Intrinsically Disordered Protein

A. Keith Dunker[1], J. David Lawson[1], Celeste J. Brown[1], Pedro Romero[2], Jeong S. Oh[1], Christopher J. Oldfield[1], Andrew M. Campen[1], Catherine M. Ratliff[1], Kerry W. Hipps[3], Juan Ausio[4], Mark S. Nissen[1], Raymond Reeves[1], ChulHee Kang[1], Charles R. Kissinger[5], Robert W. Bailey[1], Michael D. Griswold[1], Wah Chiu[6], Ethan C. Garner[1], and Zoran Obradovic[7]

[1]School of Molecular Biosciences, Washington State University, Pullman, WA 99164-4660

[2]School of Electrical Engineering and Computer Science, Washington State University, Pullman, WA 99164;

[3]Department of Chemistry, Washington State University, Pullman, WA 99164;

[4]Department of Biochemistry, University of Victoria, Victoria, B. C. V8W 3P6;

[5]Pfizer Global Research & Development, La Jolla 11099 North Torrey Pines Road La Jolla, CA 92037;

[6]National Center for Macromolecular Imaging, Baylor College of Medicine, One Baylor Plaza, Houston, TX 77030;

[7]Center for Information Science and Technology, Temple University, Philadelphia, PA 19122, USA

**The information in this paper was presented at a meeting on Protein Flexibility and Folding held in Traverse City, Michigan, from August 13-17, 2000. This review paper will appear in a special issue of Journal of Molecular Graphics in 2001.*

Abstract

Proteins can exist in a trinity of structures: the ordered state, the molten globule and the random coil. Five examples follow which suggest that native protein structure can correspond to any of the three states (not just the ordered state) and that protein function can arise from any of the three states and their transitions. 1. In a process that likely mimics infection, fd phage converts from the ordered into the disordered molten globular state. 2. Nucleosome hyperacetylation is crucial to DNA replication and transcription; this chemical modification greatly increases the net negative charge of the nucleosome core particle. We propose that the increased charge imbalance promotes its conversion to a much less rigid form. 3. Clusterin contains an ordered domain and also a native molten globular region. The molten globular domain likely functions as a proteinaceous detergent for cell remodeling and removal of apoptotic debris. 4. In a critical signaling event, a helix in calcineurin becomes bound and surrounded by calmodulin, thereby turning on calcineurin's serine/threonine phosphatase activity. Locating the calcineurin helix within a region of disorder is essential for enabling calmodulin to surround its target upon binding. 5. Calsequestrin regulates calcium levels in the sarcoplasmic reticulum by binding about 50 ions/molecule. Disordered polyanion tails at the carboxy terminus bind many of these calcium ions, perhaps without adopting a unique structure. In addition to these examples, 16 more proteins with native disorder will be discussed. These disordered regions include molecular recognition domains, protein folding inhibitors, flexible linkers, entropic springs, entropic clocks and entropic bristles. Motivated by such examples of intrinsic disorder, we are studying the relationships between amino acid sequence and order/disorder, and from this information we are predicting intrinsic order/disorder from amino acid sequence. The sequence/structure relationships indicate that disorder is an encoded property, and the predictions strongly suggest that proteins in nature are much richer in intrinsic disorder than are those in the Protein Data Bank. Recent predictions on 29 genomes indicate that proteins from eucaryotes apparently have more intrinsic disorder than those from either bacteria or archaea, with typically > 30 % of eucaryotic proteins having disordered regions of length = 50 consecutive residues.

Keywords: disordered protein, molten globules, random coils, phase changes, structure, function, genomics, proteomics

Introduction

The 20th Century saw the development and solidification of the protein structure/function paradigm, which can be represented in brief as

Amino Acid Sequence ® 3 Dimensional Structure ® Protein Function.

The key point of this paradigm is that 3D structure is taken to be an *obligatory* prerequisite for protein function, so native protein structure equates with ordered 3D structure. In this introduction, a brief history of this paradigm is presented, followed by information that is leading to a call for its reassessment.

Brief history of the current protein structure/function paradigm: Extracellular extracts of beer yeast (containing invertase) hydrolyzed α-glucosides but not β-glucosides, while emuslin hydrolyzed the latter but not the former. From these observations, in 1894 Fischer[1] wrote (as translated in[2]) "To use a picture, I would like to say that enzyme and glucoside have to fit to each other like a lock and key in order to exert a chemical effect on each other."

In 1936, Mirsky and Pauling[3] compiled the following information: loss of pepsin activity correlated with the amount of protein denatured[4]; acid, alkali and urea all increased the viscosity of protein solutions and denatured the proteins without aggregation[5]; many native proteins form characteristic crystals while denatured proteins don't crystallize; exposure of sulfhydryls and other side chain groups and denaturation typically occur concomitantly. From this information and other observations, they concluded, "The characteristic specific properties of native proteins we attribute to their uniquely defined configurations. The denatured protein molecule we consider to be characterized by the absence of a uniquely defined configuration." Slightly earlier than the work of Mirsky and Pauling, Hsien Wu culminated a series of twelve papers on protein denaturation with a thirteenth,[6] which probably contains the first statement[6a] that protein function depends on prior structure. Neither Mirsky and Pauling, nor Wu cited the still earlier work of Fischer, whose lock and key proposal provided a very strong argument supporting their main thesis.

The next decade saw increasing numbers of experiments on protein denaturation. These experiments added to the list of proteins supporting the view that function depended on 3D structure.[7] In 1952 Edsall speculated: "Is it reasonable to hope that the endless variety of proteins found in nature, and their extraordinary diverse and specific interactions with one another and with other substances, can be explained on the basis of *a few relatively rigid general patterns*, simply by varying the nature and sequence of the side chains attached to the fundamental repeating pattern?".[8] This idea took firm root and came to underlie almost all of the subsequent work and thinking.[9]

By the time RNase had been renatured *in vitro*[10,11] and by the time atomic resolution structures of myoglobin[12] and lysozyme[13] had been determined, a highly-specific 3D structure was already long accepted as the necessary prerequisite for protein function. The subsequent avalanche of more than 12,000 structures[14] has buried alternatives to the current protein structure/function

paradigm, yet most of these structures are similar to each other and have recognizable sequence similarity to only a small fraction of the proteins in nature[15,16,17,18].

Evidence for the lack of generality of the current structure/function paradigm: In 1950, Karush[19] reported that, unlike essentially every other native protein known at the time, serum albumin exhibited a nearly universal capacity for the high-affinity binding of small, hydrophobic, typically anionic molecules. Competitive binding was demonstrated for molecules of very different shapes. Using the same logic as Fischer for the lock and key, Karush inferred that albumin's binding site assumes a large number of configurations of similar energies in equilibrium with each other, and that, upon interacting with a given hydrophobic anion, the best-fitting configuration becomes selected from albumin's structural ensemble. Karush called this phenomenon *configurational adaptability.* Studies on albumin binding to the present support Karush's original proposal of an ensemble of structures, but with some added and interesting twists such as the existence of different conformers with slow rates of conversion, corresponding to slightly less than 15 seconds.[20]

Amino acid side chains differ from one to another but yet all are joined by the same peptide bond within proteins. This indicates the need for conformational changes during protein synthesis in order to accommodate the differently shaped side chains. From this insight, Koshland independently proposed configurational adaptability, but instead named this process *induced fit*.[21]

Configurational adaptability and induced fit are evidently the first suggestions of significant conformational change being responsible for protein function. At the time induced fit was proposed, no mechanism was implied, leaving open whether the process of binding induces a new conformation or whether the process of binding involves selection of the best-fit alternative from an ensemble of structures in equilibrium.

Evidence supporting a significant domain movement upon substrate binding was first presented for the association of glucose with hexokinase in 1978.[22] The bound and free enzyme structures were of different isozymes due to failure to obtain crystals of the same protein with and without ligand. The inferred domain shifts were further supported by large shape changes upon binding in solution as measured by small angle X-ray scattering.[23] This domain shift upon ligand binding was taken as support for the induced-fit hypothesis and used to explain the astonishing differences between ATPases and kinases. The latter have the specific need to prevent water from reacting at the active site; the domain movements were proposed to decrease water accessibility.[24] Thus, the induced-fit hypothesis shifted from a mechanism for binding different substrates to a way of gaining deeper understanding of domain movements upon ligand binding.

More recently, aspartate aminotransferase, aromatic amino acid transfererase, and especially a chimera of the two were shown to be capable of operating on a diverse set of substrates by means of structural accommodations within the binding pocket.[25] More than any of the domain-shifting examples, this protein fits Koshland's original description of induced fit as a means for acting on

a set of related but structurally distinct molecules. However, due to the appropriation of induced fit for explaining domain movements upon binding, the authors suggested *multi-induced fit* as a way to explain their data.

Oxygen binding by hemoglobin and catalysis by aspartate transcarbamoylase are both regulated by the binding of non-substrate ligands. For both proteins, non-substrate-ligand binding is associated with shifts between alternative 3D structures in multisubunit proteins. Several models have been developed to explain this behavior, two of which are symmetric allosterism, by Monod, Wyman and Changeux[26] and sequential allosterism, by Koshland, Nemethy and Filman.[27]

Discovery of intrinsically disordered protein segments: In contrast to the view that function depends strictly on prior 3D structure, or on structural accommodations within a prior 3D structure, or on regulatory shifts between alternate structures, examples were discovered in which non-structured segments of proteins play important roles in protein function. Indeed, more than 20 years ago, at a time when only about 20 protein crystal structures had been determined, some protein segments were discovered to yield no discernable electron density and yet to be essential for function.[28,29] A possible relationship between these studies and Karush's much earlier insights[19] has not been considered until now.

Missing electron density in protein structures can arise from failure to solve the phase problem, from crystal defects,[30] or even from unintentional proteolytic removal during protein purification. However, a common reason for missing electron density is that the unobserved atom, side chain, residue, or region fails to scatter X-rays coherently due to variation in position from one protein to the next, e.g. the unobserved atoms are *disordered*. By now, the literature contains numerous reports of disordered regions that are crucial for function.[31,32,33,34,35]

In 1978, the same year that functional disorder was indicated by X-ray crystallography, NMR revealed the highly charged, functional tail of histone H5 to be disordered.[36] Since then, NMR 3D structural determination has led to the characterization of several proteins containing functional, yet disordered regions.[37,38] A further, surprising result of NMR protein structure determinations was the discovery of functional proteins that apparently lack any discernable structure, e.g. apparently native proteins that are disordered from end-to-end.[39,40,41] Because NMR is more certain in its characterization of disorder than is X-ray diffraction, the rediscovery of native disorder by NMR had significant impact.[42,43]

The two-state model of protein folding: Denaturation by urea or guanidine can cause globular proteins to unfold from their compact shapes into extended forms often called *random coils* in deference to earlier work in polymer chemistry. Protein random coils differ from true random coils in important respects, such as a non-random population of internal bond angles along the chain.

A variety of data supported a two-state model for protein folding. In such models, denaturation, involves just the ordered and random coil states without significant populations of any intermediate forms.[8,11]

Partially folded intermediates: In contrast to the two-state view of protein folding and denaturation, partially-unfolded intermediates between the ordered state and the random coil have been observed as the major species in urea, guanidine, and pH titration studies for several (but not all) proteins.[44,45,46] These folding intermediates exhibit side chains with motional characteristics more like those of the random coil but with back-bone secondary structure more like that of the ordered state. A surprise at the time was the discovery that these folding intermediates were compact, with only slight expansion compared to their ordered states, as initially determined by intrinsic viscosity,[47] and as later confirmed by dynamic light scattering,[48,49] and small angle X-ray scattering.[49,50]

Ptitsyn and co-workers provided an interpretation for the partially unfolded state.[47] In their model, the protein converts from an ordered (they used the term *native*) state into a form having some liquid-like characteristics, for example with the side chains going from rigid to non-rigid packing, while the secondary structure remains almost unchanged and the shape remains compact. Furthermore, Ptitsyn coined the term *molten globule*[51] to describe his model for the liquid-like, partially folded state.

There has been considerable uncertainty regarding the molten globule hypothesis.[52] For example, it is uncertain whether heat, guanidine, urea, and pH extremes induce the same folding intermediate as was originally claimed.[47] For bulk liquids the properties are uniform, and, except for rare examples like super-fluidity that depend on macroscopic quantum states, transitions from one type of liquid to a second type of liquid are not observed. In contrast to the behavior of liquids, partially folded α-lactalbumin[53] displays a series of related forms and apomyoglobin[54] reveals two distinct partially-folded intermediates that appear sequentially during refolding. Some of the amide groups in partially folded α-lactalbumin show substantial protection against hydrogen exchange.[55] Indeed, as reported at this meeting, folding intermediates of a very small protein, bovine pancreatic trypsin inhibitor, exhibit both a less mobile core and a more mobile disordered region as determined by hydrogen exchange.[56] Such non-uniform protection against exchange is inconsistent with the original molten globule model,[57] which included a proposal for rapid structural fluctuations. In addition, confusion arose because researchers were not careful in distinguishing between actual protein behavior and Ptitsyn's idealized molten globular model.[57] By now, however, many of the critics have come to largely accept the term *molten globule* if not all of the details of the original proposal.[58,59]

Of course, a wide range of behaviors should be expected for molten globules as a result of variations in structure and sequence between different proteins. Furthermore, the existence of more than one liquid phase can be reconciled with Ptitsyn's molten globule hypothesis by noting that the properties of liquids become granular as the volume becomes sufficiently small. Perhaps because of their small size and sequence hetereogeneity, proteins (unlike bulk liquids) can evidently exhibit more than one liquid-like (molten globular) form. With these caveats in mind, we agree with Ptitsyn that the molten globule represents a third thermodynamic state for proteins.[60]

The molten globule was initially discovered as an equilibrium structure observed in studies on protein denaturation. Later this protein form was proposed to exist as a transient intermediate during the kinetics of protein folding.[53] Indeed, several ingenious studies have provided evidence for this proposal, at least for some proteins.[61]

Besides existing as a stable intermediate during *in vitro* denaturation and as a transient during *in vitro* folding of some proteins, the molten globule has also been proposed to provide the basis for biological function *in vivo*. The insertion of proteins into membranes[62] and the transfer of retinal from its blood-stream carrier to its cell-surface receptor[63,64] have both been suggested to depend on the molten globular state. These and other examples support the existence of the molten globule form *in vivo*.[65] Additional proteins that form molten globules rather than ordered structure under physiological conditions continue to be discovered.[66,67,68,69,70,71]

The information presented above suggests that it is appropriate to reassess the current protein structure / function paradigm. The remainder of this review attempts such a reassessment.

Outline: In the following section, the methods used to characterize intrinsic disorder are discussed briefly. Following this discussion of methods, The Protein Trinity is presented. This hypothesis relates native protein function to the three thermodynamic states of protein and to their transitions. Next, in support of The Protein Trinity, we present five proteins that utilize disorder and, for some, transitions between order and disorder, for their functions. We then present 16 more examples grouped as historically important examples, wholly unfolded proteins, and disordered regions with interesting functions. Next, our efforts to understand sequence / disorder relationships and our predictions of disorder from sequence are briefly reviewed. Protease digestion is sometimes used to argue against the existence of intrinsic disorder, so this argument is explored along with the role of chaperones. We end with a brief comment on the implications of intrinsic disorder for the structural genomics project.

Determination of Intrinsic Disorder

Several methods have been used to characterize intrinsic disorder in proteins, each having its own strong and weak points. Here the most important methods are very briefly discussed.

X-ray crystallography: As mentioned above, disorder leads to missing electron density in protein structures determined by X-ray crystallography. Two types of disorder have been recognized, static and dynamic.[30,72] If a dynamic region freezes into a single preferred structure upon cooling, then collecting data at lower temperatures distinguishes dynamic from static disorder in some cases.[73] However, from our point of view, more important than static or dynamic, is whether the missing region assumes one set of the Ramachandran Φ, Ψ angles along the backbone or whether the missing region exists as an ensemble of angles. We are calling a missing region with one set of Φ, Ψ angles, whether static or dynamic, a *wobbly domain* because such a region assumes different positions as a

rigid body, with the transitions between the different positions being slow (e.g. static disorder) or fast (dynamic disorder) on the X-ray analysis time scale. From our point of view, a region existing as an ensemble of Φ, Ψ angles, whether static or dynamic, is intrinsically disordered. The major uncertainty regarding information from X-ray diffraction is that, without additional experiments, it is unclear whether a region of missing electron density is a wobbly domain, is intrinsically disordered, or is the result of technical difficulties.

NMR spectroscopy: 3D structures can be determined for proteins in solution by NMR. The absence of the requirement for crystallization means that NMR provides a less biased estimate of the commonness of disorder as compared to crystallography. Under favorable circumstances, NMR provides motional information on a residue-by-residue basis[74] by means of a variety of different isotopic labeling and pulse sequence experiments (for an excellent review, see the paper by Braken in this volume[75]).

Especially useful is the ^{15}N-^{1}H hetereonuclear NOE measurement, which gives positive values for (more slowly tumbling) ordered residues and negative values for (more rapidly tumbling) disordered residues.[74,75,76] When these NOE data are displayed for an amino acid sequence, ordered regions are evident as a series of consecutive positive values and disordered regions as consecutive negative ones.[37]

Compared to ordered proteins, relatively few molten globules have been structurally characterized by NMR, indicating the existence of significant experimental difficulties. First, proteins with molten globular regions often aggregate at the concentrations needed for NMR. Second, molten globules are heterogeneous with structural inter-conversions on the millisecond time-scale; this leads to extreme broadening of the side-chain NMR peaks. Indeed, to obtain side-chain information for molten globules, fitting to molecular simulations data [77] and residue-specific isotopic labeling[78] were used in place of the more standard approaches. Also, NMR backbone data on molten globules are also difficult to obtain due to the opposite effects, that is because of a lack of chemical shift dispersion (rather than extreme line broadening).

Unlike most other molten globules, including that of apomyoglobulin at pH 2, the molten globule formed by apomyoglobin at pH 4 is stable up to 50º C. At this temperature, the individual peaks sharpen enough so that even poorly dispersed spectra can be used.[79] Employing this strategy, Wright and co-workers were able to determine a backbone structure for the pH 4 apomyoglobin molten globule, revealing regions of secondary structure and flexibility.[80]

Although NMR is not biased against random-coil disorder, the molten globule presents experimental difficulties that can be over-come only with great difficulty if at all. Thus, native molten globular protein domains would probably be under-represented in any collection of disordered proteins characterized by NMR.

Circular dichroism (CD) spectroscopy: Structural information for proteins in solution is also provided by circular dichroism.[81] Far UV CD spectra provide estimates of secondary structure and so distinguish ordered and molten globular forms from random coil. On the other hand, near UV CD show sharp peaks for

aromatic groups when the protein is ordered, but these peaks disappear for molten globules and random coils due to motional averaging.[47,51,82] Thus, combined use of near and far UV CD can distinguish whether a protein is ordered, molten globular or random coil. However, this method is only semi-quantitative and lacks residue-specific information and so does not provide clear information for proteins that contain both ordered and disordered regions.

Protease digestion: By the late 1940s, it was already recognized that protease digestion gave insight into protein structure and flexibility,[83] with a more detailed view emerging after more had been learned about protein structure.[84] More recent studies by Fontana, Thornton, Hubbard, and their co-workers[85,86,87,88] provide compelling evidence that flexibility, not mere surface exposure, is the major determinant for digestion of possible cut-sites.

Since a structured region has to become unfolded over more than 10 residues to be cut by typical proteases,[87] the ratio of the digestion rates, (disordered) / (ordered), should be about equal to $e^{-\Delta G/RT}$, where ΔG is the free energy of unfolding the segment. For two-state unfolding, whole protein unfolding free energies and segmental unfolding free energies would not be distinguishable. For typical proteins, unfolding free energies range from 7 to 10 Kcal/mol,[89] so a disordered region would be expected to undergo digestion ~ 10^5 to 10^7 times faster than an ordered one.

Studies by Fontana and co-workers demonstrate huge increases in digestion rates after the F helix of myoglobin is converted to a disordered state in apomyoglobin, with the cut loci for several different proteases occurring within the disordered region that arises from the F helix. The exact (disorder) / (order) digestion rate ratio was not determined, but the authors indicate[86] that the relative rates could be even larger than the 10^5 to 10^7 estimated above. Thus, hypersensitivity to proteases is sure evidence of protein disorder. Protease digestion gives position-specific information. However, the requirement for protease-sensitive residues limits the demarcation of order/disorder boundaries by this method.

Protease digestion is especially useful when used in combination with other methods. For example, protease digestion can be used with X-ray diffraction to help sort out whether a region of missing electron density is due to a wobbly domain or to intrinsic disorder.[90,91] A protein with two structured domains connected by a flexible linker converts to two lower molecular weight fragments with no digestion intermediates, while a protein with intrinsic disorder typically exhibits a set of intermediate-sized fragments (from random cleavage of the intrinsic disorder) that converge to a single fragment at longer times of digestion. Protease digestion is also useful when coupled with CD spectra, which lack position-specific information. Finally, the combination of proteolysis and mass spectrometry for fragment identification shows special promise for indicating the presence of intrinsic disorder.[92]

Stoke's radius determination: Random-coil disorder has also been detected by various methods for obtaining Stoke's radius such as small angle X-ray scattering or size exclusion chromatography.[93] Abnormally large radii for a given molecular weight provide evidence of disorder. Such methods have been

used in combination with CD spectroscopy,[94,95] which provides independent evidence of random coil structure. If the random coil is responsible for function, then protein activity is not lost by incubation at high temperatures.[95]

The Protein Trinity

As stated above, the current protein structure/function paradigm emphasizes the role of ordered 3D structure as being a necessary prerequisite to protein function. Our alternative proposal, herein called *The Protein Trinity*, is shown in Figure 1. In this view, native, intracellular proteins or functional regions of such proteins can exist in any one of the three thermodynamic states noted previously,[60] namely: ordered forms, molten globules and random coils.

