Willy Wriggers[1,*]
Sugoto Chakravarty[2,*]
Patricia A. Jennings[3,*]
[1] *School of Health Information Sciences and Institute of Molecular Medicine University of Texas, Health Science Center Houston, 7000 Fannin St., Houston, TX 77030*

[2] *Department of Biochemistry and Molecular Biology Baylor College of Medicine, One Baylor Plaza, Houston, TX 77030*

[3] *Department of Chemistry and Biochemistry, University of California, San Diego, 9500 Gilman Drive, MC 0332, La Jolla, CA 92093-0332*

Control of Protein Functional Dynamics by Peptide Linkers

Received 31 January 2005; revised 22 March 2005; accepted 19 April 2005
Published online 6 May 2005 in Wiley InterScience (www.interscience.wiley.com). DOI 10.1002/bip.20291

Abstract: *Control of structural flexibility is essential for the proper functioning of a large number of proteins and multiprotein complexes. At the residue level, such flexibility occurs due to local relaxation of peptide bond angles whose cumulative effect may result in large changes in the secondary, tertiary or quaternary structures of protein molecules. Such flexibility, and its absence, most often depends on the nature of interdomain linkages formed by oligopeptides. Both flexible and relatively rigid peptide linkers are found in many multidomain proteins. Linkers are thought to control favorable and unfavorable interactions between adjacent domains by means of variable softness furnished by their primary sequence. Large-scale structural heterogeneity of multidomain proteins and their complexes, facilitated by soft peptide linkers, is now seen as the norm rather than the exception. Biophysical discoveries as well as computational algorithms and databases have reshaped our understanding of the often spectacular biomolecular dynamics enabled by soft linkers. Absence of such motion, as in so-called molecular rulers, also has desirable functional effects in protein architecture. We review here the historic discovery and current understanding of the nature of domains and their linkers from a structural, computational, and biophysical point of view. A number of emerging applications, based on the current understanding of the structural properties*

Correspondence to: Willy Wriggers; e-mail: wriggers@biomachina.org
Contract grant sponsor: NIH, Human Frontier Science Program (HFSP), and Alfred P. Sloan Foundation (APSF)
Contract grant number: 1R01 GM62968 (NIH), RGP0026/2003 (HFSP), and BR-4297 (APSF)
*The authors contributed equally to the content of this article.

of peptides, are presented in the context of domain fusion of synthetic multifunctional chimeric proteins. Biopolymers (Pept Sci) 80: 736–746, 2005

This article was originally published online as an accepted preprint. The "Published Online" date corresponds to the preprint version. You can request a copy of the preprint by emailing the Biopolymers editorial office at biopolymers@wiley.com

Keywords: *structural flexibility; peptide linkers; biomolecular dynamics; peptide bond angles; secondary, tertiary, and quaternary structures; protein molecules; oligopeptides; multidomain proteins; molecular rulers*

INTRODUCTION

An important consideration in protein architecture and function is the flexibility of linkers that interconnect the various domains in multidomain proteins commonly found in many biological processes. These linkers are stretches of amino acid residues that establish communication between the different domains and the functional modules.[1] A number of early examples (e.g., immunoglobulin, diphtheria toxin, tomato bushy stunt virus) established a clear relationship between linker peptides and the functional dynamics they enable.

Significant efforts to understand the structural basis of such motion have been undertaken since the 1960s. X-ray crystallographic structures have provided experimental snapshots of motion in many proteins. Analyses of these structures have been crucial in determining the exact regions undergoing motion. Such analyses have shown that protein motion may occur due to conformational changes in individual residues or at the secondary, tertiary, or quaternary structural levels. Lactate dehydrogenase,[2,3] triose-phosphate isomerase,[4] as well as hemoglobin and related proteins[5] are some of the earliest examples of proteins that showed conformational changes with important functional implications. Subsequent analyses of different protein families have identified common structural domains[6] that undergo functionally relevant conformational changes.

Substantial work has been done since the 1970s in characterizing the tethered freedom of movement of protein domains. Prediction of the flexibility of a hinge region is based on an understanding of the rotational freedom of the attached moieties. For example, we will discuss that glycine-rich peptides confer flexibility, which allows the specific engineering of hinge regions into proteins to achieve desired functional motions. Other linkers that will be discussed are more defined by their ability to reliably predict and maintain end-to-end distances between attached domains. Such structurally rigid peptides, often called molecular rulers, have been conjugated to molecules to serve a metric function.

In this work, we will review the biophysical and computational techniques to detect and quantitatively characterize the geometric properties of domains and their linkers. Such techniques have helped in the design of multifunctional chimeric protein analogs that are synthesized using modern molecular biology techniques. Often, dynamic domains and entire proteins are fused together by specific linker peptides. These chimera proteins facilitate the study of protein folding, allow the crystallographic characterization of noncrystallizing components, and enable the labeling and tracking of proteins in optical microscopy. We give a comprehensive overview of such state-of-the-art technologies.

EMERGENCE OF A DYNAMICAL PICTURE OF PROTEIN ARCHITECTURE

The first X-ray crystallographic structures provided a static picture of protein architecture, but multidomain structures in multiple conformations were soon discovered, where individual domains were connected by flexible linkers. The concept of hinge-bending,[7] whereby the relative flexibility of short regions of the polypeptide chain allows significant movement of structural domains, gained widespread acceptance in the 1980s and early 1990s, after evidence for conformational transitions in identical or homologous proteins became known. According to Dobson,[7] the domains themselves are closely similar in each case, a fact that can be attributed to their motion as rigid bodies around a screw axis.[8] It was discovered that hinge regions are soft-linker regions of localized torsion angle changes in the polypeptide chain that allow the attached rigid domains to pivot. The rotation axes of these torsion angle changes are nearly parallel to the overall axis of rotation, so the local motion in the hinges can be directly related to the overall motion. A crucial feature of the hinge residues is that they have very few packing constraints on their main chain atoms.[8,9]

Protein domains can be defined as segmented portions of a polypeptide sequence that assume stable three-dimensional structure.[6] Such recurring protein motifs are significant because it is increasingly recognized that there are only a limited number of domain families in nature. These domains are duplicated and combined in different ways to form the set of proteins in genomes.[10] The importance of domains is further exemplified by the fact that multidomain proteins

play a major role in many cellular processes. Although a consensus in detail is still lacking, various effective criteria have been proposed to detect and define protein domains. These criteria rely mainly on the existence of local structural compactness arising from β-sheets or hydrophobic cores.[11,12] Based on such compactness, computational algorithms to detect structural domains have been proposed.[13–17]

