

Hierarchical Motion in Small Proteins: Implications for Protein Folding

running title:

Hierarchical Motions and Protein Folding

K. Anton Feenstra Herman J. C. Berendsen Alan E. Mark*

October 15, 2001

Bioson Research Institute and Laboratory of Biophysical Chemistry
University of Groningen, Nijenborgh 4, 9747 AG Groningen, The Netherlands
Phone: +31 50 363 4457, Fax: +31 50 363 4800, A.E.Mark@chem.rug.nl

submitted for publication to *PROTEINS: Structure, Function and Genetics*

October 2001

molecular dynamics

dynamical domains

motionally coherent elements

essential dynamics

long time-scale dynamics

*To whom correspondence should be addressed

Abbreviations

ED	Essential Dynamics	MD	Molecular Dynamics
MSF	Mean Square Fluctuation	PDB	Protein Data Bank
RMSD	Root Mean Square Deviation	SPC	Simple Point Charge

1 Summary

Extensive molecular dynamics simulations within a series of proteins have been used to identify elements which show highly concerted motion as well as below average internal mobility. Although secondary structure elements are frequently assumed to be comparatively rigid the motionally coherent elements identified often did not coincide with major secondary structure elements, within the proteins examined. A hierarchical organization of these motionally coherent elements was observed. This may be related to a hierarchy of folding with smaller elements being considered as possible protein folding nucleation sites or foldons.

2 Introduction

Experimental estimates of the lower bound of the folding time of small proteins range from 1 to 70 microseconds.^{1,2} Using atomic models and explicit solvent, the maximum length of time that a protein in solution can be simulated is currently on the order of hundreds of nanoseconds,³ or, by supreme effort, a few microseconds.⁴ This puts equilibrium simulations of protein folding well out of range for the foreseeable future. For this reason there is much interest in the development of simplified models that capture the essential features of a protein and the folding process, while reducing the computational cost.

One possible approach to this problem is based on the nucleation model of protein folding.⁵⁻⁸ This proposes that proteins are constructed in a modular fashion from independent folding units or “foldons”⁹ with a hierarchical organization. If this model is a realistic representation of protein folding, the identification of such folding elements would be an important first step in the development of simplified models to simulate protein folding. Several theoretical methods have been proposed to identify such modules, which are usually assumed to be rigid elements in a protein. Most of these methods are based on ad-hoc definitions of what will be a rigid element. Often rigid elements are simply defined as secondary structure elements.^{1,10-14}

Here, we propose and test a novel method to analyze the dynamics of a protein in its native state, as obtained from molecular dynamics (MD) simulations, in terms of motionally coherent elements. Preliminary aspects of this work have been described previously.¹⁵ In this paper we apply the approach to a series of five proteins in order to validate the approach and correlate the elements identified with folding nucleation sites proposed by others.

2.1 Outline of the approach

Central to our approach is the identification of groups of residues that show large-scale correlated motion in a number of separate simulations or in separate parts of a single long simulation. The aim is to obtain a consensus regarding the dynamics of particular regions of the proteins. From this one can then extract motionally coherent elements which are groups of residues that, on average, show correlated motional properties. The fact that such groups or elements show distinct motional properties also suggests that they are independently stable and thus possible candidates for foldons or folding nucleation sites.

To identify groups of residues that show large-scale correlated motion in a set of independent simulations a two step procedure was used. First, in order to focus only on large scale collective modes, uncorrelated local fluctuations motions were filtered out by performing a form of Principle Components Analysis and retaining only the dominant modes. Based on the predominant collective modes obtained from the Principle Components Analysis motional elements were then identified by clustering groups of residues that undergo rotation around a similar axis and also show little internal motion. That is, the residues must not display significant motion relative to other members of the cluster. This leads to a definition of such elements for each collective mode analyzed. From the global set of motional elements, a matrix can be constructed by counting the number of times a given pair of residues is found within the same element. Statistics on the element definitions can be improved by performing analysis of additional simulations or by analysis of separate parts of a large simulation. From this “consensus dynamics matrix” a set of consensus groups of residues with large scale correlated motion can be identified.

3 Methods

3.1 Simulations

Proteins

Five proteins were selected as test-cases (the PDB¹⁶ entry labels are given in parentheses):

Sh3 (**1shf**¹⁷/**1nyf**¹⁸): β barrel, 59 residues (Fyn-Sh3), Sh3 domain from human fyn proto-oncogene tyrosine kinase, crystal structure (**1shf**, 2 molecules in the asymmetric unit) and NMR structure (**1nyf**);

T4L (**21zm**¹⁹): α/β protein, 164 residues (T4-Lys), bacteriophage T4 lysozyme, crystal structure (**21zm**);

HPr (**1hdn**²⁰/**1poh**²¹): α /anti-parallel- β sandwich, 85 residues (HPr), a histidine containing phosphocarrier protein from the phosphoenolpyruvate dependent phosphotransferase system of *E. coli*, NMR ensemble (**1hdn**) and crystal structure (**1poh**);

1auz:²² α /anti-parallel- β sandwich, 116 residues (SpoIIAA), a phosphorylatable component of the system that regulates transcription factor σ^F of *B. subtilis*, NMR ensemble (**1auz**);

1ksr:²³ anti-parallel- β sandwich, immunoglobulin fold, 100 residues (abp-120), an actin-binding protein from the repeating segments (rod 4) of F-actin cross-linking gelation factor of *D. discoideum*, NMR ensemble (**1ksr**);

The proteins will be referred to by the labels as indicated above.

For the Fyn-Sh3 domain both a NMR and an X-ray structure is available. For T4 lysozyme an X-ray structure is available and for the other three proteins, which were chosen because they are small and globular, highly refined NMR structures are available. In the cases of **1hdn**, **1auz** and **1ksr**, simulations were started from each of the individual structures from the NMR derived ensemble of structures given in the respective PDB files. For the Sh3 domain (**1shf** and **1nyf**), long contiguous simulations were performed starting from the two conformations in the PDB file **1shf**, and from the single NMR-derived conformation in the PDB file **1nyf**. In addition, two long simulations of HPr (starting from the X-ray conformation given in the PDB file **1poh**) and two simulations of T4 Lysozyme (starting from the coordinates given in the PDB file **2lzm**) were provided by B. Hess. These simulations were performed in a manner similar to those described below, except for the use of a more conservative time step of 4 fs. Full details are given elsewhere.²⁴

Parameters

All MD simulations were performed using the GROMOS-96 forcefield²⁵ and the GROMACS molecular dynamics package.^{26,27} The LINCS algorithm²⁸ was used to constrain covalent bonds within the proteins. Fast degrees of freedom involving hydrogen atoms were eliminated from the system as described previously.²⁹ In addition out-of-plane motion in the aromatic rings of the sidechains of tyrosine, phenylalanine and tryptophane was removed as described in appendix A. For the hydrogen atoms with remaining degrees of freedom (rotatable groups like hydroxyl and amine, and water), the mass was increased from 1 *u* to 4 *u* while simultaneously decreasing the mass of the bonded heavy atom by the same amount, as described previously.²⁹ This allowed a time step of 7 fs to be used.

A twin-range cut-off for non-bonded interactions was employed with a short-range cut-off of 1.0 nm for the Van der Waals and Coulomb interactions, which were calculated every simulation step, and a long-range cut-off of 1.7 nm for Coulomb interactions which were calculated during neighbor-list update, every 21 fs (3 time steps). The temperature was maintained by weak coupling to an external bath³⁰ at 300 K with a time constant of 0.1 ps. Protein and water were independently coupled to the heat bath. The pressure was maintained by weak coupling with a time constant of 1.0 ps to an external bath of 1 bar. A relative dielectric constant (ϵ_r) of 1.0 was used.

Equilibration

Individual structures taken from the respective PDB files were first energy minimized using a steepest descent algorithm. The resulting structures were each solvated in a cubic box of Simple Point Charge (SPC) water,³¹ with a minimum distance of 0.85 nm between the protein and the box wall. The water box was constructed by replicating a cubic box containing 216 equilibrated SPC water molecules. All water molecules with the oxygen atom closer to any protein atom than the sum of their respective Van der Waals radii were removed. Energy minimization followed by 0.2 ps of unrestrained MD using a 2 fs time step was performed to relax the systems.

[Table 1 about here.]

Production simulations were performed starting from each of the structures in the ensemble of NMR solution structures in the respective PDB files for simulation times of 2 or 5 ns, or starting from single X-ray or NMR structures for simulation times of 600 ns. A total of 2 μ s of protein trajectories was generated. A summary of the

simulation parameters is given in Table I.

3.2 Analysis

The identification of motionally coherent elements consisted of a filtering step based on a Principal Components Analysis of the fluctuations of the coordinates of the backbone (N, C and C $_{\alpha}$ atoms) and an assignment step based on the domain identification procedure DYNDOM.^{32,33}

To identify collective motions, Principal Components Analysis (PCA) of the fluctuations in the atomic coordinates of the backbone (N, C and C $_{\alpha}$ atoms), also referred to as essential dynamics (ED) analysis,³⁴ was performed on each of the protein trajectories generated. For the multiple shorter trajectories generated for **1hdn**, **1ksr** and **1auz** separate eigenvectors were calculated from each of the trajectories. The first 0.2 ns of each trajectory was excluded from the analysis to minimize possible artefacts arising from the choice of starting structure. The longer trajectories generated in the other simulations were first fragmented before the PCA analysis was performed. This was done in several ways. For the two 40 ns simulations of **1poh** and **2lzm**, 20 contiguous non-overlapping segments of 4 ns each were generated. In addition, from the two 40 ns trajectories of **2lzm** a set of 40 segments of 1 ns each separated by 1 ns was generated. The three 600 ns trajectories of Sh3 domain were divided into a total of 180 segments of 10 ns each.

For each trajectory or fragment thereof the first six eigenvectors, corresponding to the first six generalized degrees of freedom with the largest fluctuations, were identified and used for further analysis. In effect this filters the motion in the trajectory eliminating small-scale motion while retaining the larger more collective

modes.

The identification of semi-rigid elements from the filtered trajectories was achieved by clustering residues based on the direction of their rotation axes as implemented in the program DYNDOM.^{32,33} DYNDOM is a model-free approach that defines dynamical domains based on two key criteria: i) residues within a domain must rotate around an axis with a similar orientation and ii) the “amount of motion” (measured by the mean square fluctuation normalized to the size of the domain) within a domain must be smaller than the “amount of motion” between domains, i.e. domains are more rigid than the whole. The rotation of each residue is determined from the two extremes of the projection of a MD trajectory onto a given ED eigenvector. In the case of multi-domain proteins clusters of residues identified by DYNDOM will correspond to the separate domains. In the case of small single-domain proteins the clusters of residues identified by DYNDOM are more appropriately thought of as semi-rigid motionally independent elements.

The input parameters of the DYNDOM program were set as follows: the maximum number of clusters, 20; the fitting segment window length for determination of residue rotation, 5 residues; the minimum element size, 5 residues; the minimum ratio of external to internal motion, 1.0. The results are, however, insensitive to the specific choice of parameters. This set of parameters was chosen because it is similar to that previously used in the analysis of multi-domain proteins.^{32,33,35} The maximum number of clusters, the maximum number of iterations and the minimum element size set the boundary conditions for the clustering algorithm that groups residues with similar rotation. The ratio of internal to external motion determines the acceptance criterion for a given cluster. Note, DYNDOM does not require that any or all residues are members of a cluster (domain). Frequently, only part of the

molecule can be assigned to a domain and in some cases no domain assignment is possible at all. Only the length of the fitting window and the ratio of the internal to external motion influence the number of elements assigned.

