtained by former study. As a summary, the lower modes are dominant for the convergence of second moment matrix.

Computational analysis of evolution is another important tool to clarify the molecular mechanism of biological machinery. Here I describe one such approach to the analysis of the movement mechanism of a bacterium, *Mycoplasma mobile*. *M.mobile* has an ability to glide, when attached on a solid surface. The mechanism of the motion is not understood yet, however, a few genes including Gli349 are experimentally shown to be related to the motion. The DNA sequence of the Gli349 gene has already been reported but the three-dimensional structure has not been solved yet. To obtain a clue for the mechanism of its movement, I carried out an evolutionary analysis of the Gli349 sequence. As a result, I found 21 sequential repeats within the Gli349. Each repeat consists of approximately 100 amino acids with a conserved sequence motif, "YxxxxxGF". No homologous sequence to Gli349 was found in NCBI RefSeq non-redundant sequence database. In general, repeat sequences are known to associate with other protein. This leads to a speculation that the repeat regions in Gli349 are also responsible for the interaction with other proteins. Chymotrypsin breaks the amide bond where the sequence is not within a structural domain, such as a loop region between structural domains. The structural domain regions speculated from the chymotrypsin cleavage experiment agreed well with the repeats I discovered. I predict one sequence repeat corresponds to one structural domain, and the entire structure of Gli349 is composed of the tandem domains. Electron microscopy experiment shows Gli349 takes rod shape with a couple of kinks. The size and shape of the observed shape fits very well with the predicted domain repeat structure. Immunoglobulin fold is known to form a rod structure by forming tandem repeats. However, the repeat sequences do not fit into any of know immunoglobulin fold structures. Therefore, I conclude that the repeat sequence in *M.mobile* responsible for the gliding forms a novel type of structural domain.

Keywords:

molecular dynamics, sampling, evolutionary analysis, Mycoplasma, gliding protein

Contents

1. Introduction of the thesis

Since the discovery of double-helical structure of DNA in 1953, a tremendous amount of progress has been made in the field of molecular biology. Determination of genome sequences of many organisms is one of its most important achievements. In the present post genomic era, however, one of the goals for the molecular biology is to reveal the meaning of such genome sequences, that is, to identify and reveal the characteristics of the proteins translated from the genome sequences.

There are roughly four steps to identify and reveal the characteristics of proteins. The first step is to find the regions to be transcribed from DNA sequence into messenger RNA (mRNA) sequence. To reveal the transcribed region is a difficult question because whether transcription happens or not depends on the type of cells and their environment. The second is to find how the mRNA is spliced into mature mRNA which is to be translated into proteins by ribosome in eukaryote. This is also a difficult question because alternative

splicing may happen. The third step is to reveal the 3D structure of proteins, which is one of the important characteristics of protein. This can be realized by the X-ray crystallography or NMR experiment for smaller proteins but is hardly realized for larger proteins. The fourth is to elucidate the function of proteins and the relationship between structure, motion and function of the proteins.

In this thesis, two studies about the third and fourth steps are reported. One study corresponding to the third step is the sequence analysis of Gli349. Gli349 is the protein in a motile bacteria *Mycoplasma mobile* and is suggested to be something to do with its motility. Sequence of Gli349 is known but either structure or function is not solved. Since Gli349 is too large to apply the experimental 3D structure analyses such as X-ray crystallography and NMR, we conjectured the structure of Gli349 from sequence computationally. The other study which corresponds to the fourth step is the molecular dynamics (MD) simulation study. Since MD can track the motion of every atom in the protein, MD is useful for the investigation of the relationship among 3D structure, motion of the protein and function of the protein. But MD has one disadvantage. For the

precise calculation of various physical quantities such as entropy, very long simulation, for example more than 10 nanoseconds, is necessary, which is still beyond the reach of the present computational power. In this study, we investigated the time length for precise calculation of secondary moment matrix of atomic positions by which many physical quantities are expressed.

These two studies are common in that these are computational approach of molecular biology, which is called "bioinformatics" recently. These two studies can also be regarded as two bioinformatics studies in different time scale. MD simulation deals with the motion of proteins in the order of nanoseconds. Sequence analysis is one of the analyses of molecular evolution, which deals with the events that take place during the time scale of millions of years. This thesis can also be interpreted as the approach to two biological problems of two distinct time scales, that is, nano-second order and millions of years order. In this thesis, chapters from 2 to 4 correspond to the MD study and chapters from 5 to 8 do to the sequence analysis study.

2. Introduction to the sampling problem

Molecular dynamics (MD) simulation, numerical integration of classical multi-dimensional Newton equation, has become a useful tool to study various biological and physical properties of proteins. Such physical properties as various space and/or time correlation functions can be calculated from MD trajectory. Although physical quantities are formulated as an ensemble average in statistical mechanics, a time average is usually used instead of an ensemble average in the use of MD trajectory. This substitution is valid only when MD trajectory is sufficiently long. In protein MD, it is not a trivial task to determine length of simulation necessary to calculate these physical quantities reliably.

Slow convergence of sampling in protein MD has been pointed out especially by Hess[1, 2]. He has performed a few n sec protein MD simulations and from the trajectory constructed a matrix of second moments of variations of all dynamical variables around respective means. He then diagonalized this positive-definitive symmetric matrix to identify the axes of the largest collective

variation, the second-largest collective variation, etc. (the principal component analysis). When he examined the MD trajectory by projecting it onto the axes of the first several largest eigenvalue principal components (PCs), he observed that the projected trajectories behaved like cosines as functions of time. He showed that this is a behavior expected for a random walk on a multi-dimensional flat energy surface. This means that in the conformational subspace spanned by the axes of the largest several eigenvalue principal components, the energy surface explored during a few n sec is essentially flat. The energy surface should of course go up if explored wider enough. A few n sec simulation is too short to see this.

Besides the problem pointed out by Hess, the multi-dimensionality of the protein conformational space should also be a serious problem in estimating necessary length for reliable simulation. The dimensionality M of the conformational space of a protein consisting of N_{atom} atoms is given by $3N_{atom}$ -6, where 6 comes from six translational and rotational degrees of freedom. The volume of conformational space accessible for protein is in proportion to the

power of the dimensionality M. If the state point ought to visit the whole conformational space densely and uniformly, MD simulation should be useless. Although this problem originating from multi-dimensionality appears common to any systems with a large degrees of freedom, it is less severe in often studied systems consisting of many (N) identical atoms or molecules, because the whole phase space for such systems consists of N! mutually identical subspaces due to symmetry of permutation of identical atoms or molecules. Such symmetry does not exist in protein MD.

However, in protein MD we still expect that for a certain class of physical quantities, dense and uniform visit may not be necessary. Because many important physical quantities can be expressed as a second moment of dynamical variables, we ask in this paper length of simulation necessary for sufficient convergence of second moment matrix of dynamical variables. Therefore, we ask how the multi-dimensionality affects the convergence of the second moment matrix of dynamical variables. In the next section for method and results we analyze this problem more carefully and describe how this question can be

addressed by studying a reasonable theoretical model.

3. Method and Results of sampling problem

Results from a molecular dynamics simulation

We formulate our question as follows: how does a diagonalized second moment matrix calculated from a MD trajectory approach its equilibrium value as a function of duration time T of simulation and how does this behavior depend on the multi-dimensionality of the problem?

We have performed a 1 n sec MD simulation of hen egg white lysozyme for which M is 5880. Even though the elementary time step of numerical integration was set to 0.5 f sec, the trajectory is stored at every 5 f sec. From this stored trajectory we constructed 20 different M×M second moment matrices for 20 different duration times, each starting from the start of the simulation and ending at an integral multiple of 50 p sec. These matrices were then diagonalized (principal component analysis) and diagonal elements are rank-ordered in the descending order of their magnitude.

In the following discussion we sometimes change the definition of dynamical variables from the original mass weighted atomic Cartesian coordinates to their linear combination corresponding to the principal component eigenvectors. These new dynamical variables are called collective variables and each of them is identified by a vector defining the linear combination. Their mean square fluctuations during the simulation are given by respective diagonal elements of the diagonalized second moment matrix. High and low rank-order collective variables have respectively large and small amplitude variations during the simulation.

The values of the diagonal elements of selected rank-orders are plotted in Figure 1 against duration time. The values of the lowest rank-order diagonal elements do not show indication of approaching the equilibrium values during 1 n sec. There is a serious problem as to equilibration of these elements as has been pointed out by Hess. This problem comes from the nature of the shape of the conformational energy surface in the subspace spanned by the lowest rank-order eigenvalue principal component axes. We do not ask this question in

this paper.

However, for high rank-order diagonal elements, we see that they appear to converge to seemingly equilibrium values rather readily. By analyzing the results shown in Figure 1, we see that the values of diagonalized elements of about 300th or higher in the rank-order behave well as a single exponential function of simulation time T and appear to approach within 10% of their converged values (supposedly their equilibrium values) as simulation time T becomes longer than about 400 p sec, even though low rank-order diagonal elements are still far from equilibration at such simulation time. The number of such elements accounts as many as 95% out of the totally M=5880 diagonal elements.

The results of the simulation shown in Figure 1 and summarized above imply that identification of the higher rank-order (or smaller amplitude) collective variables and equilibration of their distributions are not much influenced by (and are therefore essentially independent of) and realized earlier than identification of the lowest rank-order (or largest amplitude) collective variables and their equilibration. Equilibration of distribution of sampled points

should of course be determined by the shape of the energy surface. However, the above observation implies that there may be a rule in the order of identification of and equilibration in subspaces of a multi-dimensional space. The order of identification and equilibration appears to be determined mainly by the width of distribution in each subspace, but not by details of the shape of energy function in each subspace. If in fact distribution in a subspace with smaller amplitude of variation equilibrates faster than that in a subspace with larger amplitude of variation, this may provide a mechanism by which protein MD, even with a large dimensionality of the conformational space, is still a useful method of research.

In order to make this qualitative discussion more quantitative, we will study behavior of a well-defined theoretical model and the second moment matrices from it.

Langevin dynamics model

To proceed further we refer to the results of previous studies.[3, 4] They studied the dynamics of human lysozyme (for which M = 6117) both by the MD

simulation and the normal mode analysis. In the latter analysis a simple parabolic potential energy function is assumed with the same Hessian (second derivative matrix) as the one at one of many local minima of the potential energy for which the MD simulation is performed. They also have done the principal component analysis of the MD trajectory, and rank-ordered diagonal elements in the descending order of their magnitude. They have found that the behavior of projections of the MD trajectory onto the principal component axes of 31st and higher rank orders can be well simulated by a Langevin dynamics in a parabolic potential. The number of those principal component axes accounts for as many as 99.5% of the totally M=6117 axes.

They have further found that the values of the diagonalized second moments of 301st or higher rank-order fit well with the values theoretically expected for the normal modes of corresponding rank-order. The number of modes in this range accounts for about 95% of the totally M=6117 modes. Angular frequency ω_i of the lowest frequency normal mode in this 95% range, i.e., that of 301st normal mode, is 68 cm^{-1} (here expressed by corresponding

wave number $\omega_i/2\pi c$, c being light velocity), or its $1/\omega_i$ is 78 f sec. It has also been observed that the projection of the MD trajectory onto each of these high rank-ordered PC axes produces a single Gaussian frequency histogram with the width expected for the corresponding normal mode. These facts indicate that the conformational energy function in the 6117 dimensional space consists not exactly but essentially of a direct product between a set of 5817 dimensional simple parabolic function and a certain function in the remaining 300 dimensional space. These high frequency modes in this 95% range were called simple harmonic modes.[4]

They have also found that the projection of the MD trajectory onto each of the 31st through 300th PC axes produces a single Gaussian frequency histogram but with a width wider than that expected for the corresponding normal mode. As mentioned earlier, the projected MD trajectories onto these PC axes can also be well simulated by a Langevin dynamics in a parabolic potential. Modes in this range were called singly hierarchic modes.[4]

Thus, except for the 0.5% of the degrees of freedom, protein MD trajectory

can be well simulated by a Langevin dynamics in a multi-dimensional parabola. We are concerned in this paper with a possible difficulty due to multi-dimensionality of the conformational space, and we came to conjecture that a clue of mechanism of how this difficulty may actually be avoided lies in a sequential identification and equilibration of subspaces in the ascending order of widths of variations. Now we want to examine the validity of this conjecture. Because the main aspect of the problem here is the multi-dimensionality and a sequence of widths of variations in a series of subspaces, and is not in the details of energy function in each subspace, we can ask this question by assuming a Langevin dynamics in a multi-dimensional parabola, a model implicated by the analysis of Kitao, et. al.[4] In the remaining 0.5% degrees of freedom problem still certainly exists in the shape of the energy surface. This problem is outside the scope of this paper.