Hinge regions occur between domains, allowing them to move independently of one another while maintaining the individual domains' three-dimensional shape.[6] In that sense, hinge regions are characterized by a structural softness that enables this motion. Early structural biologists observed that hinge regions may remove steric constraints from the relative motion of the attached moieties. In tomato bushy stunt virus protein, domain hinging by 20° helps pack 180 identical proteins into the 60-fold icosahedral symmetry of the virus shell.[18] According to Harrison, the "strain" for quasiequivalent bonding is concentrated in the protein hinge region, suggesting that the energy needed to change the relative orientation of domains might not be very great. Such low-energy barriers for hinge bending have also been suggested by other early experimental[7] and computational[19] studies. In fact, it was already known in the 1960s that the low-energy barrier for hinge bending[20] that allows rotations of 180° or more in the case of immunoglobulin G[21] is in contrast to the presumably high barrier in hard linkers.[22]

Macromolecular motions encompass hinge bending as well as other types of molecular flexibility. McCammon and Harvey described hinge motions and compared this to general "single-strand motions" that involve local denaturation of the polypeptide chain, unfolding of regions of the polypeptide chain on the protein surface, and helix–coil transitions.[19] Later, in a number of structural surveys, Gerstein and coworkers further classified examples of hinge bending.[9,23–25] Their findings suggest that frequently occurring natural polypeptide linkers might be good candidates for designing soft hinge-type connectors in engineering applications. Other less frequently observed motions may be attributed to shear-like gliding at domain interfaces and denaturing or irregular folding.[9,23–25] Such complex motions frequently involve larger interfaces than provided for by the polypeptide linkers and are outside the scope of this review.

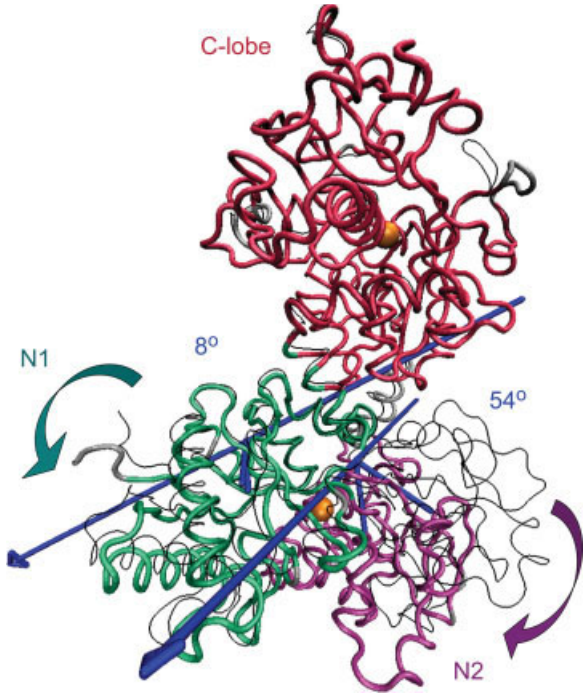

FIGURE 1 Domain motions in lactoferrin, visualized by the *Hingefind* program.[49] Backbone trace of iron-bound lactoferrin (PDB entry 1LFG) presented in color: green, lobe N1; purple, lobe N2; red, Globe; gray, unclassified remainder[49]; orange: iron ions. The iron-free form (PDB entry 1LFH) of lactoferrin is shown in black. The reference domain (C-lobe) is structurally intact, even in the absence of iron. The mobile domains N1 and N2, as identified by *Hingefind,* are considered to move as rigid bodies. The axis of rotation (blue, left) of lobe N1 passes through the interface to the C-lobe, representing a $\beta = 8°$ (Figure 2) twist motion about the axis relative to the C-lobe. In contrast, the hinge rotation of lobe N2 by $\beta = 54°$ relative to the C-lobe (blue axis, right) is significantly more pronounced than the twist motion of N1. The differences in domain movements in the C- and N-lobes have been interpreted as effects of crystal packing forces. The V-shaped blue lines point from the hinge axes to the centroids of the rotating domains to represent the magnitude of the reorientation. The arrow heads (blue) indicate the chirality (the rotations are right-handed in the direction of the arrows).

DISCOVERY OF GEOMETRIC AND BIOPHYSICAL PROPERTIES OF LINKERS

One important consequence of the flexibility afforded by soft peptide linkers is the ability of linked domains to move to and from close spatial proximity. For example, in diphtheria toxin (DT), the entire 15 kDa "R" domain rotates by 180° from a detached, open dimeric form to a closed monomeric form by changing the main-chain conformations of loop 380–386 only.[26] Such large-scale rearrangement results in a close packing of the detached "R" domain into the cleft between the "C" and the "T" domains. Similarly, domains in tomato bushy stunt virus[27] and lactoferrin (Figure 1) form close contacts upon a structural collapse that is induced by hinge rotation. A

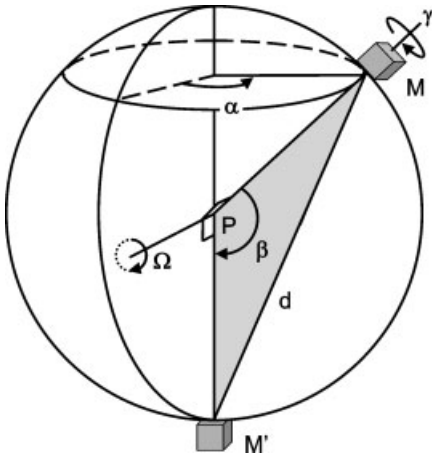

FIGURE 2 Movement between moieties M and M′ attached by a hinge.

number of linker selection studies[28–31] suggested that the flexibility and hydrophilicity of the linker are important factors in preventing the disturbance of the domain functions, thereby imparting stability to the domains.[32] A range of stability occurs depending on the rigidity of the linker peptides.[33] Soft linkers confer flexibility, whereas more rigid peptides may act to keep domains apart.[33]

Due to their ability to break or form contacts among adjacent domains, soft linkers often facilitate essential catalytic events in the overall function of a protein, as seen in the packing of tomato bushy stunt virus proteins.[18] NFκB is another example. Its "glycine-rich" hinge region is flexible enough to bring "p50" and "SWI6/ANK" repeat domains into contact to regulate intracellular transport of the transcription factor.[34] Such soft glycine "hinges" can be designed as stretches of amino acids where at least four of every six residues are glycine.[35] Limited proteolysis at high pH has also been observed in the hinge region connecting the shell and the protruding domains of noroviruses although the biological significance of such proteolysis is still not known.[36–38]

Substantial work has been done in characterizing the hinge-based freedom of movement in polypeptides. In Figure 2 we present a geometric approach in describing hinge motion. In a fully flexible hinge, the distance d between the rigid arms (attached to the pivot point P) changes as a function of hinge "latitude" β. For extended moieties, distances also change as a function of "longitude" α and twist angle γ. As the linker allows changes in the hinge angle β, the arm M rotates through space altering the distance between it and arm M′. In section 4 we will describe that $\langle d \rangle$, the thermally accessible end-to-end distance distribution between the two arms, varies with the softness of the hinge region, a property that can be exploited in the mechanical parameterization of hinges.