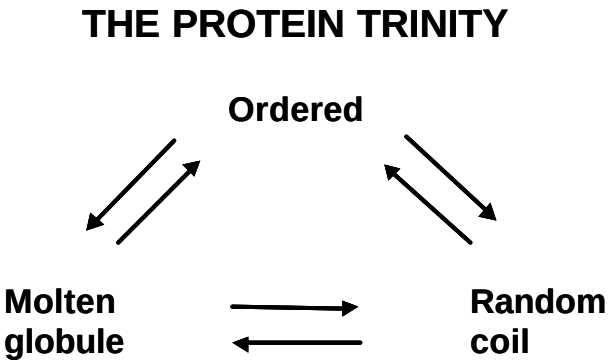

Figure 1: An alternative hypothesis for protein structure / function
As shown, The Protein Trinity suggests that *native* protein structure includes not only the ordered state, but also random coils and molten globules. Further, this hypothesis suggests that function can be carried out, not only by the native state, but *by any of the three states or their transitions.*

The key point of The Protein Trinity is that a particular function might depend on any one of these states, or a particular function might depend on a transition between two of the states. In this view, not just the ordered state is native, but any of the three states can be the native state of a protein.

These three states can be viewed as analogous to solid (ordered), liquid (molten globule) and gas (random coil) and their interconversions are analogous to phase transitions. In this analogy, the term *disordered* would be equated with the term *fluid* used as an adjective, encompassing both the molten globule and the random coil.

The transitions between the ordered state and the random coil and between the ordered state and the molten globule exhibit easily measurable cooperativity.[96,97] However, given the small sizes of proteins, the heterogeneity of the internal side chain packing, and the irregularities of the backbone structures, the transitions are not nearly so sharp as for bulk phases.

Liquids and gases usually exhibit clear phase transitions whereas the transitions between the molten globule and the random coil typically do not.[60,96] However, some artificial, synthetic helical bundles exhibit both molten globule-like, non-rigid side chain packing and cooperative unfolding.[98,99] The behavioral differences between these artificial bundles and natural molten globules appears to rest in differences in the two types of hydrophobic cores: the artificial bundles were constructed to have purely hydrophobic side chains in their cores while actual proteins in the molten globule state have mixtures of both polar and nonpolar side chains in their cores.[99] Thus, the presence or absence of a cooperative transition depends on the details. The Protein Trinity analogy to solid/liquid/gas can still hold by considering the molten globule ↔ random coil transitions without discernable cooperativity as analogous to liquid ↔ gas transitions at temperatures and pressures above their critical points.

Disorder and flexibility are often used synonymously, but the two terms are quite distinct. With regard to an ordered protein, flexibility refers to the magnitudes of the excursions of the atoms from their equilibrium positions. The Debye-Waller temperature factors (also called B-factors) obtained from X-ray data provide experimental measures of this sort of flexibility.[100,101] In addition, molecular dynamics was invented for the study of this type of motion.[101a] Indeed, almost countless numbers of papers have been published that use molecular dynamics simulations to explore local flexibility. For a disordered region, variation in flexibility refers to differences in the speed of interconversion among the various members of the structural ensemble. A variety of methods have been used to investigate the flexibilities of disordered regions and proteins, including NMR as discussed previously[74] and in this volume[75].

Examples in Support of The Protein Trinity

The formulation of The Protein Trinity proposal was derived from observations showing that numerous protein functions depend on non-ordered protein forms and also that some functions depend on transitions between two forms. Here we discuss five such proteins and their structural transitions that we have studied in our own laboratories (Table 1).

Table 1. Five Examples of Functional Disorder

Protein	Disorder	Function
fd Phage	ordered → molten globule	Membrane penetration
Histone octamer	ordered → molten globule	DNA replication, Transcription, Chromosome remodeling
Clusterin	Native molten globule	Protein detergent
Calcineurin	95 aa disordered loop	Display of calmodulin target
Calsequestrin	21 aa disordered tail	Ca^{++} Binding

Filamentous phage membrane insertion: The fd phage is comprised of 2,700 almost entirely helical coat proteins of 50 residues each. The subunit helices lie approximately parallel to the fiber axis arranged in layers with five subunits per layer. Each helix has a slight curvature that winds around the long axis of the phage. Subunits overlapping like scales on a fish lead to the formation of a helical bundle more than 8500 Å long. During infection, the coat protein helices somehow slide relative to each other while penetrating the inner membrane of the host *E. coli*. This movement eliminates the subunit overlap in the capsid, so that the subunits come to lie at a common level in the cell membrane.

As first characterized by Griffith and co-workers,[102,103,104,105] a chloroform/water interface converts the filaments into shortened rods called I-forms at lower temperatures and then the I-forms convert into spheroidal structures called S-forms at slightly higher temperatures (Figure 2). An especially important aspect of this figure is the irregularity of I-forms and S-forms when observed by typical electron microscopy, but their greater regularity when observed in the hydrated state. Regularity was observed previously for the I-forms and S-forms if they were cross-linked with glutaraldehyde prior to examination by electron microscopy.[103] The I-forms are proposed to mimic the intermediate step as the coat proteins insert into the host membranes whereas the S-forms, which have a morphology resembling vesicles yet contain no lipid, are proposed to mimic the final, membrane-inserted forms.[106,10]

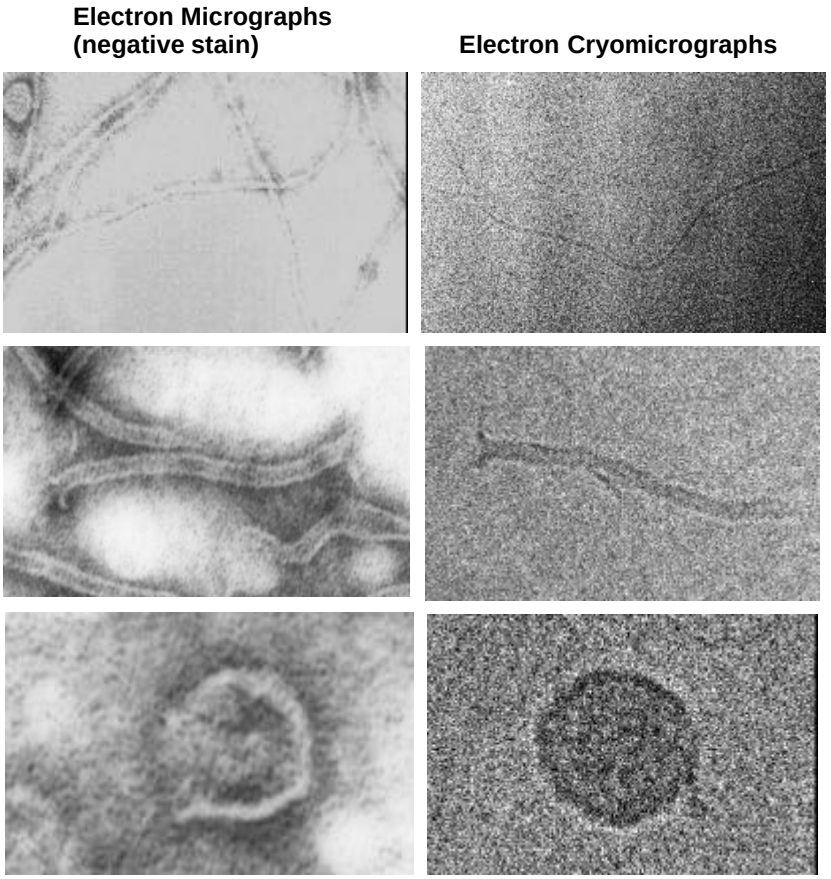

Figure 2: Contraction of filamentous phage by a chloroform/water interface Images of fd filaments (top pair), I-forms (middle pair) and S-forms (bottom pair) were prepared using either traditional negative stain electron microscopy (left set of 3) or using cryo-electron microscopy without stain (right set of 3).

Three lines of evidence suggest that both I-forms and S-forms, but not the filaments, have whole-particle molten globular characteristics: 1. The I-forms and S-forms, but not the filaments, exhibit substantial fragility when subjected to conditions for electron microscopy[108] (Figure 2). 2. The tryptophan fluorescence of I-forms and S-forms is extraordinarily sensitive to nonpolar quenchers, but normally sensitive to polar quenchers, while the tryptophan fluorescence of filaments is normally sensitive to both nonpolar and polar quenchers. These data suggest that the nonpolar quenchers dissolve into the hydrophobic interiors of the I-forms and S-forms due to non-rigid side chain packing, but do not dissolve into the interiors of the filaments due to tight packing.[106] 3. I-forms and S-forms, but not filaments, associate with the polarity-sensitive probe, 1-anilino-8-napthalene sulfonate (ANS), to greatly enhance the extrinsic fluorescence of this molecule. These data show that this probe can associate with the hydrophobic interiors of the I-forms and S-forms (as it does with known molten globules) due to non-rigid side chain packing, but that this probe does not associate with the interiors of the filaments due to tight packing.[108]

The loose side chain packing of such intermediate forms would enable the subunits to slide relative to each other as required for the phage infection process [107]. Of course the fd phage provides an interesting example of the earlier suggestion that the molten globule offers an ideal structure for the insertion of membrane proteins.[62]

Crushing typical and hyperacetylated nucleosomes: The disk-shaped nucleosome core particle contains a histone octamer wrapped by 1¾ turns of double-stranded DNA having 146 base pairs in total. The histone proteins contain positively charged tails that are sensitive to protease digestion[109] and that are unobserved in the nucleosome X-ray structure,[110,111] both of which concur in the assignment of these tails to the disordered class of protein structure.

Hyperacetylation of these tails was discovered more than 30 years ago and has been implicated in both DNA replication[112,113] and RNA transcription.[112,113,114] Solution studies comparing typical and hyperacetylated nucleosomes discern small or no changes in structure or stability due to this modification.[115,116] On the other hand, electron micrographs show that hyperacetylated nucleosomes become highly deformed on the carbon surface whereas normal nucleosomes maintain their spherical shape. This indicates that hyperacetylation causes nucleosomes to become much more sensitive to the forces encountered during sample preparation,[117] similar to the I-forms and S-forms of fd phage.

Furthermore, hyperacetylation causes the negatively charged nucleosome to become even more negatively charged. Note that increased charge imbalance, induced for example by low or high pH, is one of the primary methods for promoting transitions from the ordered state to the molten globule.[47,60,61] Ignoring bound and associated counter ions and assuming ½ charge unit for histidine due to its pKa near 7, we estimate a net charge of negative 138 for nucleosomes without acetylation, and up to as much as negative 164 for the most fully acetylated particles.

To test whether the increased charge repulsion leads to a decrease in the rigidity of nucleosomes, we turned to atomic force microscopy (AFM), which we have been using for similar investigations in fd phage.[118] The ability to control the amount of force on the particles is one of the major advantages of this method. Another is that samples can be studied in a hydrated state at atmospheric pressure (not in a vacuum as for typical electron microscopy) or even under water.

AFM images developed using tapping mode revealed very little difference between normal (Figure 3A, left) and hyperacetylated nucleosomes (Figure 3A, right). Contact mode, but not tapping mode, can be used to vary the force between the tip and the surface, so a series of experiments were carried out using this protocol. Normal nucleosome images taken at ~ 47 nN (Figure 3B, left) and retaken at the same force after exposure to 110 nN (Figure 3B, right) showed very little change. On the other hand, at even lower forces, ~ 29 nN, images of hyperacetylated nucleosomes showed considerable distortion (Figure 3C, left). Furthermore, following exposure to 38 nN and re-imaging at the lower ~ 29 nN force, the hyperacetylated nucleosomes were largely destroyed (Figure 3C, right). Note especially that the region subjected to ~38 nN is covered with irregular matter that most likely corresponds to debris from crushed nucleosomes.

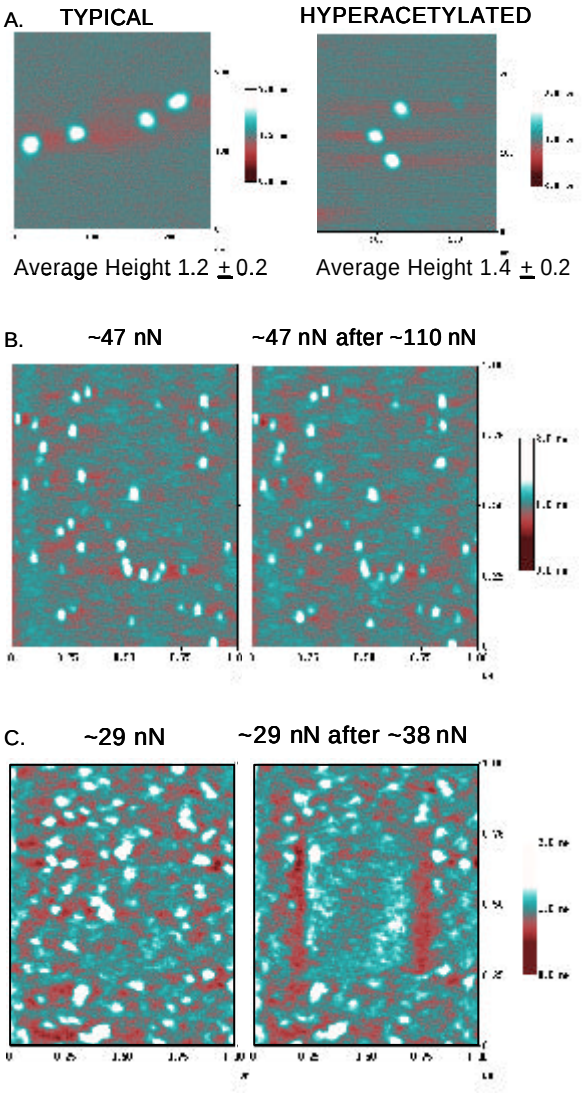

Figure 3: Atomic force microscopy of typical and hyperacetylated nucleosomes

In panel A, typical (left) and hyperacetylated nucleosomes (right) are shown for AFM images prepared using tapping mode. In panel B, an image is shown of typical nucleosomes using contact mode with a tip force of ~ 47 nN (left) and an image is shown for the same sample using the same mode at the same tip force, but after the central region had been subjected to a tip force of ~110 nN (right). In panel C, an image is shown of hyperacetylated nucleosomes using contact mode with a tip force of ~ 29 nN (left) and an image is shown for the same sample using the same mode at the same tip force, but after the central region of the image had been subjected to a tip force of ~ 38 nN (right). Note that only nucleosome debris is visible in the central region in panel C.

A large number of experiments like that shown in Figure 3 were carried out using various sample preparations and several tips having different flexibilities. The resulting curves of measured height versus force for normal and hyperacetylated nucleosomes show the hyperacetylated to be crushed by less than 40 nN while bulk nucleosomes remain intact out to 275 nN (Figure 4). Thus, hyperacetylation makes nucleosome core particles less rigid.

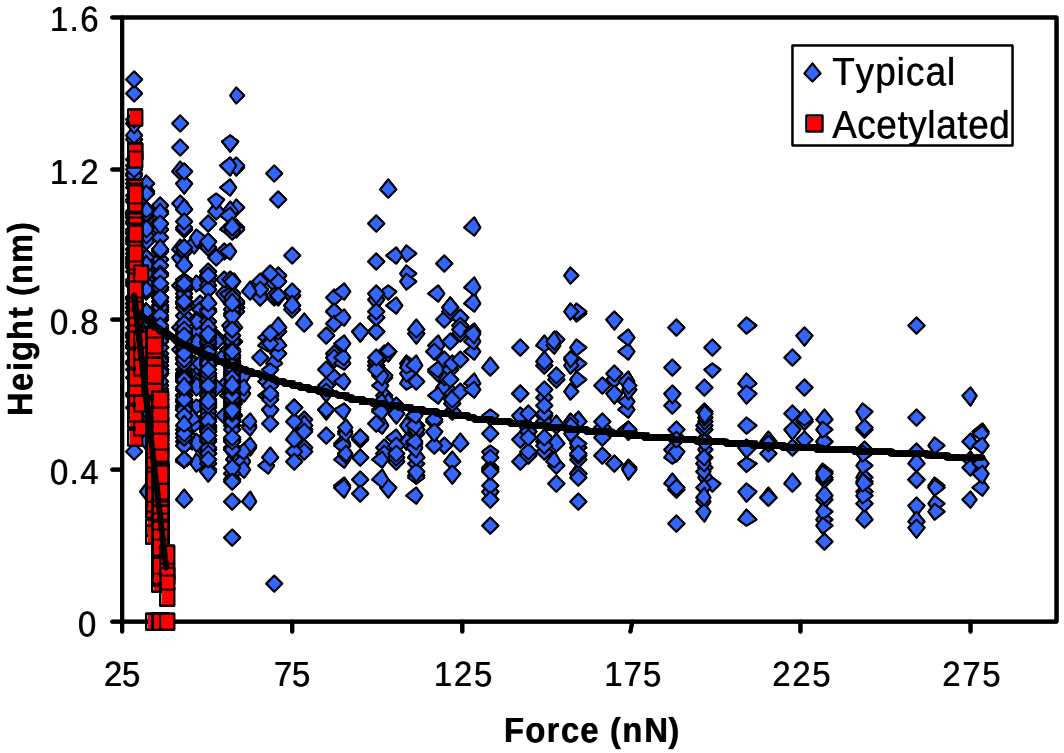

Figure 4: Height versus force for typical and hyperacetylated nucleosomes
Typical and hyperacetylated nucleosomes were imaged using AFM in the contact mode at various force values. To span such a wide range of forces, 4 different tips with different degrees of stiffness were used.

Clusterin, a protein with a native molten globular domain: Enrico Sertoli[119] first described cells, now bearing his name, that are found in close association with developing germ cells within the testis. By now it is apparent that Sertoli cells "nurse" the germ cells by providing nutrients and regulatory factors.[120] One protein, called clusterin, was found as a secretion from Sertoli cells and is presumed to play an essential role in germ cell development.[121] By sequence matching, this protein was found to be the same as other proteins found later in various tissues, especially the brain.[122] Finally, this same protein was also identified as being associated with cellular injury, lipid transport, and apoptosis.[123]

Structural characterization of clusterin has been difficult. Repeated attempts at crystallization have failed. This protein forms aggregates and sticks to hydrophobic surfaces. Although there are several possible explanations for such behavior, one is the presence of a molten globular domain.

We are developing methods for characterizing proteins that might contain both ordered and molten globular domains in the same structure. Given its behavior, clusterin provides a good model system for testing our approach. The methods focus on using protease digestion to distinguish ordered and disordered regions of a protein and ANS fluorescence to determine whether any protease sensitive regions are molten globules or random coils. In addition, as discussed below in more detail, we are developing several predictors of natural disorder regions (PONDRs) that predict order and disorder from amino acid sequence.[124,125] In essence, these predictors compare a given sequence to the set of structurally characterized ordered and disordered sequences used for training.

Protease digestion of native and denatured clusterin and predictions of order and disorder by PONDR show almost perfect agreement: in the native state, regions predicted to be disordered are sensitive to protease digestion, while regions predicted to be ordered are resistant to become sensitive to digestion when clusterin is in a non-native state (Figure 5). Furthermore, ANS binds to clusterin as indicated by 100-fold increases in extrinsic fluorescence (data not shown). Also indicitive of ANS binding is the observation that ANS inhibits trypsin digestion at one or possibly two of the sites located within the protease-sensitive, disordered region (Figure 5). The ANS binding and its protection of particular sites from trypsin digestion suggest that the disordered regions in clusterin very likely form molten globular structure.

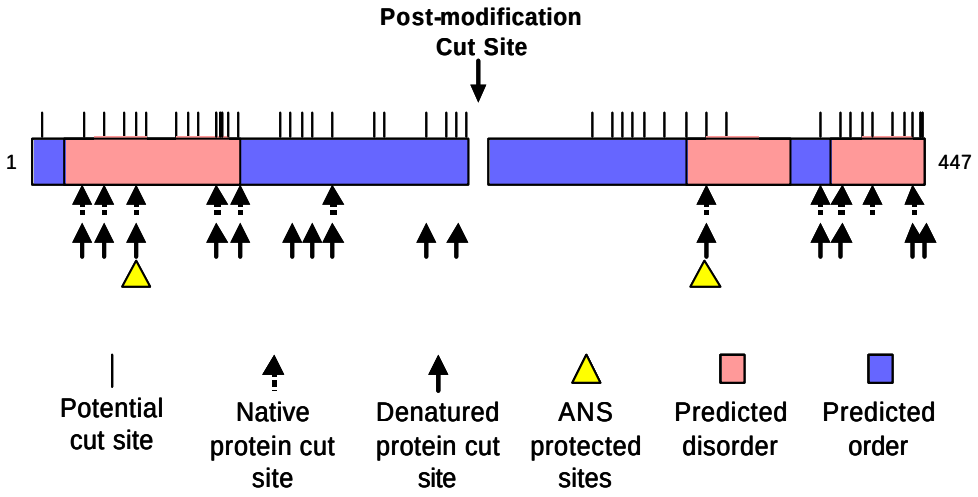

Figure 5: Protease digestion of clusterin
Purified clusterin with and without renaturation was subjected to digestion by trypsin and the fragments were separated by HPLC. The earliest appearing fragments correspond to the most sensitive cut sites, which were identified by amino acid sequence analysis. Protection against trypsin digestion by ANS is weak (if the protection exists at all) for the site on the left, while the protection is clear and unambiguous for the site on the right.