By counting the number of times residue pairs are found together inside each of the semi-rigid elements, a consensus matrix can be constructed which contains information on the relation between residues in different trajectories. From this, motionally coherent elements can be extracted. In order to determine a possible sequence dependence of the positions of the element boundaries, the number of occurrences $N_{X,\Delta}$ of residue type X at various positions Δ relative to the boundaries of the motionally coherent elements defined, and the total number of occurrences of N_X in the total sequence, were determined. From this, the relative probability $\alpha_{X,\Delta}$ of residue type X occurring at position Δ from a boundary was calculated using

$$\alpha_{X,\Delta} = \frac{N_{X,\Delta}}{p N_X} \quad \text{where} \quad p = \frac{N_\Delta}{N} \quad (1)$$

and N_Δ is the total number of residues at positions Δ and N in the whole sequence. Values observed for the relative occurrence $\alpha_{X,\Delta}$ for $|\Delta| \leq 2$ were compared with the relative probability $\alpha_{X,random}$ expected for a random distribution of residues and its standard deviation based on a binomial distribution:

$$\alpha_{X,random} = 1 \pm \sqrt{\frac{1-p}{p N}} \quad (2)$$

4 Results

4.1 Simulations

In all simulations the native fold of the protein and the overall secondary structure distribution as assigned by DSSP³⁶ remained intact. The RMSD value of the atomic positions of the backbone atoms from the starting structures, averaged over the trajectory from 0.5 ns was calculated for each simulation. These values varied for the separate simulations of, **1hdn** from 0.14 to 0.26 nm (average 0.18 nm), **1ksr** from 0.16 to 0.48 nm (average 0.35 nm) and **1auz** from 0.15 to 0.47 nm (average 0.34 nm). For the longer simulations the values, averaged starting from 4 ns, were as follows: for **1poh** 0.22 and 0.28 nm, for **2lzm** 0.28 and 0.31 nm, for **1shf** 0.32 and 0.51 nm and for **1nyf** 0.42 nm.

4.2 Essential Dynamics

[Table 2 about here.]

For each trajectory (for **1hdn**, **1auz** and **1ksr**) and all trajectory fragments (for **Sh3**, **T4L** and **1poh**), the first six ED eigenvectors were determined. The resulting total number of eigenvectors for each of the proteins is listed in the second column of Table II. The percentage of the total fluctuation contained in the subspace spanned by the first six ED eigenvectors, averaged over the trajectories of each protein is given in the third column of Table II. On average the first six ED eigenvectors contained 67% of the total atomic positional mean square fluctuation (MSF) present in each trajectory. This is sufficient for the subsequent analysis to provide information on large scale motions within the proteins.

4.3 Consensus Dynamics

Fyn-Sh3

[Figure 1 about here.]

[Figure 2 about here.]

Figs. 1 and 2 outline the results of the analysis for **Sh3**. In Fig. 1 a selection of the individual semi-rigid element assignments as made by the DYNDOM method for each of the essential modes of the **Sh3** trajectories, is represented by shaded bars. The three 600 ns trajectories were split into 60 fragments of 10 ns. For each of these trajectory fragments six eigenvectors were analyzed, yielding in total $3 \times 60 \times 6 = 1080$ eigenvectors. Element assignments for the first sixty of these eigenvectors are shown in Fig. 1. The separate semi-rigid elements are shaded differently. Blank rows correspond to eigenvectors for which no semi-rigid elements could be identified. The percentage of eigenvectors for which element could be assigned is listed in the fourth column of Table II. Blank areas which are present in most rows correspond to residues that were not assigned to any semi-rigid element for that particular eigenvector. The percentage of residues assigned to a semi-rigid element for those eigenvectors for which elements were identified, is listed in the fifth column of Table II.

Fig. 2a shows the “consensus dynamics matrix” for **Sh3**. The intensity corresponds to the number of times a residue pair was observed together in the same semi-rigid element. The most prominent feature of this plot is the hierarchical block structure. Blocks in the matrix correlate with the locations of the individual semi-rigid elements as depicted in Fig. 1. The average intensity of a block in the consensus dynamics matrix gives a qualitative measure of the apparent rigidity of the element.

Using the matrix as a guide, subdivisions of the protein can be made as shown in Fig. 2b. On the most detailed level six elements can be identified in **Sh3** that contain more than 5 residues (the minimum element size used when determining the semi-rigid elements). In Fig. 2c a histogram of the occurrence of the beginning or end of an element at a particular point along the sequence is plotted. Peaks correspond to locations where boundaries of motionally coherent elements occur frequently and are correlated with the element boundaries as defined in the hierarchy of Fig. 2b.

[Figure 3 about here.]

In Fig. 3 the motional hierarchy outlined in Fig. 2b is color coded onto the three-dimensional structure of **Sh3**. At the highest level three large partially overlapping motionally coherent elements can be recognized. In the structure, the element **Sh3:1-29** (red & pink in Fig. 3a) corresponds to the first two sets of β -strands at the N-terminus. Element **Sh3:1-40** (red, pink & green) also includes the following loop-and-strand. Element **Sh3:20-58** (pink, green & purple) contains the complete three-stranded β -sheet (to the rear in Fig. 3), with the 3_{10} -helix and C-terminal β -strand on one side and an additional β -strand on the other side. The next, most detailed, level contains six non-overlapping elements, which are indicated in Fig. 3b. Elements 2 (residues 10-19, loop, yellow in Fig. 3b), 4 (residues 29-40, strand-loop-strand, green), 5 (40-48, strand-turn-strand, purple) and 6 (48-58, strand-helix-strand, blue) form relatively compact structures within the protein.

T4 Lysozyme

[Figure 4 about here.]

The consensus dynamics matrix for T4L is shown in Fig. 4a. The upper half of the matrix was derived from the analysis of 1 ns fragments taken from the long trajectories with 1 ns intervals in between and the lower-right half from consecutive fragments of 4 ns each. A more detailed description of the procedure can be found in the Methods section (sec. 3.2). The difference between the two halves of the matrix is insignificant.

[Figure 5 about here.]

An hierarchical breakdown gives rise to three the levels depicted in Fig. 4b. The first (highest) level contains three motionally coherent elements, indicated by different colors in Fig. 5a. The three elements correspond to the N- and C-terminal domains of T4 Lysozyme (blue and green respectively in Fig. 5a) and the N-terminal α -helix that structurally is part of the C-terminal domain (element 1, red in Fig. 5a). Analysis of crystal structures, however, indicate that this fragment can be associated with either of the two domains.³⁷ The large helix that extends from one domain into the other, α -helix 3, is divided in the middle, with each end being assigned to a different element.

On the next level, shown in Fig. 5b, in the N-terminal domain two elements can be assigned and in the C-terminal domain three elements can be assigned. The protein as a whole is divided into three large parts of roughly equal size which each form relatively compact structures, and three small α -helical regions, which lie in between the larger sections. The twelve elements identified on the most detailed level are indicated in Fig. 5c.

[Figure 6 about here.]

[Figure 7 about here.]

HPr

The consensus dynamics matrix for HPr is shown in Fig. 6a. The upper half of the matrix was derived from a series of simulations started from NMR structures given in PDB file 1hdn, the lower half from a series of fragments taken from two simulations started from an X-ray structure (1poh), see Methods section sec. 3.2. The difference between the two halves of the matrix is marginal. On the most detailed level nine separate elements could be defined (Fig. 6b) most of which correspond roughly to secondary structure elements. Two notable exceptions are the β -turn between strands III and IV which forms element 5 (residues 36-41) and the half-strand, loop, half-helix that forms element 8 (residues 62-76). Both these elements form compact structures. The elements are indicated by color coding in Fig. 7.

SpoIIAA (1auz)

For 1auz the blocked structure of the consensus dynamics matrix is much less pronounced (Fig. 6a). On the first level of the motional hierarchy the protein can be split into roughly three parts (Fig. 6b). The middle section does not show a pronounced blocked structure in the matrix. In the lowest level 8 elements can be defined as indicated in Fig. 7.

Abp-120 (1ksr)

On the most detailed hierarchical level, 1ksr may be divided into 11 elements, see Figs. 6a and b. Some correspond to individual β -sheet regions. Most of the elements

are not very compact in the protein. Fig. 7 gives a graphical representation of these elements.

4.4 General properties

Certain elements contain very few or no boundaries of semi-rigid elements and correspond to minima of the demarcation counts in Figs. 2c, 4c and 6c. Most notably, β -turns are almost always correlated with such minimum (**Sh3**:30-40, **Sh3**:41-48, **T4L**:20-25, **T4L**:26-33, **1hdn**:37-41, **1auz**:8-14, **1ksr**:59-69 and **1ksr**:70-81). The only exception is the β -turn at **1ksr**:50-55. Other minima in the demarcation counts are found inside certain loop regions (**Sh3**:11-19, **Sh3**:20-29, **1auz**:67-76, **1ksr**:15-21, **1ksr**:29-36 and **1ksr**:83-92) as well as inside helical regions (**Sh3**:49-58, **T4L**:34-58, **T4L**:85-105, **T4L**:106-113, **1hdn**:42-52 and **1auz**:27-37). This suggests a relatively high (dynamical) coherence inside some loops and for certain positions within α -helices.

To a first approximation, the motionally coherent elements observed correspond to secondary structure elements, such as β -turns and α -helices (as noted above) and to a lesser extent β -strands. A number of element boundaries were, however, positioned inside secondary structure elements, dividing them into parts. Most notably, longer helices (e.g. **HPr** helix 3 and **T4L** helix 3) and β -strands are divided approximately in the middle (e.g. **HPr** strand II or **1ksr** strand AII) or near β -turns. This suggests a loss of (dynamical) coherence near the ends of secondary structure elements and a reduced coherence for some extended secondary structure elements.

[Table 3 about here.]

Table III shows the relative abundance of residue types (α_X) at positions up to two residues from the element boundaries, calculated using eqn. 1. The expected standard deviation of the relative abundance based on a random distribution of residue types ($\alpha_{X,random}$), calculated using eqn. 2, is also shown. For serine and histidine, a significant deviation from the expected occurrence is found with a statistical confidence level of more than 95 %.

No off-diagonal blocks appear in any of the consensus dynamics matrices of the proteins investigated. This is surprising as many of the motionally coherent elements identified are in close contact in the three-dimensional structure, while being sequentially distant. Possible reasons for this observation are discussed below.

5 Discussion

This work clearly demonstrates that the use of multiple starting structures for a number of relatively short protein simulations can be an efficient way to sample the accessible configurational space within a limited amount of computational time. Results of the analysis of the series of shorter simulations are essentially identical to those obtained from long contiguous simulations. Likewise, no significant differences were detected between simulations performed using a time step of 7 fs and simulations performed at 4 fs.

The analysis of the atomic motions in this study is based on the first six principle components of the motion which comprises 60 – 70 % of the total structural fluctuations. Extensive tests on HPr (data not shown) indicate that using only the first six eigenvectors results in an optimal filtering of the information available in the trajectories. Including more (higher) eigenvectors increases noise whereas using

less eigenvectors diminishes the available information. From the semi-rigid elements obtained, it is possible to infer a consensus description of the motionally coherent elements of a protein. Approximately half of the total fluctuation of the native state of the protein can be attributed to the motion of a relatively small number of structural elements. In this respect, the combination of principal component analysis on atomic coordinates and semi-rigid element analysis acts as a powerful filter for the analysis of dynamical trajectories of proteins.

Motionally coherent elements appear as blocks in the consensus dynamics matrix. Larger blocks often contain smaller ones, which suggests the motions in a protein are hierarchical in nature. The occurrence of distinct peaks and minima in the number of element boundaries lends confidence to the method. Motionally coherent elements that contain non-contiguous parts of the sequence will show up as off-diagonal blocks in the consensus dynamics matrices. Surprisingly such elements only occurred rarely in the cases tested. Nearly all motionally coherent elements consist of a single contiguous sequence. This appears to be a non-trivial property of the proteins studied. It suggests a description of the dynamics of the proteins in terms of a hierarchy of structurally and motionally coherent, sequential elements. This is consistent with models of protein folding that assume the formation of local structural elements which then assemble sequentially into the folded protein.