Prediction of the softness of a peptide linker is based on an understanding of the rotational freedom of the residues involved. In the 1960s Schimmel et al. analyzed the effects of restricted rotation in polyproline and compared this to other polymers with fewer steric restrictions about the peptide φ and ψ bonds.[39] By summing the rotational potentials of the individual bonds in the polymer, they created a predictable model of chain softness as a function of chain dimensions for polyproline. The authors found that within a certain distance (up to around 40 residues) the rotational hindrance potential of the individual residues forced the chain end-to-end separation to increase directly with the number of residues at constantly increasing rate. This is diagrammed in Figure 3.

The value $\langle r^2 \rangle_0 / x l_u^2$ (also known as the characteristic ratio) in Figure 3 represents the distance maintained by a polymer before it begins to change direction. Proline homopolymers provide a predictable end-to-end distance over a length of more than 100 residues.[39] Conversely, glycine homopolymers shown in Figure 3 do not maintain a direction past more than a few residues (i.e., they are flexible). This property of glycine-rich regions is evident in many natural systems. For example, the NFκB glycine-rich hinge allows one terminus to "fold back" on to the other.[34]

The fold-back property of polyglycine has been shown empirically in pulse-radiolysis experiments.[40] In these experiments an electron donor was separated from an electron acceptor by either a proline bridge or glycine bridge of 0–3 residues. Bobrowski et al. found the kinetic constant of intramolecular electron transfer correlated with the length of the proline bridge (a factor of 3.5 per each proline residue added). However, the kinetic constant did not correlate with the length of the glycine bridge. This suggests that the moieties attached

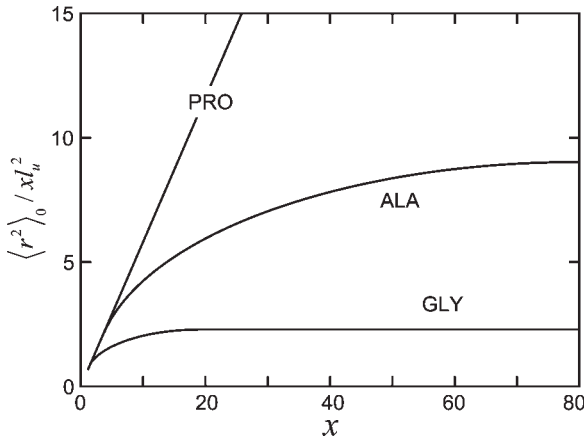

FIGURE 3 Ratio $\langle r^2 \rangle_0 / x l_u^2$ plotted against units, x, of homopolymers glycine, alanine, and proline (after Ref. 39).

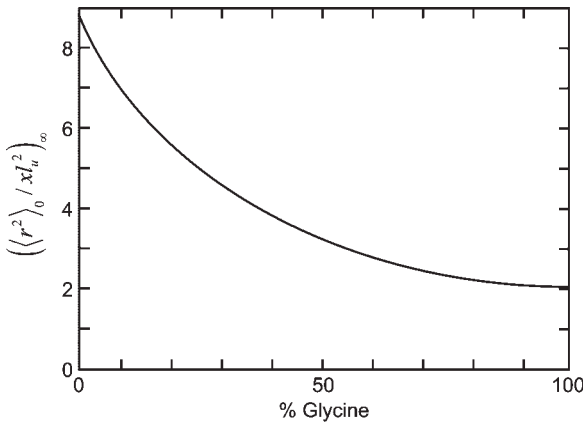

FIGURE 4 Effect of randomly distributed glycine residues on polypeptide flexibility (after Ref. 109).

to the glycine bridge must transfer energy via direct collision (rather than electron transfer through the peptide backbone). In other words, the polyglycine bridge folds back, allowing the electron donor to come into direct contact with the electron acceptor attached to the opposite end of molecule. This result is consistent with Schimmel's values for oligoglycine shown in Figure 3.

In the case of random polypeptides (Figure 4), the value of $\langle r^2\rangle_0/xl_u^2$ increases with the number of peptide residues (Figure 3). However, this value approaches the limit designated $(\langle r^2\rangle_0/xl_u^2)_\infty$, typically a value of 9.0 for random-coil polypeptides. As shown in Figure 4, even a small percentage of randomly distributed glycine residues have a large effect on the chain dimensions of random peptides. A polypeptide region with four out of six randomly distributed glycine residues (66.6% glycine) has a $(\langle r^2\rangle_0/xl_u^2)_\infty$value of less than 2.5. Thus, a polypeptide region with four out of six randomly distributed glycine residues would make a very effective hinge region. This observation provides a simplified method for identifying glycine hinges based on a polypeptide sequence. Mancebo et al. used this as a guide to identify glycine hinge regions in the predicted sequences of U1 70K snRNPs.[35] Hinge regions identified by polypeptide sequence are also found by Zahler et al.[41] Today, much more is known about detailed sequence-dependent properties of flexible peptide linkers. We will return to this in the section, Current Understanding of Sequence-Dependent Linker Properties, after discussing the control of structural softness in linker peptides.

ABSENCE OF FLEXIBILITY: MOLECULAR RULERS

As Schimmel et al.[39] pointed out, ordered chains, such as helices, are restricted throughout and can provide an increase in $\langle r^2\rangle_0/xl_u^2$ that rises without bound with increasing number of residues. The results of their polyproline calculations demonstrate that the chain dimensions should be extremely sensitive to relatively small changes in the rotational potential. The large influence of the rotational hindrance potential on the chain dimensions is due to the fact that the rotations are restricted, sterically, to a small domain. In other words, the ability of the polymer to extend at a particular distance is directly related to the physical restrictions about the Cα—C bond. This demonstrates how minor physical features, which prevent rotational freedom in each amino acid residue throughout the chain, control the overall chain dimensions. When all rotation angles are fixed at the same value (i.e., for a helix), $\langle r^2\rangle_0/xl_u^2$ rises without bound with increasing number of residues.