The function of clusterin is unclear, but recent unpublished data show that this protein dissolves bacteriorhodopsin and thus acts as a proteinaceous detergent. Such a function is commensurate with the properties of the molten globular state. Furthermore, a detergent-like function could be important

whenever cellular components need to be moved, such as for cellular remodeling in sperm maturation, in building projections from the cellular bodies in brain tissue, and in removing cell debris following apoptosis.

Calcineurin, a protein with essential disorder: This protein contains a catalytic A subunit and a calmodulin-like B subunit (Figure 6A). This model was constructed from coordinates obtained from PDB (accession code: 1aui) as originally determined by Kissinger and co-workers.[91] The calcineurin A subunit is a calcium/calmodulin-activated serine/threonine phosphatase.[126] Being involved in both calcium and phosphorylation signaling pathways, calcineurin plays important roles in many if not all eucaryotic cells. For example, calcineurin was so-named because of its high concentrations in brain tissue,[127] yet this same protein plays a key role in T-cell mitogenesis. With respect to the latter, inhibition of calcineurin's phosphatase activity is the key event in immuno-suppression by drugs such as cyclosporin and FK506 that are commonly used to prevent rejection following organ transplantation.[128,129]

Disorder plays an important role in calcineurin function. The calmodulin target helix is located within a 95 amino acid disordered segment that is unobserved in crystal structures.[91] The lack of ordered structure in this region was previously discovered due to its hypersensitivity to protease digestion.[90] Calmodulin completely surrounds its helical target upon binding[130] as shown in Figure 6B (the coordinates for this figure were taken from PDB (accession code: 1cdm). Locating this helical target within a region of disorder enables it be easily surrounded by calmodulin, so the region of disorder is essential to the regulation of calcineurin by calcium / calmodulin.

Calsequestrin, a calcium storage protein: Calcium release out of the sarcoplasmic reticulum (SR) stimulates muscle contraction while calcium uptake into the SR leads to muscle relaxation. A protein within the lumen of the SR facilitates the calcium uptake/release functions by high capacity, low affinity calcium binding. This protein, calsequestrin, binds 40-50 calcium ions with affinities of about 1 mM.[131,132,133]

Rabbit skeletal muscle calsequestrin has three regions of missing electron density in its crystal structure:[134] a glutamate dimer (EE) at the amino terminus, a seven residue disordered loop with five consecutive negative charges (MDDEEDL) from residue 327 to 333, and a 20-residue carboxy terminus having 16 glutamates + aspartates (EGEINTEDDDDEDDDDDDDD), and thus having 17 negative charges when the carboxy terminus is included. There are barely enough excess negative charges in calsequestrin to bind to the ~ 50 calcium ions sequestered by this protein. Thus, all of the disordered, negatively charged residues are almost certainly involved in calcium binding.

A calsequestrin tetramer from the crystal structure[134] is shown in Figure 6C. The coordinates used for the ordered parts of the protein in this figure can be found in PDB (accession code: 1a8y). Please note that the structural representations of the disordered regions (yellow) are completely *ad hoc*. The amino terminus and 327-333 loop each represent a hypothetical single conformation of what must be a large ensemble of native protein conformations. Furthermore, the 20 residue carboxy-terminal tail certainly does not adopt the

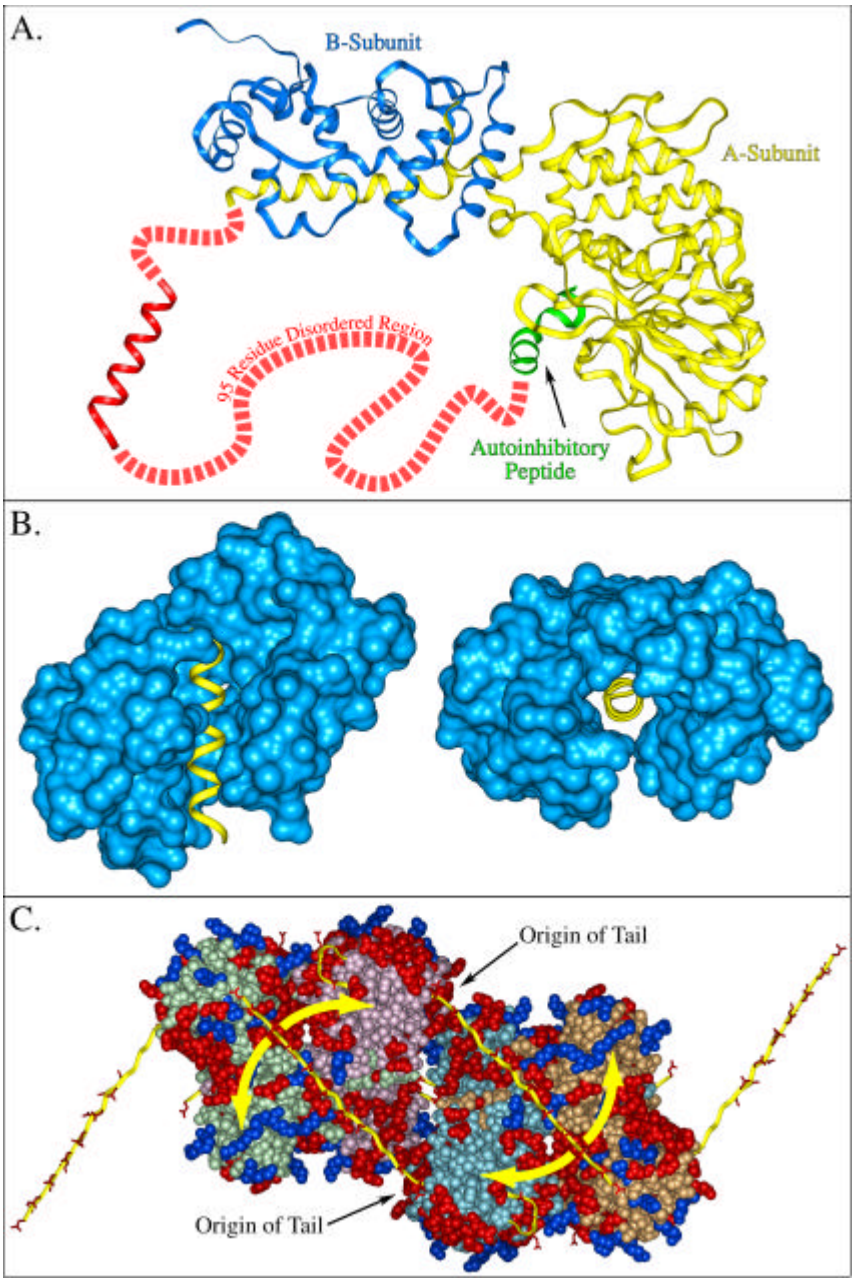

Figure 6: Structures of proteins with essential disorder

A. The A subunit of calcineurin contains a phosphatase domain (yellow), a helix (yellow extension) that binds the B subunit and an autoinhibitory peptide (green) that binds to the active site on the phosphatase domain. The B subunit (blue) binds to the A subunit helical extension and resembles calmodulin. The 95 amino acid intrinsically disordered region of the A subunit connects the end of the helical extension (residue 374) with the end of the autoinhibitory peptide (residue 470). This region of disorder was identified by its lack of electron density and so this part of the protein is unobserved.

B. Calmodulin (blue) bound to the target helix from calmodulin-dependent protein kinase II (yellow) is shown from two orientations: looking from the side of the target helix (left) and looking down the target helix (right).

C. Representations of 4 calsequestrin molecules are shown with a slightly different color for each. The charged residues are indicated, with aspartates and glutamates in red, and with lysines, arginines and histidines in blue. The 3 disordered regions are rendered in yellow with their aspartates and glutamates in red. The disordered regions, especially the polyanion tail, were rendered so as to emphasize their possible extents, not to indicate realistic possibilities for their structures.

conformation shown. The purpose of showing this polyanion tail in a highly extended conformation is to demonstrate that the tail has many options for calcium-mediated interactions with other negative side chains (in red), including possible calcium cross-bridges with each other and also with the seven-residue disordered loop.

Implications for The Protein Trinity

Disorder-to-order transitions upon binding are becoming recognized as a common occurrence (see examples below), with the ordered state end-point being crucial for function. For such proteins, it might be considered a minor detail whether the ordered state occurs prior to, or concomitant with, molecular association. The five examples given above do not follow the pattern of disorder-to-order upon binding; indeed, for these examples, an ordered state may have little or nothing to do with function.

For our first two examples above, the protein assemblies undergo order-to-disorder transitions (not the reverse) and the resulting disordered state is responsible for function. For the filamentous phage, the non-rigid side chain packing allows the subunits to move relative to each other. This facilitates the sliding motion that must occur as the coat protein penetrates the cell membrane. For the nucleosome core particle, reducing particle rigidity would facilitate local structural deformations, thereby allowing access to the DNA without a requirement for dissociating the histones from the DNA. How DNA replication and RNA transcription can occur without separating the histones from the DNA has been a great mystery for both of these processes.

In our third example, calmodulin completely wraps around its target helix in calcineurin upon binding. The target helix must be well separated from the remainder of the protein (Figure 6B). The ~ 20 residue target helix is located within a 95 residue disordered region. The disorder allows the target helix to separate from the rest of the protein, thus permitting access for binding. For the calmodulin/target helix interaction, it is unknown and possibly unimportant whether the other ~ 75 disordered residues become ordered upon binding. The important function of disorder in this example is to display and make accessible the target helix. Not only does trypsin digestion rapidly activate calcineurin, thereby confirming the flexibility of this disordered region, but trypsin digestion activates a number of other calmodulin-regulated enzymes as well,[90] suggesting that disorder adjacent to the calmodulin target helix is a common feature of these enzymes.

In our final examples, clusterin and calsequestrin both have disordered regions that bind ligands. The former contains a molten globular domain that apparently functions as a proteinaceous detergent and the latter has several negatively charged disordered segments that bind calcium. For both proteins, ligand binding does not necessarily bring about a specific disorder-to-order transition. The tails of detergents typically do not become ordered when incorporating hydrophobic molecules into their micelles; perhaps the same lack of induction of order occurs as well when the molten globular clusterin domain

associates with its ligand. As shown in Figure 6C, the polyanion tail of calsequestrin can potentially come close to a number of carboxyl groups at very different locations on the protein surface; thus, assuming that the tails are random coils in the absence of calcium (not the extended structures shown in this figure to illustrate their potential reach), addition of calcium ions could induce many different possible cross-bridges leading to many different, likely static structures. Thus, calicium binding by calsequestrin need not correspond to the induction of a particular structure, but rather most probably induces an ensemble of protein structures, all members of which exhibit low affinity calcium binding.

More Examples Supporting The Protein Trinity

There have been numerous reports of proteins that contain regions of intrinsic disorder required for function. Here we discuss examples selected for their historical importance and potential interest (Table 2).

Table 2. Important Examples of Native Disorder

Protein	Disorder Length	Function
Trypsinogen	18	Folding inhibitor
TMV capsid	26	RNA binding
Lac repressor	61	DNA binding
Calmodulin	4	Flexible linker, Target binding
p21*Waf1/Cid1/Sdi1*	164	Kinase inhibitor
HIV-1 gp120 V3 loop	24	Antibody epitope

Trypsinogen: More than 20 years ago, studies on this protein pointed to the importance of protein disorder.[29,135] Indeed, this is the first high-resolution protein crystal structure that identified a functionally important region of disorder by its missing electron density, with two independent structure determinations in basic concurrence.[29,136] In both of these studies, residues 1 – 18 were unobserved.

Every biochemical textbook explains that proteolytic digestion removes the VDDDDK amino terminus and that this cleavage converts inactive trypsinogen into active trypsin; none that we can find explains that the site of cleavage is within an 18-residue region with intrinsic disorder, the flexibility of which is important.

The structural changes that occur after the cleavage to convert inactive trypsinogen into active trypsin are less clear and in some dispute.[136,137] One side in this dispute suggests that the structure of trypsin's substrate-binding pocket, although exhibiting some disorder, can be resolved on the basis of a single structure that is the predominate one in the ensemble.[136] The other side suggests that this region, which they call the *activation domain*, is disordered.[29] Several additional experiments support the latter interpretation as discussed in detail by Bennett and Huber.[137] If the latter is true, then cleavage of the amino terminus brings about a disorder → order transition in the substrate-binding pocket.

The proteolytic cleavage of trypsinogen's amino terminal VDDDDK sequence, creates a new amino terminus in trypsin. The amino terminal VD

dipeptide in trypsinogen is converted to the more hydrophobic amino terminal IV dipeptide in trypsin. According to Huber and co-workers, this new, hydrophobic dipeptide inserts into its own binding site, thereby stabilizing the correctly folded activation domain. Indeed, added IV dipeptide can bind to uncleaved trypsinogen and stabilize the activation domain in the structure that it adopts in trypsin.[29] These data suggest that the function of the last six residues of the disordered tail is to inhibit proper folding by capping a critical dipeptide terminus and that the remaining 12 disordered residues allow the cleavage-created IV terminus to search for and find its binding site.

Tobacco Mosaic Virus (TMV) coat protein: The TMV capsid is a rod-like structure containing more than 2000 coat protein molecules and an RNA of 6,400 nucleotides long.[138] Reports on this protein were the second to identify a functionally important disordered region by its missing electron density.[28,139,140] Furthermore, NMR experiments indicate dynamic rather than static disorder for this region.[141]

The TMV coat protein's disordered region is a 32-residue loop that is positively charged. This loop undergoes a disorder-to-order transition upon RNA binding during the assembly of the capsid. As pointed out by Holmes,[138] the flexibility and disorder of this region is absolutely essential for the observed assembly pathway. That is, more than 30 TMV coat protein molecules associate to form a rodlet. The RNA joins this pre-formed protein rodlet by threading into a hole in its center. Without the flexibility of the region of disorder, there would not be enough room in the central hole to thread the RNA.

Lac repressor, a paradigm for gene regulation: This protein represents a prototype for understanding gene regulation. It binds to a specific DNA sequence and thereby prevents transcription of the associated operon.[142] The DNA binding domain is comprised of a 61 amino acid segment at the amino terminus that is unobserved in the crystal structure of the protein, but this segment is observed and bound to DNA in protein/nucleic acid co-crystals.[34] NMR studies on DNA binding by the first 56 amino acids of this fragment indicate local regions of structure, but also highly mobile loops that become immobilized upon binding to DNA.[143] Thus, like the TMV capsid protein, the lac repressor undergoes a disorder-to-order transition upon binding to nucleic acid.

Calmodulin, a protein with two disorder features: Calmodulin has two globular domains, each with 2 EF-hands that bind calcium. In the protein crystal, a helix connects these two globular domains, yielding the famous dumbbell structure.[144] The connecting helix is an artifact of crystallization. When calmodulin is in solution, part of the helical rod melts into a flexible linker,[145] thus enabling calmodulin to wrap around its target (Figure 6B).

Disorder plays a second role in this protein. Even after being saturated with calcium, the globular domains in calmodulin switch among several alternative conformations in solution, more so than many other folded proteins (Minji Zhang, personal communication). This limited disorder may account for the significant plasticity observed within the globular domains when they bind to different targets.[146,147] This plasticity is reflected in alterations in the helix/helix contacts within the globular domains.[148] That is, the binding of target helices

with different sequences probably selects out different members of an ensemble of structures as suggested by Karush for the binding of hydrophobic anions by the albumins.[19] Furthermore, the side chains in the binding surface have an abnormally high flexibility,[149] perhaps due to high proportions of methionine.[150] The absence of branching permits methionine to assume many different shapes, thereby enabling this side chain to adopt many different conformations and so adapt to different hydrophobic surfaces. In this regard, we noticed that a high proportion of methionine has been observed in two completely different proteins that bind to a variety of hydrophobic ligands.[148]

Cyclin-dependent kinase inhibitor, p21$^{Waf1/Cip1/Sdi1}$: This 164-residue protein binds to cyclin-dependent kinases thereby inhibiting their activity. The unbound inhibitor appears to be completely disordered as indicated by protease digestion, CD and NMR.[39] Upon binding to CDK2, the amino-terminal 84 amino acids adopt an ordered conformation. Thus, this is another example of a disorder-to-order transition, in this case, upon binding to another protein. The p21 molecule is an important component of cell-cycle control through its ability to inhibit many different cyclin-dependent kinases. The disordered state of native p21 is suggested to enhance this protein's ability to bind to multiple protein targets.[39]

V3 loop of HIV-1 gp120: The gp120 protein from HIV-1 is involved in binding to receptors on host cells. Deletion of the V3 loop of this protein leads to loss of infectivity. A 24 residue fragment of the V3 loop of HIV-1 strain IIIB occurs as structurally heterogeneous peptides as shown by NMR spectroscopy.[151] This and similar V3 loop fragments occur in different β-turn conformations when bound to different V3 antibodies.[152,153] These authors suggest that the V3 loop either exists as an ensemble of structures with very different conformations or that binding forces conformational changes. The importance of the disorder-to-order transition of this loop is that a variety of chemokine receptors can be bound by this heterogeneous mixture of structures, allowing different avenues into the cell. This same mechanism makes devising a vaccine against HIV very difficult; some V3 loops can escape detection by antibodies that specifically recognize a given structure.

Interestingly, other pathogens have been shown to have attachment proteins that undergo disorder-to-order transitions upon binding to their receptors, including *Staphylococcal aureus*[154,155] and Foot-and-Mouth Disease Virus.[156] Avoidance of the immune response[154] and ability to bind to different cell surface molecules[153] have both been suggested as possible adaptive advantages that derive from the local disorder in the unbound state.

Disorder and Molecular Recognition

The TMV capsid protein, the lac repressor, calmodulin, p21$^{Waf1/Cip1/Sdi1}$, and the HIV V3 loops all utilize disordered regions in molecular recognition via disorder-to-order transitions. From the lock and key perspective, the involvement of disorder in molecular recognition seems counter-intuitive, yet by now a large number of similar examples have been discovered in addition to those in Table 2,

suggesting that the involvement of disorder in molecular recognition is quite common. Here we discuss the thermodynamic and kinetic aspects of the involvement of intrinsic disorder in molecular recognition.

Thermodynamic aspects: To our knowledge, the first attempt to understand protein associations involving disorder-to-order transitions is the work of Schulz[31] as shown in (Figure 7). His simple figure, which we have named The Schulz Diagram[148] is a thought-experiment, not a proposed reaction pathway. The first step (on the left) imagines making both partners rigid, which necessarily increases the free energy of the system, and the second step (on the right) gives the sum of the free energies that arise as the result of the contacts between the two folded forms. Of course, for many reactions, this sequence of events is simply impossible for steric reasons (e.g. see Figure 6B), which re-emphasizes that this figure does not represent a reaction pathway.

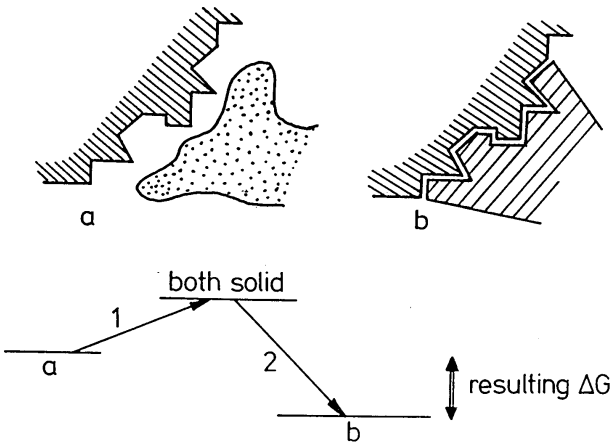

Figure 7: Schulz diagram
If a protein undergoes a disorder-to-order transition upon binding, the free energy to organize the disordered region is deducted from the free energies due to the contacts. This can potentially lead to high specificity coupled with low affinity.

From this simple diagram, Schulz suggested that one possible advantage of disorder-to-order upon binding is the possibility of high specificity coupled with low affinity. As mentioned above, another possible advantage is the binding to multiple partners by structural accommodation as observed for p21$^{Waf1/Cip1/Sdi1}$ (ref39) and the HIV VP 3 loop[151,153]. There is a certain symmetry for these two capabilities: the former depends on flexibility in the unbound state while the latter leads to structural polymorphism in the bound state[148] (Table 3).

Table 3. Disorder and Molecular Recognition

Characteristic	Structural Variability
High Specificity / Low Affinity	Flexibility in the Unbound State
High Affinity / Low Specificity	Polymorphism in the Bound State

The ability of one protein to bind to many partners has been called one-to-many signaling.[157] Calmodulin and p21$^{Waf1/Cip1/Sdi1}$ are cases in point. Wright and Dyson[43] have recently reviewed intrinsically unstructured proteins with some emphasis on examples of this type.