A trajectory of the Langevin dynamics in a multi-dimensional parabola is given by $x(t) = (x_1(t), x_2(t), \cdots, x_M(t))$, where $x_i(t)$ satisfies the following Langevin equation.

$$\frac{d^2x_i}{dt^2} = -\omega_i^2 x_i - \gamma_i \frac{dx_i}{dt} + R_i(t)$$

where x_i, ω_i, γ_i, $R_i(t)$ are a coordinate, an angular frequency, a friction constant and a random force, respectively. We use the method of numerical integration of this equation worked out by Ermak, Buckholz and McCammon.[5, 6] Temperature T, which is related to the magnitude of the random force $R_i(t)$ by the equation $\langle R_i(t)^2 \rangle = 2\gamma_i k_B T$ (k_B being the Boltzmann constant), is set to be 298 K, and we choose the time step for integration Δt to be 0.5 f sec so that the total energy is sufficiently conserved. Trajectory is stored at every 5 f sec. A set of angular frequencies defines the shape of the multi-dimensional parabola. Because, as mentioned above, the result of MD simulation and subsequent principal component analysis can be fairy well reproduced by the normal mode analysis, we assume a set of values of angular frequencies from the normal mode analysis (NMA) of BPTI (PDB code: 5PTI). The value of dimensionality M for this protein is 2712. The distribution of the values of angular frequencies for various globular proteins is known to be rather insensitive to details of three dimensional structures and thus roughly common to many proteins. Therefore by

studying just the case of BPTI we should be able to capture the general properties of globular proteins. A uniform value 47 cm^{-1} of the friction coefficient γ_i (here expressed by corresponding wave number γ_i/c or $1/\gamma_i$ being 0.71 p sec) is assumed, the value that has been deduced in the previous study.[3]

The time correlation function of the solution of equation (1) for one particular value of i becomes overdamped for $\gamma_i > 2\omega_i$ and underdamped for $\gamma_i < 2\omega_i$, and is given by[7]

$$\left\langle x_i(0)x_i(t)\right\rangle = \begin{cases} \dfrac{k_BT}{2\omega_i^2}\left(1+\dfrac{\gamma_i}{\gamma_i'}\right)\exp\left(-\dfrac{t}{\tau_i}\right)+\dfrac{k_BT}{2\omega_i^2}\left(1-\dfrac{\gamma_i}{\gamma_i'}\right)\exp\left(-\dfrac{t}{\tau_i'}\right) & (\textit{for overdamp}) \\ \dfrac{k_BT}{\omega_i^2}\left(\cos\omega_i't+\dfrac{\gamma_i}{2\omega_i'}\sin\omega_i't\right)\exp\left(-\dfrac{t}{\tau_i}\right) & (\textit{for underdamp}) \end{cases}$$

(2)

where $\gamma_i' \equiv \left(\gamma_i^2-4\omega_i^2\right)^{\frac{1}{2}}$ and $\omega_i' \equiv \left(\omega_i^2-\frac{1}{4}\gamma_i^2\right)^{\frac{1}{2}}$. There is one relaxation process for underdamped motion and two relaxation processes for overdamped motion. Relaxation times τ_i and τ_i' are given by

$$\begin{cases} \tau_i = \dfrac{2}{\gamma_i - \gamma_i'}, \quad \tau_i' = \dfrac{2}{\gamma_i + \gamma_i'} & (for\ overdamp) \\ \tau_i = \dfrac{2}{\gamma_i} & (for\ underdamp) \end{cases}$$

This means that the solution of equation (1) shows a correlated behavior within the relaxation time τ_i and becomes roughly uncorrelated beyond it. In other words the solution of equation may be regarded very roughly as an ensemble of piecewise trajectories of duration τ_i of undamped harmonic motion with angular frequency ω_i with mutual phases of piecewise trajectories being random.

For high frequency modes which satisfy the condition $\tau_i >> 1/\omega_i$, each piece of the trajectory of duration τ_i undergoes many cycles of the harmonic motion, and thus is expected to make a contribution to the second moment matrix, which is insensitive to the value of phase of the piece of trajectory and is nearly equal to the value expected for the equilibrium value.

With these simple theoretical analyses, we ask the following question by actually implementing the Langevin dynamics simulation of BPTI model and by performing the subsequent principal component analysis: How do the diagonal

elements of diagonalized second moment matrix approach their equilibrium value? For the purpose of asking this question, we treat in our Langevin dynamics simulation only 300 lowest frequency modes out of the totally 2712 modes. Simulation is done for 10 n sec. The normal mode frequencies of the 150th and 300th modes are respectively about 70 cm^{-1} and 150 cm^{-1}. Because the value of γ_i for BPTI is 47 cm^{-1}, modes above the 151st may be regarded to satisfy the condition $\tau_i >> 1/\omega_i$. These higher frequency modes above 151st belong to the simple harmonic modes.[4]

Figure 2 shows how the diagonal elements of diagonalized second moment matrix corresponding to these higher frequency modes approach their respective equilibrium values. Unlike in the case of the corresponding calculation from the MD trajectory described in section 2.1, the equilibrium values to which the diagonal elements of diagonalized second moment matrix should approach as simulation time T becomes sufficiently long are explicitly known from the definition of the model. All the diagonal elements in the range of the 151st mode to the 300th mode approach within 10% of their converged values as simulation

time T becomes longer than about 650 p sec. This coincides with the result observed in the case of calculation from the MD trajectory. Thus, we can summarize that the variances of higher frequency modes converge within 1 n sec in a way independent of possibly complex behavior involving the lower frequency modes.

A comment may be necessary about the convergence of the high frequency modes. These modes show underdamped motions with relaxation time τ_i much longer than period $1/\omega_i$ of the harmonic motion. Therefore many cycles of nearly undamped harmonic motion should take place before damping, thus making near-equilibrium contribution to the second moment matrix. This analysis suggests that these high frequency modes should converge in a way dependent to $1/\omega_i$. However, this is not what we observed in Figure 1 and 2 for these high frequency modes. The time of convergence appears to be in the range of 400 p sec in the case of Figure 1 and in the range of 650 p sec in the case of Figure 2 and is insensitive to the value of $1/\omega_i$. A very technical reason accounts for these behavior. In both of MD simulation and Langevin dynamics simulation,

trajectory is stored at every 5 f sec. The second moment matrix is constructed from such trajectories. Then, even for very high frequency modes stored trajectory does not look like a harmonic motion, but like a series of randomly sampled points. This is the reason of the apparently slow convergence of the very high frequency modes. Even with such a technical problem, which exists also in usual MD simulations, we came to a conclusion that 1 n sec simulation is enough for convergence of variances of higher frequency modes.

Figure 3 shows how variances of low frequency modes approach their respective equilibrium values. Modes from 31st to 150th and those from 16th to 30th approach within 2% of their equilibrium values within 1 n sec and 4 n sec respectively. For modes from 3rd to 15th more than 10 n sec is necessary for this approach. In the lowest and second lowest modes approach to the respective equilibrium values appears to take even longer time.

What we have seen in Figure 3 is again sequential identification and equilibration of subspaces in the ascending order of widths of variation. In the BPTI model simulation of 4 n sec is found enough to see equilibration in

subspaces accounting for 99.5% out of the totally 2712 dimensional space. Equilibration in the remaining 0.5% dimensional space takes more than 10 n sec. This is the space for which Hess pointed out a problem from a different point of view.

A key found in this paper for the problem of multi-dimensionality is sequential identification and equilibration of subspaces. We found that the effect of lower frequency motions to those of the higher frequency is small, and therefore the latter can be identified and equilibrated even before the former is identified and equilibrated. We then ask; Can high frequency motions be neglected to study the lower frequency motions? To examine this question, we compare eigenvalues obtained by diagonalization of the full 300×300 second moment matrix with those obtained by diagonalization of only 150×150 second moment matrix pertaining only to lower frequency 150 modes. In Figure 4, differences between corresponding eigenvalues are shown. We see that the differences are less than 0.1% within 1 n sec for 1st to 100th modes. This means that the higher frequency motions hardly affect the convergence of lower

frequency motions. We also find that the difference of 130th mode is 1% within 1 n sec and 0.12% within 10 n sec. This slightly slow convergence is due to strong mixing with modes higher than 151st in the full 300 dimensional treatment. In summary it is found that the variances of lower modes hardly affect that of higher modes and vice versa. This means that the subspaces composed of higher frequency motions and lower frequency motions respectively are nearly independent.

4. Summary and conclusion of the sampling problem

In this paper we studied the problem of convergence of sampling in protein molecular dynamics simulation occurring from the multi-dimensionality of their conformational space. Because many important physical quantities can be expressed as a second moment of dynamical variables, we asked in this paper length of simulation necessary for sufficient convergence of second moment matrix of dynamical variables. At first we performed a molecular dynamics simulation of a protein, lysozyme, and constructed a series of second moment

matrices as a function of simulation time T. These matrices were then diagonalized (principal component analysis) and diagonal elements are rank-ordered in the descending order of their magnitude. We observed that, as simulation time T becomes longer, higher rank-ordered principal component vectors with smaller diagonal elements are identified and their amplitudes of variation (diagonalized second moments) are equilibrated before low rank-ordered vectors are identified and equilibrated. This sequential identification and equilibration mechanism does not depend much on the details of the conformational energy function, but depends mainly on the spectrum of distribution of magnitudes of diagonalized second moments. Therefore detailed mechanism of sequential identification and equilibration was studied in terms of a simplified theoretical model, Langevin dynamics model.

It has been known from previous works that higher rank-ordered principal modes are actually characterized by simple parabolic energy surface governing them, and correspond well with normal modes with frequencies 70 cm^{-1} or higher.[3, 4] It is now shown in this work that these modes can be identified and

equilibrated by a trajectory of 1 n sec molecular dynamics simulation, even when lower rank-ordered modes are not yet identified and equilibrated.

In the case of BPTI consisting of 58 amino acid residues, identification and equilibration of modes from 16th (with frequency of 14 cm^{-1}) to 150th (with frequency 70 cm^{-1}) require simulation time of 4 n sec. The remaining 0.5% dimensional space takes more than 10 n sec to equilibrate. However, in this relatively small dimensional space there is also a severe problem pointed out by Hess.

In summary we found that the high dimensional conformational space of protein consists of a series of subspaces which are mutually relatively independent. Subspaces with smaller amplitudes of variation are identified and equilibrated during the molecular dynamics simulation before those with larger amplitudes of variation are identified and equilibrated. Due to this sequential identification and equilibration mechanism protein molecular dynamics simulation is a useful tool even with its intrinsic multi-dimensionality.

5. Introduction to the sequence analysis of gliding protein in *Mycoplasma mobile*

Mycoplasmas are gram-positive bacteria and some species such as *Mycoplasma mobile*, *M. pulmonis*, *M. gallisepticum*, *M. pneumoniae* and *M. genitalium* have an ability to glide on solid surface[8]. The mechanism for such gliding is thought to differ from other known mechanisms of movement, such as the flagella motor in bacteria or actin-myosin complexes in myocytes and other cell types, since no protein homologous to flagellin, myosin, actin or any other known motor protein has ever been found in mycoplasmas[9-12].

M. mobile, which glides at an average velocity of about 2.0 to 4.5 μm/s, or about ten times faster than the other four species mentioned above[13, 14], expresses two large proteins, Gli349 and Gli521, that are known to be responsible for the gliding: destructive mutation of either *gli349* or *gli521* eliminates the ability to glide from the organisms[15, 16]. Gli349 is required for *M. mobile* to adhere to glass, and it is believed that it forms a spike that protrudes from the cell surface and in some way transduces the energy needed for motion[17,

[18]. Beyond this, however, little is known about the gliding mechanism of *M. mobile*.

In the present study, therefore, we carried out sequence analysis of Gli349 from *M. mobile* and its homologue MYPU2110 from *M. pulmonis* to characterize those sequences and decipher the structures of these proteins, which we found not to be homologous to any other known protein. Based on our findings, we propose a structural model in which Gli349 is composed mostly of tandem repeats of homologous domains.

6. Materials and methods of sequence analysis

Hidden Markov model for repeat sequence searches

Comparison of the sequence of Gli349 with itself in a dot matrix plot suggested that several weak repeats exist and that each contains the motif YxxxxxGF (where x denotes any amino acid residue, and hereafter referred to as the YGF motif). To further analyze the structures, we then manually extracted subsequences of 120 amino acid residues containing the YGF motif from Gli349

and MYPU2110 to construct a hidden Markov model (HMM) [19-21], which was then applied to search for new repeats within Gli349 and MYPU2110 that were similar to input training data (i.e., repeat subsequences containing the YGF motif).

We used the HMMER package (http://hmmer.wustl.edu/)[22] to carry out HMM, which took as input a multiple sequence alignment (MSA), which serves as training data, together with the entire sequence of Gli349 or MYPU2110. The output was comprised of subsequences that match the profile obtained from the MSA. The HMM is composed of one "begin state," several "match states" (i.e., matches to one of the amino acid residues), several "insert states," several "delete states" and one "end state." The transition probabilities between states are trained by the input MSA. Once the model is trained, we can use it to detect new repeats that match the profile within a given sequence, the best starting and end points of each repeat, and the reliability of each repeat (E-value).