It is expected in protein modeling that a repeating oligopeptide with each monomer having the same stable relationship to its predecessor will generate a helix.[42] An important physical property of relatively hard linkers, such as structurally stable helices, is the ability to reliably predict its end-to-end distances. Such polypeptides have been termed "molecular rulers" because of their ability to measure the effect of different depths of binding pockets on various properties of proteins of interest.[43] Consequently, Stryer and Haugland used a poly-L-proline "spectroscopic ruler" to measure distances of energy transfer between attached moieties. Poly-L-proline peptides were of defined length to separate an energy donor and acceptor by distances ranging from 12 to 46 Å.[22]

Although short stretches of hard linker sequences are located between functionally relevant regions of protein structure, mutations within such sequences may have no effect on the function.[44] Such linkers are therefore necessary to keep the other amino acid interactions in register, but the nature of the side chain is often unimportant. The question that arises is then, How is the length of the ruler controlled? Many molecular rulers are made up of repeating molecules of distinct, rigid monomer units. Because of their rigidity, predictable changes in length are observed (up to a limiting point) as additional monomers are added between functional groups.[43] For instance, in electron transfer experiments, proline peptides provide predictable separation distance based on the number of proline residues.[45]

The major concern in the design of a molecular ruler is the possibility of softening and structural failure that arises when the ruler is unable to provide a predictable separation distance between its bound moieties. An adequate cushion distance is often required when designing the linkers. In affinity chro-

matography, structurally stable linkers between the solid matrix and ligand keep the polymer-bound ligand sufficiently distant from the polymer surface.[33] Examples of such linkers include polylysine, and multipoly-DL-alaninepolylysine. Similarly, hard linkers are often used to conjugate antigenic moieties to carriers in biomedical applications.

The question that often arises is whether the control of peptide flexibility through molecular rulers is purely an engineering enterprise or whether functionally active molecular rulers are found in nature. Recently, a naturally functioning molecular ruler was discovered in bacterial transcription elongation factors GreA and GreB that promote transcription elongation by stimulating an endogenous, endonucleolytic transcript cleavage activity of the RNA polymerase. The structure of *Escherichia coli* core RNA polymerase bound to GreB was determined using cryoelectron microscopy and image processing of helical crystals to a nominal resolution of 15 Å, allowing a fitting of high-resolution RNA polymerase and GreB structures.[46] In the resulting model, the GreB N-terminal coiled-coil domain extends 45 Å through a channel directly to the RNA polymerase active site. The model leads to detailed insights into the mechanism of Gre factor activity especially in reference to its behavior as a molecular ruler. GreB measures the length of any backtracked RNA and cleaves it precisely such that the arrested RNA polymerase can resume its forward processing of DNA after the release of the excised polynucleotide. In transcription, the molecular ruler model may explain a wide range of experimental observations.[46]

COMPUTATIONAL DETECTION OF DOMAINS, LINKERS, AND THEIR MOVEMENTS

In the mid-1990s computer algorithms were developed that characterized the geometric properties of hinge regions quantitatively and that allowed the prediction and classification of linker peptides. These techniques are nowadays routinely used in databases of domain motions of biopolymers and macromolecular assemblies. The algorithms and databases have reshaped our understanding of biomolecular dynamics especially because large-scale structural heterogeneity of proteins, facilitated by linker peptides, is now seen as the norm rather than the exception.

Historically, the early structures of immunoglobulins, lysozyme and alcohol dehydrogenase, along with the structures of their mutants and ligand complexes, clearly indicated that domain motions are one of the critical links between the structure and the function of macromolecules.[6] With an increasing number of protein structures showing domain movements, computational methods were reported that automatically detected the domains, their connectivity, and the axes of motion from available structures. Procedures for optimizing the least-squares superposition of protein domains by excluding poor-fitting residues have been used for a long time.[47] Lesk[48] has formulated a sieving routine that minimizes the root mean square deviation of a domain by subsequent elimination of atoms that lie far apart in the superposition. Because the method intrinsically assumed the existence of only one domain, parts of the molecule that did not belong to the rigid core were assumed flexible. The method, therefore, could not directly detect multiple flexible domains within the same protein molecule. Using a variant of this method in which even the apparently flexible regions were sieved after identifying the rigid core, Wriggers and Schulten[49] were able to detect domains in a number of multidomain structures. Alternative approaches that did not require rigid cores in defining domains have also been reported. A method based on the clustering of interresidue vectors and their rotational properties correctly identified domain movements in T4 lysozyme and citrate synthase.[50] An Ising model approach, in which the structural elements of the model changed states according to the state of their neighbors, successfully detected domains in multidomain structures using the Cα coordinates.[51] Such domain detection techniques, when applied to initially 24 multiple conformer X-ray structures, have yielded a database of macromolecular motions[23] and valuable insights into the structural elements that are important in domain movements.[52]

Rigid-body movements of the domains can be described by six rigid-body degrees of freedom. The classic screw axis of rigid-body movement[53–55] is a shifted rotation axis but it retains all freedom of the rigid-body model: The mathematical theorem by Chasles[54] states that an optimal position for the rotation axis can be found for which any residual translation vector is parallel to the rotation axis.[55] Any rigid-body movement can then be then described as a helical twist about the axis, accompanied by a helical rise along the axis.[55] However, the hinge-bending motion depicted in Figure 2 requires no such helical rise, because the moieties M and M′ can be brought into register by means of a rotation about the hinge axis alone. In other words, the three angles α, β, and γ are no longer independent for hinge rotations. Therefore, as an alternative to the widely used screw axis, Wriggers and Schulten defined an "effective

rotation axis" Ω for hinge motions that is perpendicular to the legs PM and PM′ (Figure 2). This effective rotation axis approximates the unconstrained rigid-body rotation α, β, and γ, but it affords an exact registration of the moieties M and M′.[49] A computer application of both domain recognition and hinge axis determination is demonstrated in Figure 1. The effective rotation axes (including arrows that indicate chirality and V-shaped angles that illustrate the magnitude of rotation) are shown in blue in Figure 1, whereas the found rigid body domains are shown in red, green, and purple colors.