Little preference could be detected for the occurrence of specific amino acids at the boundaries between motionally coherent elements. Although the statistics are insufficient to draw general conclusions, the hydrogen-bonding amino acids His, Ser and Thr seem to be preferred at the boundary locations. From the available data it is not possible to extract statistics about combinations of residues or about sequence patterns.

There also does not appear to be a strong relation between the location of the motionally coherent elements and the location of secondary structure elements in the protein. Many α -helices and β -sheets contain element boundaries and belong to more than one element. On the other hand, many elements consist of parts of two secondary structure elements together with the connecting turn or loop region, e.g. strand-turn-strand or helix-loop-strand. This suggests that loop regions can undergo concerted motion. It also suggests a much more dynamic and flexible character of secondary structure elements than is usually assumed. In part, this difference might simply reflect ambiguities resulting from the use of a discrete algorithm (such as DSSP) for the classification of secondary structure elements when conformations fall near the classification boundaries. Comparison with a continuous classification scheme might lead to more consistent results.³⁸

A similar conclusion on the nature of secondary structure elements can be drawn from the comparison between the consensus dynamics matrices and the overall secondary structure of the proteins. **T4L**, **Sh3** and **HPx** consist mainly of relatively large secondary structure elements and feature sharply demarcated blocks in the matrix. **1auz** has a similarly well-ordered secondary structure, but a more blurred appearance in the consensus dynamics matrices. **1ksr** in contrast has fewer large secondary structure elements (many loop and coil regions), but does show sharply demarcated blocks in the matrix. No definitive relationship between the nature of the secondary structure of a protein and its dynamical behavior is evident.

5.1 Comparison with Experimental Studies

Despite the wealth of protein folding studies, there is little experimental information at an atomic level that can be directly compared to the results of the work we have presented. The most direct data on the mechanisms of protein folding, that is data involving the least number of assumptions, is NMR data. This includes relaxation data (NOESY/ROESY intensities or build-up curves and T1 relaxation times) and exchange data (hydrogen/deuterium exchange protection factors). The other main source of information is mutational studies, primarily the so-called Φ_f analysis which attempts to relate the effects of single amino acid substitutions on the rates of folding and unfolding of proteins to the nature of the folding transition state, and studies involving the stability or dynamics of isolated peptide fragments taken from different parts of the protein.

Fyn-Sh3

[Figure 8 about here.]

The folding and stability of Fyn-Sh3 has been extensively studied by Riddle *et al.*³⁹ who have performed Φ_f analysis and applied an *ab-initio* folding method. The effects of single alanine and glycine mutations at many positions along the sequence on the rates of folding and unfolding (Φ_f analysis) were analyzed.³⁹ The results of this work are schematically depicted in Fig. 8b. Blocks correspond to regions of the protein considered to be important in the transition state of folding and hence are formed early in the folding process. The size of the blocks corresponds to their relative importance. Note that the least important (number five) region spans two non-consecutive regions, residues 3-6 and 50-57. The positions of these regions

correlate fairly well with the elements identified from the consensus dynamics matrix (Fig. 8a). Not surprisingly, the two most important folding regions are β -turns.

The *ab-initio* folding method ROSETTA was used to predict a progression of folding events³⁹ (Fig. 8c). ROSETTA ranks the elements according to the probability they will adopt a particular conformation based on an analysis of protein structural databases. The order in which separate regions of the protein fold according to this algorithm, roughly corresponds to the order observed for the five fragments found in the Φ_f analysis (Fig. 8b). Regions of the protein that display distinct behavior in the ROSETTA analysis correlate well with the motionally coherent elements identified from our analysis. It is interesting that the most important and early folding segment comprising residues 40-48 is also the element with the clearest demarcation in our analysis (see Fig. 2c, element 5). The early folding of element 5, followed by folding of segments 3-6 (residues 19-56) in the next hierarchical level, is compatible with all available data.

Overall, a reasonable correspondence is observed between the positions of the motionally coherent elements and the various regions identified in the analyses performed by Riddle *et al.*³⁹

T4 Lysozyme

[Figure 9 about here.]

The folding and stability of T4 Lysozyme has been extensively studied. Hilser & Freire used Φ_f analysis.⁴⁰ Lu & Dahlquist used pulsed hydrogen exchange experiments.⁴¹ Najbar *et al.* analyzed the stability of isolated peptide fragments taken from T4 Lysozyme.⁴²

The effect of mutations at many positions along the sequence on the rate of folding was measured and expressed as the logarithm of the folding rate ($\ln \kappa_f$)⁴⁰ is shown in Fig. 9b. High values of $\ln \kappa_f$ correspond to more stable regions of the protein. Qualitatively, regions with higher $\ln \kappa_f$ correlate with regions with lower demarcation counts (Fig. 4c), most notably for the peaks of $\ln \kappa_f$ inside helices 1, 5, 7 and 10.

From pulsed hydrogen exchange experiments, the protection factors for 84 individual residues in T4L were determined by Lu & Dahlquist⁴¹ (see Fig. 9c). This procedure can identify sites in the protein that are protected early in folding (around 80 ms). Several regions show significant protection (more than a factor of 20):⁴¹ two parts in the N-terminal β -sheet region, α -helix 1 and 5 as well as some residues in helix 3. These regions correspond to relatively low demarcation counts in Fig. 4c.

For several peptide fragments from T4L the importance to the folding was determined by Najbar *et al.* from analysis of stability and other properties of the isolated peptides in solution.⁴² Three peptide fragments were identified as potential folding initiation sites. Two of these, residues 1-13 and 92-107, correspond roughly to elements identified in our analysis. The third potential folding initiation site identified by Najbar *et al.* corresponds to α -helix 3, which bridges the two domains of T4L and corresponds to two consecutive elements in our analysis.

5.2 Conclusions

[Figure 10 about here.]

It is generally accepted that the separate domains in multi-domain proteins fold independently. Furthermore, for some proteins (e.g. lysozyme⁴³) it has been sug-

gested that one of the domains must fold before another can complete its folding. Other studies suggest local patterns, such as β -turns or one turn of an α -helix initiate folding. This leads to a concept of a hierarchy of folding, with the assembly of protein structural domains at the top level, and a series of independent folding units of several residues in length at the most basic level. Our hypothesis was that such a folding hierarchy may be reflected in a hierarchy of motions in the protein in its native state. What we have found is that there is a hierarchical organization of motionally coherent elements within proteins. The motional hierarchy observed most likely reflects the folding hierarchy and as such is a guide to construct a hierarchy of folding. Within this framework it is still possible to devise *multiple* folding pathways that describe all elements observed as is illustrated in Fig. 10. Nevertheless the procedure described in this paper provides an objective means to identify boundaries between potential folding elements.

5.3 Acknowledgments

We would like to thank B. Hess (RuG, Groningen) for providing the long simulations of HPr and T4 Lysozyme. K. A. F. acknowledges support from the Netherlands Foundation for Life Sciences (SLW) with financial aid from the Netherlands Organization for Scientific Research (NWO).

A Out-of-plane vibrations in aromatic groups

When using dummy-atom constructions for non-rotating explicit hydrogens together with a time step of 7 fs at 300 K as described in Feenstra *et al.*,²⁹ occasional instabilities in the aromatic groups occur as a result of out-of-plane vibrations. The construction for the conjugated planar group in arginine was described previously.²⁹ These can be eliminated by using additional dummy-atom constructions. The position of a dummy atom is calculated every time step from the position of three real atoms. The total mass of the whole group is redistributed to three of the heavy atoms. This is done such that the center of mass and moments of inertia are preserved as closely as possible. Constraints between the three atoms fix the geometry and all other atoms are constructed as dummy atoms from them.

[Figure 11 about here.]

For phenylalanine and histidine this is trivial, the heavy atoms at γ and ϵ positions are kept as normal atoms. For tyrosine O_η and H_η also remain as normal atoms, with constraints between $C_{\epsilon 1}$, $C_{\epsilon 2}$ and O_η . C_ζ is constructed from these three atoms. The bond angle in the hydroxyl group (C_ζ - O_η - H_η) is constrained by a constraint between C_γ and H_η (constraints involving a dummy atom are not implemented). The planarity of the whole group (which could still fold around $C_{\epsilon 1}$ - $C_{\epsilon 2}$) is preserved by the original improper dihedral angles, but without the instability of the separately moving atoms. For tryptophan two interaction-less masses are created at the center of mass of each of the rings and with constraints between them and with C_β . In Fig. 11 a schematic representation of these constructions is shown.

References

- [1] Eaton, W. A., Muñoz, V., Thompson, P. A., Henry, E. R., Hofrichter, J. Kinetics and dynamics of loops, α -helices, β -hairpins and fast-folding proteins. *Acc. Chem. Res.* 31:745–753, 1998.
- [2] Jamin, M., Yeh, S.-R., Rousseau, D. L., Baldwin, R. L. Submillisecond unfolding kinetics of apomyoglobin and its pH 4 intermediate. *J. Mol. Biol.* 292:731–740, 1999.
- [3] Daura, X., Jaun, B., Seebach, D., van Gunsteren, W. F., Mark, A. E. Reversible peptide folding in solution by molecular dynamics simulation. *J. Mol. Biol.* 280:925–932, 1998.
- [4] Duan, Y., Kollman, P. Pathways to a protein folding intermediate observed in a 1-microsecond simulation in aqueous solution. *Science* 282:740–744, 1998.
- [5] Fersht, A. R. Nucleation mechanisms in protein folding. *Curr. Opin. Struct. Biol.* 7:3–9, 1997.
- [6] Ptitsyn, O. B. Protein folding and protein evolution: Common folding nucleus in different subfamilies of c-type cytochromes? *J. Mol. Biol.* 278:655–666, 1998.
- [7] Poupon, A., Mornon, J.-P. Predicting the protein folding nucleus from a sequence. *FEBS Lett.* 452:283–289, 1999.
- [8] Dokholyan, N. V., Buldyrev, S. V., Stanley, H. E., Shakhnovich, E. I. Identifying the protein folding nucleus using molecular dynamics. *J. Mol. Biol.* 296:1183–1188, 2000.