Statistical significance of motif occurrences in one sequence

The statistical significance of the number of YGF motif occurrences in a

sequence was evaluated as follows. Suppose that the amino acid residues within a sequence were shuffled to make any possible sequence under a fixed residue composition. If the number of chance occurrences of the motif in the shuffled sequence was sufficiently smaller than N_{motif} times, then the number of the motif occurrences in one sequence would be considered statistically significant. For example, the probability of the occurrence of the motif AxxBxC is written

$$P(motif) \equiv f(\mathrm{A}) f(\mathrm{B}) f(\mathrm{C}) \tag{4}$$

where A, B and C are particular amino acid residues that characterize the motif, and $f(\mathrm{A})$ is the frequency of the amino acid residue A in the original sequence. When the motif length is much smaller than the sequence length, the probability that a shuffled sequence contains the motif N_{motif} times is given by

$$\left(P(motif)\right)^{Nmotif} \times \left(1 - P(motif)\right)^{L-(l-1)-Nmotif} \times \binom{L-(l-1)}{i} \tag{5}$$

where L is the sequence length, l is the motif length and $\binom{a}{b}$ is $a!/((a-b)! \times b!)$. When a motif length is considered, the probability becomes smaller than that given by equation (5)[23], so that the probability P that one shuffled sequence has the motif more than N_{motif} times is given by the equation

$$P < \sum_{i=N\,motif+1}^{L-(l-1)} \left(P(motif)\right)^{i} \times \left(1-P(motif)\right)^{L-(l-1)-i} \times \binom{L-(l-1)}{i}$$

$$= 1 - \sum_{i=0}^{Nmotif} \left(P(motif)\right)^{i} \times \left(1-P(motif)\right)^{L-(l-1)-i} \times \binom{L-(l-1)}{i}$$

Here, we should consider two kinds of multiplicities. One is the multiplicity of amino acids. The probability we should calculate is not the probability of the occurrence of "YxxxxxGF" but the probability of the occurrence of any motifs "AxxxxxBC" where A, B, C denote one of the amino acids. We must also consider the multiplicity of position of triplets. We should rather calculate the probability of any motifs including triplets "ABC", "ABxC", "AxBC", "ABxxC", "AxBxC", "AxxBC", ... , "AxxxxxxxxxxxxxxxxxxBC" where we regarded that the length of a motif is limited to 20 amino acids.

By Bonferroni correction, we can consider the multiplicity of amino acids easily. As there are 8000 (= 20 x 20 x 20) combination of triplets, the probability of occurrence of any motifs which is about as few as YGF motif is given by $8000 \times P$ where P is given by equation (6). The multiplicity of position can also be considered by the Bonferroni correction. Since the number of motifs

including triplets whose length is L is given by $L-2$, the number of motifs whose length is over than 2 and less than 21 is 176. Finally, the probability we should calculate is given by the following equation.

$$176 \times 8000 \times P \tag{7}$$

7. Results of sequence analysis

Finding repeat sequences in Gli349 and MYPU2110

Through visual inspection, we found that the YGF motif appears 11 times in Gli349 and 16 times in MYPU2110, which are significantly greater numbers of occurrences than one would expect from chance. Using the amino acid residue fractions for Y, G and F in Gli349 (1.8%, 5.0% and 5.6%, respectively), the approximate probability that the YGF motif would appear at least 11 times within the 3183 amino acid residues of Gli349 was calculated to be 4.0×10^{-17} using equation (6):

$$1-\sum_{i=0}^{10}(0.018\times 0.05\times 0.056)^{i}\times(1-0.018\times 0.05\times 0.056)^{3176-i}\times\binom{3176}{i} \tag{8}$$

Considering the multiplicity, the probability of the occurrence of any motifs

including any triplets is 5.8×10^{-12} given by equation (6) and (7). Similar analysis showed the probability for the YGF motif to occur 16 times by chance within MYPU2110 to be 1.5×10^{-26} and the probability for any motifs with any triplets to be 2.1×10^{-21}. These low probabilities show the frequency of the occurrence of the YGF motif to be significant. Notably, the positions of the Y within the YGF motif in Gli349 are 655, 867, 979, 1087, 1588, 1694, 1801, 2009, 2123, 2322 and 2653; those in MYPU2110 are 138, 434, 634, 737, 962, 1066, 1363, 1464, 1679, 1775, 1872, 1977, 2081, 2289, 2415 and 2640. Thus each YGF motif is separated by a multiple of about 100 (100, 200, 300 ...) amino acid residues, which implies the existence of a repeat whose length is a multiple of 100. When we took the greatest common measure (i.e., 100) as the repeat length, we found there to be 11 subsequences of about 100 residues in Gli349 and 16 subsequences in MYPU2110 that were well aligned and had several conserved sites in addition to the YGF motif. Figure 1 shows the MSA of the 11 subsequences in Gli349 generated using clustalW[24, 25]. Note that the region around the YGF motif is conserved best.

It is difficult to determine the start and end of a repeat when the amino acid residues are so weakly conserved among the repeats[26]. Furthermore, most of the repeats in Gli349 and MYPU2110 appear in tandem. Bearing those in mind, suppose repeats composed of 100 amino acid residues were situated contiguously at residue positions 1 to 100, 101 to 200 and 201 to 300 and so on; or they could be shifted 50 residues and start at positions 51 to 150, 151 to 250, 251 to 350, and 351 to 450 and so on. In this way, any position can be the starting point of a repeat if contiguous tandem repeats exist. Here, we assumed the boundaries (the start and end) of the repeats to be one of the residues in the least conserved region. We named this repeat set as Set #1 (Table 1). To exclude starting-point dependency in the alignment, we also employed another alignment in which the starting points were shifted by 50 residues. In this case, the YGF motif was again best conserved, and the least conserved region also agreed with the least conserved region in the prior alignment, demonstrating that the least conserved region is unaffected by the starting point of the alignments.

We then searched for subsequences of Gli349 and MYPU2110 that are

distantly homologous to the repeats in Set #1. Using the knowledge that Gli349 is orthologous to MYPU2110[11, 27, 28], we conjectured that there might be a distantly homologous subsequence in Gli349 that could be aligned to one of the repeats in Set #1 of MYPU2110 and vice versa. In this way, we found an additional seven repeats within Gli349 and two within MYPU2110. Then using the MSA of the 36 (27 + 9) subsequences prepared using clustalW, which we call Set #2, we searched for additional repeat subsequences, and a profile was built using the hidden Markov model with HMMER[15]. When a repeat search using this profile found new repeats, they and the repeats in Set #2 were aligned to build a new profile based on Set #2 (updating the profile), and then the procedure returned to the starting point in the cycle of iterations (see Fig. 2). The cycle was repeated until the alignment at the i-th iteration and the new alignment at the $(i+1)$-th iteration had the same alignment score. After seven iterations, we finally obtained additional repeats, three in Gli349 and four in MYPU2110, whose positions are shown in Table 2. This last repeat set containing 43 repeat sequences was called Set #3. The similarities of all the repeats in Set #3 were

statistically significant (E-value of each repeat in Set #3 was smaller than 2.2×10^{-14} against the profile). And as shown in Fig. 3, the peaks of the alignment scores correspond well to the positions of the repeats. The scores were calculated using a 120-residue long window so that the window would contain the entire repeat. Hereafter, "repeat" denotes repeat sequences in Set #3.

Characteristics of the repeat

We calculated sequence identities among the repeats using pairwise alignments generated with clustalW[24, 25] and found them to fall in a range of 13.0~35.1% for Gli349 and 12.7~37.1% for MYPU2110. The degree of sequence conservation in each column of the MSA of the 43 repeats is shown in the form of a sequence logo[29] in Fig. 4. The YGF motif is the most conserved region within repeats, and there are three regions having high information values: 1 to 7, 27 to 39 and 55 to 60 (Fig. 4). These regions all have a binary pattern of hydrophobic and hydrophilic residues, suggesting that the regions form amphipathic β strands and are located on the surface of the protein. Note that, in addition to these regions, there are several conserved amino acid residues: Gly at

45, Asn at 72, Tyr at 79, Leu, Ile or Val at 118, Phe at 124 and Ile at 126.

The predicted secondary structure of each repeat, determined using the NPS server (http://npsa-pbil.ibcp.fr/cgi-bin/npsa_automat.pl?page=/NPSA/npsa_seccons.html)[30], was found to be ββββα,with the second β-strand being well conserved among all repeats (Fig. 5). The locations of the predicted β-strands are particularly consistent with the binary pattern, which suggests the existence of three β-strands. The most highly conserved residues in the YGF motif, Gly and Phe, are found in the region extending from the middle of the third β-strand to the beginning of the following α-helix. Both the primary and secondary structures of the region are well conserved, which may suggest that those regions are important for determining the structure of Gli349. The secondary structure near the C-terminus is less well conserved (Fig. 5), suggesting this region may be a linker between domains. The characteristics of the predicted secondary structures of the repeats in MYPU2110 are similar to those of Gli349.

Gli349 is comprised of 3183 amino acid residues; the last repeat starts at

position 2,607 and ends at 2,729. The N-terminal 86% of Gli349 is mostly composed of repeats, suggesting the C-terminal region takes a different structure than the rest of the molecule. Repeats C to I, J to R and S to V form three contiguous tandem repeats (Fig. 3). Interestingly, the gap lengths between repeats I and J and between repeats R and S are about 100 amino acid residues long (90 between repeats I and J, and 91 between repeats R and S), or about as long as one repeat. This suggests to us that the gaps are also repeats, but have become too divergent for detection of sequence similarity using currently available methods.

Search for homologous proteins using repeat profiles

Using HMMER with a profile based on the MSA of repeat Set #3, we searched for sequences homologous to both Gli349 and MYPU2110 repeats in the NCBI reference sequence (RefSeq) database[31, 32] (http://www.ncbi.nlm.nih.gov/RefSeq/), which is a non-redundant protein sequence dataset, but no homologous sequences with E-value less than 0.1 were

found. We also checked whether either Gli349 or MYPU2110 matches any profile of known repeat sequences compiled in the REP database[26]. Neither Gli349 nor MYPU2110 matched any profile in the REP database, suggesting that the repeat in Gli349 and MYPU2110 is novel.

Chymotrypsin treatment

Chymotrypsin breaks the C-terminal peptide bonds of aromatic residues that are exposed to solvent and are flexible and could, therefore, be used to discover exposed flexible regions in the Gli349. Knowledge about such regions would provide a hint about the structure of the entire protein, as well as about the structure of each repeat. Based on the experimentally obtained molecular weights of the fragments of Gli349 (unpublished data) and our understanding of the peptide bonds targeted by chymotrypsin, we estimated the positions (shown in Fig. 3(a) and Table 3) of 20 cleavage sites. Of those, 15 fell within either the non-repeat region or the boundary region of the repeats (Fig. 3(a)), suggesting that these areas are exposed to solvent and are flexible. We suggest that these

regions might be the linkers between domains, which is consistent with our assumption that each repeat sequence folds into a structural domain. The remaining five sites fell within the repeats: one was on the second β-strand, one was on the third β-strand, two were on the α-helix, and one was in the region just after the α-helix. We expect that these areas are exposed to the solvent, as their sequences appear to form an amphipathic β-strand or α-helix.

8. Discussion for sequence analysis

Proteins with tandem repeat sequences

We have found that Gli349 is a protein comprised of tandem repeats, most of which are marked by a YGF motif. Structures with tandem repeats are known to occur either in linear arrays or superhelical structures with repeats arranged around a common axis, as is seen in the β-propeller structure[33]. Either of the structures presents an extensive surface that is well suited for interaction with other molecules. Indeed, so far the best-known function of the proteins with known repeats is the binding of other proteins[33]. We suggest that Gli349 also

provides an extensive surface with which it interacts with other molecules.

This is the first description of the YGF motif, which was situated within each repeat. When we searched the SWISS-PROT database[34] for proteins containing repeat sequences with the YGF motif, we only found kelch-like protein 10 from *Homo sapien*[35]. The kelch repeat has the superhelical structure of a six-bladed β-propeller in which the repeats each consist of about 40 amino acid residues and are arranged around a common axis[36]. Gli349 is unlikely to assume a structure similar to kelch, because 1) the YGF motif is not conserved in any kelch repeats, 2) the length of repeats differs from the kelch repeat, and 3) electron microscopy (EM) images of the Gli349 (discussed later) show an overall rod-shaped structure.

Proteins with an Ig-fold repeat are often found on cell surfaces and mediate interactions with other cells, as is the case with Filamin, which contains six Ig-folds in a chain and functions to cross-link pairs of F-actin chains[37]. The size of the Ig-fold is about 100 residues, which resembles the size of the YGF-containing repeat in Gli349, and is also an all-β structure. We therefore

used PSIPRED[38] and FORTE[39] to thread the repeat sequences of Gli349 and MYPU2110 on the known Ig-fold structures to test whether the repeat sequences are compatible with the Ig-fold. No repeat in Gli349 or MYPU2110 fit into any known Ig-fold structure, however. Apparently, the YGF-containing repeat does not assume an Ig-fold structure, but we will need further structural determination to confirm this speculation.

Implication for richness of Asparagine

One of the characteristics of the Gli349 sequence is that it is rich in Asn residues, which accounts for 12.0% of the residues making up Gli349, or about three times more than the average fraction (4.3%) of Asn residues in all the protein sequences in SWISS-PROT[34]. There are also Asn-rich proteins in *Plasmodium falciparum*, which is responsible for malaria in humans. It has been suggested that the Asn-rich proteins in *P. falciparum* may be useful for avoiding host immunogenic responses[40]. Gli349 may have similar features. *M. mobile* lives as a parasite in the gill organ cells of freshwater fish, which are exposed to

the water. Almost all of Gli349 is predicted to be outside the cell membrane by SOSUI[41]. In addition, P1 adhesin from *M. pneumoniae*, which has been known to be responsible for the binding to animal cells and glass surfaces[42-45] and to be related with avoiding immunogenic responses, was recently suggested to directly participate in the gliding[46]. Taken together, these results suggest that Gli349 may also play a role in enabling *M. mobile* to escape host immunogenic responses.