In general, computationally detected hinge axes[23] appear to pass through only a limited number of structural elements, even though in complex cases the flexible domains may interact among each other through more than a dozen contact regions. Most protein domains, however, interact mainly through two linker regions, thereby giving rise to the "double-hinged" type of domain movements.[52] Domain movements, in general, appear to be spread uniformly over the interdomain contact regions. In the case of a large number of contact regions between the domains, domain movements are a result of small individual conformational changes spread over all the contact areas.[56] In contrast, domain movements in molecules having relatively few interdomain contacts arise from larger but localized conformational changes within polypeptide linker regions.

Although X-ray structures have yielded a wealth of information about the structural elements that participate in domain motions, details of such motion cannot be deduced from such structures alone as most X-ray structures are known only in single conformations. Computer simulations offer an alternative method to study domain motions in the absence of additional experimental isoforms. Normal mode analyses and molecular dynamics (MD) simulations of the crystal structures are providing valuable insights into the mechanisms of domain flexibility. For example, normal mode analysis of the biomolecular motor Fl-ATPase has revealed the predominance of low-frequency modes that results in the inherent flexibility of the α, β, and γ subunits in isolation and as part of the functioning $\alpha_3\beta_3\gamma_3$ complex that is in accord with the motor function.[57] A similar analysis has also been reported for microtubule components,[58] Trp repressor, calmodulin, calbindin, HIV-1 protease, and troponin C[59] in which localized large conformational changes have been shown to arise due to collective normal mode oscillations of the entire protein at specific frequencies. MD simulations of acetylcholinesterase,[60] DNA binding LacI and PurR proteins,[61] integrins,[62] calmodulin,[63,64] metalloproteins,[65] and the folding of the SH3 domain[66] clearly indicate the importance of domain flexibility in the functioning of these systems. With the rapidly increasing number of structures becoming available due to structural genomics initiatives, automated methods to identify domains and linkers[67,68] are expected to become increasingly important in domain flexibility studies.

Theoretical techniques for detecting and classifying domain motion can also be used in conjunction with biophysical techniques to parameterize mechanical models of flexible systems. In cryoelectron tomography experiments, one may take several snapshots of a given biological macromolecule.[69] In principle, a large enough collection of snapshots of the molecule may then be used to calculate its equilibrium configuration in terms of the experimentally accessible degrees of freedom, and hence to estimate its potential energy[69] and distribution of conformational states at room temperature. Skoglund et al. recently analyzed the results of cryoelectron tomography experiments performed on monoclonal murine immunoglobulin antibodies. The authors introduced a statistical description of the immunoglobulin configuration, which yielded the probability distributions of the Fab–Fab and Fab–Fc hinge angles. The final mechanical model allowed them to calculate quantitative estimates of the relevant frequencies of the immunoglobulin hinge-bending motion.[69]

CURRENT UNDERSTANDING OF SEQUENCE-DEPENDENT LINKER PROPERTIES

As described above, it has been well known for decades that the softness and the length of linkers affect protein stability, folding rates, and domain–domain interactions.[70–72] The detailed amino acid propensity and the preferred geometry of naturally occurring linkers have been described in a recent survey,[73] which is summarized here.

Linker residues were determined from the existing structural domain databases using a domain identification method modified from Ref. 51. The length of the linkers varies between 2 and 18 residues with an average linker length of 5.15 residues, a value that is much higher than was earlier believed.[52] The linkers are predominantly α-helical with progressively decreasing occurrences in coiled, β-strand, or turn conformation. Both the helical and nonhelical linkers have similar hydrophobicity. Pro is the most common terminal linker residue followed by Arg, Phe, Thr, Glu, and Gln in decreasing order of preference. The probable reason why proline is favored over other

residues in linking different domains is the inability of proline to donate hydrogen bonds or participate comfortably in any regular secondary structure conformation. This ensures a relatively rigid separation of the domains, thereby preventing unfavorable contacts between them.[73] That most of the proline residues are in the *trans* conformation further helps maintain fairly rigid interdomain separation. However, the main-chain conformation around proline is neighbor dependent, and there are cases where the neighbors of proline favor *cis–trans* isomerization, thereby making the linkers more flexible.

The maintenance and curation of linker databases[73] is facilitated by computational methods that automatically detect domain linkers.[67,74] Structural domain databases,[75,76] the domain linker database,[52] recent analyses of individual domains,[10,77] and domain searches[78] from sequence homology based domain databases[76,79–82] all are useful in designing linkers. The available empirical data is of particular importance for the design of multifunctional chimeric domain assemblies.

DESIGN OF CHIMERIC PROTEINS WITH ENGINEERED DOMAINS AND LINKERS

Nowadays, gene fusion techniques are indispensable tools in a variety of biochemical research areas.[32] Recombinant chimeric fusion proteins are routinely constructed to increase the expression of soluble proteins and to facilitate protein purification.[32,83] Other engineering approaches that link two proteins or protein domains by a peptide linker include immunoassays (e.g., using chimeras between antibody fragments and proteins[84,85]), selection and production of antibodies,[86] and engineering of bifunctional enzymes.[87]

In the respiratory chain, electron transfer protein domains of flavodoxin and cytochrome c553 from *Desulfovibrio vulgaris* and the heme domain of P450 BM3 from *Bacillus megaterium* have been used as molecular "Lego"-type building blocks in different combinations to build artificial redox chains having variable redox potentials.[88] Multidomain bacterial protein toxins have been used in designing potential carriers for targeted delivery of biomolecules.[89] Catalytically functional flavocytochrome chimeras[90] and modified cellular signaling circuits through modular recombination of domains[91] are some of the other recently reported chimeric proteins having additional functions introduced into them through engineered domains and linkers.

The selection of the linker sequence is particularly important for the construction of functional chimeric proteins.[32] A recent study on the streptococcal protein G-Vargula luciferase chimera suggested that the spatial separation of the heterofunctional domains of a chimeric protein by an appropriate linker peptide is important for the domains to work independently.[92] In a similar study, Arai et al. designed linkers to effectively separate the two domains of a chimeric protein.[83] The authors introduced helix-forming peptide linkers, $(EAAAK)_n$, between two green fluorescent protein (GFP) variants. Circular dichroism (CD) spectroscopic analysis suggested that the introduced linkers form an α-helix, and that the α-helical contents increase as the lengths of the linkers increase. Fluorescence resonance energy transfer (FRET) observed between the two GFP variants also suggested that the distance between the two GFP domains increases as the lengths of the linkers' increase.