- [9] Panchenko, A., Luthey-Schulten, Z., Wolynes, P. Foldons as independently folding units of proteins. *Physica* 107:312–315, 1997.
- [10] Furois-Corbin, S., Smith, J. C., Kneller, G. R. Picosecond timescale rigid-helix and side-chain motions in deoxymyoglobin. *PROTEINS: Struct. Funct. Gen.* 16:141–154, 1993.
- [11] Bruscolini, P. A coarse-grained, realistic model for protein folding. *J. Chem. Phys.* 107:7512–7529, 1997.
- [12] Bruscolini, P. Testing the helix model for protein folding on four simple proteins. *Modern Phys. Lett. B* 11:691–702, 1997.
- [13] Fischer, K., Marqusee, S. A rapid test for identification of autonomous folding units in proteins. *J. Mol. Biol.* 302:701–712, 2000.
- [14] Yue, K., Dill, K. A. Constraint-based assembly of tertiary protein structures from secondary structure elements. *Prot. Sci.* 9:1935–1946, 2000.
- [15] Feenstra, K. A., Berendsen, H. J. C. The domain decomposition of a single-domain protein. In *Monte Carlo approach to biopolymers and protein folding* (London, 1998). Grassberger, P., Barkema, G., Nadler, W., eds. Höchstleistungsrechenzentrum Jülich, Germany. World Scientific.
- [16] Berman, H., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T., Weissig, H., Shindyalov, I., Bourne, P. The protein data bank. *Nucl. Acids Res.* 28:235–242, 2000.
- [17] Noble, M. E., Musacchio, A., Saraste, M., Courtneidge, S. A., Wierenga, R. K. Crystal structure of the SH3 domain in human Fyn; comparison of the three-

- dimensional structures of SH3 domains in tyrosine kinases and spectrin. *EMBO J.* 12:2617, 1993.
- [18] Morton, C. J., Pugh, D. J., Brown, E. L., Kahmann, J. D., Renzoni, D. A., Campbell, I. D. Solution structure and peptide binding of the SH3 domain from human Fyn. *Structure* 4:705, 1996.
- [19] Weaver, L. H., Matthews, B. W. Structure of bacteriophage T4 lysozyme refined at 1.7 Å resolution. *J. Mol. Biol.* 193:189, 1987.
- [20] van Nuland, N., Hangyi, I. W., van Schaik, R. C., Berendsen, H. J. C., van Gunsteren, W. F., Scheek, R. M., Robillard, G. T. The high-resolution structure of the histidine-containing phosphocarrier protein HPr from *Escherichia coli* determined by restrained molecular dynamics from NMR nuclear overhauser effect data. *J. Mol. Biol.* 237:544–559, 1994.
- [21] Jia, Z., Quail, J. W., Waygood, E. B., Delbaere, L. T. J. The 2.0 Å-resolution structure of *Escherichia coli* histidine-containing phosphocarrier protein HPr. A redetermination. *J. Biol. Chem.* 268:22490–22501, 1993.
- [22] Kovacs, H., Comfort, D., Lord, M., Campbell, I. D., Yudkin, M. D. Solution structure of spoIIAA, a phosphorylatable component of the system that regulates transcription factor σ^F of *Bacillus subtilis*. *Proc. Natl. Acad. Sci. USA* 95:5067–5071, 1998.
- [23] Fucini, P., Renner, C., Herberhold, C., Noegel, A. A., Holak, T. A. The repeating segments of the f-actin cross-linking gelation factor (abp-120) have an immunoglobulin-like fold. *Nature Struct. Biol.* 4:223, 1997.

- [24] Hess, B. Convergence of sampling in protein simulations. *Phys. Rev.* **E**. (Submitted Aug 2001).
- [25] van Gunsteren, W. F., Billeter, S. R., Eising, A. A., Hünenberger, P. H., Krüger, P., Mark, A. E., Scott, W. R. P., Tironi, I. G. *Biomolecular simulation: GROMOS96 manual and user guide*. BIOMOS b.v. Zürich, Groningen 1996.
- [26] Berendsen, H. J. C., van der Spoel, D., van Drunen, R. *GROMACS: A message-passing parallel molecular dynamics implementation*. *Comp. Phys. Comm.* 91:43–56, 1995.
- [27] van der Spoel, D., Hess, B., Feenstra, K. A., Lindahl, E., Berendsen, H. J. C. *GROMACS User Manual version 2.0*. Nijenborgh 4, 9747 AG Groningen, The Netherlands. Internet: <http://md.chem.rug.nl/~gmx> 1999.
- [28] Hess, B., Bekker, H., Berendsen, H. J. C., Fraaije, J. G. E. M. *LINCS: A linear constraint solver for molecular simulations*. *J. Comp. Chem.* 18:1463–1472, 1997.
- [29] Feenstra, A. K., Hess, B., Berendsen, H. J. C. *Improving efficiency of large time-scale molecular dynamics simulations of hydrogen-rich systems*. *J. Comp. Chem.* 20:786–798, 1999.
- [30] Berendsen, H. J. C., Postma, J. P. M., van Gunsteren, W. F., DiNola, A., Haak, J. R. *Molecular dynamics with coupling to an external bath*. *J. Chem. Phys.* 81:3684–3690, 1984.
- [31] Berendsen, H. J. C., Postma, J. P. M., van Gunsteren, W. F., Hermans, J. *Interaction models for water in relation to protein hydration*. In *Intermolecular Forces* (Dordrecht, 1981). Pullman, B., ed. . D. Reidel Publishing Company.

- [32] Hayward, S., Kitao, A., Berendsen, H. J. C. Model-free methods of analyzing domain motions in proteins from simulation: A comparison of a normal mode analysis and a molecular dynamics simulation of lysozyme. *PROTEINS: Struct. Funct. Gen.* 27:425–437, 1997.
- [33] Hayward, S., Berendsen, H. J. C. Systematic analysis of domain motions in proteins conformational change; new results on citrate synthase and T4 lysozyme. *PROTEINS: Struct. Funct. Gen.* 30:144–154, 1998.
- [34] Amadei, A., Linssen, A. B. M., Berendsen, H. J. C. Essential dynamics of proteins. *Proteins: Struct. Funct. Gen.* 17:412–425, 1993.
- [35] Hayward, S. Structural principles governing domain motions in proteins. *PROTEINS: Struct. Funct. Gen.* 36:425–435, 1999.
- [36] Kabsch, W., Sander, C. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* 22:2577–2637, 1983.
- [37] de Groot, B. L., Hayward, S., van Aalten, D. M. F., Amadei, A., Berendsen, H. J. C. Domain motions in bacteriophage t4 lysozyme; a comparison between molecular dynamics and crystallographic data. *Proteins: Struct. Funct. Gen.* 31:116–127, 1998.
- [38] Taylor, W. R. Defining linear segments in protein structure. *J. Mol. Biol.* 310:1135–1150, 2001.
- [39] Riddle, D., Grantcharova, V., Santiago, J., Alm, E., Ruczinski, I., Baker, D. Experiment and theory highlight role of native state topology in SH3 folding. *Nature Struct. Biol.* 6:1016–1024, 1999.

- [40] Hilser, V. J., Townsend, B. D., Freire, E. Structure-based statistical thermodynamic analysis of T4 lysozyme mutants: Structural mapping of cooperative interactions. *Bioph. Chem.* 64:69–79, 1997.
- [41] Lu, J., Dahlquist, F. W. Detection and characterization of an early folding intermediate of T4 lysozyme using pulsed hydrogen exchange and two-dimensional NMR. *Biochemistry* 31:4749–4756, 1992.
- [42] Najbar, L. V., Craik, D. J., Wade, J. D., McLeish, M. J. Identification of initiation sites for T4 lysozyme folding using CD and NMR spectroscopy of peptide fragments. *Biochemistry* 39:5911–5920, 2000.
- [43] Chen, L., Hodgson, K. O., Doniach, S. A lysozyme folding intermediate revealed by solution x-ray scattering. *J. Mol. Biol.* 261:658–671, 1996.
- [44] Kraulis, P. J. MOLSCRIPT: a program to produce both detailed and schematic plots of protein structures. *J. Appl. Cryst.* 24:946–950, 1991.
- [45] Esnouf, R. M. An extensively modified version of MOLSCRIPT that includes greatly enhanced coloring capabilities. *J. Mol. Graphics* 15:132–134, 1997.

List of Tables

- I Proteins, system sizes and overall simulation properties. From left to right: label by which the system is referred to; PDB entry label; number of structures in the PDB file; number of residues in the protein; number of atoms in the protein and in the water; number, length and total length of simulations; simulation box size and radius of gyration (r_g) of the protein. The number of water atoms, box size and radius of gyration are averaged over the simulations. 32
- II Essential dynamics (ED) and semi-rigid element assignments. From left to right: total number of eigenvectors analyzed; percentage of the total mean square fluctuation (MSF) in the atomic positions contained in the first six eigenvectors, percentage of ED eigenvectors with a semi-rigid element assignment and percentage of residues assigned to a semi-rigid element. Long trajectories (≥ 40 ns, see Table I) were fragmented for analysis into adjacent 10 ns parts (*), 1 ns parts with a 1 ns separation (†) or adjacent 4 ns parts (‡). . . . 33
- III Statistics of amino acid residue composition relative to element boundaries in absolute and relative numbers. From left to right: residue type; total number of that type in all sequences (N_X); and close to boundaries ($N_{X,\Delta}$ for $|\Delta| \leq 2$); relative occurrence of that type observed (α_X) and expected for random ($\alpha_{X,random}$) and statistical confidence level for deviation of observed value from random. Relative occurrences (α_X) deviating more than one σ from random are indicated in bold. 34

protein label	PDB entry	# Str.	# Res.	# atoms		# Sim.	length (ns)	total (ns)	box (nm)	r_g (nm)
Sh3	1shf	2	59	627	7968	2	600	1200	4.53	1.05
	1nyf	1	58	618	7983	1	600	600	4.53	1.03
T4L	2lzm	1	164	1731	21484	2	40	80	6.33	1.63
HPr	1hdn	30	85	821	7957	30	2	60	4.84	1.17
	1poh	1	85	821	12225	2	40	80	5.04	1.15
1auz	1auz	24	116	1146	14152	24	2	48	5.51	1.31
1ksr	1ksr	20	100	931	14801	20	5	100	5.53	1.37

Table I: Proteins, system sizes and overall simulation properties. From left to right: label by which the system is referred to; PDB entry label; number of structures in the PDB file; number of residues in the protein; number of atoms in the protein and in the water; number, length and total length of simulations; simulation box size and radius of gyration (r_g) of the protein. The number of water atoms, box size and radius of gyration are averaged over the simulations.



Protein		total # of eigenvectors	% of MSF in eigenvectors	% of eigenv. with s.-rigid elements	% of residues in s.-rigid elements
Sh3	1shf/1nyf *	1 080	67	58	78
T4L	2lzm †	240	63	45	80
	2lzm ‡	120	67	52	83
HPr	1hdn	180	65	56	63
	1poh ‡	120	58	51	73
1auz	1auz	144	76	44	61
1ksr	1ksr	120	73	29	66

Table II: Essential dynamics (ED) and semi-rigid element assignments. From left to right: total number of eigenvectors analyzed; percentage of the total mean square fluctuation (MSF) in the atomic positions contained in the first six eigenvectors, percentage of ED eigenvectors with a semi-rigid element assignment and percentage of residues assigned to a semi-rigid element. Long trajectories (≥ 40 ns, see Table I) were fragmented for analysis into adjacent 10 ns parts (*), 1 ns parts with a 1 ns separation (†) or adjacent 4 ns parts (‡).

res. type	N_X	$N_{X,\Delta}$ $ \Delta \leq 2$	α_X	$\alpha_{X,random}$	confidence level
Gly	43	18	0.91	1.00±0.17	
Ala	36	12	0.72	1.00±0.18	
Val	40	17	0.92	1.00±0.17	
Ile	27	13	1.05	1.00±0.29	
Leu	44	22	1.09	1.00±0.16	
Pro	16	7	0.95	1.00±0.27	
Phe	20	10	1.09	1.00±0.24	
Trp	5	1	0.43	1.00±0.48	
Met	10	5	1.09	1.00±0.34	
Cys	5	2	0.87	1.00±0.48	
Ser	27	18	1.45	1.00±0.29	> 95%
Thr	36	22	1.33	1.00±0.18	
Asn	20	7	0.76	1.00±0.24	
Gln	19	5	0.57	1.00±0.25	
Tyr	14	7	1.09	1.00±0.29	
His	9	8	1.93	1.00±0.36	> 95%
Asp	34	13	0.83	1.00±0.19	
Glu	37	15	0.88	1.00±0.18	
Lys	36	19	1.15	1.00±0.18	
Arg	22	9	0.89	1.00±0.23	
total	500	230			

Table III: Statistics of amino acid residue composition relative to element boundaries in absolute and relative numbers. From left to right: residue type; total number of that type in all sequences (N_X); and close to boundaries ($N_{X,\Delta}$ for $|\Delta| \leq 2$); relative occurrence of that type observed (α_X) and expected for random ($\alpha_{X,random}$) and statistical confidence level for deviation of observed value from random. Relative occurrences (α_X) deviating more than one σ from random are indicated in bold.