Disorder-to-order upon binding facilitates not only one-to-many signaling as for p21$^{Waf1/Cip1/Sdi1}$ and calmodulin, but also many-to-one. An example is

provided by interactions between nuclear localization signals (NLSs) and their receptor. The NLS does not depend on a precise sequence, but rather upon having a total of 4 lysines plus arginines within a segment of 6 amino acids that is flanked by negatively charged groups and helix breakers such as proline or glycine. The flanking regions help to insure "exposure" of the signal hexapeptide.[158] Intrinsic disorder would provide the need for exposure of the basic groups and, if long enough, would enable the NLS region to reach its receptor even if localized within the nuclear pore. An important feature of this system is the poor sequence conservation for the NLS regions, and yet these various regions apparently bind to the same receptor.

Disorder-to-order upon binding can overcome steric restrictions and thereby enable larger interaction surfaces in the complex than could be obtained for rigid partners. Two of many possible examples serve to illustrate the point. First, as discussed above, a flexible linker allows calmodulin to almost completely surround its target helix[130] as shown in Figure 6B. Second, multiple linkers between zinc finger domains (eight linkers between nine domains in the case of TFIIIA) enable such poly-domain proteins to wrap up their DNA partners.[159,160]

When a protein undergoes a conformational change upon binding to a ligand, it is often unclear whether the change results from selection of one member of an ensemble by the best-fit with the ligand, or whether the change results from induction of specific structure by interaction with the ligand. For special cases such as calmodulin/target helix interactions and TFIIIA/DNA interactions, steric clashes make it impossible to associate directly with one member from an ensemble, so the structure of the bound form must truly be induced by association with the ligand.

In addition to the qualitative reasoning discussed above, there have been studies aimed at making quantitative estimates of the free energy contributions associated with disorder-to-order transitions upon binding. The work by DeLisi and coworkers[163,164] is especially germane to this problem. Of interest here are two sets of complexes that they studied: 1. A set of 9 serine endopeptidase-inhibitor complexes; and 2. A set of 13 flexible nonameric peptides interacting with the class I major histocompatibility complex (MHC) receptor. The first set involves association between basically rigid molecules while the second involves a disorder-to-order transition upon binding.

Rather than attempting to estimate the binding free energies from first principles, DeLisi and coworkers[163] relieved some restrictions from an earlier semi-empirical method developed by Novotny.[165] In addition to the terms typically included in molecular mechanics energy functions, this method also includes terms for solvation, and entropy associated with side chain conformations, translational/rotational motion, and dilutional effects.

For the endoserine protease-inhibitor complexes, the observed and calculated values were within 2 kcal for 7 of the 9 complexes,[163] a result we find very impressive. The largest error, where the binding affinity is over-estimated by 9.1 kcal, is for the trypsinogen-BPTI complex. The explanation for this large over-estimate of course is in Figure 7. Due to the fact that trypsinogen is

disordered before binding[30] (e.g. see Table 2), the energy required for bringing about order, in this case perhaps ~ 9 kcal, takes away from the overall binding energy.

The results for the peptide-MHC receptor complexes are extremely interesting. To represent the ligand disorder, the authors created an ensemble of ligand structures using conformational searching. This ensemble was then analyzed to generate the various free energy terms for the receptor-ligand association. In their studies, DeLisi and co-workers don't specifically calculate the free energy change for the disordered ensemble converting an ordered state that is the same as the bound state but in the absence of receptor (thus corresponding to step 1 (on the left) in Figure 7). However, this free energy change can be estimated from their work by collecting terms dealing with the peptide only and leaving out all terms involving interactions between the peptide and MHC receptor, giving $?G = E_l^b - E_l^f - T?S_{bb}$, if the side chain entropy contributions are ignored, and giving $?G = E_l^b - E_l^f - T?S_{bb} - T?S_{sc}$, if the side chain contributions are included, where the superscripts b and f mean bound and free, respectively, and where the subscripts bb and sc mean backbone and sidechains, respectively. When the values of the various terms estimated by DeLisi and co-workers[163] are plugged into these free energy equations, a slightly positive ?G - when the side chain contributions are ignored - becomes a very positive ?G - when the side chain contributions are included. Thus, these calculations add considerable insight into Figure 7, providing evidence that side chain free energy contributions for disorder-to-order transitions are very important.

The readers should be aware that some references to the work of DeLisi and co-workers conclude that the disorder-to-order transition upon binding leads to an increase in the overall binding affinity compared to the same binding by a rigid molecule. From the much simpler Schulz Diagram and from the analysis of the previous paragraph, a conclusion of increased affinity is simply incorrect.

Kinetic Aspects: The rate of association of a protein with its partner is generally expressed as $k_{assoc} \sim 4D\alpha\kappa f$, where D is the relative translational diffusion coefficients and a is the sum of the radii (as in the Smoluchowski equation[164]), and where κ and f are dimensionless parameters to account for orientational and electrostatic effects, respectively.[165] This leads to a modified Smoluchowski equation. The electrostatic factor, f, can be very large. For example, the oppositely charged barnase and barstar exhibit association rates at high salt of about 10^5 $M^{-1}s^{-1}$ that increase to about 5×10^9 $M^{-1}s^{-1}$ at low salt.[166] In the absence of other effects, the orientational factor, κ, would be expected to be very small since the binding orientation represents a small fraction of all possible orientations.

Association rate constants frequently exceed the limits calculated from the modified Smoluchowski equation even when electrostatic effects are small. A model that explains the unexpectedly high association rates involves formation of an "encounter complex"[167] or a "transition state"[168] that can be thought of an ensemble of conformations undergoing microcollisions. Even in the absence of

attractive forces, Brownian motion predicts several microcollisions before the molecules diffuse apart.

Given the above framework for assocation rates, disorder can have several effects. First, the relative diffusion coefficient, D, would become smaller. Countering this effect, the effective the effective size, a, would markedly increase. But the really large effects would occur for the orientational parameter, κ, where disorder could lead to productive collision for essentially any relative orientation of the two molecules. Thus, conversion from the ordered to the disordered state can lead to large increases in the rates of assocation.[169] On the other hand, one could imagine cases in which the folding step becomes rate-limiting, in which case increasing the disorder would slow down rather than speed up the on-rates.

The effects of disorder on dissociation rates are also complex. In cases where the two molecules cannot come apart because of steric restrictions, disorder would clearly have the effect of drastically speeding up the off-rate. The formation of disorder could also enable dissociation of part of the interface, like unzippering, which might also speed up dissociation.

Advantages of Disorder: From the above discussion, flexibility can either raise or lower binding affinities and either raise or lower on-rates (and most likely off-rates as well) compared to binding between rigid partners. Also, the binding affinities of rigid partners can be increased or decreased through mutation. Thus, for rigid partners, it surely would be possible to raise or lower specificity and to increase or decrease on- and off-rates via suitable mutations. Finally, as demonstrated by moonlighting proteins,[170] rigid proteins can evolve to bind to more than one partner. These comparisons of the various properties describing molecular associations suggest that, despite earlier comments to the contrary, including some by us,[148,157] there might be no intrinsic advantage for disorder-to-order transitions upon binding as compared to association by rigid partners.

The paper by Goldstein and collaborators presented at this meeting is especially germane to this point.[171] The argument is frequently made that meta-stable proteins exist because such instability is needed to carry out function. In contrast, Goldstein and his collaborators argue that metastable proteins are favored during evolution because there are vastly larger numbers of sequences coding for these proteins as compared to very rigid ones. Of course, as was pointed out in discussions at this meeting, functions that depend on intrinsically disordered proteins represent the extreme case of this argument. Thus, the involvement of intrinsic disorder in protein function may relate more to history and evolution than to advantage and necessity.

Native Proteins Without Secondary or Tertiary Structure

With the advent of structure-determination methods based on NMR, protein crystallization was no longer required to obtain a protein's 3D structure. To the surprise of nearly everyone, several proteins that carried out function in their respective *in vitro* assays were intrinsically disordered from end-to-end as revealed by NMR spectroscopy. Four of these wholly disordered, yet native proteins are discussed herein (Table 4). Prior to or concurrent with these NMR

findings, several workers had suggested the existence of native random coils on the basis of CD spectroscopy coupled with other methods;[93,95,172] these CD-based studies failed to gain as much attention as the NMR-based work, in part because of the limitations associated with structural characterization by CD spectroscopy and in part because these CD-characterized, apparently native random coils lacked well-understood biological functions.

Table 4. Wholly Unfolded Native Proteins

Protein	Disorder Length	Function
Flgm	97	Transcription promoter
4E-binding protein	118	Translation inhibitor
HMGI(Y)	106	Architectural transcription factor
Neurogranin	78	Calmodulin storage protein

FlgM, a protein that regulates flagella assembly: NMR experiments indicate that FlgM is entirely disordered under physiological conditions and that, upon binding to σ^{28}, an anti-termination transcription factor, the carboxy–terminal half becomes ordered.[40] This molecular interaction regulates flagella assembly in an interesting way. In *Salmonella typhimurium,* flagella are constructed from a transmembrane basal body attached to a polymer of the major protein subunit. More than 35 proteins assemble to form the basal body. In a puzzling regulatory mechanism, a mutation in any one of the basal body proteins shuts down transcription of the flagella subunit mRNA. This regulatory puzzle is solved by the FlgM protein.

FlgM is exported via a channel in the center of the basal body assembly, thus maintaining a low level of FlgM. However, if basal body assembly is defective as for any of the mutants, FlgM cannot exit via the channel and so its levels rise. At these elevated levels, FlgM binds to σ^{28} and thereby shuts down synthesis of the mRNA for the flagella subunit. Also, as flagella assembly nears completion, export of FlgM is also slowed due to the length of channel. This again shuts down synthesis of the flagella proteins.

FlgM almost certainly must be intrinsically disordered to migrate through the small diameter channel in the basal body assembly. If this scenario is correct, FlgM's regulatory mechanism depends on its own intrinsic disorder. This argument is reminiscent of the requirement for disorder in the TMV coat protein loop due to steric restrictions within the small central hole as originally discussed more than 15 years ago.[138]

4E-binding protein 1 (4E-BP1): This protein, which contains 118 amino acids, inhibits translation by binding to the initiation factor eIF4E. NMR studies show 4E-BP1 to have little or no folded structure under physiological conditions,[41,173] and yet this protein is able to inhibit translation in reticulocyte lysates under similar conditions. A short region in 4E-BP1, namely from residues 49 to 68, is responsible for binding to eIF4E.[174] This region undergoes a disorder-to-order transition upon binding to eIF4E, forming a helix followed by an extended loop on the surface of eIF4E.[175] Indeed, this region of 4E-BP1 acts as a molecular mimic of a helix plus loop in eIF4G, which is the normal docking

site of eIF4E, thus explaining the competitive inhibition of the docking of eIF4E by 4E-BP1.

High Mobility Group Proteins I and Y, HMg-I(Y): These alternatively spliced transcripts from the HMG-I(Y) gene produce the DNA-binding isoform proteins HMG-I(106 residues) and HMG-Y (95 residues). They were first identified by their high electrophoretic mobility among the nuclear proteins and are founding members of a new class of regulatory elements called 'architectural transcription factors' that control the expression of a large number of human genes.[176] Various physical studies, including NMR spectroscopy, have demonstrated that, as free molecules in solution, the HMG-I(Y) proteins have no detectable secondary or tertiary structure. Discrete sub-regions of the proteins assume a planar, crescent-shaped configuration called the 'A.T-hook' when bound to the minor groove of A.T-rich DNA substrates.[177]

Neurogranin, a calmodulin storage protein, possibly involved in memory: There exists several similar proteins, including neurogranin and neuromodulin, that are evidently unstructured proteins as determined by NMR.[178,179,180,181] Calmodulin associates with these proteins at low, not high, calcium levels via a specific sequence called the IQ motif.[180,182,183] In addition to calcium concentration, a second level of regulation in this system depends on phosphorylation.[184] Using membrane anchors for localization, neuromodulin/ neurogranin may maintain calmodulin near the growing tips of the nerve cell projections when calcium levels drop. Because of this localization function, chains having the IQ motif have been called calmodulin storage proteins.[185]

Interesting Functions Associated with Disorder

Although much of the focus on proteins centers on molecular recognition, protein molecules do have other functions that are important for maintaining life. The following examples are presented in which proteins with native disorder have interesting functions not directly involving molecular recognition (Table 5).

Table 5. Native Disorder with Interesting Functions

Protein	Disorder Length	Function
fd pIII	21, 40	Flexible linkers
Bcl-x_L	63	Display of protease and phosphorylation sites
Kinesin	~15	Stepping motor
Titin	2 174	Entropic spring
K+ Channel	64	Entropic clock
Neurofilament H	158	Entropic bristle

The fd pIII protein: This phage protein, which is encoded by gene 3, is located on one tip of the fd filament and is responsible for attachment to the host *E. coli*.[186,187] The existence and importance of a flexible linker for this protein was suggested long ago.[188,189] By now we know that this protein contains three distinct domains: one integrated into one end of the filament and two that bind to

a pair of co-receptors on the host cell.[190,191] This double attachment is coordinated in an interesting way suggesting the need for movement between the domains.[192,193]

Two flexible linkers connect these three domains. The linker between the two attachment domains contains ~ 21 residues and the linker between the other two domains contains ~ 40 residues. Both are constructed from approximate repeats of the pentamer, EGGGS. As expected for such a sequence, the linker between the two attachment domains is unobserved in crystal structures. [194]

The ability to easily remove the phage attachment domains by proteolysis[188] indicates that the conformation of the linkers is likely quite mobile. These linkers uncouple the two attachment domains from the body of the phage particle and allow motional freedom so that the attachment domains can reorient independently. This motional uncoupling speeds up searches for productive docking orientations between the attachment protein domains and their receptors. The flexible linker may also act as a shock absorber after attachment has occurred.[189] A shock-absorber function has special importance in this system due to the extreme length of the phage.

Whenever a protein binds to two or more partners and the partners vary in spacing or orientation over time, just as discussed above for the fd attachment protein, flexible linkers can be essential. Thus, flexible linkers are quite common components of multi-functional enzymes, topoisomerases, transcription factors, and several other classes of proteins.[159,195,196,197,198,199,200,201,202]

Bcl-x_L and its homologue, Bcl-2: These apoptosis inhibitors contain long, flexible loops characterized both by missing density in X-ray crystal structures and as being highly mobile by NMR in Bcl-x_L and characterized by similar location in the related protein Bcl-2. In Bcl-2, this disordered loop plays a key role in programmed cell death.[203,204] Furthermore, protease digestion [205,206] and phosphorylation[207,208] within the disordered loops of these two proteins play critical roles in modulating programmed cell death.

We suggest that the intrinsic disorder of these loops plays an important role in the display of the protease digestion and phosphorylation sites. Each of these possibilities is discussed in turn.

As discussed above, protease cut sites within disordered regions are digested orders of magnitude faster than are potential cut sites within ordered regions. Molecular modeling provides insight into this hypersensitivity. Even though only 6 residues form close association with trypsin during proteolysis, molecular modeling indicates that a minimum of about 13 residues have to be unfolded in order for trypsin to be able to bind with high affinity. Without such local unfolding, binding is inhibited sterically.[87] In a folded protein, a 13 amino acid segment would typically be folded nearly all the time and therefore would be inaccessible to digestion as discussed above. Thus, an unfolded local region of sequence (e.g. a disordered region) digests many orders of magnitude faster than when folded.[86]

While trypsin binds six residues in the vicinity of the cleavage site, the cAMP-dependent protein kinase binds more than 20. The X-ray crystal structure of cAMP-dependent protein kinase shows the protein immobilizing a 20-residue

peptide that spans the phosphorylation site.[209] This peptide was derived from a natural 8 kD kinase inhibitor. Furthermore, fluorescence anisotropy measurements on probes attached to site-directed cysteine mutants suggests a disorder-to-order transition as this peptide binds to the kinase.[210] Other protein kinases appear to bind similarly large segments of their substrates. By analogy to the studies on trypsin, these results suggest that a phosphorylation site within a disordered region would react orders of magnitude faster than the same site within an ordered region. In support of this hypothesis, of the five phosphorylation sites structurally characterized before and after phosphorylation, four are disordered prior to phosphorylation and become ordered after the modification.[211]

In addition to roles in proteolysis and phosphorylation, the disordered loop of Bcl-2 also appears to be involved in binding small molecules. The anticancer drug taxol induces apoptosis and this induction is overcome by over-expression of Bcl-2.(refs [212,213]) Mobility shift experiments[214] and selection of taxol-binding peptides from a random library by phage display[215,216] both suggest that taxol binds to the disordered loop of Bcl-2. To our knowledge, these are the first experiments demonstrating the binding of a drug to an intrinsically disordered region of a protein.

Kinesin: The kinesin family of microtubule-based motor proteins is involved in oragnelle transport, mitosis, meiosis, and several other cellular processes. As shown by Rice and co-workers using EPR spectroscopy and cryo-electron microscopy, a region linking kinesin's two 'head' domains undergoes a nucleotide- and microtubule-dependent disorder-to-order transition during the force-generating cyle.[217] More specifically, the disorder-to-order transition of the bound head's linker region positions the unbound head over the next binding site on the microtubule, thus facilitating the stepping motion of the motor.

The region in question appears as both ordered and disordered in different crystal structures.[218,219,220] This structural variation is likely due to crystal packing forces as the various structures all have ADP bound in the active site. The corresponding region in myosin (the protein motor responsible for muscle contraction) also appears to undergo an analogous disorder-to-order transition during its force generating cycle[221,222] although it is not clear exactly what role this transition plays.

The field of molecular motors has traditionally been approached using a mechanical analogy, i.e. the idea that a motor protein is just a very small assembly of levers, dashpots, and springs. However, it has recently become clear from the results of Rice and co-workers.[217] as well as others,[223,224] that disorder, dynamic flexibility, and even phase changes between order and disorder are important in both motor function and regulation. Thus, even for motor proteins, models incorporating thermodynamics (e.g. including phase transitions) may be more appropriate than models based solely on mechanical analogies.

Titin: This gigantic protein extends from mid-line to Z-line in muscle cells and is suggested to fix the length of the muscle cell. This protein has a region rich in P + E + V + K that, in some isoforms, is greater than 1000 amino acids in length. The PEVK region is very likely to be structurally disordered and may

function as a rubber-like entropic spring that helps to restore over-stretched muscle cells to their natural, relaxed length.[225]

K^+ Channel ball and chain: Voltage-gated K^+ channels in nerve axons exist in three main states: closed, open and inactive. The closed state converts to the open state in response to an appropriate voltage gradient, while the inactive state is non-responsive to voltage gradients. Conversion from the open state to the inactive state happens when a "ball" on the end of a flexible "chain" plugs the open channel.[226,227,228] The timing of this closure is critical to proper nerve cell function and evidently depends on the length and flexibility of the chain. Here we suggest that this flexiblity be recognized as an entropic clock due to its important function as a timing device.

Neurofilament H and M: Fibrous assemblies of neurofilaments L, H and M are major components of motor neuron axons. The ends of the H and M filament proteins have long regions of disorder (300 and 600 residues, respectively). These disordered ends were proposed to be entropic bristles that occupy space by thermally driven motion and thereby maintain the separation of neighboring filaments.[229] The separation of filaments also maintains the axonal bore, possibly allowing the movement of small molecules and maintaining the shape of the axon against compression. Numerous other proteins contain such disordered regions that are proposed to function as entropic bristles for modulating protein structure and function.[230]

Implications of Intrinsic Disorder for the Protein Structure Hierarchy

Why has it taken so long for concepts related to functional disorder to make it into the textbook world of biochemistry? First, there is an approximately 100-year history based on the lock and key model[1] and an approximately 70-year history equating native proteins with 3D structure.[3,4] One term proposed to describe non-folded yet functional proteins, namely *natively denatured* [93] is an oxymoron, but nevertheless serves to illustrate just how deeply ingrained is the perceived equivalence between ordered 3D structure and native protein structure. Indeed, in common speech, the terms *native* and *(ordered) structure* are used interchangeably.

In addition to *natively denatured*, the terms *natively unfolded*[95] and *intrinsically unstructured*[43] have been suggested to describe these proteins. These two terms seem to imply lack of any secondary structure and so would exclude native molten globules. We are therefore using the term *intrinsically disordered* to describe such proteins.

Structural ensembles in equilibrium (e.g. intrinsic disorder) were associated with protein function 50 years ago.[19] Furthermore, prominent examples such as activation of trypsinogen and DNA recognition by the lac repressor likely involve disordered regions. Yet none of the widely used biochemical texts mentions even a single example of the role of disorder in protein function.

In our view, a second contributor to the slowness to recognize native protein disorder has been the current structural hierarchy taught in all the

biochemistry texts: e.g. that protein structure can be explained as primary, secondary, tertiary, and quaternary. In this hierarchy, there is no category called *intrinsic disorder*. As has been discussed from the time of the early Greek philosophers Plato and Aristotle and as reviewed and discussed from a more modern perspective by Lakoff,[231] categorization plays a key role in the development of knowledge. Since intrinsic disorder has not been included as a standard category of protein structure, textbook writers could not fit intrinsic disorder into their examples used to explain protein structure/function relationships.

The first three levels of the protein structural hierarchy (e.g., primary, secondary, tertiary) were developed in efforts to understand proteolytic digestion of proteins, several years before any 3D structures were known.[232] To our surprise, a formalized description of this hierarchy[233] included a category of proteins called *partially folded*. Furthermore, about 30 years ago there were several interesting discussions of such partially unfolded proteins,[31,32,135,141,234] but these concepts never made it into textbook biochemistry perhaps because of the rapid increase in the number of crystal structures being determined over the same time span.