It is difficult to determine the orientation of the domains and the conformation of the linker of the chimeric proteins by FRET and CD analyses alone. Therefore, a direct visualization of the three-dimensional (3D) structures in solution is desirable. To gain information about the general shape, it is not necessary to solve structures at atomic resolution. Single particle cryo-EM routinely yields low-resolution (10–30 Å) shapes of particles larger than about 300 kDa molecular weight.[93] Synchrotron X-ray small-angle scattering (SAXS) has emerged in recent years as a complementary low-resolution alternative for the investigation of smaller chimeric proteins. Computational techniques[94–97] allow one to deduce the 3D shape from one-dimensional (1D) SAXS scattering profiles that correspond to the isotropically averaged reflections of X-rays on the randomly oriented particles in solution. In Ref. 32, the shapes and sizes of the chimeric proteins, consisting of GFP variants with the helical linkers $(EAAAK)_n$ ($n = 2$–5) and flexible linkers $(GGGGS)_n$ ($n = 3$, 4), were deduced from the SAXS diffraction pattern with an ab initio modeling procedure. Figure 5 shows two of the resulting models.

The SAXS experiments demonstrated that short helical linkers ($n = 2$, 3) cause multimerization, while the longer linkers ($n = 4$, 5) solvate monomeric chimeric proteins.[32] Also, chimeric proteins with a helical linker assumed a more elongated conformation compared to those with a flexible linker. The elongation depends on the length of the helical linker element in agreement with the molecular ruler hypothesis (Absence of Flexibility section). The chimeric proteins with the flexible linker exhibited an elongated conformation as well, rather than the most compact side-by-side conformation expected from

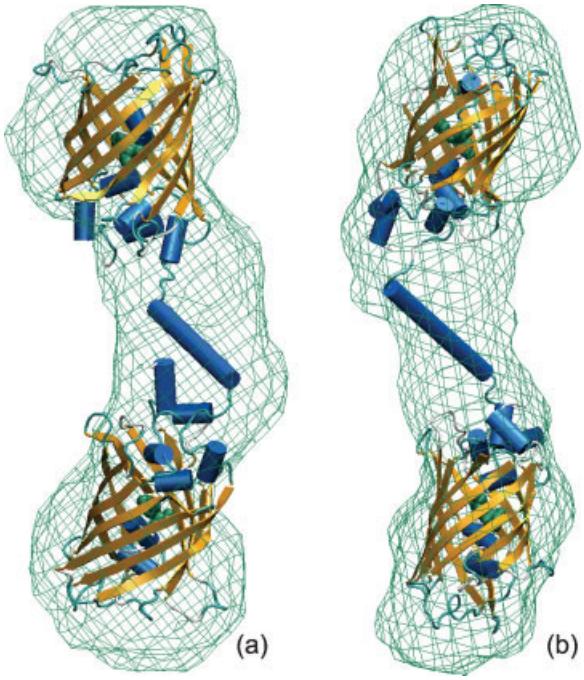

FIGURE 5 Modeling of atomic structures into 3D shapes from SAXS.[32,97] Variants of green and blue fluorescent protein[32] connected by (a) $(EAAAK)_4$ and (b) $(EAAAK)_5$ linkers (cartoon representation). The SAXS envelopes and fluorophores are shown in green. The β-sheets are shown in yellow and α-helices in blue.

FRET analysis. Information about the global conformation of the chimeric protein is thus necessary for optimization of the linker design.[32]

Linker engineering, with the aim to control the distance, orientation, and relative motion of two functional domains, will increase in importance with increasing emphasis on the de novo design of multidomain proteins. A number of recent databases and surveys aid in the design of linkers for chimeric proteins.[73] However, despite many empirical surveys, very little is known about the structural factors that govern interdomain flexibility. Such lack of knowledge is a limiting factor in de novo chimera design. Therefore, a number of recent studies focused on the structural principles governing the domain architecture and their assembly.[98–107] The emerging concepts, along with the bioinformatics tools that attempt to detect domains and their motions from sequence information alone,[108] may one day lead to a precise de novo engineering of interdomain flexibility, thereby helping achieve the desired functioning of synthetic chimeras. In the mean time, the successful use of feedback techniques such as CD, FRET, and SAXS suggests that gene fusion applications should be accompanied by geometric analysis for appropriate biophysical validation of the linker design process.

We thank Tetsuro Fujisawa for kindly providing the SAXS data of chimeric GFP proteins.

REFERENCES

1. Gokhale, R. S.; Khosla, C. Curr Opin Chem Biol 2000, 4, 22–27.
2. Rossmann, M. G.; Adams, M. J.; Buehner, M.; Ford, G. C.; Hackert, H. L.; Lentz, P. J., Jr.; McPherson, A., Jr.; Schevitz, R. W.; Smiley, I. E. Cold Spring Harbor Symp Quant Biol 1971, 36, 179–191.
3. White, J. L.; Hackert, M. L.; Buehner, M.; Adams, M. J.; Ford, G. C.; Lentz, P. J., Jr.; Smiley, I. E.; Steindel, S. J.; Rossmann, M. G. J Mol Biol 1976, 102, 759–779.
4. Phillips, D. C.; Rivers, P. S.; Sternberg, M. J. E.; Thornton, J. M.; Wilson, I. A. Biochem Soc Trans 1977, 5, 642–647.
5. Perutz, M. F. Quart Rev Biophys 1989, 22, 139–236.
6. Richardson, J. S. Adv Protein Chem 1981, 34, 167–339.
7. Dobson, C. M. Nature 1990, 348, 198–199.
8. Gerstein, M.; Anderson, B. F.; Norris, G. E.; Baker, E. N.; Lesk, A. M.; Chothia, C. J Mol Biol 1993, 234, 357–372.
9. Gerstein, M.; Lesk, A. M.; Chothia, C. Biochemistry 1994, 33, 6739–6749.
10. Apic, G.; Huber, W.; Teichmann, S. A. J Struct Funct Genomics 2003, 4, 67–78.
11. Janin, J.; Wodak, S. J. Prog Biophys Mol Biol 1983, 42, 21–78.
12. Janin, J.; Chothia, C. Methods Enzymol 1985, 115, 420–430.
13. Rose, G. D. J Mol Biol 1979, 134, 447–470.
14. Swindells, M. B. Protein Sci 1995, 4, 93–102.
15. Swindells, M. B. Protein Sci 1995, 4, 103–112.
16. Sowdhamini, R.; Rufino, S. D.; Blundell, T. L. Fold Des 1996, 1, 209–220.
17. Islam, S. A.; Luo, J.; Sternberg, M. J. Protein Eng 1995, 8, 513–525.
18. Harrison, S. C. Trends Biochem Sci 1978, 3, 3–7.
19. McCammon, J. A.; Harvey, S. C. Dynamics of Proteins and Nucleic Acids; Cambridge University Press: Cambridge, 1987.
20. Green, N. M. Adv Immunol 1969, 11, 1–30.
21. Wrigley, N. G.; Brown, E. B.; Skehel, J. J. J Mol Biol 1983, 169, 771–774.
22. Stryer, L.; Haugland, R. P. Proc Natl Acad Sci USA 1967, 58, 719–726.
23. Gerstein, M.; Krebs, W. Nucleic Acids Res 1998, 26, 4280–4290.
24. Krebs, W. G.; Gerstein, M. Nucleic Acids Res 2000, 28(8), 1665–1675.
25. Echols, N.; Milburn, D.; Gerstein, M. Nucleic Acids Res 2003, 31, 478–482.
26. Bennett, M. J.; Choe, S.; Eisenberg, D. Proc Natl Acad Sci USA 1994, 91, 3127–3131.