List of Figures

1	Schematic representation of a selection from all semi-rigid element assignments for Sh3 (60 out of a total of 1080 are shown). On the horizontal axis the residues are plotted. On the vertical axis a cumulative index is plotted of all essential modes for the trajectories that were selected for display. Different shades correspond to different semi-rigid elements assigned by DYNDOM for that particular essential mode, shades do not correlate between different essential modes.	37
2	Outline of consensus dynamics results for Sh3 . (a) Consensus dynamics matrix as derived from the individual element assignments (see Fig. 1a). Residue number is indicated on both axes. Intensities correspond to the number of times both residues of a pair can be found together in the same semi-rigid element. (b) Schematic view of the hierarchy of motionally coherent elements as observed in the matrix of a , organized into hierarchical levels (numbered from 1) and secondary structure elements (labeled ‘S’,  =α-helix,  =β-strand, labeled according to the PDB file). Horizontal lines indicate the elements. The vertical dotted lines indicate the boundaries and the correlation between hierarchical levels. Elements in the lowest hierarchical level are numbered along the sequence. (c) Count of semi-rigid element boundaries. Vertical dotted lines indicate significant peaks and correspond to element demarcations as defined in b	38
3	Representation of the two levels of motional hierarchy for Sh3 depicted schematically in Fig. 2b, color coded onto a cartoon representation of the crystal structure of the protein. Colors indicate different elements. The last residue of each element, as well as residue 1, are labeled. Plots were generated using a modified version of MolScript. ^{44,45}	39
4	Outline of results for T4L . (a) Consensus dynamics matrix (cf. Fig. 2a). Results for fragmenting the long trajectories into 1 ns parts with a 1 ns separation or into adjacent 4 ns parts are plotted in the upper left and the lower right half of the matrix respectively. (b) Hierarchy of elements (cf. Fig. 2b). (c) Demarcation counts (cf. Fig. 2c) summed over the results corresponding to both halves of the matrix.	40
5	Representation of the three levels of motional hierarchy for T4L as depicted schematically in Fig. 4b (cf. Fig. 3).	41
6	Results for HP_r , 1auz and 1ksr . (a) Consensus dynamics matrix (cf. Fig. 2a). For HP_r , results for the simulations started from 1hdn and 1poh are plotted in the upper left and the lower right half of the matrix respectively. (b) Hierarchy of elements (cf. Fig. 2b). (c) Demarcation counts (cf. Fig. 2c). For HP_r only the total demarcation counts corresponding to both halves of the matrix are plotted.	42
7	Representation of the lowest level of motional hierarchy for HP_r , 1auz and 1ksr depicted schematically in Fig. 6b (cf. Fig. 3).	43
8	Comparison with experimental data for Sh3 from Riddle <i>et al.</i> ³⁹ (a) Hierarchy of motionally coherent elements (see Fig. 2b). (b) Five structural regions in order of importance in the folding transition state. ³⁹ Block height corresponds to relative importance. (c) Sequence of folding events as determined by the <i>ab-initio</i> folding method ROSETTA ³⁹ (plot modified from original paper with permission), folding progresses from the bottom upwards. Black corresponds to residues without native contacts and white to residues with all native contacts formed.	44

9	Comparison with experimental data for T4L. (a) Hierarchy of elements (see Fig. 4b). (b) logarithm of predicted folding rate ($\ln \kappa_f$) as determined by mutational analysis performed by Hilser & Freire ⁴⁰ (plot manually extracted from original paper). (c) protection factors determined by pulsed hydrogen exchange performed by Lu & Dahlquist ⁴¹ (plot manually extracted from original paper). The horizontal dotted line corresponds to the lower bound of 20 mentioned in the text. (d) Proposed folding initiation sites as determined from analysis of peptide fragments performed by Najbar <i>et al.</i> ⁴² Boxes indicate investigated fragments, filled boxes correspond to proposed folding initiation sites.	45
10	Non-unique examples of tentative folding pathways for Sh3 and T4L where folding can be thought to progress from the outer branches inwards. These pathways can account for the occurrence of all motionally coherent elements observed and correspond to the hierarchy as observed in Figs. 2 and 4.	46
11	Schematic view of the different types of dummy atom constructions used for aromatic sidechains. Legend: ● atoms and masses used in the construction of the dummy atom(s); ● dummy atoms; — chemical bonds; - - constraints. Hydrogens are smaller than heavy atoms. Note, the hydroxyl hydrogen is not a dummy atom in tyrosine, and the constraint between C _{e1} and C _{e2} in histidine.	47

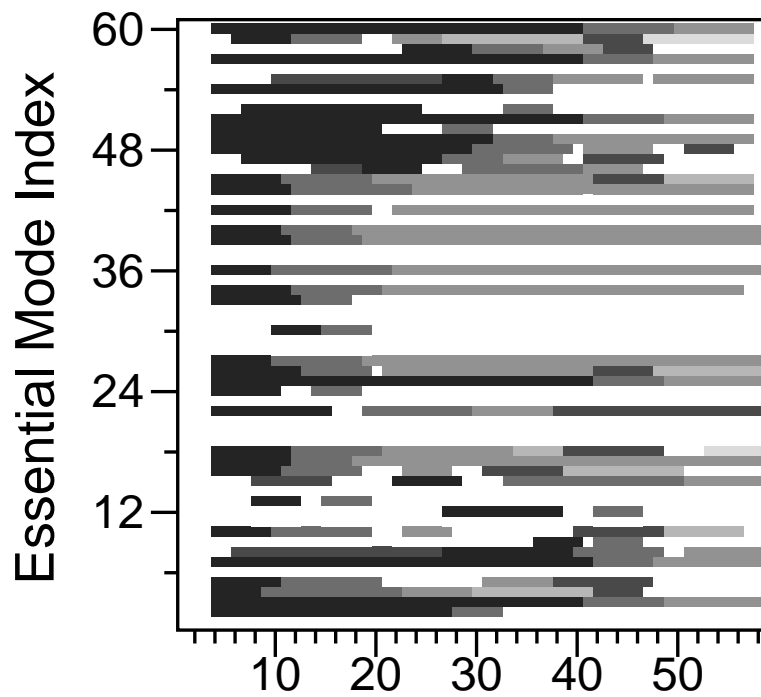


Figure 1: Schematic representation of a selection from all semi-rigid element assignments for Sh3 (60 out of a total of 1080 are shown). On the horizontal axis the residues are plotted. On the vertical axis a cumulative index is plotted of all essential modes for the trajectories that were selected for display. Different shades correspond to different semi-rigid elements assigned by DYNDOM for that particular essential mode, shades do not correlate between different essential modes.

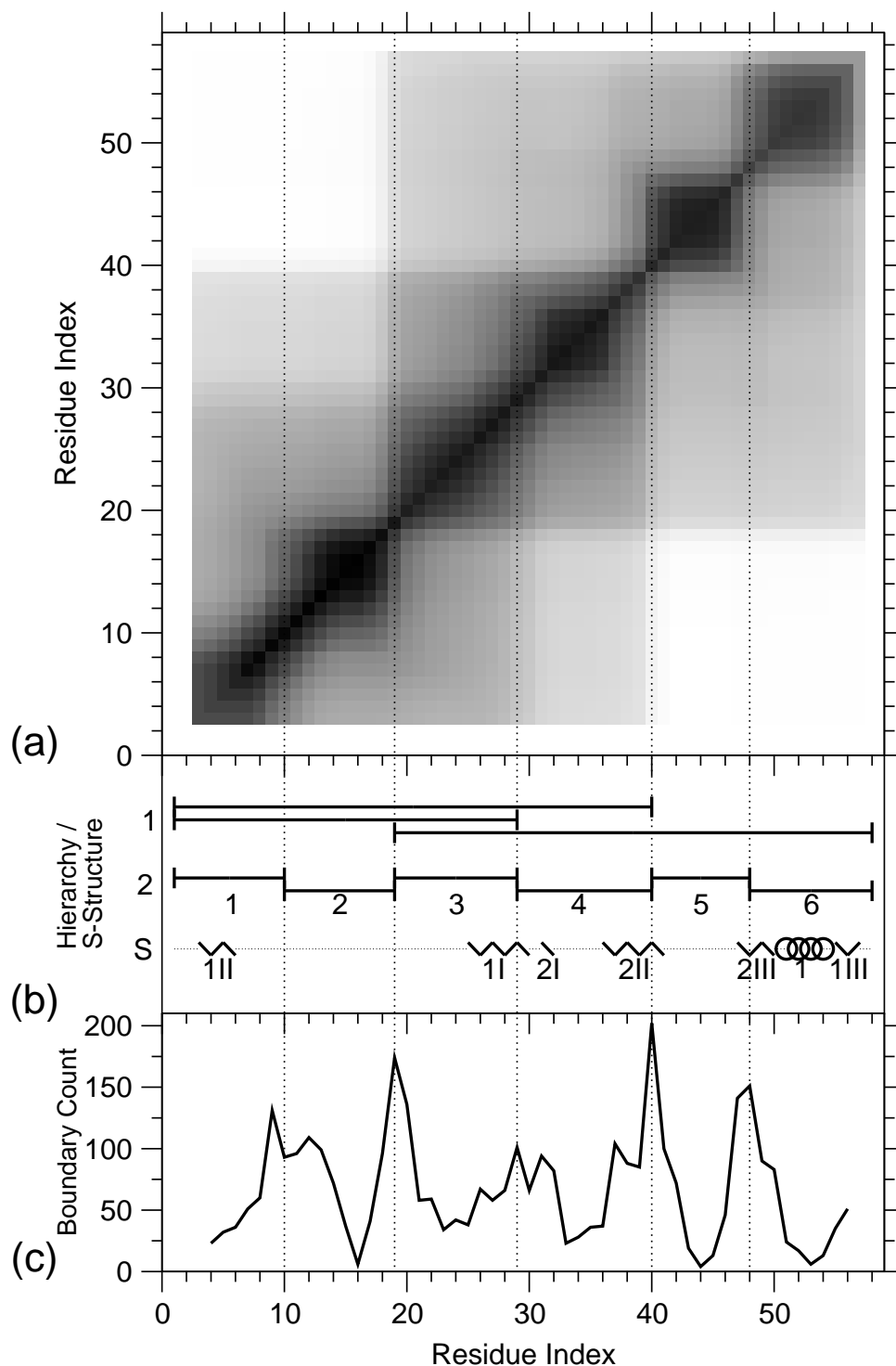


Figure 2: Outline of consensus dynamics results for Sh3. (a) Consensus dynamics matrix as derived from the individual element assignments (see Fig. 1a). Residue number is indicated on both axes. Intensities correspond to the number of times both residues of a pair can be found together in the same semi-rigid element. (b) Schematic view of the hierarchy of motionally coherent elements as observed in the matrix of a, organized into hierarchical levels (numbered from 1) and secondary structure elements (labeled 'S', α -helix, β -strand, labeled according to the PDB file). Horizontal lines indicate the elements. The vertical dotted lines indicate the boundaries and the correlation between hierarchical levels. Elements in the lowest hierarchical level are numbered along the sequence. (c) Count of semi-rigid element boundaries. Vertical dotted lines indicate significant peaks and correspond to element demarcations as defined in b.