Structural model based on sequence analysis and EM imaging

Under the assumption that each repeat sequence folds into a structural domain, we determined the tertiary structure of Gli349 predicted using the ROBETTA server (http://robetta.bakerlab.org) and the ROSETTA algorithm[47, 48] to estimate the size of the domain. ROSETTA is one of the most successful methods for *ab initio* tertiary structure prediction. We predicted the structures of repeats K and N because they had the highest scores in the alignment used for the HMM profile. Ten predicted tertiary structures per input sequence were obtained with ROBETTA. The sizes of the predicted structures were similar (the average

size was 4.2 ± 0.5 nm, where the length is defined as the farthest distance between two Cα atoms), though they exhibited a large variety of folds.

A preliminary EM image of Gli349 shows the shape of Gli349 to have an inverted Z-like structure and to be composed of at least four parts (Fig. 6). It also showed that the joint between rod 1 and rod 2 is very flexible, whereas the joint between rod 2 and rod 3 is very rigid (Fig. 6). We then tried to place the predicted structure of the repeat sequences on the image of Gli349, taking into account the rough estimation of the length of Gli349, and assuming that the lengths of the three rods are proportional to the number of residues and that the entire structure is a string-like filament. We found that there are two ways to place the repeats on the image: the N-terminus of Gli349 can be assigned to either the tip (model 1, Fig. 6(a)) or the base to the body of *M. mobile* (model 2, Fig. 6(b)). In both assignments, there are nine repeats and two non-repeat regions within the 46-nm rod 1 (Fig. 6). We can estimate that the length of one repeat is shorter than 5.1 nm (= 46/9), which agrees well with the average size of the predicted repeat structures. Because Gli349 is predicted to have a transmembrane

region near the N-terminus and the mutation of Ser to Leu at 2770 in Gli349, where the mutation site is to be located in the oval region in model 1, cannot adhere to glass (unpublished data), we propose that model 1 more accurately depicts the true structure of Gli349, though in both models the non-repeat regions correspond well to the flexible joints.

9. Summary of the thesis

In the study of sampling problem on MD simulation, we found that the conformational space of protein consists of a series of subspaces which are mutually quite independent and that subspaces with smaller amplitudes of variation are identified and equilibrated during the molecular dynamics simulation before those with larger amplitudes of variation are identified and equilibrated. We can say that protein molecular dynamics simulation is a useful tool even with its intrinsic multi-dimensionality.

In the study of the sequence analysis of *M.mobile*, we found the repeat including YGF motif for 21 and 22 times for Gli349 and MYPU2110,

respectively. No sequence homologous to the repeat could be found in the RefSeq non-redundant sequence database. No compatible fold structure was found among known protein structures. These facts suggest this repeat is novel and takes a new fold structure. We also found that the cleavage positions were often located between the repeats by the proteolysis experiment of Gli349 using chymotrypsin. This implies that regions connecting repeats are unstructured and the repeat regions are structured and non-flexible. As we assumed each repeat folds into a structural domain, we constructed a model of Gli349 that fits well into the shape and size of images obtained by electron microscopy.

These two studies are included in the category of bioinformatics and are particular cases of the basic problem in molecular biology that is to identify and find the characteristics of the proteins. These works may shed light on basic problems in bioinformatics, which are structure prediction from sequence and solution of the relationship between structure and function of protein.

Acknowledgements

We warmly thank excellent researchers Nobuhiro Go, Hidetoshi Kono and Kei Yura at Japan Atomic Energy Research Institute and Professor Akio Kitao at University of Tokyo for their help. We were supported by Professor Makoto Miyata and students Atsuko Uenoyama and Jun Adan for experimental data at Osaka City University. We also thank Professors Shin Ishii, Takeshi Kawabata and Gautam Basu at NAIST for their support. We thank Dr. Kentaro Tomii at AIST for helping us using FORTE. Computations reported in this work were carried out at JAERI using ITBL computer. This work was supported in part by Special Coordination Funds Promoting Science and Technology from MEXT (Ministry of Education, Culture, Sports, Science and Technology, Japan).

Figure Captions

Fig. 1: Convergence of diagonalized second moments calculated for MD simulation of HEW lysozyme plotted against length of simulation. (a) 1st mode. (b) 31st mode. (c) 301st mode. (d) 5000th mode. The ordinate indicating second moments originally has a physical dimension of atomic mass times atomic coordinate squared. But the values scaled by mass of carbon atom are shown, so that the values are in a more familiar range of Angstrom squared.

Fig. 2: Convergence of diagonalized second moments calculated for Langevin simulation of BPTI plotted against length of simulation. The ordinate indicates the difference of the value of diagonalized second moment from its equilibrium value scaled by the equilibrium value.

Fig. 3: Fractional convergence of diagonalized second moments calculated for Langevin simulation of BPTI plotted against length of simulation. (a) 1st and 2nd modes. (b) Square root of average of squared fractional convergence for modes of 3rd to 15th, of 16th to 30th, and 31st to 150th.

Fig. 4: Difference between diagonalized second moments calculated for Langevin simulation in the 300 dimensional space and in the lower 150 dimensional space. Ordinate indicates the difference scaled by its theoretical equilibrium value.

Fig. 5: Multiple sequence alignment of 11 subsequences of Gli349 containing a YGF motif is shown. The start and end positions of each repeat are denoted in the first column. Colors on the sequences denote as follows: yellow, 100% conserved residues; cyan, >50% conserved residues; orange, sites that are >70% conserved for D, N, S and T; green, sites that are >70% conserved for A, I, L and V.

Fig. 6: Procedure for detecting repeats using the hidden Markov model.

Fig. 7: Alignment scores of subsequences of 120 residues against the profile of repeat Set #3. Plotted are the scores at the center position of the subsequences of Gli349 (a) and MYPU2110 (b). Scores were calculated using HMMER[22]. The unit on the vertical axis is the negative logarithm of the E-value of the alignment. The bars above the line denote repeats

detected by HMMER[22]. Most of the repeats were found to be in tandem form. For Gli349, experimentally determined chymotrypsin susceptible sites are shown by asterisks (a).

Fig. 8: The degree of residue conservation at each position in the repeat is shown as information content. The information content of amino acid residue "a" at position "i" is calculated by the equation, $I(a,i) = -\log_2 p(a,i)$, where $p(a,i)$ is the fractional content comprised by an amino acid residue. $\sum_a p(a,i) = 1$ at each position. As $\sum_a I(a,i)$ becomes larger, the position "i" is regarded as more conserved. The three black bars indicate well conserved regions.

Fig. 9: MSA of repeat Set #3. Residues are colored according to the predicted secondary structures: red, α-helix; blue, β-strand; yellow, coil; and white, ambiguous. The consensus secondary structure is βββα shown in the bottom. The subsequences are listed in order of their E-value. The YGF motif is denoted by a black bar on the top line. GLI and MYPU denote Gli349 and MYPU2110, respectively, and the following number denotes

the repeat start position in each subsequence. Residue numbering is the same as in Fig. 4.

Fig. 10: Model of Gli349. Low resolution image of Gli349 obtained with electron microscopy (EM) is shown in gray shade. Repeat regions shown in ovals connected by lines are assigned into the EM image of Gli349. The N-terminus is placed at the far right side in (a), and at the far left side in (b). The length of each rod and angles between two rods are the average values over EM images (unpublished data).

References

1. Hess, B. Similarities between principal components of protein dynamics and random diffusion. *Phys. Rev. E.* **62**, 8438-8448 (2000).
2. Hess, B. Convergence of sampling in protein simulations. *Phys. Rev. E.* **65**, 031910 (2002).
3. Hayward, S., Kitao, A., Hirata, F. & Go, N. Effect of solvent on collective motions in globular protein. *J. Mol. Biol.* **234**, 1207-1217 (1993).
4. Kitao, A., Hayward, S. & Go, N. Energy landscape of a native protein: jumping-among-minima model. *Proteins.* **33**, 496-517 (1998).
5. Ermak, D. L. & McCammon, J. A. Brownian dynamics with hydrodynamic interactions. *J. Chem. Phys.* **69**, 1352 (1978).
6. Ermak, D. L. & Buckholz, H. Numerical Integration of the Langevin Equation: Monte Carlo Simulation. *J. Comp. Phys.* **35**, 169 (1980).
7. Chandrasekhar, S. Stochastic Problems in Physics and Astronomy. *Rev. Mod. Phys.* **15**, 1-89 (1943).
8. Shimizu, T. & Miyata, M. Electron microscopic studies of three gliding Mycoplasmas, Mycoplasma mobile, M. pneumoniae, and M. gallisepticum, by using the freeze-substitution technique. *Curr Microbiol.* **44**, 431-434 (2002).
9. Fraser, C. M., Gocayne, J. D., White, O., Adams, M. D., Clayton, R. A., Fleischmann, R. D., Bult, C. J., Kerlavage, A. R., Sutton, G., Kelley, J., M.J. L. Fritchman, J. F. Weidman, K. V. Small, M. Sandusky, J. Fuhrmann, D. Nguyen, T. R. Utterback, D. M. Saudek, C. A. Phillips, J. M. Merrick, J.-F. Tomb, B. A. Dougherty, K. F. Bott, P.-C. Hu & Lucier, T. S. The minimal gene complement of Mycoplasma genitalium. *Science.* **270**, 397-403 (1995).
10. Himmelreich, R., Hilbert, H., Plagens, H., Pirkl, E., Li, B. C. & Herrmann, R. Complete sequence analysis of the genome of the bacterium Mycoplasma pneumoniae. *Nucleic Acids Res.* **24**, 4420-4449 (1996).

11. Chambaud, I., Heilig, R., Ferris, S., Barbe, V., Samson, D., Galisson, F., Moszer, I., Dybvig, K., Wroblewski, H., Viari, A., Rocha, E. P. & Blanchard, A. The complete genome sequence of the murine respiratory pathogen Mycoplasma pulmonis. *Nucleic Acids Res.* **29**, 2145-2153 (2001).
12. McBride, M. J. Bacterial gliding motility: multiple mechanisms for cell movement over surfaces. *Annu Rev Microbiol.* **55**, 49-75 (2001).
13. Miyata, M. & Uenoyama, A. Movement on the cell surface of the gliding bacterium, Mycoplasma mobile, is limited to its head-like structure. *FEMS Microbiol Lett.* **215**, 285-289 (2002).
14. Miyata, M., Ryu, W. S. & Berg, H. C. Force and velocity of mycoplasma mobile gliding. *J Bacteriol.* **184**, 1827-1831 (2002).
15. Miyata, M., Yamamoto, H., Shimizu, T., Uenoyama, A., Citti, C. & Rosengarten, R. Gliding mutants of Mycoplasma mobile: relationships between motility and cell morphology, cell adhesion and microcolony formation. *Microbiology.* **146 (Pt 6)**, 1311-1320 (2000).
16. Kusumoto, A., Seto, S., Jaffe, J. D. & Miyata, M. Cell surface differentiation of *Mycoplasma mobile* visualized by surface protein localization. *Microbiology.* **150**, 4001-4008 (2004).
17. Uenoyama, A., Kusumoto, A. & Miyata, M. Identification of a 349-kilodalton protein (Gli349) responsible for cytadherence and glass binding during gliding of Mycoplasma mobile. *J Bacteriol.* **186**, 1537-1545 (2004).
18. Miyata, M. & Petersen, J. D. Spike structure at the interface between gliding Mycoplasma mobile cells and glass surfaces visualized by rapid-freeze-and-fracture electron microscopy. *J Bacteriol.* **186**, 4382-4386 (2004).
19. Krogh, A., Brown, M., Mian, I. S., Sjolander, K. & Haussler, D. Hidden Markov models in computational biology. Applications to protein modeling. *J Mol Biol.* **235**, 1501-1531 (1994).
20. Eddy, S. R. Hidden Markov models. *Curr Opin Struct Biol.* **6**, 361-365 (1996).
21. Eddy, S. R. Multiple alignment using hidden Markov models. *Proc Int Conf Intell Syst Mol Biol.* **3**, 114-120 (1995).

22. Eddy, S. R. Profile hidden Markov models. *Bioinformatics.* 14, 755-763 (1998).

23. Ewens, W. J. & Grant, G. R. *Statistical Methods in Bioinformatics.* (Springer, New York, 2002).

24. Thompson, J. D., Higgins, D. G. & Gibson, T. J. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* 22, 4673-4680 (1994).

25. Higgins, D. G., Thompson, J. D. & Gibson, T. J. Using CLUSTAL for multiple sequence alignments. *Methods Enzymol.* 266, 383-402 (1996).

26. Andrade, M. A., Ponting, C., Gibson, T. & Bork, P. Identification of protein repeats and statistical significance of sequence comparisons. *J. Mol. Biol.* 298, 521-537 (2000).

27. Miyata, M. Gliding motility of mycoplasma--a mechanism cannot be explained by today's biology. *Nippon Saikingaku Zasshi.* 57, 581-595 (2002).

28. Miyata, M. Gliding motility of mycoplasmas - the mechanism cannot be explained by current biology. in *Mycoplasmas: pathogenesis, molecular biology, and emerging strategies for control.* (G. Browning ed.) vol. pp. in press (Horizon Scientific Press, 2005).

29. Schneider, T. D. & Stephens, R. M. Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res.* 18, 6097-6100 (1990).

30. Perriere, G., Combet, C., Penel, S., Blanchet, C., Thioulouse, J., Geourjon, C., Grassot, J., Charavay, C., Gouy, M., Duret, L. & Deleage, G. Integrated databanks access and sequence/structure analysis services at the PBIL. *Nucleic Acids Res.* 31, 3393-3399 (2003).