Since intrinsic disorder and the transitions between protein states play important roles in native protein function, we have modified the first three steps of the current structural (Figure 8). In this figure, τ is used for turns and amphomorphic is used for segments that switch between α and β structure, and which may also represent a distinct category of secondary structure.[234a]

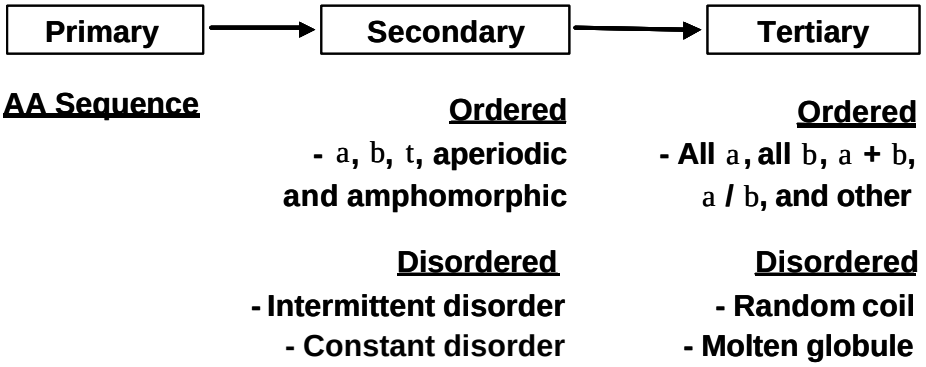

Figure 8: Implications for the protein structural hierarchy
The primary-secondary-tertiary structural hierarchy for proteins as found in all textbooks, has been to include intrinsic disorder as a category of structure

The proposed additions in figure 8 to include intrinsic disorder all involve time-dependent fluctuations or time-dependent changes in protein structure. Thus, Clare Woodward (discussion at this meeting) pointed out that these additions represent one way of adding a fourth dimension, time, to the currently static protein structure hierarchy.

First, at the secondary structural level we suggest two categories of disorder: intermittent and constant. The former describes protein molecular recognition domains and other regions that undergo order/disorder transitions to carry out binding or other functions. The latter describes regions such as

proteinaceous detergents, flexible linkers, entropic springs, entropic clocks and entropic bristles that remain disordered while carrying out function.

At the tertiary level, we suggest two categories: random coil and molten globule. With this altered hierarchy, it becomes possible to explain protein structure/function in terms of a diagram that includes not just *Amino Acid Sequence ® 3D Structure ® Function*, but also other structural forms and their transitions (Figure 9).

Shown in Figure 9 is conformational switching, where a segment of protein switches from α → β or from β → α; this idea was represented in figure 8 by the term *amphomorphic*. Secondary structure switching has been observed over a long period of time.[235,236,237,238] Recently, from a collection of such examples, a predictor of switch sequences was developed[239] and applied to myosin[240] in an effort to better understand the conformational changes in this molecule. Such conformational switching fits within The Protein Trinity's solid/liquid/gas analogy. Many solids have different phases and cooperatively switch from one to another under appropriate conditions.

Also included in Figure 9 are the already-discussed transitions from order → disorder and from disorder → order. From the examples given above, these transitions have been implicated as underlying important structure/function relationships. Figure 9 is essentially a re-drawing of Figure 1, The Protein Trinity, with many more details included.

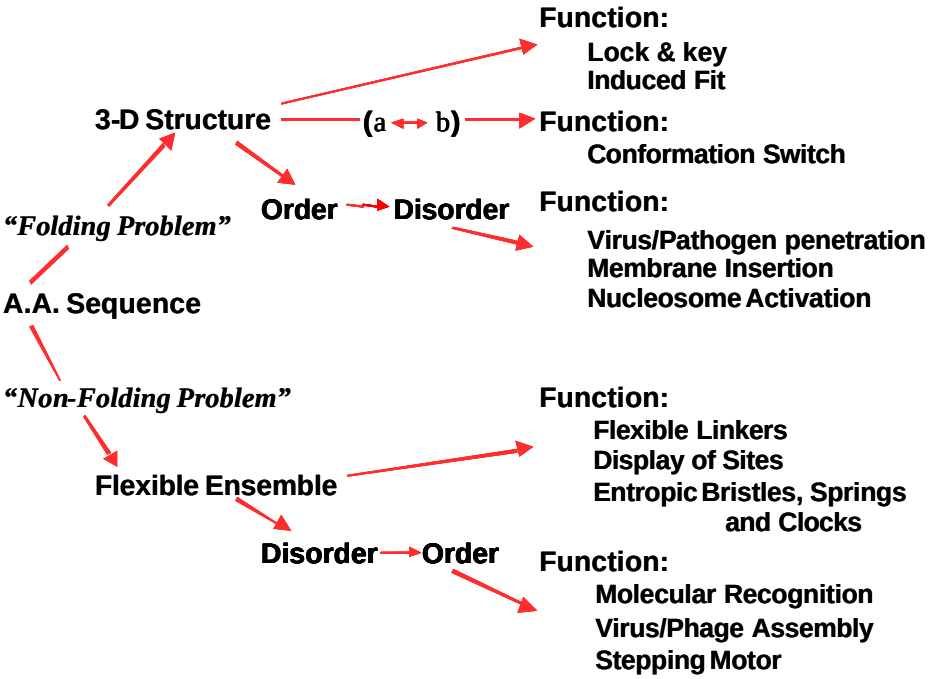

Figure 9: Protein structure / function relationships
The structure/function relationships presented in this figure are based on actual examples, many of which are given in Tables 1, 2, 4 and 5, but some of which have not been discussed in this review.

Databases of Intrinsic Order and Disorder

Given The Protein Trinity Hypothesis and the many examples of intrinsic disorder, it appears worthwhile to generate a database of proteins that have been characterized as intrinsically disordered. For purposes of comparison, it is useful to have a database of intrinsically ordered protein segments in the same format.

Databases: Table 6 summarizes ordered and disordered data accumulated to date. The ordered and disordered databases will be discussed in turn.

Table 6. Summary of Databases of Intrinsic Order and Disorder

Database	Number of Segments	Number of Residues
O_PDB_Select_25	1 111	220 668
dis XRAY	56	2 844
dis NMR	41	4 019
dis CD	53	10 554
dis ALL	150	17 417
dis Fam32	572	52 688

To obtain a set of ordered protein segments, we started with a non-redundant set of proteins from PDB called PDB_Select_25.[241] Sander and co-workers developed this dataset by grouping the proteins in PDB such that, within a group, every protein has > 25% identity to at least one other member of the group, and between any two groups, no pair of proteins has > 25% identity. Once the groups were determined, one member having the highest quality structure was chosen to represent each group.[241] For each of the representative proteins, every disordered (unobserved) residue was removed, yielding the ordered data set called *O_PDB_Select_25* (O for ordered). These ordered segments span all of the 3D structures known at the time this database was constructed.

The disordered sets are indicated as dis_X-ray, dis_NMR, and dis_CD according to the method used for the identification of disorder. The last group, dis_Fam32 was obtained by homology. A set of 32 proteins with disorder identified by a variety of means provided the starting points. Families of similar proteins were identified by sequence alignments using standard methods. Putative disorder in these proteins was then identified by alignment with the structurally characterized regions of disorder.

All of the disordered databases certainly contain segments having local tendencies for order. For example, segments with a high tendency for order are sometimes found to be binding sites within regions characterized as disordered.[242] Likewise, the ordered databases certainly contain disordered segments misclassified as ordered. For example, several of the proteins in the ordered dataset are from co-crystals with DNA; such DNA-binding proteins often have significant regions of disorder when separated from their ligand. We are attempting to develop methods based on sequence analysis to identify ordered and disordered regions that are misclassified.

Comparison of Ordered and Disordered Protein Segments

Given the set of ordered and disordered amino acids, it is useful to compare them. We have carried out two comparisons. First, we have compared the ordered and disordered segments with respect to their amino acid compositions. Second, we have compared these segments with respect to their sequence attribute values (e.g. hydropathy, hydrophobic moment, etc.).

Amino acid compositions: To compare the compositions of the four disordered datasets with each other and with ordered data, we expressed the composition of each amino acid in a given disordered dataset as (Disordered - Ordered) / (Ordered). Thus, negative peaks indicate that disordered segments are depleted compared to the ordered ones in the indicated amino acids and positive peaks indicate the reverse (Figure 10).

In Figure 10, the largest possible magnitude for a negative peak is minus 1 (e.g. for any amino acid that is completely missing from the given disordered region). In contrast, if a disordered region were composed exclusively of one amino acid, the positive peak value for that amino acid would be (100 – Ordered) / (Ordered). For an amino acid that comprised 5% of the residues in the protein, the positive peak would have a value of 19. Despite this potential for a very large asymmetry, the positive and negative peaks in Figure 10 have similar magnitude, with positive peak values ranging from 0.025 to 0.61 and negative values from -0.007 to -0.70.

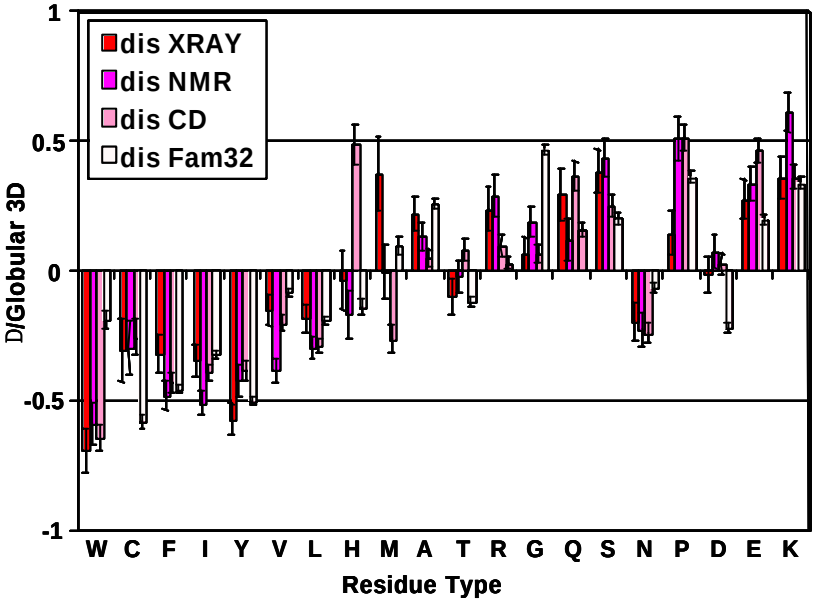

Figure 10: Amino acid composition profiles
The amino acid compositions of each disordered dataset in Table 6 are plotted as their differences from the compositions of the ordered dataset in this same table. Details are given in the text, but, briefly, a negative peak indicates that the given disorder dataset is depleted in that amino acid compared to the ordered set, while a positive peak indicates enrichment. The amino acids are on the basis of B-factor estimates of residue flexibility, with the more rigid on the left and the more flexible on the right. Such B-factor estimates are not purely related to intrinsic flexibility but also include environmental effects.

The arrangement of the amino acids from least to most flexible in this figure was based on the scale of Vihinen and co-workers,[243] where their scale was defined by the average residue B-factors of the backbone atoms for 92 unrelated proteins. For the proteins in their study, the backbone atoms of W had the lowest B-factors on average and the backbone atoms of K had the highest. As the developers of this scale pointed out, the ranking does not reflect intrinsic flexibility, in which case G would have the highest rank. Rather the ranking depends on the degree to which a given side chain tends to be buried (low ranking) or exposed (high ranking) in the crystal structure of globular proteins.

Overall, the four databases of intrinsic disorder are similar to each other for most of the amino acids.[125,244] All four are substantially depleted in W, C, F, I, Y, V, L and N. We proposed that these be called *order-promoting* amino acids. All four are substantially enriched in A, R, G, Q, S, P, E and K. We propose that these be called *disorder-promoting* amino acids. H, M, T and D are not consistently enriched or depleted among the four sets of intrinsically disordered proteins. So these are neither order-promoting nor disorder-promoting.

T, N and D stand out. All three are in a disorder-promoting neighborhood on the flexibility scale, yet T and N clearly promote order, not disorder, and D is ambivalent. These amino acids contain polar branches at the β-carbon and can readily form strong hydrogen bonds with the backbone. This side-chain/backbone interaction could lower the entropy of the disordered state, thus increasing the order-promoting tendencies of these residues compared to their neighbors in Figure 10.

Attributes: In addition to amino acid composition, the disordered segments have also been compared with the ordered ones by various attributes such as hydropathy, net charge, flexibility index, helix propensities, strand propensities, and compositions for groups of amino acids such as W + Y + F (aromaticity). The values for these attributes were estimated by simple averages over windows of 21 amino acids. The disordered data was balanced by an equal amount of randomly selected ordered data, and then conditional probability plots were constructed (Figure 11). The ability of the various attributes can be compared by dividing the area between the two curves by the total area of the graphs giving an area ratio (A.R.) value.[245]

Attribute ranking: Using the area ratio method to assess the ability to discriminate order and disorder, we have ranked 265 property-based attribute scales[244] plus more than 6,000 composition-based attributes[246] (e.g. all possible combinations having one to four amino acids in the group).

Many of the 265 scales and amino acids groupings are very highly correlated with each other. To find a set of such attributes having the best overall ability to distinguish between order and disorder, one needs to find a set of attributes that might not be the best individually, but that work best in combination. We have investigated various methods for obtaining optimal combinations of attributes, such as feedforward selection, branch and bound, and principal component analysis. Our previous studies[124,247,248,249,250] were carried out using much smaller sets of disorder than those we recently accumulated. Studies are in progress to find such optimal sets for the current, much larger data

sets of disorder. Presented in Table 7 is a representative, ranked set of attributes that provide fairly good discrimination between order and disorder.

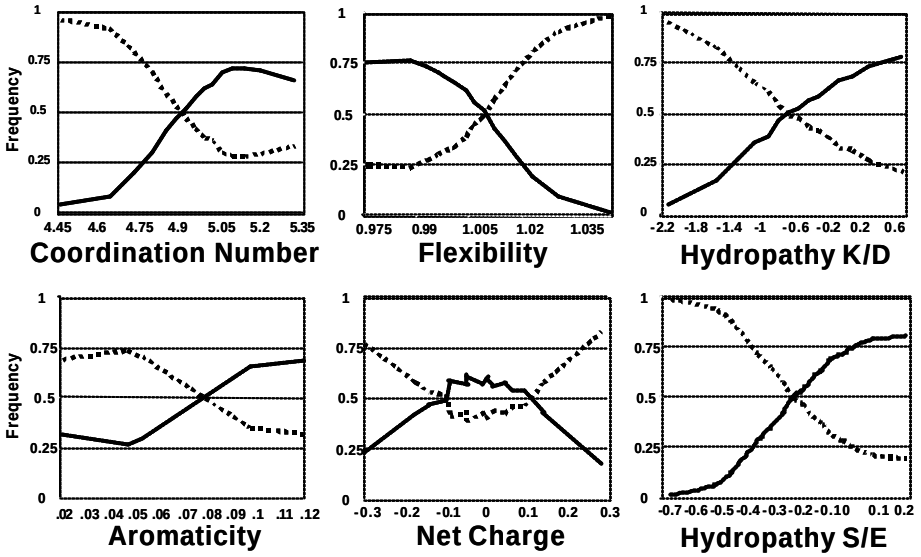

Figure 11: Sequence Attributes
Balanced datasets of ordered and disordered residues were used to construct conditional probability curves that indicate the ability of a given attribute to discriminate between order (solid line) and disorder (dashed line). The greater the separation between the two curves, the better the discrimination. The coordination number relates to how many side chains can pack around a given side chain. Flexibility is the scale used for figure 14. The two hydropathy scales are from Kyte and Doolittle[281] (K/D) and from Sweet and Eisenberg[280] (S/E). Aromaticity is simply the sum of W + Y + F in a given region. Net Charge is simply (K + R) minus (D + E); including H had essentially no effect on these curves.

Table 7. Ranking Properties for O / D Discrimination

Attribute	Area Ratio
14 Å contact number	0.542
Hydropathy[280]	0.535
Flexiblity[243]	0.521
β-sheet	0.514
Coordination number	0.478
Hydropathy[281]	0.421
RESP	0.334
Bulkiness	0.314
CFYW	0.285
Volume[282]	0.246
Refractivity	0.242
Net Charge	0.236

Commonness of Disordered Regions as Estimated by Sequence Analysis

From the current set of experimentally characterized proteins with disorder, experimental biases preclude any meaningful estimate of the commonness of intrinsic disorder and of the relative amounts of the random coil and molten globule varieties. PDB is biased against disordered protein, mainly because of two factors: 1. At the protein purification step, intrinsic disorder could be underrepresented due to hypersensitivity to protease digestion. 2. Once purified, proteins with significant fractions of disorder don't crystallize. On the other hand, as discussed above, NMR is biased against molten globular disorder. Finally, although use of both near and far UV CD in combination differentiates between random coil and molten globular disorder, the failure to routinely use near UV CD means that, in practice, disorder identified by CD is almost all of the random coil variety.

Structural characterization is much more tedious and expensive than determination of amino acid sequence. Furthermore, as discussed above, the structural characterizations are biased. Thus, the most efficient and perhaps least biased estimates of the commonness of intrinsic disorder are likely to come from sequence analysis (bioinformatics) methods as discussed below.

Computational methods for estimating intrinsic disorder or non-globularity: Below we will discuss two methods: the widely used SEG analysis by Wootton and co-workers[251,252,253,254] and our own efforts to develop predictors of disorder from amino acid sequence.[124,246,247,248,249,250]

Sequence complexity using Shannon's entropy: Both Karlin[255,256] and Wootton[251,252] realized that some protein segments have statistically significant deviations from the average amino acid composition, with fewer than the typical number of amino acids. Wootton and co-workers paid attention to protein structural characteristics of the unusual sequences, while Karlin and co-workers focused on the statistical aspects. We are following the approach of Wootton and co-workers because of their efforts to relate low complexity sequences to protein structure.

Specifically, Wootton and co-workers recognized that fibrous proteins have fewer amino acids per unit length of sequence than do ordered, globular proteins, and they suggested that low complexity sequences would be non-globular if not fibrous. To identify such sequences, they proposed use of Shannon's entropy, K2, which is given by $K2 = \Sigma (n_i/L)\log_2 (n_i/L)$, where the summation is over the 20 amino acids and where $n_i = 1$ if the ith amino acid is present in a window of length L and $n_i = 0$ if the ith amino acid is absent.

For windows of any length, the lower limit for K2 is 0, corresponding to 1 amino acid type at every position in the window. For an alphabet of 20 characters and windows with $L \leq 20$, the upper limit increases for longer windows, reaching a plateau at ~ 4.32 for $L \geq 20$. Using this measure, silks, collagens and coiled-coils exhibit lower complexity values than does a non-redundant database of protein segments derived from PDB (Figure 12), where all non-globular and disordered regions have been removed. This set of ordered segments is called Globular-3D.

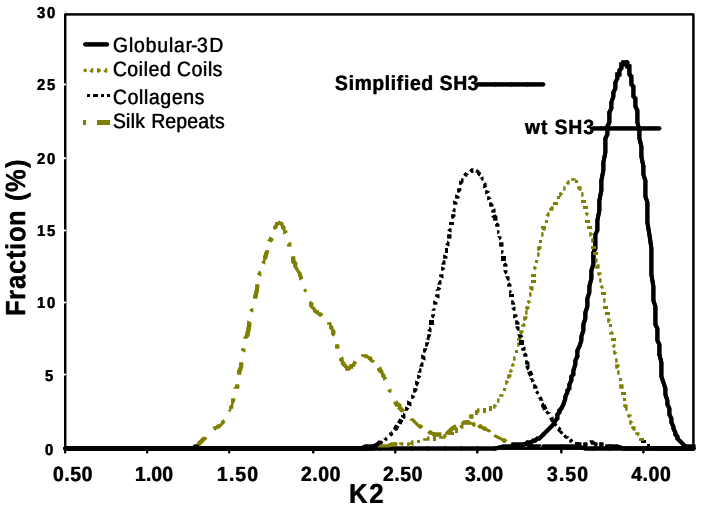

Figure 12: Sequence complexities
Smoothed histograms of K2 values for 4 sets of proteins: silks, collagens, coiled coils and globular proteins. The identities of the proteins in each of these sets can be found at our website: www.disorder.wsu.edu.

For complexity analysis, short windows exhibit large statistical variations and very long windows converge to the average complexity of the database. We, and Wootton and co-workers before us using different methods of analysis, both found that windows of 45 residues provide a reasonable compromise between the two extremes.

Examination of over 2.5 x 10^6 windows of 45-residue ordered segments[257] shows that no currently known ordered segment of this length from a globular protein contains fewer than ~ 10 different amino acids nor a K2 lower than ~ 2.9. Furthermore, using laboratory evolution to select low complexity sequences of the SH3 domain while maintaining its binding function as evidenced by its selection, Baker and co-workers developed a highly simplified version of this protein.[258] The complexity ranges of wild type SH3 and of simplified SH3 are indicated in Figure 12 as horizontal bars, showing that the laboratory evolution experiments succeeded in markedly simplifying this protein. Nevertheless, the simplified protein has nearly identical lower bounds for number of amino acids and sequence complexity as compared to the 2.5 x 10^6 windows for the current set of all known ordered proteins. The agreement between the laboratory selection and the sampling of natural selection suggests that about 10 amino acids and a K2 of about 2.9 may represent the lower bound for the complexity of globular, ordered protein structure.[257] Ordered helical bundles have been constructed with even lower sequence complexities, but these in our opinion are more like short segments of coiled coil rather than like true globular proteins.