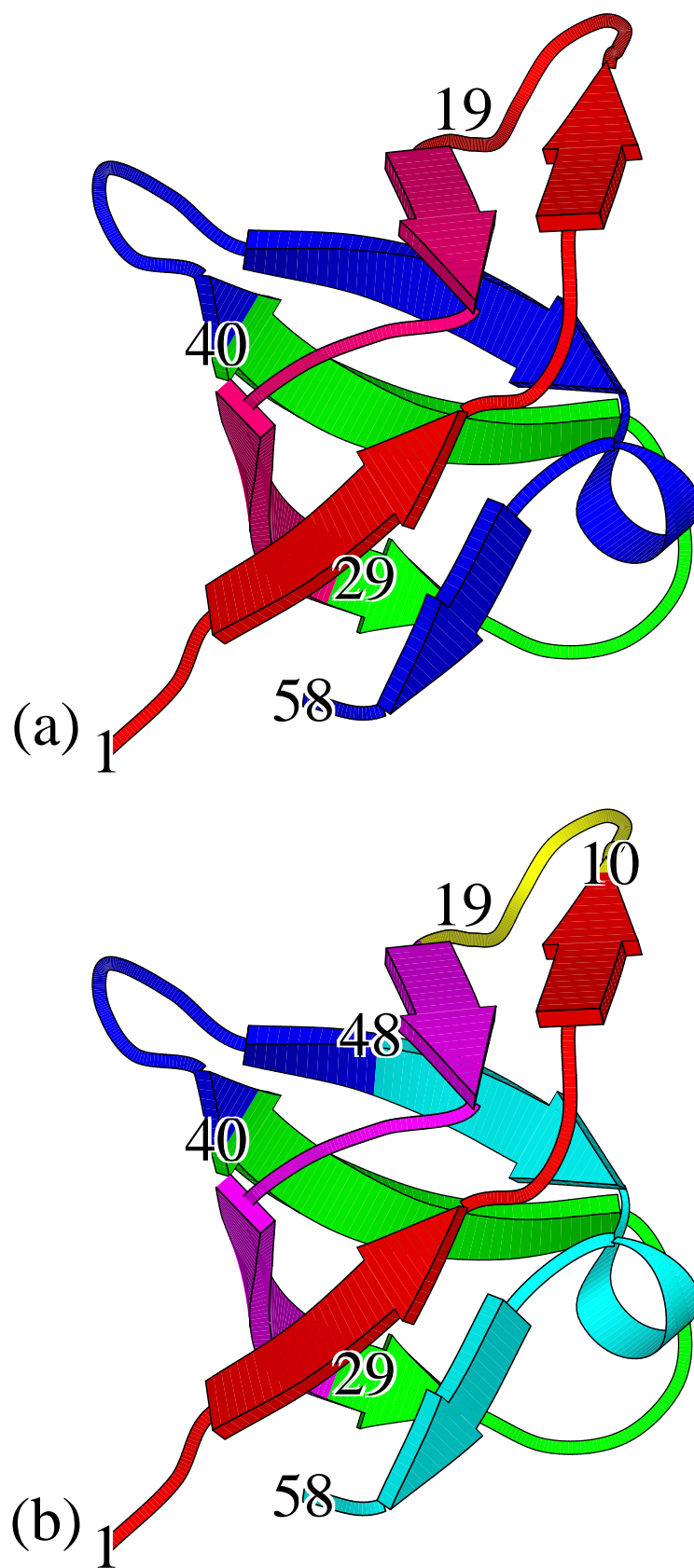


Figure 3: Representation of the two levels of motional hierarchy for Sh3 depicted schematically in Fig. 2b, color coded onto a cartoon representation of the crystal structure of the protein. Colors indicate different elements. The last residue of each element, as well as residue 1, are labeled. Plots were generated using a modified version of MolScript.^{44,45}

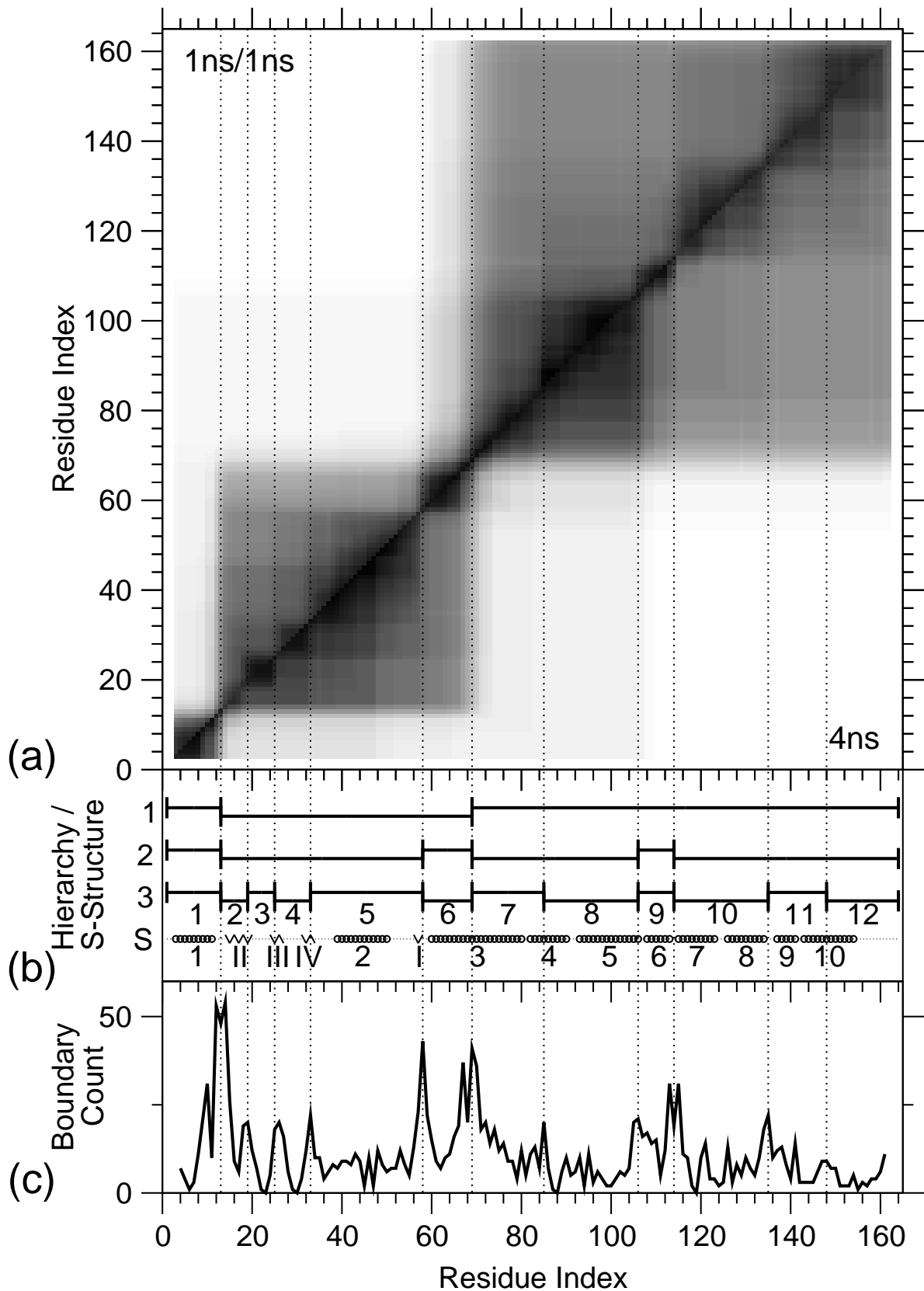


Figure 4: Outline of results for T4L. (a) Consensus dynamics matrix (cf. Fig. 2a). Results for fragmenting the long trajectories into 1 ns parts with a 1 ns separation or into adjacent 4 ns parts are plotted in the upper left and the lower right half of the matrix respectively. (b) Hierarchy of elements (cf. Fig. 2b). (c) Demarcation counts (cf. Fig. 2c) summed over the results corresponding to both halves of the matrix.

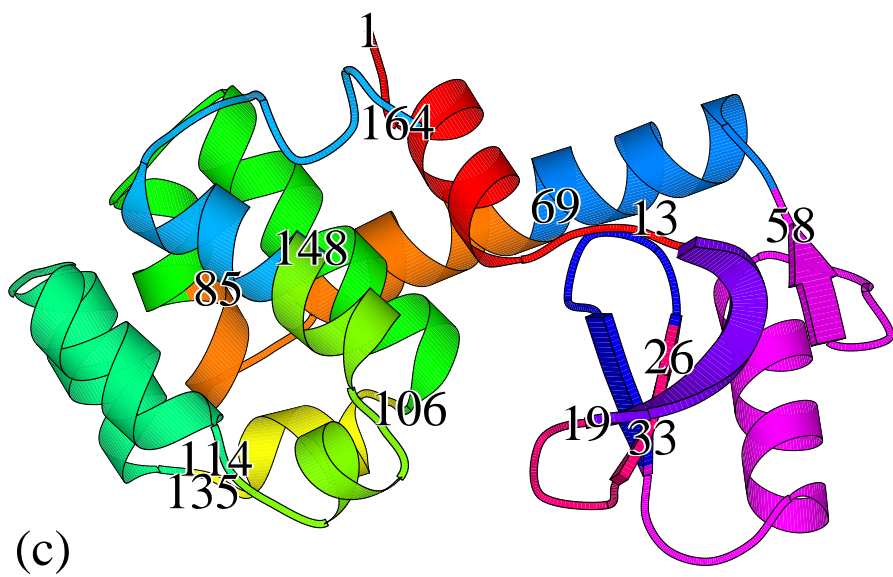
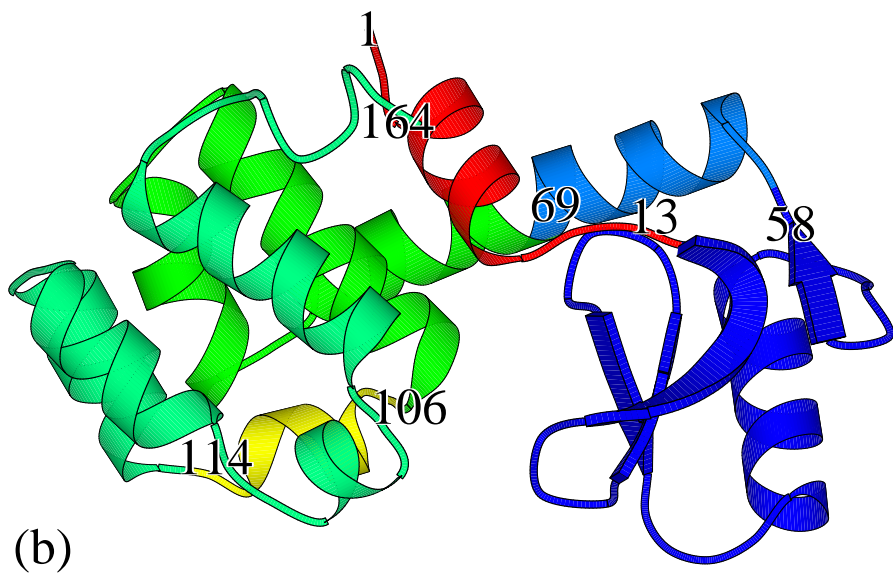
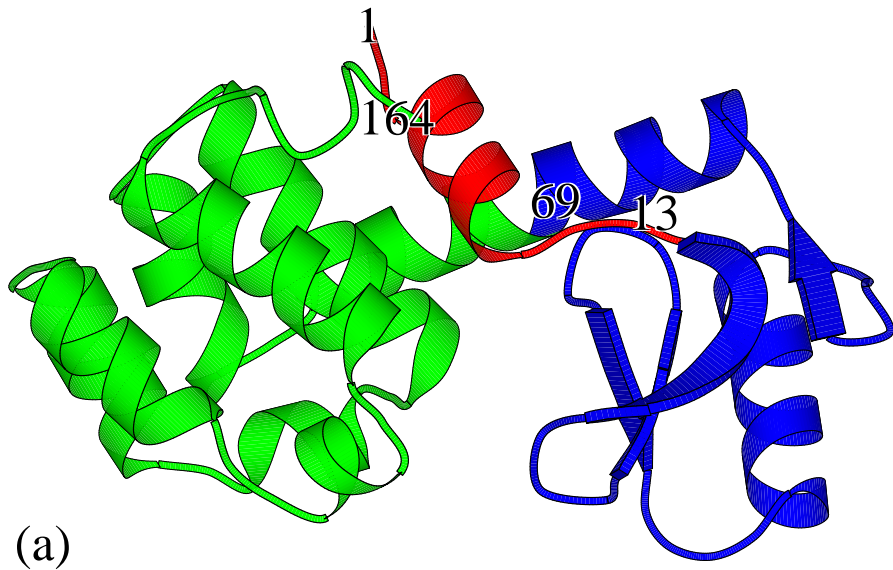


Figure 5: Representation of the three levels of motional hierarchy for T4L as depicted schematically in Fig. 4b (cf. Fig. 3).