31. Pruitt, K. D., Katz, K. S., Sicotte, H. & Maglott, D. R. Introducing RefSeq and LocusLink: curated human genome resources at the NCBI. *Trends Genet.* 16, 44-47 (2000).

32. Pruitt, K. D. & Maglott, D. R. RefSeq and LocusLink: NCBI gene-centered resources. *Nucleic Acids Res.* 29, 137-140 (2001).

33. Andrade, M. A., Perez-Iratxeta, C. & Ponting, C. P. Protein repeats: structures, functions, and evolution. *J Struct Biol.* 134, 117-131 (2001).

34. Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M.-C., Estreicher, A., Gasteiger, E., Martin, M. J., Michoud, K., O'Donovan, C., Phan, I., Pilbout, S. & Schneider, M. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.* **31**, 365-370 (2003).

35. Yan, W., Ma, L., Burns, K. H. & Matzuk, M. M. Haploinsufficiency of kelch-like protein homolog 10 causes infertility in male mice. *Proc Natl Acad Sci U S A.* **101**, 7793-7798 (2004).

36. Li, X., Zhang, D., Hannink, M. & Beamer, L. J. Crystal structure of the Kelch domain of human Keap1. *J Biol Chem.* (2004).

37. Popowicz, G. M., Muller, R., Noegel, A. A., Schleicher, M., Huber, R. & Holak, T. A. Molecular structure of the rod domain of dictyostelium filamin. *J Mol Biol.* **342**, 1637-1646 (2004).

38. McGuffin, L., Bryson, K. & Jones, D. The PSIPRED protein structure prediction server. *Bioinformatics.* **16**, 404-405 (2000).

39. Tomii, K. & Akiyama, Y. FORTE: a profile-profile comparison tool for protein fold recognition. *Bioinformatics.* **20**, 594-595 (2004).

40. Brocchieri, L. Low-complexity regions in Plasmodium proteins: in search of a function. *Genome Res.* **11**, 195-197 (2001).

41. Hirokawa, T., Boon-Chieng, S. & Mitaku, S. SOSUI: classification and secondary structure prediction system for membrane proteins. *Bioinformatics.* **14**, 378-379 (1998).

42. Feldner, J., Gobel, U. & Bredt, W. *Mycoplasma pneumoniae* adhesin localized to tip structure by monoclonal antiboy. *Nature.* **298**, 765-767 (1982).

43. Hu, P. C., Cole, R. M., Huang, Y. S., Graham, J. A., Gardner, D. E., Collier, A. M. & Clyde, J. W. A. Mycoplasma pneumoniae infection: role of a surface protein in the attachment organelle. *Science.* **216**, 313-315 (1982).

44. Razin, S. & Jacobs, E. Mycoplasma adhesion. *J. Gen. Microbiol.* **138**, 407-422 (1992).

45. Svenstrup, H. F., Nielsen, P. K., Drasbek, M., Birkelund, S. & Christiansen, G. Adhesion and inhibition assay of Mycoplasma genitalium and M. pneumoniae by immunofluorescence microscopy. *J. Bacteriol.* **51**, 361-373 (2002).

46. Seto, S., Kenri, T., Tomiyama, T. & Miyata, M. Involvement of P1 adhesin in gliding motility of *Mycoplasma pneumoniae* as revealed by the inhibitory effects of antibody under optimized gliding conditions. *J. Bacteriol.* in press. (2005).

47. Simons, K. T., Bonneau, R., Ruczinski, I. & Baker, D. Ab initio protein structure prediction of CASP III targets using ROSETTA. *Proteins.* **Suppl 3**, 171-176 (1999).

48. Simons, K. T., Kooperberg, C., Huang, E. & Baker, D. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. *J Mol Biol.* **268**, 209-225 (1997).

Tables

Table 1. Positions of the repeats in repeat Sets #1 and #2

Set #1		Set #2	
Gli349	MYPU2110	Gli349	MYPU2110
	96-215	101-220	96-215
	391-510	386-506	391-510
601-720	591-710	601-720	591-710
	691-810	706-825	691-810
816-935		816-935	796-916
926-1045	921-1040	926-1045	921-1040
1036-1155	1021-1140	1036-1155	1021-1140
	1321-1440	1331-1450	1321-1440
	1421-1540	1431-1550	1421-1540
1536-1655		1536-1655	1541-1660
1641-1760	1641-1760	1641-1760	1641-1760
1751-1870	1741-1860	1751-1870	1741-1860
	1839-1958	1851-1970	1839-1958
1956-2075	1941-2060	1956-2075	1941-2060
2071-2190	2041-2160	2071-2190	2041-2160
2271-2390	2241-2360	2271-2390	2241-2360
	2381-2500	2381-2500	2381-2500
2601-2720	2601-2720	2601-2720	2601-2720

Table 2. Positions of the repeats in repeat Set #3

repeat ID	Gli349	MYPU2110	repeat ID	Gli349	MYPU2110
A	109-222	101-218	L	1452-1548	1428-1543
B		296-393	M	1552-1658	1544-1641
C	395-503	394-503	N	1662-1764	1642-1740
D	504-595	504-593	O	1765-1873	1742-1845
E	611-717	597-695	P	1876-1965	1847-1946
F	721-831	700-801	Q	1966-2079	1947-2047
G	834-941	803-917	R	2083-2193	2049-2161
H	943-1043	925-1028	S	2284-2387	2246-2355
I	1046-1157	1030-1146	T	2395-2496	2379-2498
J	1247-1342	1230-1322	U	2505-2605	2502-2593
K	1343-1451	1326-1427	V	2607-2729	2597-2716

Table 3. Positions cleaved by chymotrypsin

molecular weight	position	molecular weight	position
32.0	292	69.4	631
34.7	316	73.4	667
35.8	326	98.5	893
37.1	337	115.0	1052
42.3	392	131.6	1206
44.7	405	181.4	1665
48.5	438	197.1	1816
49.8	450	215.6	1982
65.5	594	241.3	2219
67.5	613	245.0	2253

The positions were calculated based on the molecular weight (MW) and the known target sites of chymotrypsin.

Fig. 1

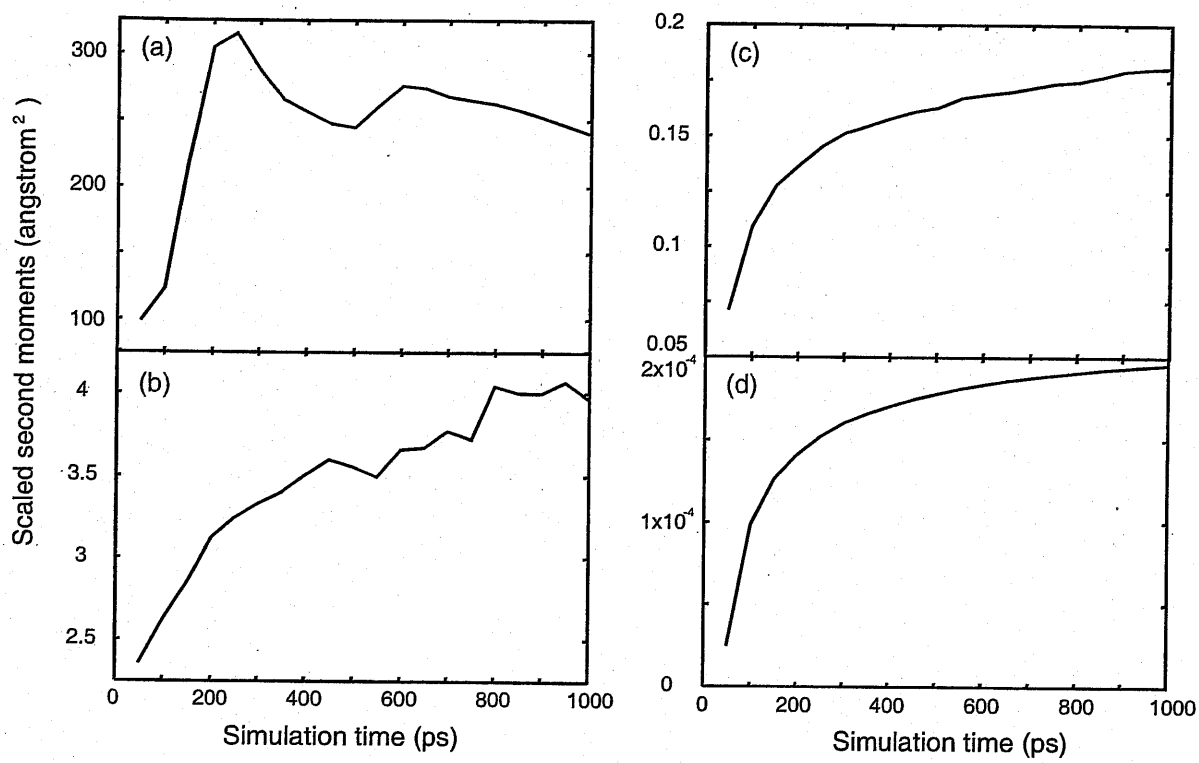

Scaled second moments (angstrom^2)
(a)
(b)
(c)
(d)
300
200
100
4
3.5
3
2.5
0.2
0.15
0.1
0.05
2x10^{-4}
1x10^{-4}
0
0
200
400
600
800
1000
Simulation time (ps)
Simulation time (ps)

Fig. 2

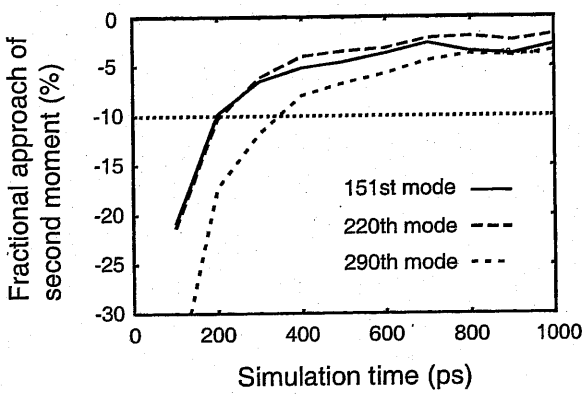

Fractional approach of second moment (%)
0
-5
-10
-15
-20
-25
-30
0
200
400
600
800
1000
Simulation time (ps)
151st mode
220th mode
290th mode

Fig. 3

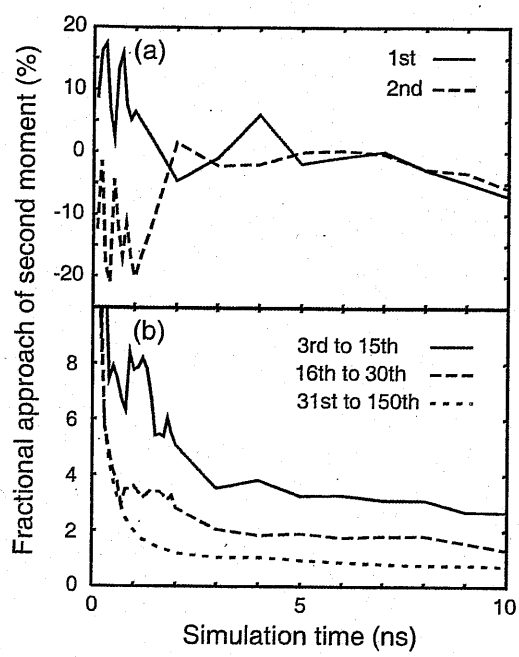

(a)
1st
2nd
(b)
3rd to 15th
16th to 30th
31st to 150th
Fractional approach of second moment (%)
Simulation time (ns)
20
10
0
-10
-20
8
6
4
2
0
0
5
10

Fig. 4

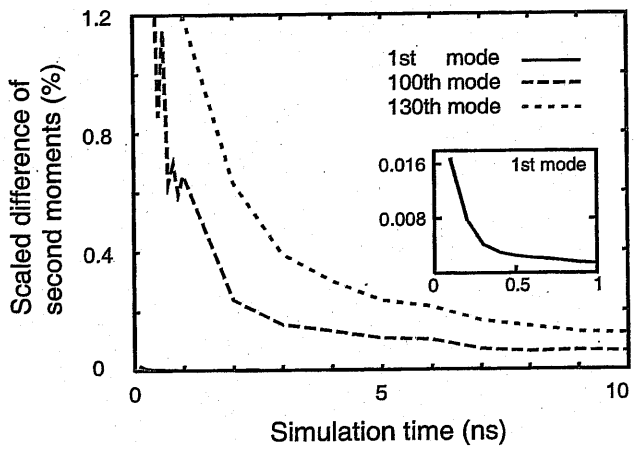

1.2
0.8
0.4
0
0
5
10
Scaled difference of second moments (%)
Simulation time (ns)
1st mode
100th mode
130th mode
0.016
0.008
1st mode
0
0.5
1