Application of complexity analysis to the Swiss Protein database of protein sequences shows about 7% of all sequences have 45 residue windows with K2 values lower than 2.9 compared to 0.0% for ordered protein structure.[125] Furthermore, only a small fraction of these have sequence periodicities appropriate for one of the fibrous protein structures. Thus, there exists a class of almost certainly disordered protein structures having very low sequence complexity. We will show below that there is another class of disordered segments having high complexity. We speculate that sequences with low entropy

disorder may loosely correlate with random coils and that sequences with high entropy disorder with molten globules.

Prediction of Disorder: As another means for estimating the commonness and understanding the functions of intrinsically disordered proteins, we are developing predictors of natural disordered regions (PONDRs). As shown in Figure 13, predictor development and application consist of several steps: 1. Construct databases of order and disorder for training. 2. Compare sequence attributes of ordered and disordered segments to find those to be used as inputs for the predictors. 3. Partition ordered and disordered data into training and test sets. 4. Train the predictors using the training sets and test using the test sets (e.g., cross validation). 5. Validate the predictors on additional ordered and disordered data. 6. Apply the predictors to single protein sequences, to databases of both structure and sequence, and to genomic sets of sequences.

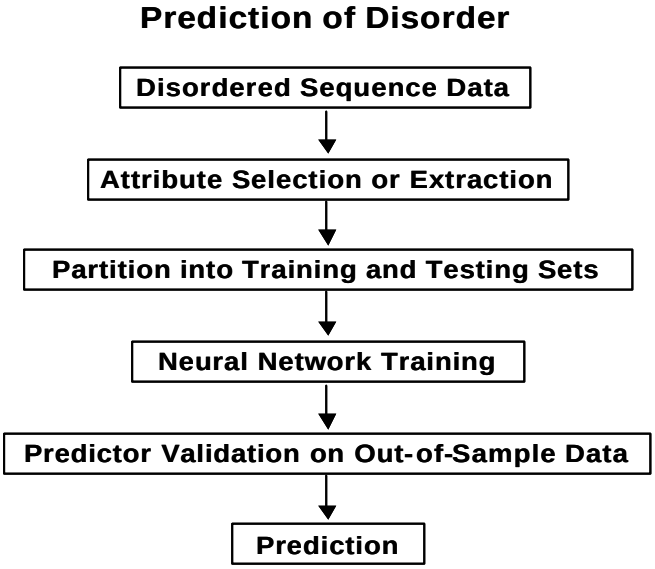

Figure 13: Steps for the development and application of PONDRs

The development and application of our predictors of natural protein disorder is shown schematically in this figure.

PONDRs: Given sets of ordered and disordered segments and their sequence characteristics as discussed above (steps 1 and 2), we have tried both linear (logistic regression and discriminant analysis) and non-linear (neural network) models as the basis for predictors of order and disorder. The nonlinear models are just slightly better than the linear ones. In our original work,[124] 10 attributes were selected for predictor development. The ordered and disordered data were randomly divided into five ordered and disordered subsets of equal size. Predictor training was carried out on 4/5 of the data and testing on 1/5. Over-prediction was avoided by dividing the 4/5 set into a training set (80% of the 4/5) and a validation set (the remaining 20%). When testing accuracy on the validation set started to diminish, training was stopped. It was then that testing was performed on the testing set. The data was then recombined and a new predictor was trained to the same level of accuracy. This original predictor is herein called PONDR XL1 (X for X-ray characterized disorder and L for long regions of disorder).

A more recent predictor was developed using the similar methods, but included a larger training set, with examples characterized by various methods.[125]

This predictor is called PONDR VL1 (V for variously characterized, L for long regions of disorder).

Since we used windows of 21 and then smoothed the outputs by averaging over windows of nine, our initial predictor failed to give outputs for the first and last 14 amino acids in a given sequence.[124] Therefore, we developed two predictors, one using training data from the amino terminal region and one from the carboxyl terminal regions.[249] These end-specific predictors, herein called XT (X for X-ray characterized data and T for their location at the termini) were later merged with the more recent internal regions predictor to give our current PONDR VL-XT that has outputs from the first to the last residue in a sequence.[125]

PONDR Accuracies: The datasets listed in Table 6 have not been studied for possible misclassification of ordered sequences as disordered. For the PONDRs developed to date, efforts were made to use disordered regions verified by more than one method and thereby reduce the amount of noise from misclassification. The disadvantage of such an approach is that the training sets have been small.

Table 8 gives the accuracies of the various PONDRs on several of the ordered and disordered databases. As mentioned above, PONDR XL1 used only X-ray-characterized disorder and the training set was very small (only about 500 disordered amino acids),[124] while PONDR VL1 used X-ray- and NMR-characterized disorder and the training set was larger but still relatively small (about 1300 disordered amino acids).[125] PONDR VL-XT involves a merger of the VL1 predictor with specific predictors for the two ends of the proteins.[249] For this predictor, two training accuracies are given. The value on the left is an average of the accuracies estimated for the first and last 10 residues; the value on the right is for the VL1 predictor. Finally, PONDR CaN was trained using the disordered regions from 12 calcineurin proteins.[247]

Table 8. PONDR Accuracies

	Predictors			
	XL1	VL1	VL-XT1	CaN
Residues	930	2298	8086	1720
Training	73 $\pm$ 4%	83.5 $\pm$ 2%	75.3 - 83.5%	83 $\pm$ 5%
O_PDB_S25	68	82	80	83
D_PDB_S25	49	44	63	33
NMR	52	59	64	31
CD	49	50	54	39
CaN Test	86	78	80	85

Note that all of the predictors exhibit similar accuracies on the ordered test set as observed during the training where 5-cross validation was used to estimate the accuracies. For the original VL1 predictor, both of these values are near 75% and are at 80% or above for the other predictors. The improved accuracy for VL1 as compared to XL1 probably relates to the increased size of the training set and the use of a better set of inputs.

Although the predictor accuracies on a large dataset of order matches the accuracies during training fairly well, there is a large drop in accuracy for each of

the predictors when applied to out-of-sample disordered data. Our interpretation of these results is that all types of order, whether helices, strands, turns, or aperiodic, are fairly similar to each other, whereas the disordered regions show a much larger variation. Thus, a predictor trained on one type of disorder does poorly when applied to another type.[250] This point is discussed more fully below in terms of a specific prediction example.

We are currently attempting to classify the various types of disorder into groups, or *flavors*, and then construct order/disorder predictors for each group. Although preliminary, this approach appears to have promise. Partitioning our disordered set into three flavors leads to prediction accuracies above 80% for both order and disorder within a given flavor, but with very low out-of-flavor accuracies (as expected). For an unknown protein, its membership in one of the three flavors can be established by comparing amino acid compositions.

Application to single proteins: One use is to PONDR single sequences, as shown in Figure 14A for the prion amino acid sequence and in Figure14B for 4E-BP1. For prion, two PONDRs were used, one based on a training set containing a collection of different proteins, e.g. VL-XT and one based on a training set containing long disordered regions from one family, calcineurin, with the disordered regions identified by homology (called CaN). For 4E-BP1, three PONDRs were used, VL-XT, CaN, and XL1.

Although PONDR CaN does as well as PONDR VL-XT for some proteins, for prion the CaN predictor misses the disordered region entirely. From several studies of this type, it is becoming clear that there are indeed a variety of different flavors of disorder.[242,250] Much of our current work is devoted to determining methods to classify disordered regions into flavors, developing predictors for each flavor, and using such predictors in attempts to understand flavor / function relationships.

Application to 4E-BP1 reveals predictions of order for all three PONDRs, where the predicted region of order is within a region characterized by NMR to be disordered (Figure 14B). The region predicted to be ordered by all three predictors corresponds approximately to a region of this protein responsible for binding to elongation factor EIF4E as indicated by the bar in the figure. As discussed above, this region undergoes a disorder-to-order transition upon binding.[174,175] PONDR identified a binding site in 4E-BP1 by predicting order within a region structurally characterized to be disordered. This has been observed for a few other proteins as well, but the data are currently too limited to propose this result as a generality.[248] Much more work is needed, but the prediction of order within a long, structurally characterized region of disorder has the promise of being a signal for the presence of a binding region that is independent of any particular sequence motif.

Commonness of disorder and low complexity: Although our predictors have a 20% error rate in the prediction of disorder for an ordered residue, the error rate drops significantly, to less than 0.01% for 40 consecutive predictions of disorder on segments known to be ordered. Thus, we are focusing our initial attention on such long disordered regions. To estimate the commonness of disordered regions of length L = 40 residues, we applied PONDR VL-XT to a

non-redundant set of PDB proteins and to the sequences in Swiss Protein. Sequence complexity analysis was also applied to these same sets of proteins to find segments with K2 = 2.9. The results are shown in Table 9.

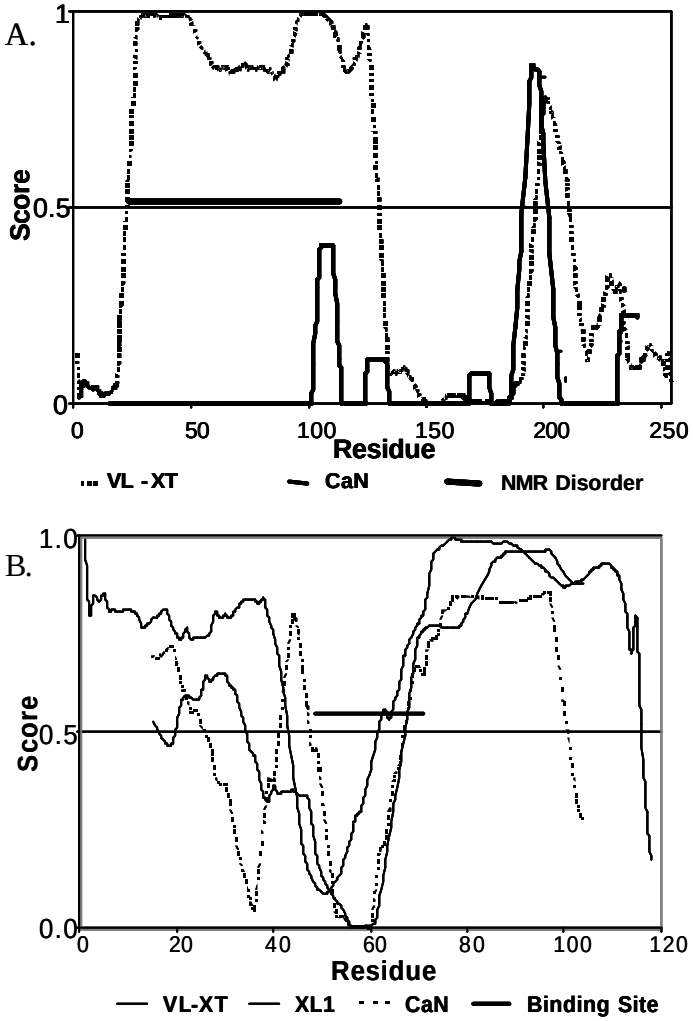

Figure 14: Application of various PONDRs to single sequences
In panel A, two PONDRs as indicated were applied to the mouse prion sequence and in B, three PONDRs as indicated were applied to 4E-BP1. For both panels, the ordinate shows the predictor output values (normalized to lie between 0 and 1) while the abscissa gives the residue number. Using the back-propagation learning algorithm, the PONDRs were taught to predict disorder by finding the weights that maximized the number of training-set disordered residues giving outputs > 0.5 while simultaneously maximizing the number of training-set ordered residues giving outputs < 0.5.

Table 9. Commonness of Intrinsic Disorder

Category*	PDB_Select_25 No. chains	PDB_Select_25 % chains	SwissProt No. chains	SwissProt % chains
Proteins >45 aa	920		81 005	
Low K2	6	0.7	5 748	7.1
Predicted LDRs	101	11.0	23 570	29.1
Extreme LDRs	5	0.5	6 291	7.8
Low K2, no LDR	2	0.2	881	1.1
LDR, no Low K2	97	10.5	18 703	23.1
Low K2, or LDR	103	11.2	24 451	30.2

For both data sets, there are fewer low complexity segments than segments predicted as disordered and nearly all the low complexity segments are also predicted as disordered.

Application of PONDR VL-XT suggests that 11% of the sequences in PDB_S_25 contain disordered segments. This value of 11% is much higher than the actual percentage of sequences in this structural database having disordered regions of L = 40 residues, suggesting that many of these predictions are errors. However, further examination shows that many of the apparent prediction errors are from segments that are involved with binding to DNA or to some other ligand. Such segments might undergo disorder-to-order transitions upon binding, in which case the prediction of disorder is correct, not an error. Thus, in the end, it is difficult to give an accurate estimate of errors in this analysis.

The current analysis predicts 29% of the sequences from Swiss Protein to contain at least one disordered region of 40 residues as compared to 11% for PDB. This difference is so large that, despite the various uncertainties, it is safe to conclude that PDB is a biased set of proteins with fewer regions of disorder as compared to the proteins in nature as represented by Swiss Protein. The obvious likely explanation is that the filter imposed by crystallization leads to exclusion of proteins with significant amounts of intrinsic disorder from the PDB. The protein with the largest amount of intrinsic disorder per unit cell in the current PDB is Bcl-x_L, with ~ 30% of this molecule being unobserved. Incidentally, the structure of this protein has been determined by both NMR and X-ray crystallography, with good overall agreement between the two methods in the assignment of disorder.[37]

Genomic Disorder

Probably the best way to estimate the commonness of intrinsic disorder is to predict disorder for the amino acids of whole genomes, including both the known and putative protein sequences. Such predictions could then be verified by laboratory experimentation.

We have carried out such predictions on 31 genomes, most of which are complete (Table 10). These data used two measures of disorder. The first measure of disorder was the percentage of sequences in each genome with segments predicted by PONDR VL-XT to have = 50 consecutive disordered residues. The second measure was the percentage of sequences in each genome to be wholly disordered, using a method based on cumulative distribution functions described in more detail elsewhere.[259]

To our surprise, the amount of predicted disorder shows a wide range among organisms from the 3 kingdoms, with the eucaryia exhibiting more disorder by these two measures than either the procaria or the archaea. For the first measure, the four eucaryia were predicted to have ~ 30-40% (!) of their proteins with disordered regions of length = 50 consecutive residues. As for the second measure, the four eucaryia were predicted to have ~ 6-17% of their chains to be wholly disordered like FlgM, 4E-BP1, HMG-I(Y) or neurogranin/neuromodulin. *Drosophila melanogaster* appears to have much more disorder than the other 3 eucaryia; this may be due to the incomplete nature of the Drosophila database, or to the physiology of an organism that undergoes several developmental stages.

Table 10. Genomic Disorder

Kingdom	Species	# seqs	Length ≥ 50		CDF*	
Archaea	*Methanococcus jannaschii*	1 714	71	4%	26	2%
Archaea	*Pyrococcus horikoshii*	2 062	164	8%	70	3%
Archaea	*Pyrococcus abyssi*	1 764	157	9%	62	4%
Archaea	*Archaeoglobus fulgidus*	2 402	244	10%	93	4%
Archaea	*Methanobacterium thermoautotrophicum*	1 869	365	20%	140	7%
Archaea	*Aeropyrum pernix K1*	2 694	637	24%	490	18%
Eubacteria	*Ureaplasma urealyticum*	611	14	2%	9	1%
Eubacteria	*Rickettsia prowazekii*	834	23	3%	5	1%
Eubacteria	*Borrelia burgdorferi*	845	26	3%	14	2%
Eubacteria	*Campylobacter jejuni*	2 309	80	3%	21	1%
Eubacteria	*Mycoplasma genitalium*	480	20	4%	10	2%
Eubacteria	*Helicobacter pylori*	1 532	69	5%	24	2%
Eubacteria	*Aquifex aeolicus*	1 522	94	6%	29	2%
Eubacteria	*Haemophilus influenzae*	1 708	126	7%	27	2%
Eubacteria	*Bacillus subtilis*	4 093	323	8%	87	2%
Eubacteria	*Escherichia coli*	4 281	363	8%	107	2%
Eubacteria	*Vibrio cholerae*	3 815	333	9%	93	2%
Eubacteria	*Mycoplasma pneumoniae*	675	60	9%	14	2%
Eubacteria	*Xylella fastidiosa*	2 761	246	9%	103	4%
Eubacteria	*Thermotoga maritima*	1 842	165	9%	53	3%
Eubacteria	*Neisseria meningitidis MC58*	2 015	190	9%	64	3%
Eubacteria	*Chlamydia pneumoniae*	1 052	100	10%	40	4%
Eubacteria	*Synechocystis sp*	3 167	338	11%	104	3%
Eubacteria	*Chlamydia trachomatis*	894	99	11%	42	5%
Eubacteria	*Treponema pallidum*	1028	115	11%	37	4%
Eubacteria	*Mycobacterium tuberculosis*	3916	747	19%	293	7%
Eubacteria	*Deinococcus radiodurans chr I*	2 580	534	21%	212	8%
Eucaryote	*Caenorhabditis elegans*	17 049	4 636	27%	1 322	8%
Eucaryote	*Arabodiopsis thaliana chrII, IV*	7 849	2 248	29%	653	8%
Eucaryote	*Saccharomyces cerevisiae*	6 264	1 858	30%	356	6%
Eucaryote	*Drosphila melanogaster*	13 885	5 651	41%	2 403	17%

*Sequences in each genome judged to be wholly disordered, using a method based on cumulative distribution functions.[262]

Neither of these estimates is corrected for false negative or false positive predictions. From Table 8, PONDR VL-XT seems to yield much larger false negative than false positive predictions. If this trend carries over to predictions on genomic sequences, then the 30% and 7% values for proteins with disordered regions of length = 50 and for proteins that are wholly disordered, respectively, are substantial under-estimates. These data suggest that intrinsic disorder is, indeed, a very common element of protein structure.

When considering a region of intrinsic disorder of length = 50, Figure 6C should be kept in mind. The calsequestrin monomer in this figure has 367 amino acids, yet an extended disordered tail of just 20 residues more than crosses the

entire diameter of the protein. An unfolded disordered segment of 50 residues would be long enough to occlude a very large fraction of the surface of a protein of this size.

Disorder and Intracellular Protease Digestion

Sensitivity to protease digestion is sometimes used as an argument against the existence of disordered proteins *in vivo*. To the contrary, there are so many cellular events controlled by protease digestion that intrinsic disorder is almost certainly a crucial (but, for the most part, unrecognized) component of proteolytic regulation.

Cells regulate their resident proteins through synthesis and degradation; together these two processes are called *turnover*. Schoenheimer[260] first recognized this dynamic behavior in 1942. In eucaryotic cells, individual proteins show characteristic half-lives, and for different intracellular proteins these half-lives vary over a wide range, from a few minutes to twenty days or more.[261] Studies such as these suggest a great deal of control and regulation.

A number of specific mechanisms for regulating intracellular proteolysis have come to light in recent years. A few examples include: regulation by the presence of certain sequences, such as regions rich in PEST;[262,263] regulation by the ubiquitin/proteosome system for digestion of unfolded protein;[264,265] and regulation by calcium levels via the calpain/calpastatin system.[266,267,268]

Specific protease digestion events utilizing one or more of the protease systems mentioned above are linked by a very large number of regulatory processes. To give a small sample from a very long list, events regulated by protease digestion include the well-known activation of various inactive precursors including peptide hormones,[269] the control of cell cycle progression,[270,271] the activation programmed cell death[205,272] and even the regulation of cholesterol content.[273]

Given the importance of flexibility for protease digestion,[87,274] we suggest another obvious point of control for proteolytic regulation, namely by modulating the order/disorder status of the protease sensitive site. Such modulation could be accomplished by phosphorylation/dephosphorylation, protein/ligand interactions, and protein/protein interactions, for example. With respect to the last possibility, the current view is that chaperone proteins exist for the purpose of mediating protein folding into the correct 3D structure.[275,276] From the point of view of intrinsic disorder, the true functions of some of these proteins might be to modulate accessibility to loci important for protease control. For example, calmodulin is substantially disordered in the absence of calcium and so might be expected to be digested via the ubiquitin/proteosome system. However, as discussed above, a collection of proteins with IQ motifs bind to calmodulin at low calcium. Perhaps the function of this association is not only for storage as was suggested previously[185] but also to protect the unfolded calmodulin from protease digestion.

Implications for Structural Genomics and Proteomics

The first grants to support consortia for high throughput protein structure determination have now been awarded, thus marking the actual start of structural genomics.[277,278] This effort is squarely based on the paradigm: *Amino Acid Sequence ® 3D Structure ® Function*. Of course the goal of these efforts is to use structural information to help catalogue the functions of the protein in the human and other genomes. Given the information and discussions provided in this review, which are from a similar perspective as information and discussions published previously,[279] it is evident that the structural genomics project will be very seriously incomplete unless a more systematic effort is made to discover and study intrinsically disordered proteins.

Summary

Intrinsic protein disorder is very likely encoded by the amino acid sequence and represents a common feature of native proteins. Thus, disorder ought to be recognized as a distinct category of protein structure. Proteins with intrinsic disorder can bind permanently with a partner and thereby exist as a structured protein throughout their lifetimes. Proteins with intrinsic disorder can undergo disorder-to-order transitions upon binding with one or more partners and thus be structured part of the time and unstructured part of the time. Finally, proteins with intrinsic disorder can in some cases carry out function without ever becoming ordered and thus remain disordered throughout their existence. In order to account for all of these possibilities, a new paradigm for protein structure/function is needed; we therefore propose The Protein Trinity as illustrated and explained in Figure 1.