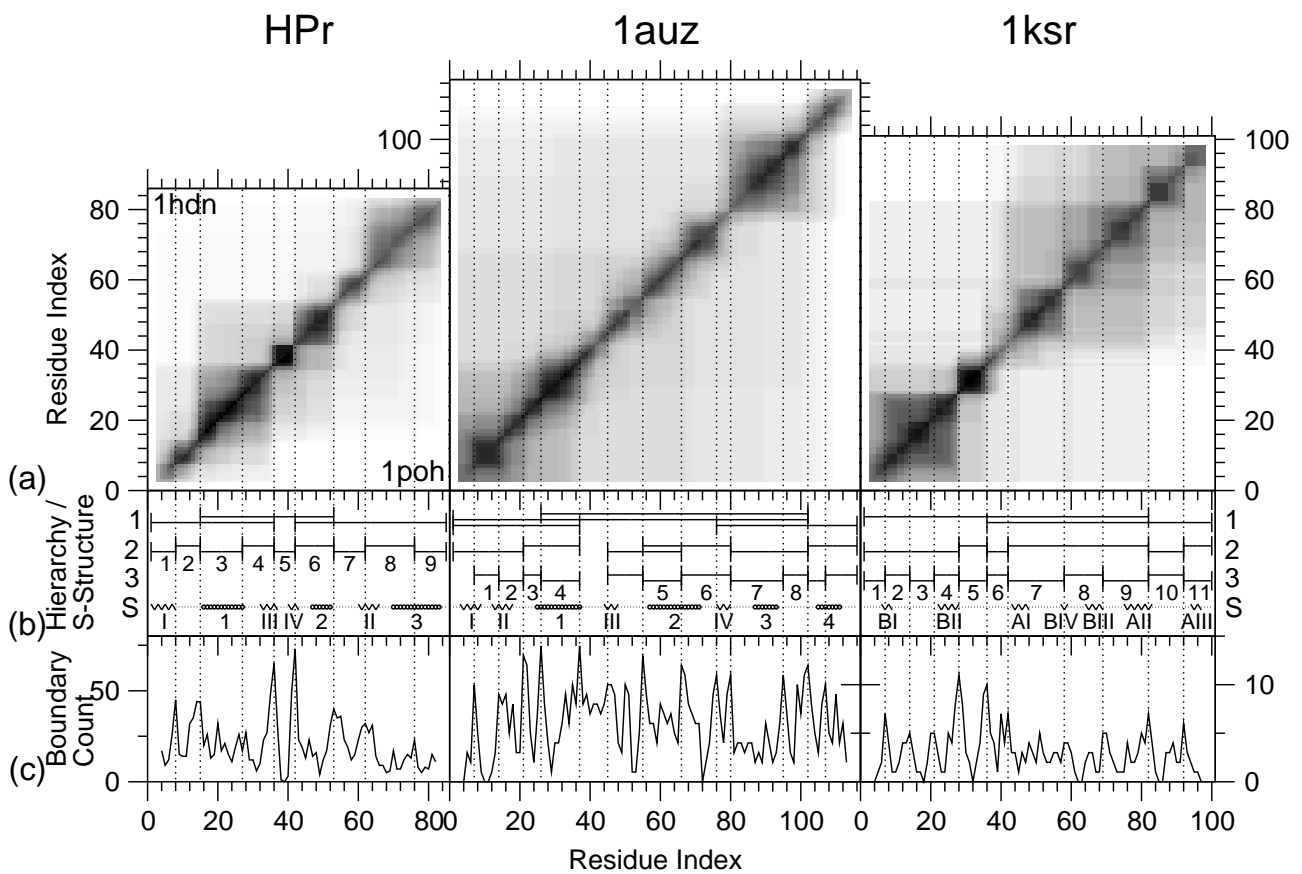


Figure 6: Results for HPr, 1auz and 1ksr. (a) Consensus dynamics matrix (cf. Fig. 2a). For HPr, results for the simulations started from 1hdn and 1poh are plotted in the upper left and the lower right half of the matrix respectively. (b) Hierarchy of elements (cf. Fig. 2b). (c) Demarcation counts (cf. Fig. 2c). For HPr only the total demarcation counts corresponding to both halves of the matrix are plotted.

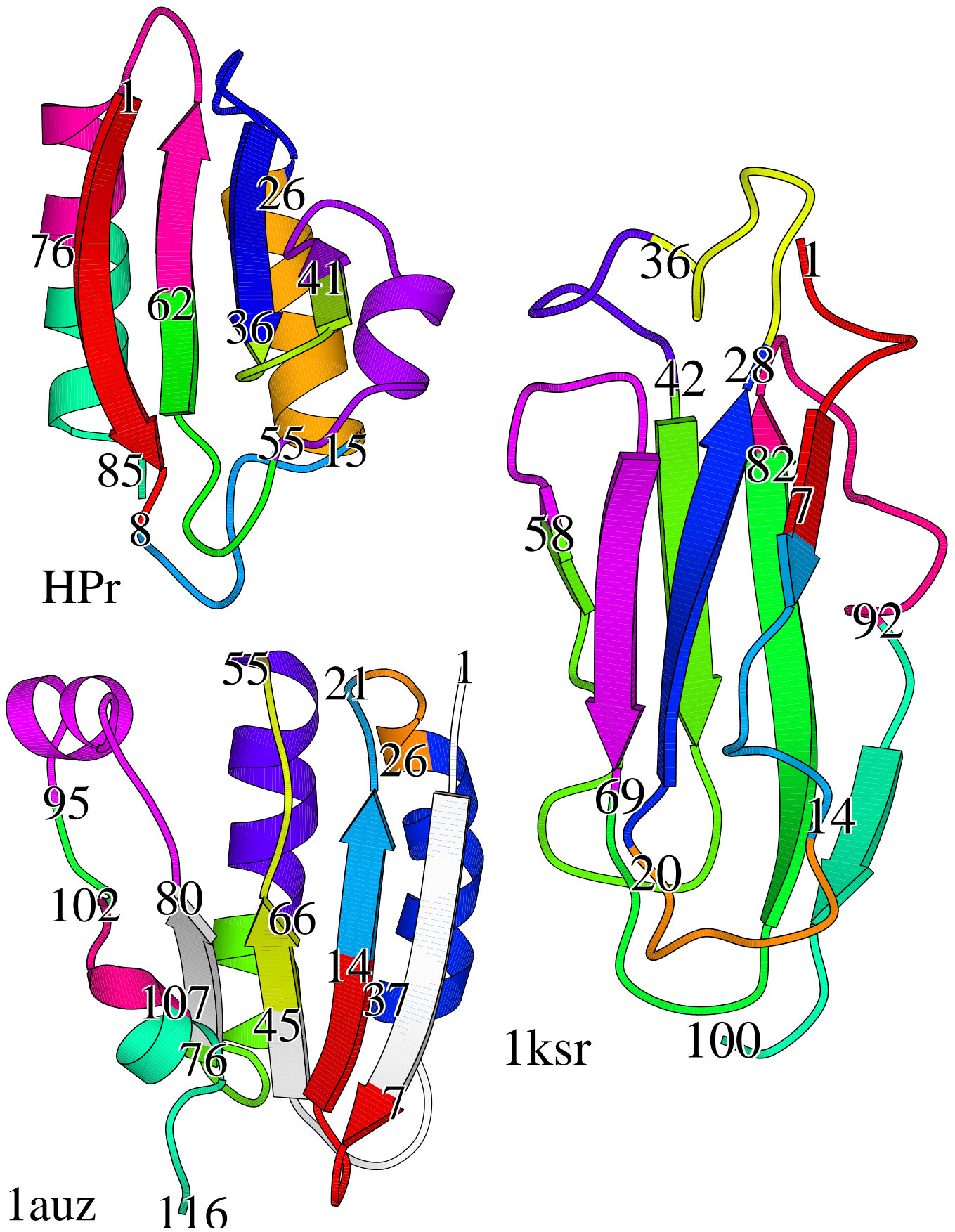


Figure 7: Representation of the lowest level of motional hierarchy for HPr, 1auz and 1ksr depicted schematically in Fig. 6b (cf. Fig. 3).

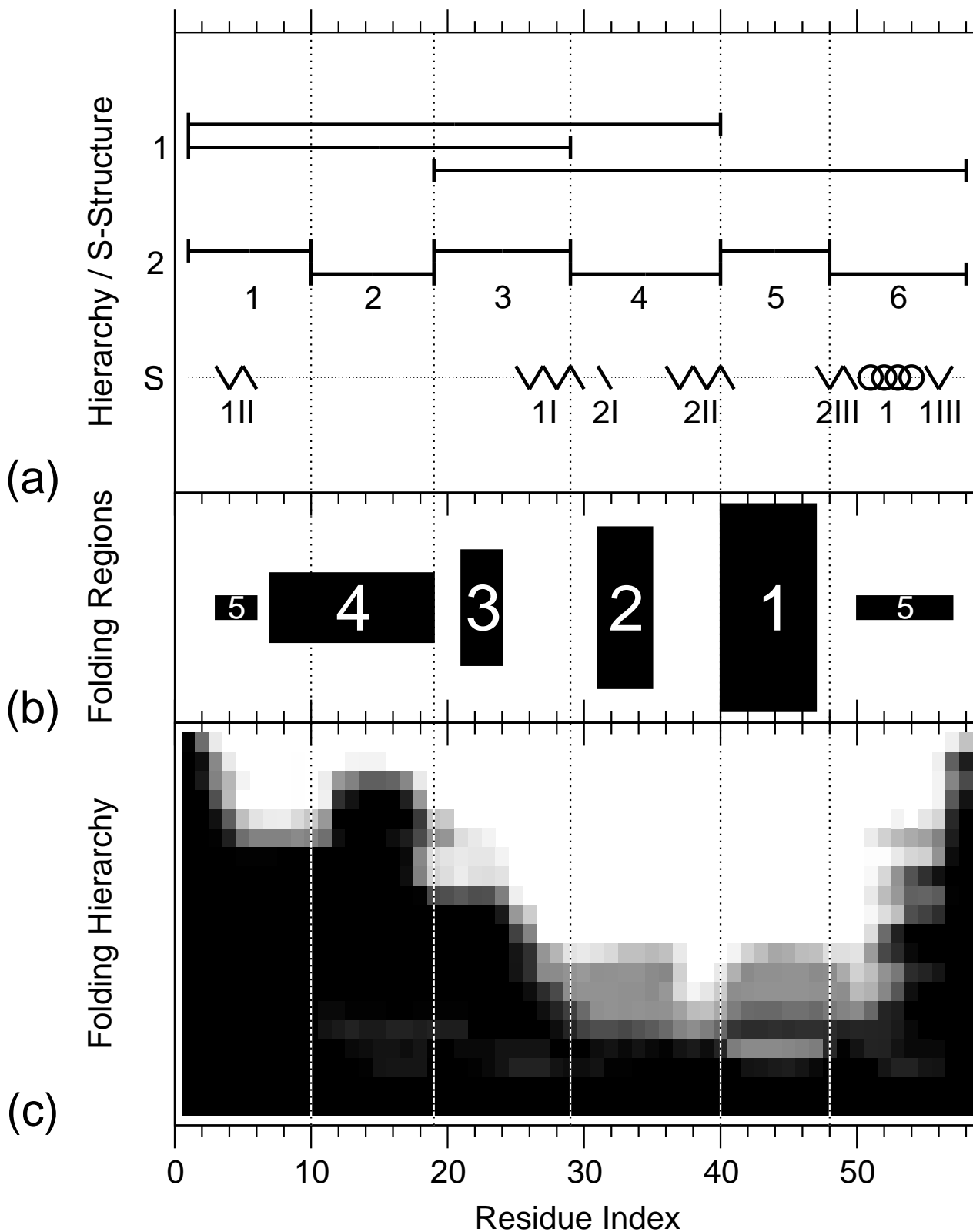


Figure 8: Comparison with experimental data for Sh3 from Riddle et al.³⁹ (a) Hierarchy of motionally coherent elements (see Fig. 2b). (b) Five structural regions in order of importance in the folding transition state.³⁹ Block height corresponds to relative importance. (c) Sequence of folding events as determined by the *ab-initio* folding method ROSETTA³⁹ (plot modified from original paper with permission), folding progresses from the bottom upwards. Black corresponds to residues without native contacts and white to residues with all native contacts formed.