27. Winkler, F. K.; Schutt, C. E.; Harrison, S. C.; Bricogne, G. Nature 1977, 265, 509–513.
28. Alfthan, K.; Takkinen, K.; Sizmann, D.; Soderlund, H.; Teeri, T. T. Protein Eng 1995, 8, 725–731.
29. Argos, P. J Mol Biol 1990, 211, 943–958.
30. Crasto, J. C.; Feng, J. Protein Eng 2000, 13, 309–312.
31. Robinson, C. R.; Sauer, R. T. Proc Natl Acad Sci USA 1998, 95, 5929–5934.
32. Arai, R.; Wriggers, W.; Nishikawa, Y.; Nagamune, T.; Fujisawa, T. Proteins Struct Funct Bioinformatics 2004, 57, 829–838.
33. Wilchek, M.; Miron, T. Methods Enzymol 1974, 34, 72–76.
34. Henkel, T.; Zabel, U.; van Zee, K.; Müller, J. M.; Fanning, E.; Baeuerle, P. A. Cell 1992, 68, 1121–1133.
35. Mancebo, R.; Lo, P. C.; Mount, S. M. Mol Cell Biol 1990, 10, 2492–2502.
36. Hardy, M.; White, L.; Ball, J.; Estes, M. J Virol 1995, 69, 1693–1698.
37. Prasad, B. V.; Hardy, M. E.; Dokland, T.; Bella, J.; Rossmann, M. G.; Estes, M. K. Science 1999, 286, 287–290.
38. Chakravarty, S.; Hutson, A. M.; Estes, M. K.; Prasad, B. V. V. J Virol 2005, 79, 554–568.
39. Schimmel, P. R.; Flory, P. J. Proc Natl Acad Sci USA 1967, 58, 52–59.
40. Bobrowski, K.; Wierzchowski, K. L.; Holcman, J.; Ciurak, M. Int J Radiat Biol 1990, 57, 919–932.
41. Zahler, A. M.; Lane, W. S.; Stolk, J. A.; Roth, M. B. Genes Dev 1992, 6, 837–847.
42. Matsushima, N.; Creutz, C. E.; Kretsinger, R. H. Proteins Struct Funct Bioinformatics 1990, 7, 125–155.
43. Johnson, J. L.; Cusack, B.; Hughes, T. F.; McCullough, E. H.; Fauq, A.; Romanovskis, P.; Spatola, A. F.; Rosenberry, T. L. J Biol Chem 2003, 278, 38948–38955.
44. Bottema, C. D.; Ketterling, R. P.; Ii, S.; Yoon, H. S.; Phillips, J. A., 3rd; Sommer, S. S.; Am J Hum Genet 1991, 49, 820–838.
45. Isied, S. S.; Ogawa, M. Y.; Wishart, J. F. Chem Rev 1992, 92, 381–394.
46. Opalka, N.; Chlenov, M.; Chacón, P.; Rice, W. J.; Wriggers, W.; Darst, S. Cell 2003, 114, 335–345.
47. Freer, S. T.; Kraut, J.; Robertus, J. D.; Wright, H. T.; Xuong, Ng. H. Biochemistry 1970, 9, 1997–2009.
48. Lesk, A. M. Protein Architecture; Oxford University Press: New York, 1991.
49. Wriggers, W.; Schulten, K. Proteins Struct Funct Bioinformatics 1997, 29(1), 1–14.
50. Hayward, S.; Berendsen, H. J. C. Proteins Struct Funct Bioinformatics 1998, 30, 144–154.
51. Taylor, W. R. Protein Eng 1999, 12, 203–216.
52. Hayward, S. Proteins Struct Funct Bioinformatics 36, 425–435, 1999.
53. Goldstein, H. Classical Mechanics; Addison-Wesley: Reading, MA, 1980.
54. Chasles, M. Férusac, Bull Sci Math 1830, 14, 321–326.
55. Babcock, M. S.; Pednault, E. P. D.; Olson, W. K. J Mol Biol 1994, 237, 125–156.
56. Gerstein, M.; Chothia, C. J Mol Biol 1991, 20, 133–149.
57. Cui, Q.; Li, G.; Ma, J.; Karplus, M. J Mol Biol 2004, 340, 342–372.
58. Keskin, O.; Durell, S. R.; Bahar, I.; Jernigan, R. L.; Covell, D. G. Biophys J 2002, 83, 663–680.
59. Cao, Z. W.; Chen, X.; Chen, Y. Z. J Mol Graph Model 2002, 21, 309–319.
60. Shen, T.; Tai, K.; Henchman, R. H.; McCammon, J. A. Acc Chem Res 2002, 35, 332–340.
61. Swint-Kruse, L.; Larson, C.; Pettitt, B. M.; Matthew, K. S. Protein Sci 11, 778–794, 2002.
62. Jin, M.; Andricioaei, I.; Springer, T. A. Structure 2004, 12, 2137–2147.
63. Wriggers, W.; Mehler, E.; Pitici, F.; Weinstein, H.; Schulten, K. Biophys J 1998, 74, 1622–1639.
64. Shepherd, C. M.; Vogel, H. J. Biophys J 2004, 87, 780–791.
65. van den Bosch, M.; Swart, M.; van Gunsteren, W. F.; Canters, G. W. J Mol Biol 2004, 344, 725–738.
66. Ollerenshaw, J. E.; Kaya, H.; Chan, H. S.; Kay, L. E. Proc Natl Acad Sci USA 2004, 101, 14748–14753.
67. Tanaka, T.; Kuroda, Y.; Yokoyama, S. J Struct Funct Genomics 2003, 4, 79–85.
68. Liu, J.; Rost, B. Proteins Struct Funct Bioinformatics 2004, 55, 678–688.
69. Bongini, L.; Fanelli, D.; Piazza, F.; De Los Rios, P.; Sandin, S.; Skoglund, U. Proc Natl Acad Sci USA 2004, 101, 6466–6471.
70. Briggs, S. D.; Smithgall, T. E. J Biol Chem 1999, 274, 26579–26583.
71. Gokhale, R. S.; Tsuji, S. Y.; Cane, D. E.; Khosla, C. Science 1999, 284, 482–485.
72. Ikebe, M.; Kambara, T.; Stafford, W. F.; Sata, M.; Katayama, E.; Ikebe, R. J Biol Chem 1998, 273, 17702–17707.
73. George, R. A.; Heringa, J. Protein Eng 2002, 15, 871–879.
74. Miyazaki, S.; Kuroda, Y.; Yokoyama, S. J Struct Funct Genomics 2002, 2, 37–51.
75. Holm, L.; Sander, C. Proteins Struct Funct Bioinformatics 1998, 33, 88–96.
76. George, R. A.; Spriggs, R. V.; Thornton, J. M.; Al-Lazikani, B.; Swindells, M. B. Bioinformatics 2004, 20(Suppl 1), I130–I136.
77. Gu, J.; Gu, X. Gene 2003, 317, 49–57.
78. Heger, A.; Holm, L. J Mol Biol 2003, 328, 749–767.
79. Corpet, F.; Servant, F.; Gouzy, J.; Kahn, D. Nucleic Acids Res 2000, 28, 267–269.
80. Sonnhammer, E. L.; Eddy, S. R.; Durbin, R. Proteins Struct Funct Bioinformatics 1997, 28, 405–420.
81. Murvai, J.; Vlahovicek, K.; Barta, E.; Cataletto, B.; Pongor, S. Nucleic Acids Res 2000, 28, 260–262.
82. Schultz, J.; Copley, R. R.; Doerks, T.; Ponting, C. P.; Bork, P. Nucleic Acids Res 2000, 28, 231–234.
83. Arai, R.; Ueda, H.; Kitayama, A.; Kamiya, N.; Nagamune, T. Protein Eng 14, 2001, 529–532.