Fig. 5

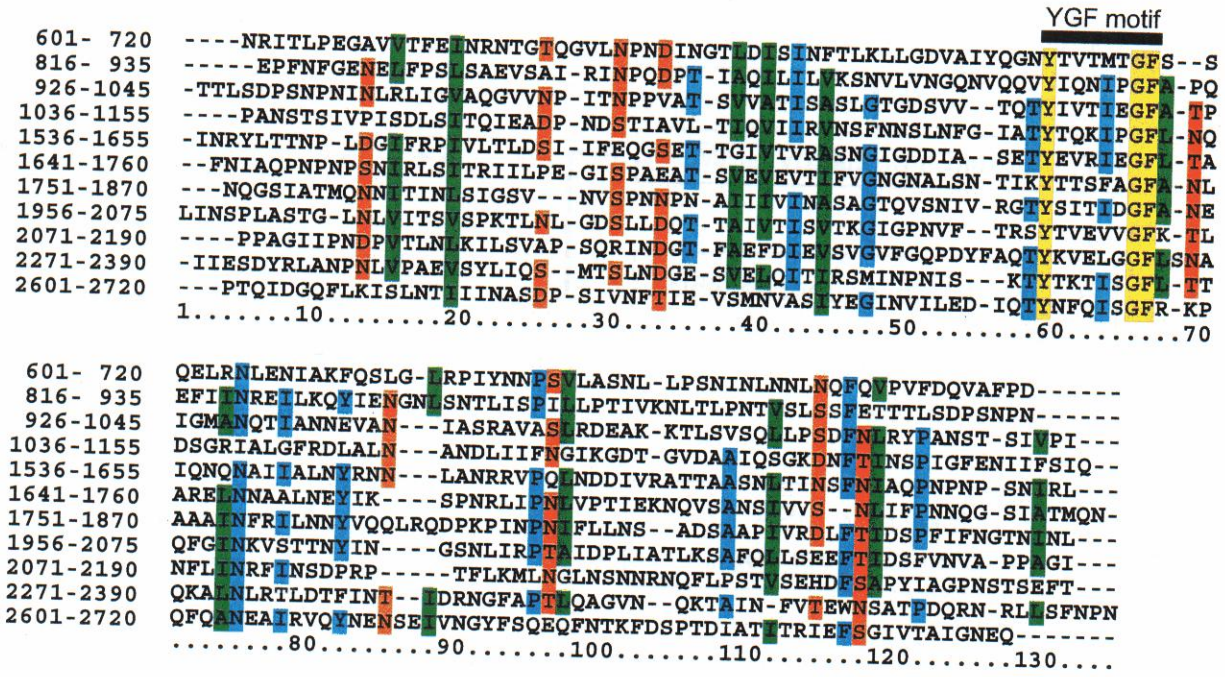

YGF motif
601- 720 ----NRITLPEGAVVTFEINRNTGIQGVLNPNDINGTLDISINFTLKLLGDVAIYQGNYTVTMTGFS--S
816- 935 -----EPFNFGENELFPSLSAEVSAI-RINPQDPT-IAQILILVKSNVLVNGQNVQQVYIQNIPGFA-PQ
926-1045 -TTLSDPSNPNINLRLIGVAQGVVNP-ITNPPVAT-SVVATISASLGTGDSVV--TQTYIVTIEGFA-TP
1036-1155 ----PANSTSIVPISDLSITQIEADP-NDSTIAVL-TIQVIIRVNSFNNSLNFG-IATYTQKIPGFL-NQ
1536-1655 -INRYLTTNP-LDGIFRPIVLTLDSI-IFEQGSET-TGIVTVRASNGIGDDIA--SETYEVRIEGFL-TA
1641-1760 --FNIAQPNPNPSNIRLSITRIILPE-GISPAEAT-SVEVEVTIFVGNGNALSN-TIKYTTSFAGFA-NL
1751-1870 ---NQGSIATMQNNTTINLSIGSV---NVSPNNPN-AIIIVTNASAGTQVSNIV-RGTYSITIDGFA-NE
1956-2075 LINSPLASTG-LNLVITSVSPKTLNL-GDSLLDQT-TAIVTISVTKGIGPNVF--TRSYTVEVVGFK-TL
2071-2190 ----PPAGIIPNDPVTLNLKILSVAP-SQRINDGT-FAEFDIEVSVGVFGQPDYFAQTYKVELGGFLSNA
2271-2390 -IIESDYRLANPNLVPAEVSYLIQS--MTSLNDGE-SVELQITTRSMINPNIS---KTYTKTISGFL-TT
2601-2720 ---PTQIDGQFLKISLNTIIINASDP-SIVNFTIE-VSMNVASIYEGINVILED-IQTYNFQISGFR-KP
1.....10.....20.....30.....40.....50.....60.....70
601- 720 QELRNLENIAKFQSLG-LRPIYNNPSVLASNL-LPSNINLNNLNQFQVPVFDQVAFPD------
816- 935 EFIINREILKQYIENGNLSNTLISPILLPTIVKNLTLPNTVSLSSFETTTLSDPSNPN------
926-1045 IGMANQTIANNEVAN---IASRAVASLRDEAK-KTLSVSQLLPSDFNLRYPANST-SIVPI---
1036-1155 DSGRIALGFRDLALN---ANDLIIFNGIKGDT-GVDAAIQSGKDNFTINSPIGFENIIFSIQ--
1536-1655 IQNQNAIIALNYRNN---LANRRVPQLNDDIVRATTAASNLTINSFNIAQPNPNP-SNIRL---
1641-1760 ARELNNAALNEYIK----SPNRLIPNLVPTIEKNQVSANSIVVS--NLIFPNNQG-SIATMQN-
1751-1870 AAAINFRILNNYVQQLRQDPKPINPNIFLLNS--ADSAAPIVRDLFTIDSPFIFNGTNINL---
1956-2075 QFGINKVSTTNYIN----GSNLIRPTAIDPLIATLKSAFQLLSEEFTIDSFVNVA-PPAGI---
2071-2190 NFLINRFINSDPRP-----TFLKMLNGLNSNNRNQFLPSTVSEHDFSAPYIAGPNSTSEFT---
2271-2390 QKALNLRTLDTFINT--IDRNGFAPTLQAGVN--QKTAIN-FVTEWNSATPDQRN-RLLSFNPN
2601-2720 QFQANEAIRVQYNENSEIVNGYFSQEQFNTKFDSPTDIATYTRIEFSGIVTAIGNEQ-------
.....80.....90.....100.....110.....120.....130....

Fig. 6

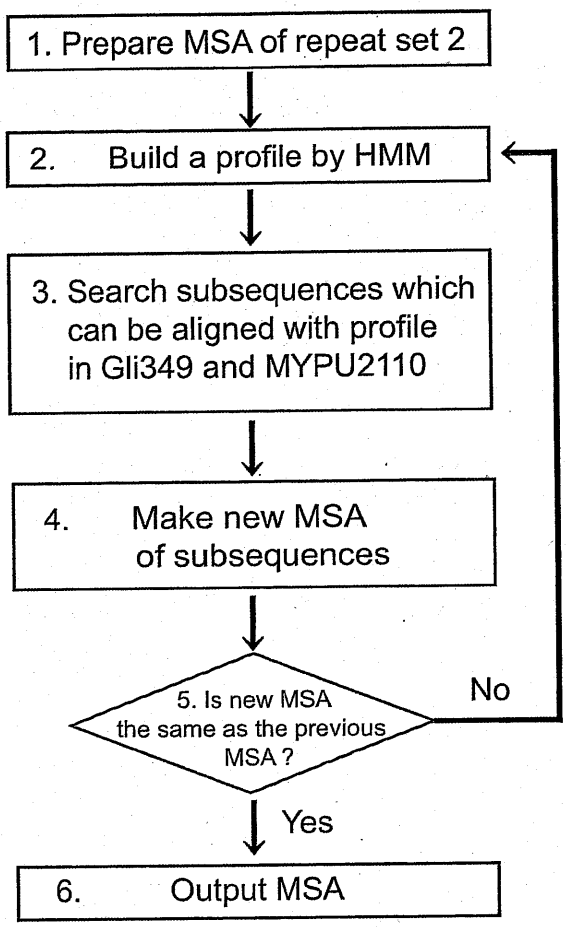

1. Prepare MSA of repeat set 2
2. Build a profile by HMM
3. Search subsequences which can be aligned with profile in Gli349 and MYPU2110
4. Make new MSA of subsequences
5. Is new MSA the same as the previous MSA ?
No
Yes
6. Output MSA

Fig. 7 (a)

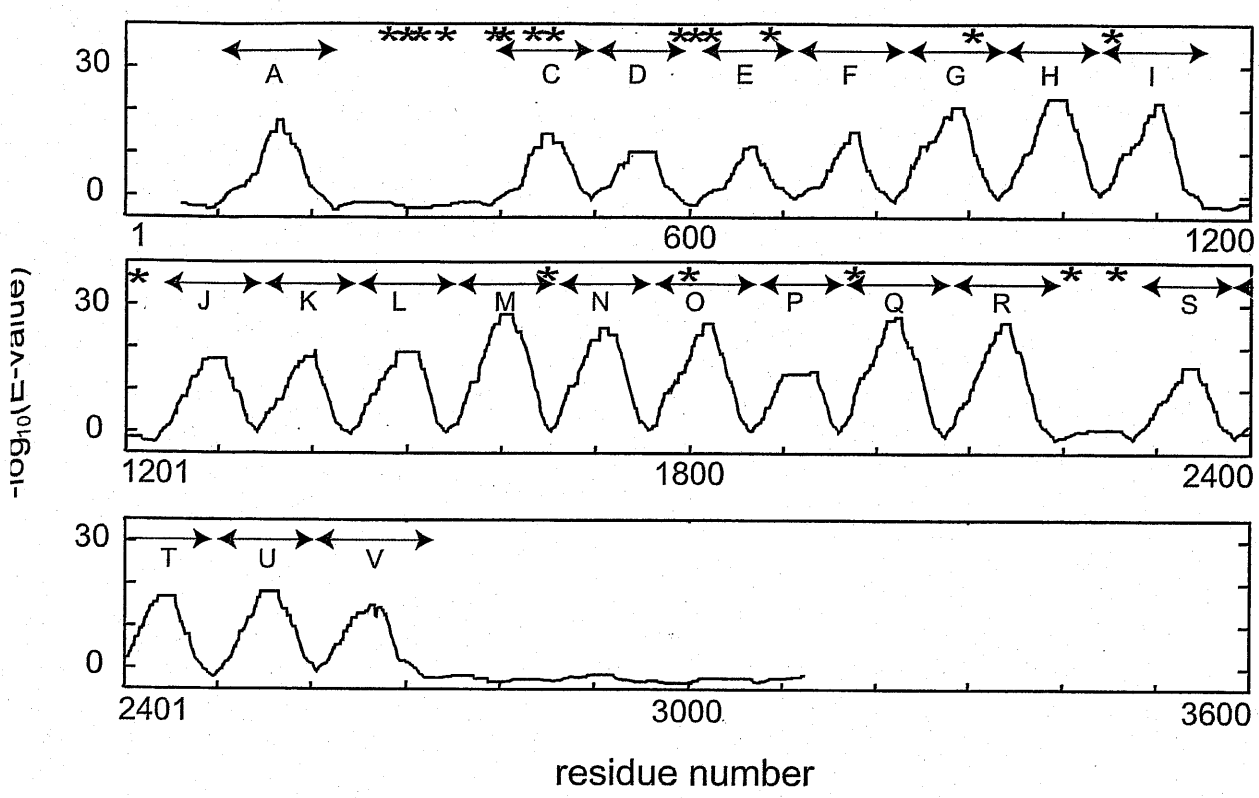

30
0
A
C
D
E
F
G
H
I
1
600
1200
J
K
L
M
N
O
P
Q
R
S
1201
1800
2400
T
U
V
2401
3000
3600
-log10(E-value)
residue number

Fig. 7 (b)

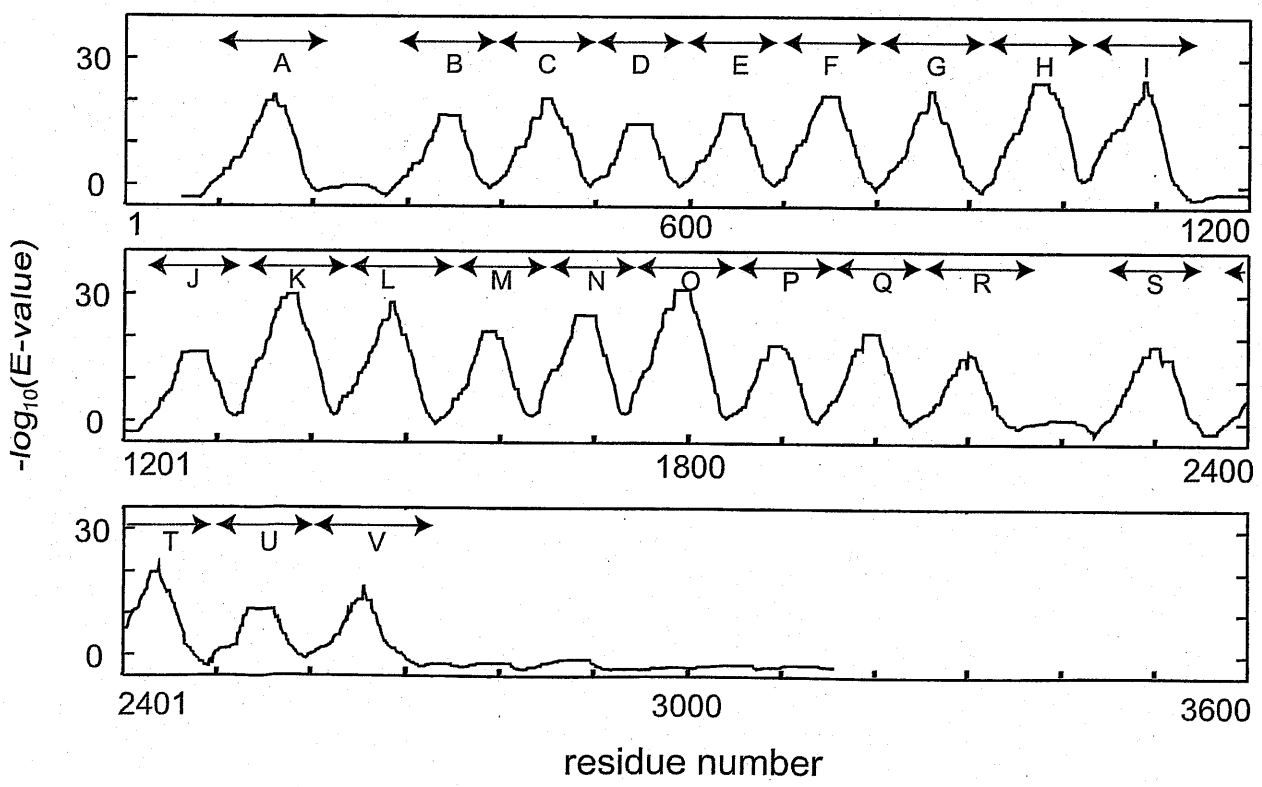

30
0
A
B
C
D
E
F
G
H
I
1
600
1200
J
K
L
M
N
O
P
Q
R
S
1201
1800
2400
T
U
V
2401
3000
3600
-log10(E-value)
residue number