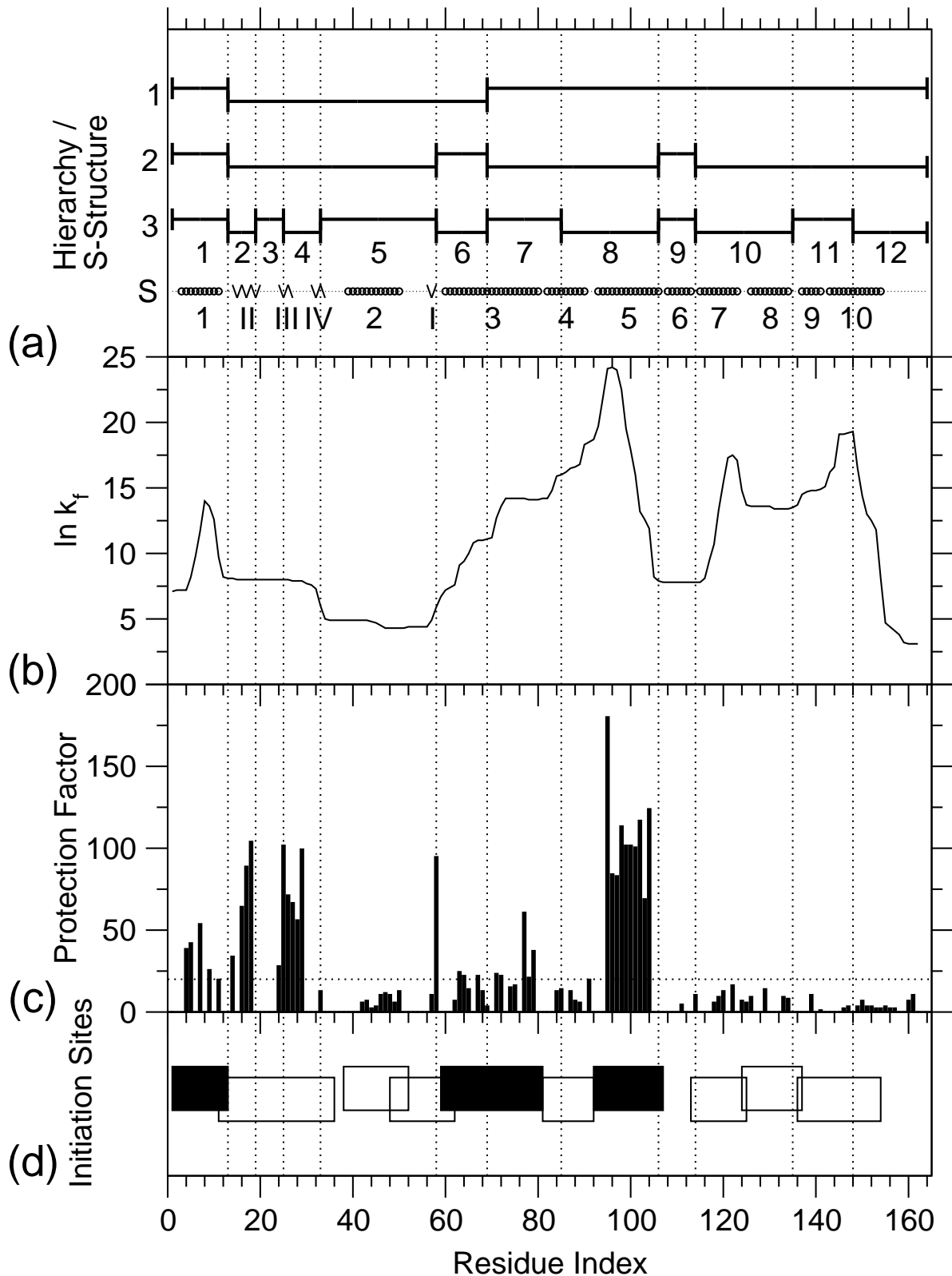


Figure 9: Comparison with experimental data for T4L. (a) Hierarchy of elements (see Fig. 4b). (b) logarithm of predicted folding rate ($\ln \kappa_f$) as determined by mutational analysis performed by Hilser & Freire⁴⁰ (plot manually extracted from original paper). (c) protection factors determined by pulsed hydrogen exchange performed by Lu & Dahlquist⁴¹ (plot manually extracted from original paper). The horizontal dotted line corresponds to the lower bound of 20 mentioned in the text. (d) Proposed folding initiation sites as determined from analysis of peptide fragments performed by Najbar et al.⁴² Boxes indicate investigated fragments, filled boxes correspond to proposed folding initiation sites.

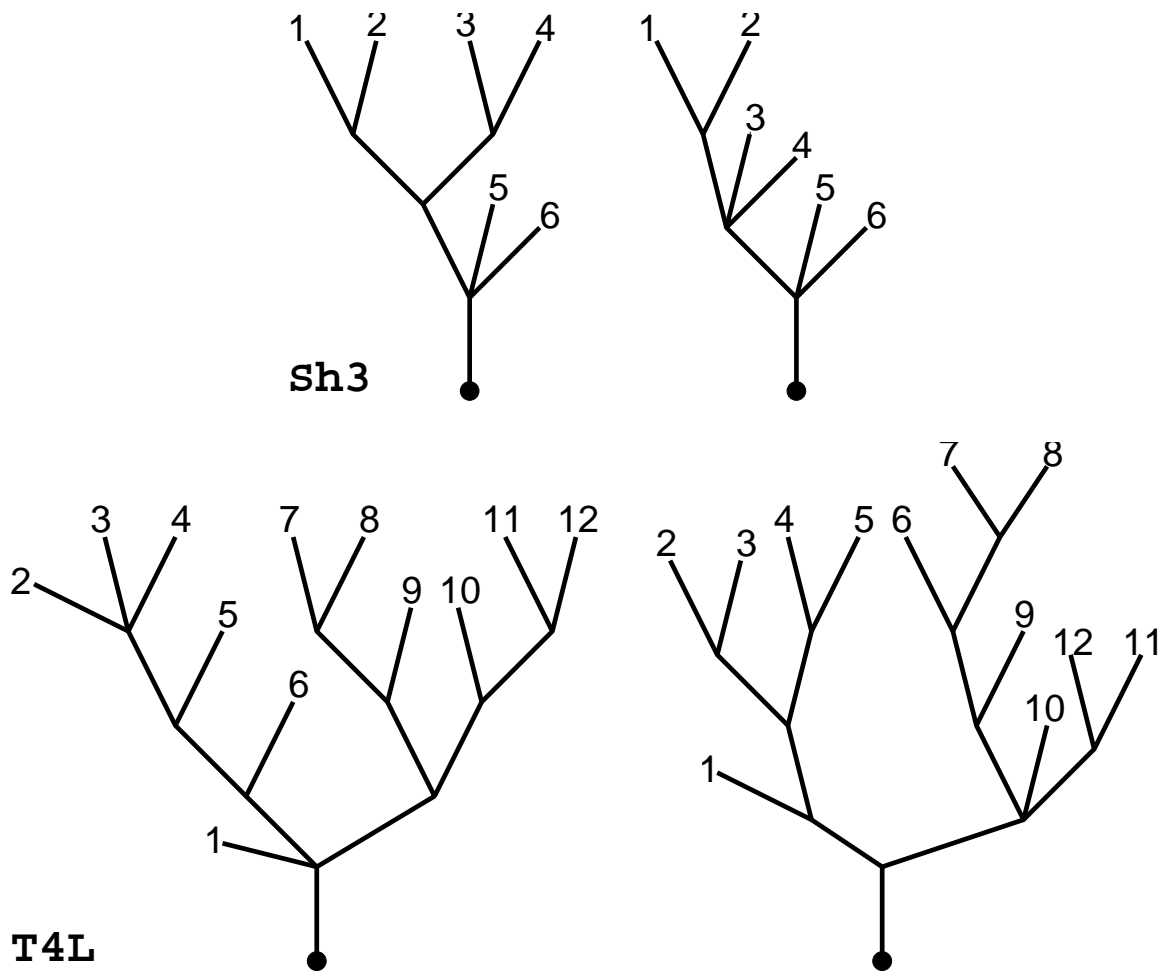


Figure 10: Non-unique examples of tentative folding pathways for Sh3 and T4L where folding can be thought to progress from the outer branches inwards. These pathways can account for the occurrence of all motionally coherent elements observed and correspond to the hierarchy as observed in Figs. 2 and 4.

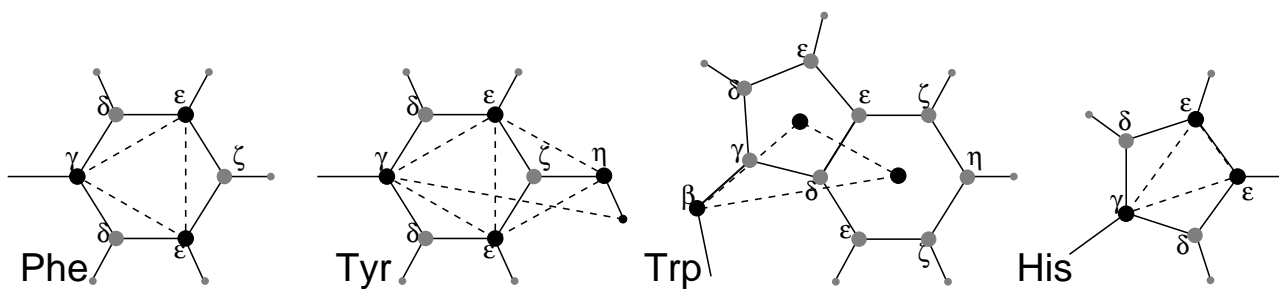


Figure 11: Schematic view of the different types of dummy atom constructions used for aromatic sidechains. Legend: ● atoms and masses used in the construction of the dummy atom(s); ● dummy atoms; — chemical bonds; - - constraints. Hydrogens are smaller than heavy atoms. Note, the hydroxyl hydrogen is not a dummy atom in tyrosine, and the constraint between $C_{\epsilon 1}$ and $C_{\epsilon 2}$ in histidine.