84. Arai, R.; Ueda, H.; Nagamune, T. J Ferment Bioeng 1998, 86, 440–445.
85. Arai, R.; Ueda, H.; Tsumoto, K.; Mahoney, W. C.; Kumagai, I.; Nagamune, T. Protein Eng 2000, 13, 369–376.
86. Bird, R. E.; Hardman, K. D.; Jacobson, J. W.; Johnson, S.; Kaufman, B. M.; Lee, S. M.; Lee, T.; Pope, S. H.; Riordan, G. S.; Whitlow, M. Science 1988, 242, 423–426.
87. Bulow, L. Eur J Biochem 1987, 163, 443–448.
88. Sadeghi, S. J.; Meharenna, Y. T.; Fantuzzi, A.; Valetti, F.; Gilardi, G. Faraday Discuss 2000, 116, 135–153; discussion: 171–190.
89. Bade, S.; Rummel, A.; Reisinger, C.; Karnath, T.; Ahnert-Hilger, G.; Bigalke, H.; Binz, T. J Neurochem 2004, 6, 1461–1472.
90. Fuziwara, S.; Sagami, I.; Rozhkova, E.; Craig, D.; Noble, M. A.; Munro, A. W.; Chapman, S. K.; Shimizu, T. J Inorg Biochem 2002, 91, 515–526.
91. Dueber, J. E.; Yeh, B. J.; Chak, K.; Lim, W. A. Science 2003, 301, 1904–1908.
92. Maeda, Y.; Ueda, H.; Kazami, J.; J.; Kawano, G.; Suzuki, E.; Nagamune, T. Anal Biochem 1997, 249, 147–152.
93. Frank, J. Three-Dimensional Electron Microscopy of Macromolecular Assemblies; Academic Press: San Diego, 1996.
94. Chacón, P.; Morán, F.; Díaz, J. F.; Pantos, E.; Andreu, J. M. Biophys J 1998, 74, 2760–2775.
95. Svergun, D. I. Biophys J 1999, 76, 2879–2886.
96. Chacón, P.; Díaz, J. F.; Morán, F.; Andreu, J. M. J Mol Biol 2000, 299, 1289–1302.
97. Wriggers, W.; Chacón, P. J Appl Cryst 2001, 34, 773–776.
98. Vogel, C.; Berzuini, C.; Bashton, M.; Gough, J.; Teichmann, S. A. J Mol Biol 2004, 336, 809–823.
99. Aroul-Selvam, R.; Hubbard, T.; Sasidharan, R. J Mol Biol 2004, 338, 633–641.
100. Selvam, R. A.; Sasidharan, R. Nucleic Acids Res 2004, 32, D193–D195.
101. Bhaduri, A.; Sowdhamini, R. Protein Eng 2003, 16, 881–888.
102. Linder, J .U.; Schultz, J. E. Cell Signal 2003, 12, 1081–1089.
103. van Ham, M.; Hendriks, W. Mol Biol Rep 2003, 30, 69–82.
104. Harmer, N. J.; Chirgadze, D.; Hyun Kim, K.; Pellegrini, L.; Blundell, T. L. Biophys Chem 2003, 100, 545–553.
105. Leys, D.; Basran, J.; Talfournier, F.; Sutcliffe, M. J.; Scrutton, N. S. Nature Struct Biol 2003, 10, 219–225.
106. Sinars, C. R.; Cheung-Flynn, J.; Rimerman, R. A.; Scammell, J. G.; Smith, D. F.; Clardy, J. Proc Natl Acad Sci USA 2003, 100, 868–873.
107. Vassylyev, D. G.; Sekine, S.; Laptenko, O.; Lee, J.; Vassylyeva, M. N.; Borukhov, S.; Yokoyama, S. Nature 2002, 417, 712–719.
108. Gouzy, J.; Corpet, F.; Kahn, D. Comput Chem 1999, 23, 333–340.
109. Miller, W. G.; Brant, D. A.; Flory, P. J. J Mol Biol 1967, 23, 67–80.