Fig. 8

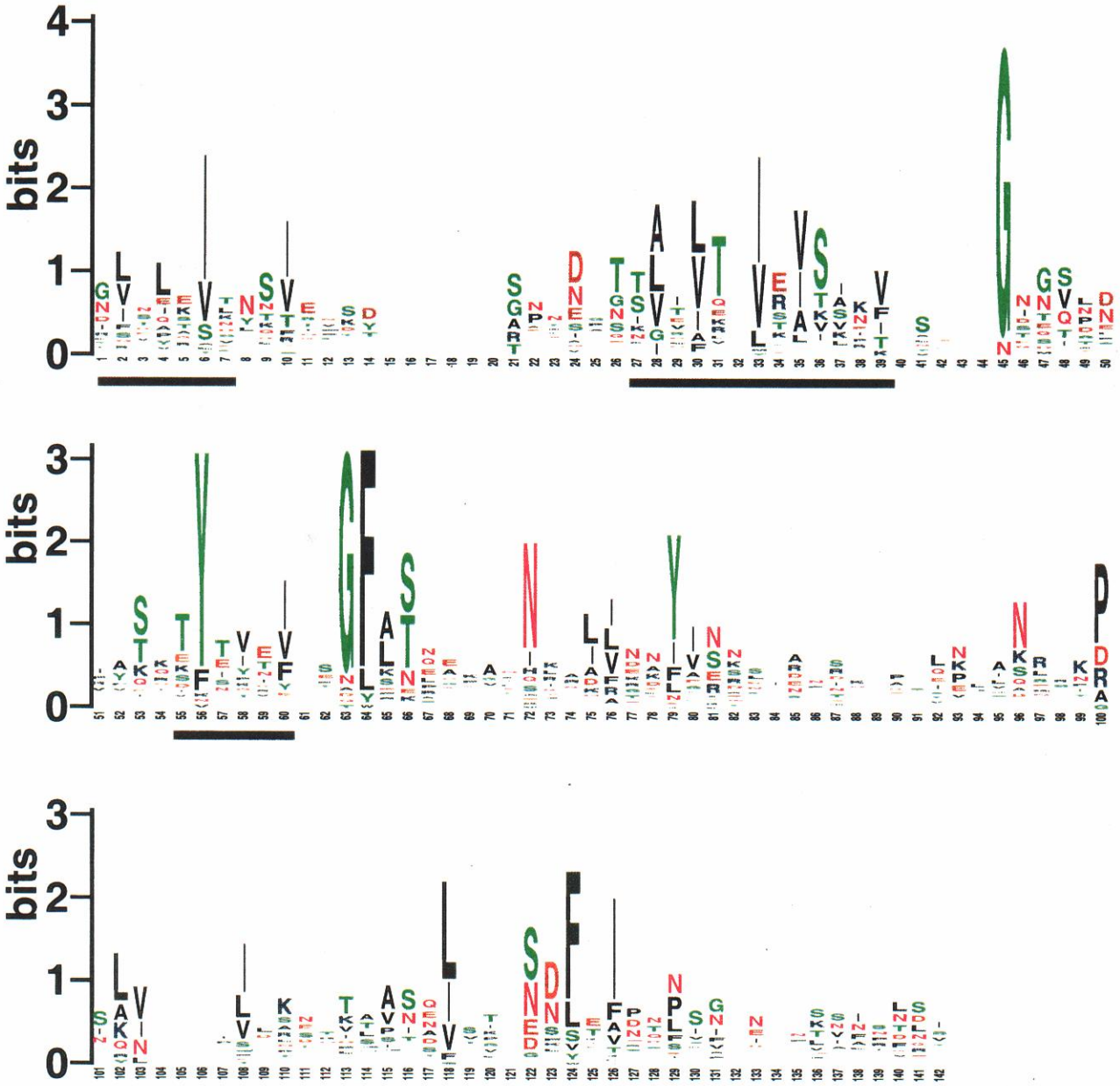


Fig. 9

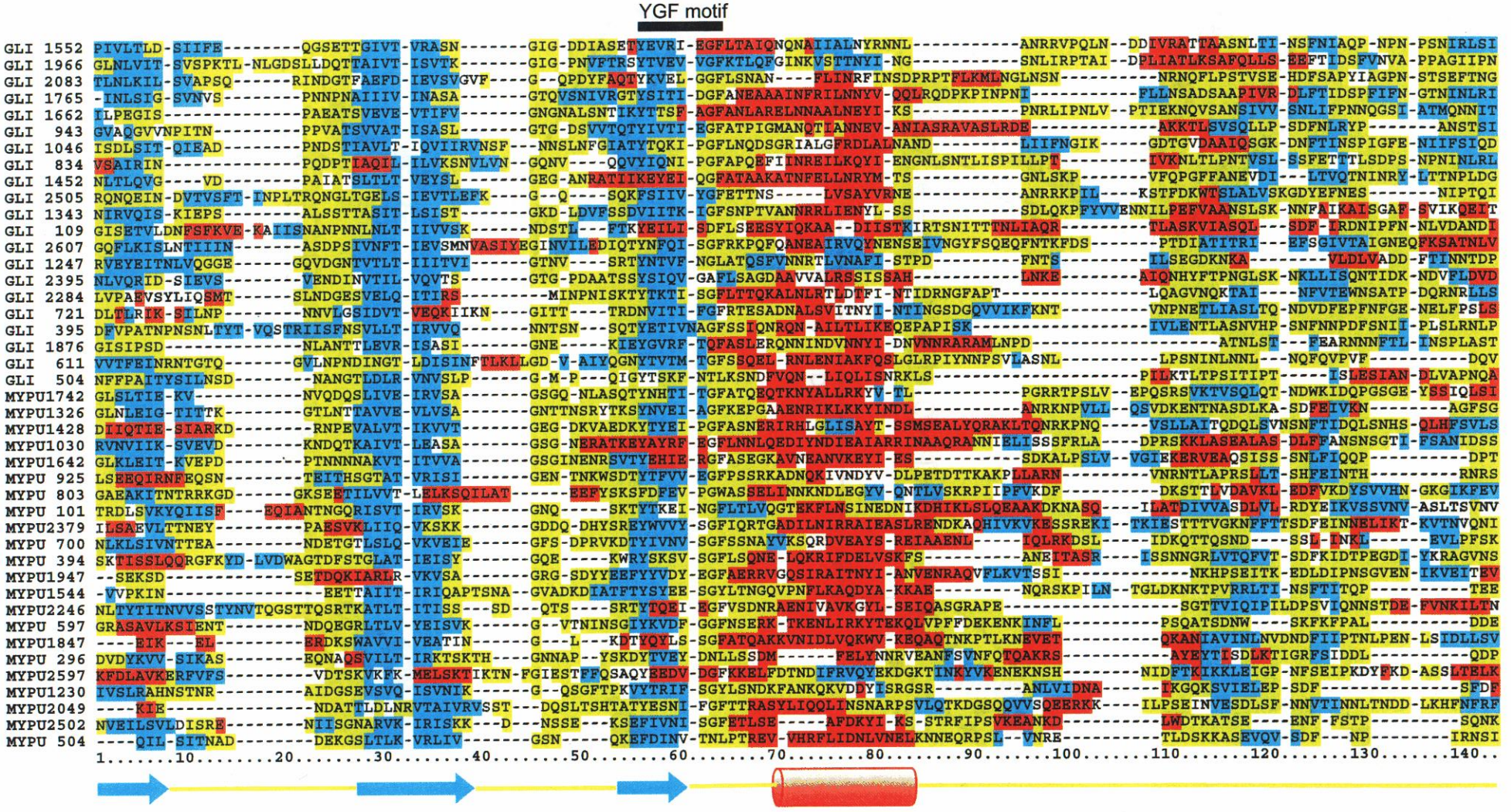
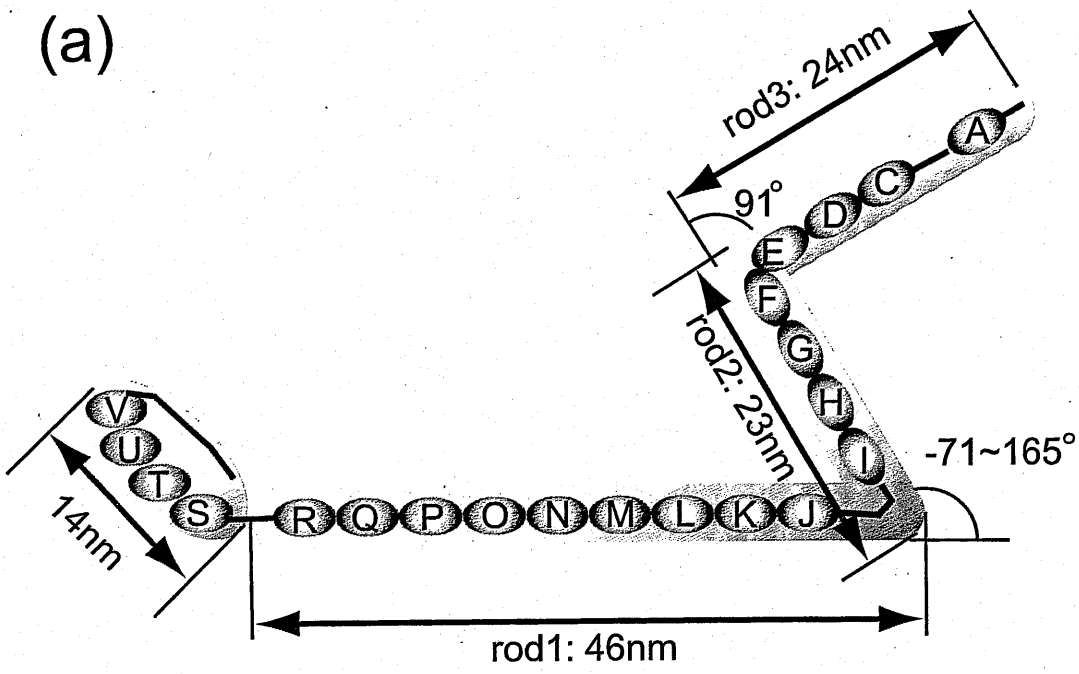
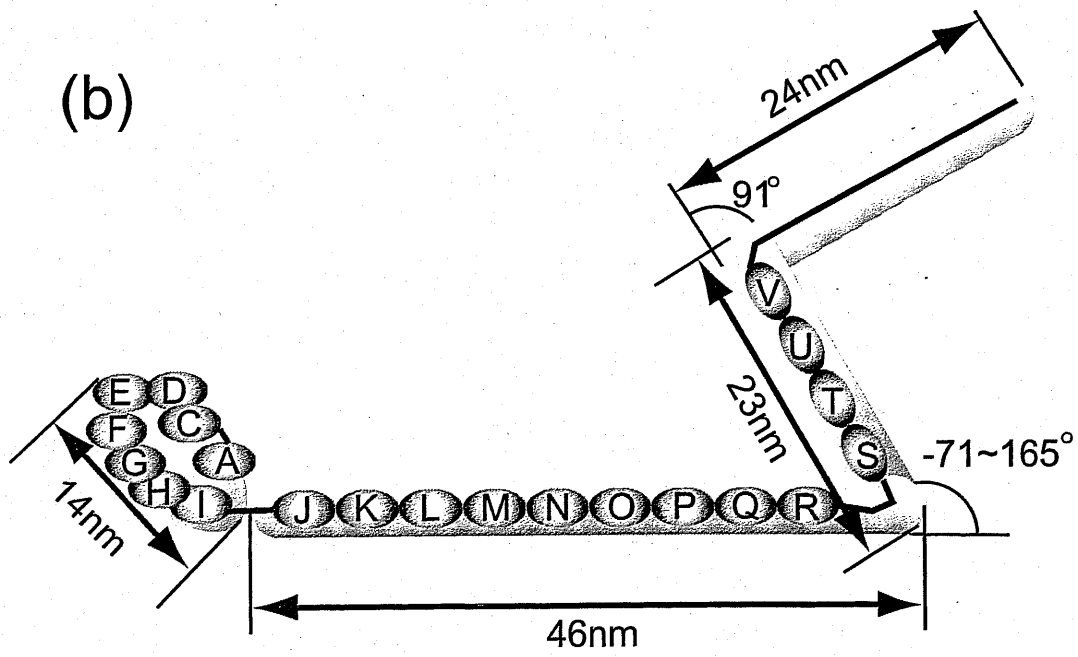