Acknowledgements

We first wish to thank Leslie Kuhn and Michael Thorpe of Michigan State University for organizing such an interesting workshop that provided an outstanding context for this paper, and also Irwin Kuntz, Clare Woodward, Clay Bracken, and Richard Goldstein whose comments at the meeting helped improve our understanding of our own work. Next, Charles DeLisi and co-workers Sandor Vajda, Rakefet Rosenfeld, and, especially, Zhiping Wang, are thanked for helping us avoid mistakes in the interpretation of their very important paper on the energetics of flexible ligand binding. Ya-Yue Van, CEO of Molecular Kinetics, is thanked for suggesting "Protein Trinity" to replace "Proposal Protein Structure/Function Diagrams." Finally, we want to thank NIH, NSF, DOE and the MRC of Canada for the various grants that enabled the authors to meet and collaborate on the study of intrinsic protein disorder, and especially for NIH-R01-LM06916, NSF-CSE-II-9711532 and NSF-REU-IID-9711532, which were awarded expressly for the purpose of studying intrinsic disorder.

REFERENCES

1. Fischer, E. Einfluss der configuration auf die wirkung der enzyme. *Ber. Dt. Chem. Ges.* 1894, **27**, 2985-2993.
2. Lemieux, U.R. and Spohr, U. How Emil Fischer was led to the lock and key concept for enzyme specificity. *Adv. Carbohydrate Chem. Biochem.* 1994, **50**, 1-20.
3. Mirsky, A.E. and Pauling, L. On the structure of native, denatured and coagulated proteins. *Proc. Natl. Acad. Sci. USA* 1936, **22**, 439-447.
4. Northrop, H.J. Crystalline Pepsin. I. Isolation and tests of purity. *J. Gen. Physiol.* 1930, **13**, 739-766.
5. Anson, M.L. and Mirsky, A.E. The effect of denaturation on the viscosity of protein systems. *J. Gen. Physiol.* 1932, **15**, 341-350.
6. Wu, H. Studies on denaturation of proteins XIII A theory of denaturation. *Chinese J. Physiol.* 1931, **1**, 219-234.

6A. Edsall, J.T. Hsien Wu and the First Theory of Protein Denaturation (1931). *Advances in Protein Chemistry.* 1995, **46**, 1-5.

7. Anson, M.L., Protein denaturation and the properties of protein groups. In *Advances in Protein Chemistry*, Anson, M.L. and Edsall, J.T., Eds., Academic Press, New York, 1945, pp. 361-384.
8. Edsall, J.T. Some comments on proteins and protein structure. *Proc. R. Soc. Lond.* 1952, **B147**, 97-103.
9. Phillips, D.C. Development of concepts of protein structure. *Perspectives in Biology and Medicine* 1986, **29**, S124-S130.
10. Sela, M., White, F.H., and Anfinsen, C.B. Reductive cleavage of disulfide bridges in ribonuclease. *Science* 1957, **125**, 691-692.
11. Anfinsen, C.B. Principles that govern the folding of protein chains. *Science* 1973, **181**, 223-230.
12. Kendrew, J.C., Dickerson, R.E., and Strandberg, B.E. Structure of myoglobin: a three-dimensional Fourier synthesis at 2 Å resolution. *Nature* 1960, **206**, 757-763.
13. Blake, C.C., Koenig, D.F., Mair, G.A., North, A.C., Phillips, D.C., and Sarma, V.R. Structure of hen egg-white lysozyme. A three-dimensional Fourier synthesis at 2 Å resolution. *Nature* 1965, **206**, 757-761.
14. Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., and Bourne, P.E. The protein data bank. *Nucleic Acids Res.* 2000, **28**, 235-242.
15. Frishman, D. and Mewes, H.W. Protein structural classes in five complete genomes. *Nat. Struct. Biol.* 1997, **4**, 626-628.
16. Frishman, D. and Mewes, H.-W. PEDANTic genome analysis. *Trends Genetics* 1997, **13**, 416-417.
17. Gerstein, M. How representative are the known structures of the proteins in a complete genome? A comprehensive structural census. *Fold. Des.* 1998, **3**, 497-512.
18. Gerstein, M. and Hegyi, H. Comparing genomes in terms of protein structure: surveys of a finite parts list. *FEMS Microbiol. Rev.* 1998, **22**, 277-304.

19. Karush, F. Heterogeneity of the binding sites of bovine serum albumin. *J. Am. Chem. Soc.* 1950, **72**, 2705-2713.
20. Lund, M., Bjerrum, O.J., and Bjerrum, M.J. Structural heterogeneity of the binding sites of HSA for phenyl-groups and medium-chain fatty acids. Demonstration of equilibrium between different binding conformations. *Eur. J. Biochem.* 1999, **260**, 470-476.
21. Koshland Jr., D.E. Application of a theory of enzyme specificity to protein synthesis. *Proc. Natl. Acad. Sci. USA* 1958, **44**, 98-104.
22. Bennett, W.S., Jr. and Steitz, T.A. Glucose-induced conformational change in yeast hexokinase. *Proc. Natl. Acad. Sci. USA* 1978, **75**, 4848-4852.
23. McDonald, R.C., Steitz, T.A., and Engelman, D.M. Yeast hexokinase in solution exhibits a large conformational change upon binding glucose or glucose 6-phosphate. *Biochemistry* 1979, **18**, 338-342.
24. Koshland Jr., D.E., The key-lock theory and the induced fit theory. *Angew. Chem. Int. Ed. Engl.* 1994, **33**, 2375-2378.
25. Kawaguchi, S., Nobe, Y., Yasuoka, J., Wakamiya, T., Kusumoto, S., and Kuramitsu, S. Enzyme flexibility: a new concept in recognition of hydrophobic substrates. *J. Biochem. (Tokyo)* 1997, **122**, 55-63.
26. Monod, J., Wyman, J., and Changeux, J.P. On the nature of allosteric transitions: A plausible model. *J. Mol. Biol.* 1965, **12**, 88-118.
27. Koshland Jr, E.D., Nemethy, G., and Flmer, D. Comparison of experimental binding data and theoretical models in proteins containing subunits. *Biochemistry* 1966, **5**, 365-385.
28. Bloomer, A.C., Champness, J.N., Bricogne, G., Staden, R., and Klug, A. Protein disk of tobacco mosiac virus at 2.8Å resolution showing the interactions within and between subunits. *Nature* 1978, **276**, 362-368.
29. Bode, W., Schwager, P., and Huber, R. The transition of bovine trypsinogen to a trypsin-like state upon strong ligand binding. The refined crystal structures of the bovine trypsinogen-pancreatic trypsin inhibitor complex and of its ternary complex with Ile-Val at 1.9Å resolution. *J. Mol. Biol.* 1978, **118**, 99-112.
30. Huber, R. Conformational flexibility and its functional significance in some protein molecules. *TIBS* 1979, **4**, 271-276.
31. Schulz, G.E., Nucleotide Binding Proteins, in *Molecular Mechanism of Biological Recognition*, ed. Balaban, M. (New York: Elsevier/North-Holland Biomedical Press, 1979), 79-94.
32. Alber, T., Gilbert, W.A., Ponzi, D.R., and Petsko, G.A. The role of mobility in the substrate binding and catalytic machinery of enzymes. *Ciba Found. Symp.* 1982, **93**, 4-24.
33. Spolar, R.S. and Record II, M.T. Coupling of local folding to site-specific binding of proteins to DNA. *Science* 1994, **263**, 777-784.
34. Lewis, M., Chang, G., Horton, N.C., Kercher, M.A., Pace, H.C., Schumacher, M.A., Brennan, R.G., and Lu, P. Crystal structure of the lactose operon repressor and its complexes with DNA and inducer. *Science* 1996, **271**, 1247-1254.

35. Dunker, A., Obradovic, Z., Romero, P., Kissinger, C., and Villafranca, E. On the importance of being disordered. *PDB Newsletter* 1997, **81**, 3-5.
36. Aviles, F.J., Chapman, G.E., Kneale, G.G., Crane-Robinson, C., and Bradbury, E.M. The conformation of histone H5. Isolation and characterisation of the globular segment. *Eur. J. Biochem.* 1978, **88**, 363-371.
37. Muchmore, S.W., Sattler, M., Liang, H., Meadows, R.P., Harlan, J.E., Yoon, H.S., Nettesheim, D., Chang, B.S., Thompson, C.B., Wong, S.L., Ng, S.L., and Fesik, S.W. X-ray and NMR structure of human Bcl-x_L, an inhibitor of programmed cell death. *Nature* 1996, **381**, 335-341.
38. Riek, R., Hornemann, S., Wider, G., Billeter, M., Glockshuber, R., and Wuthrich, K. NMR structure of the mouse prion protein domain PrP(121-321). *Nature* 1996, **382**, 180-182.
39. Kriwacki, R.W., Hengst, L., Tennant, L., Reed, S.I., and Wright, P.E. Structural studies of $p21^{Waf1/Cip1/Sdi1}$ in the free and Cdk2-bound state: conformational disorder mediates binding diversity. *Proc. Natl. Acad. Sci. USA* 1996, **93**, 11504-11509.
40. Daughdrill, G.W., Chadsey, M.S., Karlinsey, J.E., Hughes, K.T., and Dahlquist, F.W. The C-terminal half of the anti-sigma factor, FlgM, becomes structured when bound to its target, sigma 28. *Nat. Struct. Biol.* 1997, **4**, 285-291.
41. Fletcher, C.M. and Wagner, G. The interaction of eIF4E with 4E-BP1 is an induced fit to a completely disordered protein. *Protein Sci.* 1998, **7**, 1639-1642.
42. Plaxco, K.W. and Gross, M. The importance of being unfolded. *Nature* 1997, **386**, 657, 659.
43. Wright, P.E. and Dyson, H.J. Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. *J. Mol. Biol.* 1999, **293**, 321-331.
44. Tanford, C. Protein denaturation. *Adv. Protein Chemistry* 1968, **23**, 121-282.
45. Kuwajima, K., Nitta, K., Yoneyama, M., Surgai, S. Three-state denaturation of alpha-lactalbumin by guanidine hydrochloride. *J. Mol. Biol.* 1976, **106**, 359-373.
46. Myer, Y.P. Conformation of cytochromes. III. Effect of urea, temperature, extrinsic ligands, and pH variation on the conformation of horse heart ferricytochrome c. *Biochemistry* 1968, **7**, 765-776.
47. Dolgikh, D.A., Gilmanshin, R.I., Brazhnikov, E.V., Bychkova, V.E., Semisotnov, G.V., Venyaminov, S., and Ptitsyn, O.B. Alpha-Lactalbumin: compact state with fluctuating tertiary structure? *FEBS Lett.* 1981, **136**, 311-315.
48 Goto, Y. and Fink, A.L. Conformational states of beta-lactamase: molten-globule states at acidic and alkaline pH with high salt. *Biochemistry* 1989, **28**, 945-952.
49. Gast, K., Damaschum, H., Misselwitz, R., Muller-Frohne, M., Zirwer, D., and Damaschen, G. Compactness of protein molten globules:

Temperature-induced structural changes of the apomyoglobin folding intermediate. *Eur. Biophys. J.* 1994, **23**, 297-305.

50. Nishii, I., Kataoka, M., Tokunaga, F., and Goto, Y. Denaturation of the moltern globule states of apomyoglobin and a profile for protein folding. *Biochemistry* 1994, **33**, 4903-4909.
51. Ohgushi, M. and Wada, A. 'Molten-globule state': a compact form of globular proteins with mobile side-chains. *FEBS Lett.* 1983, **164**, 21-24.
52. Kim, P.S. and Baldwin, R.L. Specific Intermediates in the folding reactions of small proteins and the mechanism of protein folding. *Annu. Rev. Biochem.* 1982, **51**, 459-489.
53. Kuwajima, K. The molten globule state as a clue for understanding the folding and cooperativity of globular-protein structure. *Proteins* 1989, **6**, 87-103.
54. Loh, S.N., Kay, M.S., and Baldwin, R.L. Structure and stability of a second molten globule intermediate in the apomyoglobin folding pathway. *Proc. Natl. Acad. Sci. USA* 1995, **92**, 5446-5450.
55. Baum, J., Dobson C.M, Evans P.A., Hanley C. Characterization of a partly folded protein by NMR methods: studies on the molten globule state of guinea pig alpha-lactalbumin. *Biochemistry* 1989, **28**, 7-13.
56. Woodward, C., Barbar, E., Carrula, N., Battiste, J., and Barany, G. Hydrogen exchange and protein folding. *J. Mol. Graphics* 2001, This volume.
57. Kim, P.S. and Baldwin, R.L. Intermediates in the folding reactions of small proteins. *Ann. Rev. Biochem.* 1990, **59**, 631-660.
58. Barrick, D. and Baldwin, R.L. Stein and Moore Award address. The molten globule intermediate of apomyoglobin and the process of protein folding. *Protein Sci* 1993, **2**, 869-876.
59. Peng, Z.Y. and Kim, P.S. A protein dissection study of a molten globule. *Biochemistry* 1994, **33**, 2136-2141.
60. Ptitsyn, O.B. and Uversky, V.N. The molten globule is a third thermodynamical state of protein molecules. *FEBS Lett.* 1994, **341**, 15-18.
61. Kuwajima, K. The molten globule state of alpha-lactalbumin. *FASEB J.* 1996, **10**, 102-109.
62. Bychkova, V.E., Pain, R.H., and Ptitsyn, O.B. The 'molten globule' state is involved in the translocation of proteins across membranes? *FEBS Lett.* 1988, **238**, 231-234.
63. Bychkova, V.E., Berni, R., Rossi, G.L., Kutyshenko, V.P., and Ptitsyn, O.B. Retinol-binding protein is in the molten globule state at low pH. *Biochemistry* 1992, **31**, 7566-7571.
64. Bychkova, V.E., Dujsekina, A.E., Fantuzzi, A., Ptitsyn, O.B., and Rossi, G.L. Release of retinol and denaturation of its plasma carrier, retinol-binding protein. *Fold Des* 1998, **3**, 285-291.
65. Bychkova, V. and Ptitsyn, O. The molten globule in vitro and in vivo. *Chemtracts - Biochem. Mol. Biol.* 1993, **4**, 133-163.

66. Seeley, S.K., Weis, R.M., and Thompson, L.K. The cytoplasmic fragment of the aspartate receptor displays globally dynamic behavior. *Biochemistry* 1996, **35**, 5199-5206.
67. Gursky, O. and Atkinson, D. Thermal unfolding of human high-density apolipoprotein A-1: implications for a lipid-free molten globular state. *Proc. Natl. Acad. Sci. USA* 1996, **93**, 2991-2995.
68. Carroll, A.S., Gilbert, D.E., Liu, X., Cheung, J.W., Michnowicz, J.E., Wagner, G., Ellenberger, T.E., and Blackwell, T.K. SKN-1 domain folding and basic region monomer stabilization upon DNA binding. *Genes Dev*. 1997, **11**, 2227-2238.
69. Zurdo, J., Sanz, J.M., Gonzalez, C., Rico, M., and Ballesta, J.P. The exchangeable yeast ribosomal acidic protein YP2β shows characteristics of a partly folded state under physiological conditions. *Biochemistry* 1997, **36**, 9625-9635.
70. Zhang, J. and Matthews, C.R. Ligand binding is the principal determinant of stability for the p21(H)- ras protein. *Biochemistry* 1998, **37**, 14881-14890.
71. Quintas, A., Saraiva, M.J., and Brito, R.M. The tetrameric protein transthyretin dissociates to a non-native monomer in solution. A novel model for amyloidogenesis. *J. Biol. Chem.* 1999, **274**, 32943-32949.
72. Huber, R. and Bennett, W.S., Jr. Functional significance of flexibility in proteins. *Biopolymers* 1983, **22**, 261-279.
73. Douzou, P. and Petsko, G.A. Proteins at work: "stop-action" pictures at subzero temperatures. *Adv. Protein Chem.* 1984, **36**, 245-361.
74. Ishima, R. and Torchia, D.A. Protein dynamics from NMR. *Nat Struct Biol* 2000, **7**, 740-743.
75. Bracken, C. NMR spin relaxation methods for characterization of disorder and folding in proteins. *J. Mol. Graphics* 2001, This volume.
76. Evans, J.N.S., *Biomolecular NMR Spectroscopy* Oxford University Press, Oxford, 1995.
77. Smith, L.J., Dobson, C.M., and van Gunsteren, W.F. Side-chain conformational disorder in a molten globule: molecular dynamics simulations of the A-state of human alpha-lactalbumin. *J. Mol. Biol.* 1999, **286**, 1567-1580.
78. Bai, P., Luo, L., and Peng, Z. Side chain accessibility and dynamics in the molten globule state of alpha-lactalbumin: a (19)F-NMR study. *Biochemistry* 2000, **39**, 372-380.
79. Eliezer, D., Yao, J., Dyson, H.J., and Wright, P.E. Structural and dynamic characterization of partially folded states of apomyoglobin and implications for protein folding. *Nat. Struct. Biol.* 1998, **5**, 148-155.
80. Eliezer, D., Chung, J., Dyson, H.J., and Wright, P.E. Native and non-native secondary structure and dynamics in the pH 4 intermediate of apomyoglobin. *Biochemistry* 2000, **39**, 2894-2901.
81. Fasman, G.D., *Circular Dichroism and the Conformational Analysis of Biomolecules* Plenem Press, New York, 1996.

82. Kuwajima, K. A folding model of alpha-lactalbumin deduced from the three-state denaturation mechanism. *J. Mol. Biol.* 1977, **114**, 241-258.
83. Linderstrom-Lang, K. Structure and enzymatic break-down of proteins. *Cold Spring Harbor Symp. Quant. Biol.* 1949, **14**, 117-126.
84. Markus, G. Protein Substrate Conformation and Proteolysis. *Proc. Natl. Acad. Sci. USA* 1965, **54**, 253-258.
85. Fontana, A., Polverino de Laureto, P., and De Filippis, V., Molecular Aspects of Proteolysis of Globular Proteins. In: *Protein Stability and Stabilization*, van den Tweel, W., Harder, A., and Buitelear, M., Eds., Elsevier Science Publ., Amsterdam, 1993, pp. 101-110.
86. Fontana, A., Zambonin, M., Polverino de Laureto, P., De Filippis, V., Clementi, A., and Scaramella, E. Probing the conformational state of apomyoglobin by limited proteolysis. *J. Mol. Biol.* 1997, **266**, 223-230.
87. Hubbard, S.J., Eisenmenger, F., and Thornton, J.M. Modeling studies of the change in conformation required for cleavage of limited proteolytic sites. *Protein Sci.* 1994, **3**, 757-768.
88. Hubbard, S.J., Beynon, R.J., and Thornton, J.M. Assessment of conformational parameters as predictors of limited proteolytic sites in native protein structures. *Protein Eng.* 1998, **11**, 349-359.
89. Honig, B. and Yang, A.S. Free energy balance in protein folding. *Adv. Protein Chem.* 1995, **46**, 27-58.
90. Manalan, A.S. and Klee, C.B. Activation of calcineurin by limited proteolysis. *Proc. Natl. Acad. Sci. USA* 1983, **80**, 4291-4295.
91. Kissinger, C.R., Parge, H.E., Knighton, D.R., Lewis, C.T., Pelletier, L.A., Tempczyk, A., Kalish, V.J., Tucker, K.D., Showalter, R.E., Moomaw, E.W., Gastinel, L.N., Habuka, N., Chen, X., Maldanado, F., Barker, J.E., Bacquet, R., and Villafranca, J.E. Crystal structures of human calcineurin and the human FKBP12-FK506- calcineurin complex. *Nature* 1995, **378**, 641-644.
92. Iakoucheva, L.M., Kimzey, A.L., Masselon, C.D., Bruce, J.E., Garner, E.C., Brown, C.J., Dunker, A.K., Smith, R.D., and Ackerman, E.J. Identification of intrinisc order and disorder in the DNA damage-recognition protein XPA by time-resolved proteolysis, prediction from sequence, and Fourier transform ion cyclotron resonance mass spectrometry. *Protein Science* Submitted.
93. Schweers, O., Schonbrunn-Hanebeck, E., Marx, A., and Mandelkow, E. Structural studies of tau protein and Alzheimer paired helical filaments show no evidence for beta-structure. *J. Biol. Chem.* 1994, **269**, 24290-24297.
94. Kriwacki, R.W., Wu, J., Tennant, L., Wright, P.E., and Siuzdak, G. Probing protein structure using biochemical and biophysical methods. Proteolysis, matrix-assisted laser desorption/ionization mass spectrometry, high-performance liquid chromatography and size-exclusion chromatography of $p21^{Waf1/Cip1/Sdi1}$. *J. Chromatogr. A* 1997, **777**, 23-30.