YGF motif
GLI 1552 PIVLTLD-SIIFE--------QGSETTGIVT-VRASN-------GIG-DDIASETYEVRI-EGFLTAIQNQNAIIALNYRNNL-----------ANRRVPQLN--DDIVRATTAASNLTI-NSFNIAQP-NPN-PSNIRLSI
GLI 1966 GLNLVIT-SVSPKTL-NLGDSLLDQTTAIVT-ISVTK-------GIG-PNVFTRSYTVEV-VGFKTLQFGINKVSTTNYI-NG-----------SNLIRPTAI--DPLIATLKSAFQLLS-EEFTIDSFVNVA-PPAGIIPN
GLI 2083 TLNLKIL-SVAPSQ------RINDGTFAEFD-IEVSVGVF-----G--QPDYFAQTYKVEL-GGFLSNAN----FLINRFINSDPRPTFLKMLNGLNSN----------NRNQFLPSTVSE-HDFSAPYIAGPN-STSEFTNG
GLI 1765 -INLSIG-SVNVS--------PNNPNAIIIV-INASA-------GTQVSNIVRGTYSITI-DGFANEAAAINFRILNNYV-QQLRQDPKPINPNI----------FLLNSADSAAPIVR-DLFTIDSPFIFN-GTNINLRI
GLI 1662 ILPEGIS--------------PAEATSVEVE-VTIFV-------GNGNALSNTIKYTTSF-AGFANLARELNNAALNEYI-KS-----------PNRLIPNLV--PTIEKNQVSANSIVV-SNLIFPNNQGSI-ATMQNNIT
GLI 943 GVAQGVVNPITN---------PPVATSVVAT-ISASL-------GTG-DSVVTQTYIVTI-EGFATPIGMANQTIANNEV-ANIASRAVASLRDE----------AKKTLSVSQLLP-SDFNLRYP--------ANSTSI
GLI 1046 ISDLSIT-QIEAD--------PNDSTIAVLT-IQVIIRVNSF---NNSLNFGIATYTQKI-PGFLNQDSGRIALGFRDLALNAND---------LIIFNGIK-----GDTGVDAAIQSGK-DNFTINSPIGFE-NIIFSIQD
GLI 834 VSAIRIN--------------PQDPTIAQIL-ILVKSNVLVN--GQNV----QQVYIQNI-PGFAPQEFIINREILKQYI-ENGNLSNTLISPILLPT---------IVKNLTLPNTVSL-SSFETTTLSDPS-NPNINLRL
GLI 1452 NLTLQVG----VD---------PAIATSLTLT-VEYSL-------GEG-ANRATTIKEYEI-QGFATAAKATNFELLNRYM-TS-----------GNLSKP-------VFQPGFFANEVDI---LTVQTNINRY-LTTNPLDG
GLI 2505 RQNQEIN-DVTVSFT-LNPLTRQNGLTGELS-IEVTLEFK-----G--Q----SQKFSIIV-YGFETTNS-----LVSAYVRNE----------ANRRKPIL----KSTFDKWTSLALVSKGDYEFNES--------NIPTQI
GLI 1343 NIRVQIS-KIEPS---------ALSSTTASIT-LSIST-------GKD-LDVFSSDVIITK-IGFSNPTVANNRRLIENYL-SS----------SDLQKPFYVVENNILPEFVAANSLSK-NNFAIKAISGAF-SVIKQEIT
GLI 109 GISETVLDNFSFKVE-KAIISNANPNNLNLT-IIVVSK------NDSTL---FTKYEILI-SDFLSEESYIQKAA--DIISTKIRTSNITTTNLIAQR---------TLASKVIASQL---SDF-IRDNIPFN-NLVDANDI
GLI 2607 GQFLKISLNTIIIN-------ASDPSIVNFT-IEVSMNVASIYEGINVILEDIQTYNFQI-SGFRKPQFQANEAIRVQYNENSEIVNGYFSQEQFNTKFDS-------PTDIATITRI----EFSGIVTAIGNEQFKSATNLV
GLI 1247 RVEYEITNLVQGGE------GQVDGNTVTLT-IIITVI------GTNV----SRTYNTVF-NGLATQSFVNNRTLVNAFI-STPD---------FNTS---------ILSEGDKNKA--------VLDLVADD-FTINNTDP
GLI 2395 NLVQRID-SIEVS--------VENDINVTIL-VQVTS-------GTG-PDAATSSYSIQV-GAFLSAGDAAVVALRSSISSAH----------LNKE--------AIQNHYFTPNGLSK-NKLLISQNTIDK-NDVFLDVD
GLI 2284 LVPAEVSYLIQSMT------SLNDGESVELQ-ITIRS---------MINPNISKTYTKTI-SGFLTTQKALNLRTLDTFI-NTIDRNGFAPT----------LQAGVNQKTAI----NFVTEWNSATP-DQRNRLLS
GLI 721 DLTLRIK-SILNP--------NNVLGSIDVT-VEQKIIKN----GITT----TRDNVITI-FGFRTESADNALSVITNYI-NTINGSDGQVVIKFKNT---------VNPNETLIASLTQ-NDVDFEPFNFGE-NELFPSLS
GLI 395 DFVPATNPNSNLTYT-VQSTRIISFNSVLLT-IRVVQ-------NNTSN---SQTYETIVNAGFSSIQNRQN-AILTLIKEQEPAPISK----------IVLENTLASNVHP-SNFNNPDFSNII-PLSLRNLP
GLI 1876 GISIPSD--------------NLANTTLEVR-ISASI-------GNE-----KIEYGVRF-TQFASLERQNNINDVNNYI-DNVNNRARAMLNPD----------ATNLST---FEARNNNFTL-INSPLAST
GLI 611 VVTFEINRNTGTQ-------GVLNPNDINGT-LDISINFTLKLLGD-V-AIYQGNYTVTM-TGFSSQEL-RNLENIAKFQSLGLRPIYNNPSVLASNL----------LPSNINLNNL---NQFQVPVF----------DQV
GLI 504 NFFPAITYSILNSD--------NANGTLDLR-VNVSLP------G-M-P---QIGYTSKF-NTLKSNDFVQN--LIQLISNRKLS----------PILKTLTPSITIPT------ISLESIAN-DLVAPNQA
MYPU1742 GLSLTIE-KV-----------NVQDQSLIVE-IRVSA-------GSGQ-NLASQTYNHTI-TGFATQEQTKNYALLRKYV-TL-----------PGRRTPSLV--EPQSRSVKTVSQLQT-NDWKIDQPGSGE-YSSIQLSI
MYPU1326 GLNLEIG-TITTK--------GTLNTTAVVE-VLVSA-------GNTTNSRYTKSYNVEI-AGFKEPGAAENRIKLKKYINDL----------ANRKNPVLL--QSVDKENTNASDLKA-SDFEIVKN---------AGFSG
MYPU1428 DIIQTIE-SIARKD--------RNPEVALVTI-IKVVT-------GEG-DKVAEDKYTYEI-PGFASNERIRHLGLISAYT-SSMSEALYQRAKLTQNRKPNQ-----VSLLAITQDQLSVNSNFTIDQLSNHS-QLHFSVLS
MYPU1030 RVNVIIK-SVEVD--------KNDQTKAIVT-LEASA-------GSG-NERATKEYAYRF-EGFLNNLQEDIYNDIEAIARRINAAQRANNIELISSSFRLA----DPRSKKLASEALAS-DLFFANSNSGTI-FSANIDSS
MYPU1642 GLKLEIT-KVEPD--------PTNNNNAKVT-ITVVA-------GSGINENRSVTYEHIE-RGFASEGKAVNEANVKEYI-ES----------SDKALPSLV--VGIEKERVEAQSISS-SNLFIQQP----------DPT
MYPU 925 LSEEQIRNFEQSN--------TEITHSGTAT-VRISI-------GEN-TNKWSDTYTFVV-EGFPSSRKADNQKIVNDYV-DLPETDTTKAKPLLARN---------VNRNTLAPRSLLT-SHFEINTH---------RNRS
MYPU 803 GAEAKITNTRRKGD------GKSEETILVVT-LELKSQILAT------EEFYSKSFDFEV-PGWASSELINNKNDLEGYV-QNTLVSKRPIIPFVKDF----------DKSTTLVDAVKL-EDFVKDYSVVHN-GKGIKFEV
MYPU 101 TRDLSVKYQIISF----EQIANTNGQRISVT-IRVSK-------GNQ-----SKTYTKEI-NGFLTLVQGTEKFLNSINEDNIKDHIKLSLQEAAKDKNASQ----ILATDIVVASDLVL-RDYEIKVSSVNV-ASLTSVNV
MYPU2379 ILSAEVLTTNEY---------PAESVKLIIQ-VKSKK-------GDDQ-DHYSREYWVVY-SGFIQRTGADILNIRRAIEASLRENDKAQHIVKVKESSREKI--TKIESTTTVGKNFFTTSDFEINNELIKT-KVTNVQNI
MYPU 700 NLKLSIVNTTEA---------NDETGTLSLQ-VKVEIE------GFS-DPRVKDTYIVNV-SGFSSNAYVKSQRDVEAYS-REIAAENL-----IQLRKDSL-----IDKQTTQSND----SSL-INKL------EVLPFSK
MYPU 394 SKTISSLQQRGFKYD-LVDWAGTDFSTGLAT-IEISY-------GQE-----KWRYSKSV-SGFLSQNE-LQKRIFDELVSKFS----------ANEITASR----ISSNNGRLVTQFVT-SDFKIDTPEGDI-VKRAGVNS
MYPU1947 --SEKSD--------------SETDQKIARLR-VKVSA-------GRG-SDYYEEFYYVDY-EGFAERRVGQSIRAITNYI-ANVENRAQVFLKVTSSI----------NKHPSEITK-EDLDIPNSGVEN-IKVEITEV
MYPU1544 -VVPKIN----------------EETTAIIT-IRIQAPTSNA--GVADKDIATFTYSYEE-SGYLTNGQVPNFLKAQDYA-KKAE----------NQRSKPILN--TGLDKNKTPVRRLTI-NSFTITQP----------TEE
MYPU2246 NLTYTITNVVSSTYNVTQGSTTQSRTKATLT-ITISS---SD----QTS----SRTYTQEI-EGFVSDNRAENIVAVKGYL-SEIQASGRAPE----------SGTTVIQIPILDPSVIQNNSTDE-FVNKILTN
MYPU 597 GRASAVLKSIENT--------NDQEGRLTLV-YEISVK------G--VTNINSGIYKVDF-GGFNSERK-TKENLIRKYTEKQLVPFFDEKENKINFL----------PSQATSDNW----SKFKFPAL----------DDE
MYPU1847 ----EIK---EL----------ERDKSWAVVI-VEATIN------G---L---KDTYQYLS-SGFATQAKKVNIDLVQKWV-KEQAQTNKPTLKNEVET----------QKANIAVINLNVDNDFIIPTNLPEN-LSIDLLSV
MYPU 296 DVDYKVV-SIKAS--------EQNAQSVILT-IRKTSKTH----GNNAP--YSKDYTVEY-DNLLSSDM------FELYNNRVEANFSVNFQTQAKRS----------AYEYTISDLKTIGRFSIDDL----------QDP
MYPU2597 KFDLAVKERFVFS---------VDTSKVKFK-MELSKTIKTN-FGIESTFFQSAQYEEDV-DGFKKELFDTNDIFRVQYEKDGKTINKYVKENEKNSH--------NIDFTKIKKLEIGP-NFSEIPKDYFKD-ASSLTELK
MYPU1230 IVSLRAHNSTNR---------AIDGSEVSVQ-ISVNIK------G--QSGFTPKVYTRIF-SGYLSNDKFANKQKVDDYISRGSR---------ANLVIDNA------IKGQKSVIELEP-SDF---------------SFDF
MYPU2049 ----KIE---------------NDATTLDLNRVTAIVRVSST---DQSLTSHTATYESNI-FGFTTRASYLIQQLINSNARPSVLQTKDGSQQVVSQEERKK----ILPSEINVESDLSF-NNVTINNLTNDD-LKHFNFRF
MYPU2502 NVEILSVLDISRE--------NIISGNARVK-IRISKK---D---NSSE---KSEFIVNI-SGFETLSE-----AFDKYI-KS-STRFIPSVKEANKD----------LWDTKATSE----ENF-FSTP---------SQNK
MYPU 504 ----QIL-SITNAD--------DEKGSLTLK-VRLIV-------GSN------QKEFDINV-TNLPTREV-VHRFLIDNLVNELKNNEQRPSL--VNRE----------TLDSKKASEVQV-SDF---NP---------IRNSI
1.....10.....20.....30.....40.....50.....60.....70.....80.....90.....100.....110.....120.....130.....140..

Fig. 10

(a)


rod3: 24nm
A
C
D
91°
E
F
G
H
I
rod2: 23nm
-71~165°
V
U
T
S
14nm
R
Q
P
O
N
M
L
K
J
rod1: 46nm

(b)


24nm
91°
V
U
T
S
23nm
-71~165°
E
D
F
C
G
A
H
I
14nm
J
K
L
M
N
O
P
Q
R
46nm