95. Weinreb, P.H., Zhen, W., Poon, A.W., Conway, K.A., and Lansbury, P.T., Jr. NACP, a protein implicated in Alzheimer's disease and learning, is natively unfolded. *Biochemistry* 1996, **35**, 13709-13715.
96. Ptitsyn, O.B. Protein Folding: Hypotheses and Experiments. *J. Protein Chem.* 1987, **6**, 273-293.
97. Ptitsyn, O. Molten globule and protein folding. *Adv. Protein Chem.* 1995, **47**, 83-229.
98. Handel, T.M., Williams, S.A., and DeGrado, W.F. Metal ion-dependent modulation of the dynamics of a designed protein. *Science* 1993, **261**, 879-885.
99. Betz, S.F. and DeGrado, W.F. Controlling topology and native-like behavior of de novo-designed peptides: design and characterization of antiparallel four-stranded coiled coils. *Biochemistry* 1996, **35**, 6955-6962.
100. Ringe, D. and Petsko, G. Study of protein dynamics by X-ray diffraction. *Methods Enzymol.* 1986, **131**, 389-433.
101. Parthasarathy, S. and Murthy, M.R. Protein thermal stability: insights from atomic displacement parameters (B values). *Protein Eng.* 2000, **13**, 9-13.
101A. McCammon, J.A., Gelin, B.R., and Karplus, M. Dynamics of folded proteins. *Nature* 1977, **267**, 585-90.
102. Griffith, J., Manning, M., and Dunn, K. Filamentous bacteriophage contract into hollow spherical particles upon exposure to a chloroform-water interface. *Cell* 1981, **23**, 747-753.
103. Manning, M., Chryosogelos, S., and Griffith, J. Mechanism of coliphage M13 contraction: intermediate structures trapped at low temperatures. *J. Virol.* 1981, **40**, 912-991.
104. Manning, M., Chrysogelos, S., and Griffith, J. Insertion of bacteriophage M13 coat protein into membranes. *Biophys. J.* 1982, **37**, 28-30.
105. Manning, M. and Griffith, J. Association of M13 I-forms and spheroids with lipid vesicles. *Arch. Biochem. Biophys.* 1985, **236**, 297-303.
106. Roberts, L.M. and Dunker, A.K. Structural changes accompanying chloroform-induced contraction of the filamentous phage fd. *Biochemistry* 1993, **32**, 10479-10488.
107. Dunker, A.K., Ensign, L.D., Arnold, G.E., and Roberts, L.M. A model for fd phage penetration and assembly. *FEBS Lett.* 1991, **292**, 271-274.
108. Dunker, A.K., Ensign, L.D., Arnold, G.E., and Roberts, L.M. Proposed molten globule intermediates in fd phage penetration and assembly. *FEBS Lett.* 1991, **292**, 275-278.
109. Ausio, J., Dong, F., and van Holde, K.E. Use of selectively trypsinized nucleosome core particles to analyze the role of the histone "Tails" in the stabilization of the nucleosome. *J. Mol. Biol.* 1989, **206**, 451-463.
110. Luger, K., Mader, A.W., Richmond, R.K., Sargent, D.F., and Richmond, T.J. Crystal structure of the nucleosome core particle at 2.8 A resolution. *Nature* 1997, **389**, 251-260.
111. Luger, K. and Richmond, T.J. The histone tails of the nucleosome. *Curr. Opin. Genet. Dev.* 1998, **8**, 140-146.

112. Waterborg, J.H. and Matthews, H.R. Intranuclear localization of histone acetylation in *Physarum polycephalum* and the structure of functionally active chromatin. *Cell Biophys.* 1983, **5**, 265-279.
113. Chahal, S., S., Matthews, H.R., and Bradbury, E.M. Acetylation of histone H4 and its role in chromatin structure and function. *Nature* 1980, **287**, 76-79.
114. Mizzen, C.A. and Allis, C.D. Linking histone acetylation to transcriptional regulation. *Cell Mol. Life Sci.* 1998, **54**, 6-20.
115. Ausio, J. and van Holde, K.E. Histone hyperacetylation: its effects on nucleosome conformation and stability. *Biochemistry* 1986, **25**, 1421-1428.
116. Imai, B., S., yau, P., Baldwin, J., P., Ibel, K., May, R., P., and Bradbury, E., Morton Hyperacetylation of core histones does not cause unfolding of nucleosomes. *J. Biol. Chem.* 1986, **261**, 8784-8792.
117. Oliva, R., Bazett-Jones, D.P., Locklear, L., and Dixon, G.H. Histone hyperacetylation can induce unfolding of the nucleosome core particle. *Nucleic Acids Res.* 1990, **18**, 2739-2747.
118. Ji, X., Oh, J., Dunker, A.K., and Hipps, K.W. Effects of relative humidity and applied force on atomic force microscopy images of the filamentous phage fd. *Ultramicroscopy* 1998, **72**, 165-176.
119. Sertoli, E. e lesistenza di particulari cellule remificate nei canalicoli seminiferi dell'testicolo umano. *Morgangni* 1865, **7**, 31-40.
120. Griswold, M.D. Interactions between germ cells and Sertoli cells in the testis. *Biology of Reproduction* 1995, **52**, 211-216.
121. Kissinger, C., Skinner, M.K., and Griswold, M.D. Analysis of Sertoli cell-secreted proteins by two-dimensional gel electrophoresis. *Biol. Reprod.* 1982, **27**, 233-240.
122. Collard, M.W., Sylvester, S.R., Tsuruta, J.K., and Griswold, M.D. Biosynthesis and molecular cloning of sulfated glycoprotein 1 secreted by rat Sertoli cells: sequence similarity with the 70-kilodalton precursor to sulfatide/GM1 activator. *Biochemistry* 1988, **27**, 4557-4564.
123. Bailey, R. and Griswold, M.D. Clusterin in the male reproductive system: localization and possible function. *Mol. Cell. Endocrinol.* 1999, **151**, 17-23.
124. Romero, P., Obradovic, Z., Kissinger, C.R., Villafranca, J.E., and Dunker, A.K. Identifying disordered regions in proteins from amino acid sequences. *Proc. I.E.E.E. International Conference on Neural Networks* 1997, **1**, 90-95.
125. Romero, P., Obradovic, Z., Li, X., Garner, E.C., Brown, C.J., and Dunker, A.K. The Complexity of Disorder. *Proteins: Struc., Funct., Gen.* 2001, **42**, In press.
126. Manalan, A.S. and Klee, C.B. Calcineurin, a calmodulin-stimulated protein phophatase. In: *Calcium in Biological Systems*, Rubin, R.P. Weiss, G.B., and Putney, J.W., Eds., Plenum Press, New York, 1985, pp. 307-315.

127. Klee, C.B., Crouch, T.H., and Krinks, M.H. Calcineurin: a calcium- and calmodulin-binding protein of the nervous system. *Proc. Natl. Acad. Sci. USA* 1979, **76**, 6270-6273.
128. Liu, J., Farmer, J.D., Jr., Lane, W.S., Friedman, J., Weissman, I., and Schreiber, S.L. Calcineurin is a common target of cyclophilin-cyclosporin A and FKBP- FK506 complexes. *Cell* 1991, **66**, 807-815.
129. McKeon, F. When worlds collide: immunosuppressants meet protein phosphatases. *Cell* 1991, **66**, 823-826.
130. Meador, W.E., Means, A.R., and Quiocho, F.A. Target Enzymes Recognition by Calmodulin: 2.4 Å Structure of a Camodulin-Peptide Complex. *Science* 1992, **257**, 1251-1255.
131. MacLennan, D.H. and Wong, P.T. Isolationof a calcium sequestering protein from sarcoplasmic reticulum. *Proc. Natl. Acad. Sci. USA* 1971, **68**, 1231-1235.
132. Krause, K.H., Milos, M., Luan-Rilliet, Y., Lew, D.P., and Cox, J.A. Thermodynamics of cation binding to rabbit skeletal muscle calsequestrin. Evidence for distinct Ca(2+)- and Mg(2+)-binding sites. *J. Biol. Chem.* 1991, **266**, 9453-9459.
133. MacLennan, D.H. and Reithmeier, R.A.F. Ion Tamers. *Nat. Struct. Biol.* 1998, **5**, 409-411.
134. Wang, S., Trumble, W.R., Liao, H., Wesson, C.R., Dunker, A.K., and Kang, C.H. Crystal structure of calsequestrin from rabbit skeletal muscle sarcoplasmic reticulum. *Nat. Struct. Biol.* 1998, **5**, 476-483.
135. Huber, R. Conformational flexibility in protein molecules. *Nature* 1979, **280**, 538-539.
136. Kossiakoff, A.A., Chambers, J.L., Kay, L.M., and Stroud, R.M. Structure of bovine trypsinogen at 1.9 A resolution. *Biochemistry* 1977, **16**, 654-664.
137. Bennett, W.S. and Huber, R. Structural and functional aspects of domain motions in proteins. *Crit. Rev. Biochem.* 1984, **15**, 291-384.
138. Holmes, K.C. Flexibility in tobacco mosaic virus. *Ciba Found. Symp.* 1983, **93**, 116-138.
139. Champness, J.N., Bloomer, A.C., Bricogne, G., Butler, P.G., and Klug, A. The structure of the protein disk of tobacco mosaic virus to 5Å resolution. *Nature* 1976, **259**, 20-24.
140. Stubbs, G., Warren, S., and Holmes, K. Structure of RNA and RNA binding site in tobacco mosaic virus from 4-Å map calculated from X-ray fibre diagrams. *Nature* 1977, **267**, 216-221.
141. Jardetzky, O., Akasaka, K., Vogel, D., Morris, S., and Holmes, K.C. Unusual segmental flexibility in a region of tobacco mosaic virus coat protein. *Nature* 1978, **273**, 564-566.
142. Jacob, F. and Monod, J. Genetic regulatory mechanisms in the synthesis of proteins. *J. Mol. Biol.* 1961, **3**, 318-356.
143. Slijper, M., Boelens, R., Davis, A., Konings, R., van der Marel, G., van Boom, J., and Kaptein, R. Backbone and side chain dynamics of lac

repressor headpiece (1-56) and its complex with DNA. *Biochemistry* 1997, **36**, 249-254.

144. Babu, Y., Bugg, C.E., and Cook, W.J. Structure of calmodulin refined at 2.2 Å resolution. *J. Mol. Biol.* 1988, **203**, 191-204.
145. Wriggers, W., Mehler, E., Pitici, F., Weinstein, H., and Schulten, K. Structure and dynamics of calmodulin in solution. *Biophys. J.* 1998, **74**, 1622-1639.
146. Ikura, M., Clore, G.M., Gronenborn, A.M., Zhu, G., Klee, C.B., and Bax, A. Solution structure of a calmodulin-target peptide complex by multidimensional NMR. *Science* 1992, **256**, 632-638.
147. Meador, W.E., Means, A.R., and Quiocho, F.A. Modulation of calmodulin plasticity in molecular recognition on the basis of X-ray structures. *Science* 1993, **262**, 1718-1721.
148. Dunker, A.K., Garner, E., Guilliot, S., Romero, P., Albrecht, K., Hart, J., Obradovic, Z., Kissinger, C., and Villafranca, J.E. Protein disorder and the evolution of molecular recognition: theory, predictions and observations. *Pacific Symp. Biocomputing* 1998, **3**, 473-484.
149. Lee, A.L., Kinnear, S.A., and Wand, A.J. Redistribution and loss of side chain entropy upon formation of a calmodulin-peptide complex. *Nat. Struct. Biol.* 2000, **7**, 72-77.
150. Gellman, S.H. On the role of methionine residues in the sequence-independent recognition of nonpolar protein surfaces. *Biochemistry* 1991, **30**, 6633-6636.
151. Weliky, D.P., Bennett, A.E., Zvi, A., Anglister, J., Steinbach, P.J., and Tycko, R. Solid-state NMR evidence for an antibody-dependent conformation of the V3 loop of HIV-1 gp120. *Nat. Struct. Biol.* 1999, **6**, 141-145.
152. Stanfield, R., Cabezas, E., Satterthwait, A., Stura, E., Profy, A., and Wilson, I. Dual conformations for the HIV-1 gp120 V3 loop in complexes with different neutralizing fabs. *Structure Fold Des.* 1999, **7**, 131-142.
153. Balbach, J.J., Yang, J., Weliky, D.P., Steinbach, P.J., Tugarinov, V., Anglister, J., and Tycko, R. Probing hydrogen bonds in the antibody-bound HIV-1 gp120 V3 loop by solid state NMR REDOR measurements. *J. Biomol. NMR* 2000, **16**, 313-327.
154. House-Pompeo, K., Xu, Y., Joh, D., Speziale, P., and Hook, M. Conformational changes in the binding MSCRAMMs are induced by ligand binding. *J. Biol. Chem.* 1996, **271**, 1379-1384.
155. Penkett, C.J., Redfield, C., Dodd, I., Hubbard, J., McBay, D.L., Mossakowska, D.E., Smith, R.A., Dobson, C.M., and Smith, L.J. NMR analysis of main-chain conformational preferences in an unfolded fibronectin-binding protein. *J. Mol. Biol.* 1997, **274**, 152-159.
156. Lea, S., Hernandez, J., Blakemore, W., Brocchi, E., Curry, S., Domingo, E., Fry, E., Abu-Ghazaleh, R., King, A., Newman, J., Stuart, D., and Mateu, M.G. The structure and antigenicity of a type C foot-and-mouth disease virus. *Structure* 1994, **2**, 123-139.

157. Romero, P., Obradovic, Z., Kissinger, C.R., Villafranca, J.E., Guilliot, S., Garner, E., and Dunker, A.K. Thousands of proteins likely to have long disordered regions. *Pacific Symp. Biocomputing* 1998, **3**, 437-448.
158. Boulikas, T. Nuclear localization signals (NLS). *Crit. Rev. Eukaryot Gene Expr.* 1993, **3**, 193-227.
159. Nolte, R.T., Conlin, R.M., Harrison, S.C., and Brown, R.S. Differing roles for zinc fingers in DNA recognition: structure of a six- finger transcription factor IIIA complex. *Proc. Natl. Acad. Sci. USA* 1998, **95**, 2938-2943.
160. Choo, Y. and Schwabe, J.W. All wrapped up. *Nat. Struct. Biol.* 1998, **5**, 253-255.
161. Vajda, S., Weng, Z., Rosenfeld, R., and DeLisi, C. Effect of conformational flexibility and solvation on receptor-ligand binding free energies. *Biochemistry* 1994, **33**, 13977-13988.
162. Rosenfeld, R., Vajda, S., and DeLisi, C. Flexible docking and design. *Annu Rev Biophys Biomol Struct* 1995, **24**, 677-700.
163. Novotny, J., Bruccoleri, R.E., and Saul, F.A. On the attribution of binding energy in antigen-antibody complexes McPC 603, D1.3, and HyHEL-5. *Biochemistry* 1989, **28**, 4735-4749.
164. Noyes, R.M. Effects of diffusion rates on chemical kinetics. *Prog. React. Kinet.* 1961, **1**, 129-160.
165. Von Hippel, P.H. and Berg, O.G. Facilitated Target Location in Biological Systems. *J. Biol. Chem.* 1989, **264**, 675-678.
166. Schreiber, G. and Fersht, A.R. Rapid, electrostatically assisted association of proteins. *Nat. Struct. Biol.* 1996, **3**, 427-431.
167. DeLisi, C. The biophysics of ligand-receptor interactions. *Quart. Rev. Biophysics* 1980, **13**, 201-230.
168. Berg, O.G. and Hippel, v. Diffusion-controlled macromolecular interactions. *Ann. Rev. Biophys. Biophys. Chem.* 1985, **14**, 131-160.
169. Pontius, B.W. Close encounters: why unstructured, polymeric domains can increase rates of specific macromolecular association. *Trends Biochem. Sci.* 1993, **18**, 181-186.
170. Jeffery, C.J. Moonlighting proteins. *Trends Biochem. Sci.* 1999, **24**, 8-11.
171. Williams, P.D., Pollock, D.D., and Goldstein, R.A. Evolution of functionality in lattice proteins. *J. Mol. Graphics* 2001, This volume.
172. Gast, K., Damaschun, H., Eckert, K., Schulze-Forster, K., Maurer, H.R., Muller-Frohne, M., Zirwer, D., Czarnecki, J., and Damaschun, G. Prothymosin alpha: a biologically active protein with random coil conformation. *Biochemistry* 1995, **34**, 13211-13218.
173. Fletcher, C.M., McGuire, A.M., Gingras, A.C., Li, H., Matsuo, H., Sonenberg, N., and Wagner, G. 4E binding proteins inhibit the translation factor eIF4E without folded structure. *Biochemistry* 1998, **37**, 9-15.
174. Mader, S., Lee, H., Pause, A., and Sonenberg, N. The translation initiation factor eIF-4E binds to a common motif shared by the translation factor eIF-4 gamma and the translational repressors 4E-binding proteins. *Mol. Cell. Biol.* 1995, **15**, 4990-4997.

175. Marcotrigiano, J., Gingras, A.C., Sonenberg, N., and Burley, S.K. Cap-dependent translation initiation in eukaryotes is regulated by a molecular mimic of eIF4G. *Mol Cell* 1999, **3**, 707-716.
176. Bustin, M. and Reeves, R. HMG chromosomal proteins: Architectural components that facilitate chromatin function. *Prog. Nucl. Acids Res. Molec. Biol.* 1996, **54**, 35-100.
177. Huth, J.R., Bewley, C.A., Nissen, M.S., Evans, J.N., Reeves, R., Gronenborn, A.M., and Clore, G.M. The solution structure of an HMG-I(Y)-DNA complex defines a new architectural minor groove binding motif. *Nat. Struct. Biol.* 1997, **4**, 657-665.
178. Zhang, M., Vogel, H.J., and Zwiers, H. Nuclear magnetic resonance studies of the structure of B50/neuromodulin and its interaction with calmodulin. *Biochem. Cell Biol.* 1994, **72**, 109-116.
179. Smith, M.L., Johanson, R.A., Rogers, K.E., Coleman, P.D., and Slemmon, J.R. Identification of a neuronal calmodulin-binding peptide, CAP-19, containing an IQ motif. *Brain. Res. Mol. Brain Res.* 1998, **62**, 12-24.
180. Alexander, K.A., Wakim, B.T., Doyle, G.S., Walsh, K.A., and Storm, D.R. Identification and characterization of the calmodulin-binding domain of neuromodulin, a neurospecific calmodulin-binding protein. *J. Biol. Chem.* 1988, **263**, 7544-7549.
181. Wertz, S.L., Savino, Y., and Cafiso, D.S. Solution and membrane bound structure of a peptide derived from the protein kinase C substrate domain of neuromodulin. *Biochemistry* 1996, **35**, 11104-11112.
182. Gerendasy, D. Homeostatic tuning of Ca2+ signal transduction by members of the calpacitin protein family. *J. Neurosci. Res.* 1999, **58**, 107-119.
183. Rhoads, A.R. and Friedberg, F. Sequence motifs for calmodulin recognition. *FASEB J.* 1997, **11**, 331-340.
184. Sheu, F.S., Huang, F.L., and Huang, K.P. Differential responses of protein kinase C substrates (MARCKS, neuromodulin, and neurogranin) phosphorylation to calmodulin and S100. *Arch. Biochem. Biophys.* 1995, **316**, 335-342.
185. Chakravarthy, B., Morley, P., and Whitfield, J. Ca2+-calmodulin and protein kinase Cs: a hypothetical synthesis of their conflicting convergences on shared substrate domains. *Trends. Neurosci.* 1999, **22**, 12-16.
186. Marvin, D.A. Filamentous phage structure, infection and assembly. *Curr. Opin. Struct. Biol.* 1998, **8**, 150-158.
187. Marvin, D.A. and Hohn, B. Filamentous bacterial viruses. *Bacteriol. Rev.* 1969, **33**, 172-209.
188. Rossomando, E.F. Studies on the structural polarity of bacteriophage f1. *Virology* 1970, **42**, 681-687.
189. Gray, C.W., Brown, R.S., and Marvin, D.A. Adsorption complex of filamentous fd virus. *J. Mol. Biol.* 1981, **146**, 621-627.
190. Stengele, I., Bross, P., Garces, X., Giray, J., and Rasched, I. Dissection of functional domains in phage fd adsorption protein. Discrimination

between attachment and penetration sites. *J. Mol. Biol.* 1990, **212**, 143-149.

191. Holliger, P. and Riechmann, L. A conserved infection pathway for filamentous bacteriophages is suggested by the structure of the membrane penetration domain of the minor coat protein g3p from phage fd. *Structure* 1997, **5**, 265-275.
192. Deng, L.W., Malik, P., and Perham, R.N. Interaction of the globular domains of pIII protein of filamentous bacteriophage fd with the F-pilus of Escherichia coli. *Virology* 1999, **253**, 271-277.
193. Chatellier, J., Hartley, O., Griffiths, A.D., Fersht, A.R., Winter, G., and Riechmann, L. Interdomain interactions within the gene 3 protein of filamentous phage. *FEBS Lett* 1999, **463**, 371-374.
194. Lubkowski, J., Hennecke, F., Pluckthun, A., and Wlodawer, A. The structural basis of phage display elucidated by the crystal structure of the N-terminal domains of g3p. *Nat. Struct. Biol.* 1998, **5**, 140-147.
195. Sutrina, S.L., Reddy, P., Saier, M.H., Jr., and Reizer, J. The glucose permease of *Bacillus subtilis* is a single polypeptide chain that functions to energize the sucrose permease. *J. Biol. Chem.* 1990, **265**, 18581-18589.
196. Wagenknecht, T., Grassucci, R., Berkowitz, J., and Forneris, C. Configuration of interdomain linkers in pyruvate dehydrogenase complex of *Escherichia coli* as determined by cryoelectron microscopy. *J. Struct. Biol.* 1992, **109**, 70-77.
197. Turner, S.L., Russell, G.C., Williamson, M.P., and Guest, J.R. Restructuring an interdomain linker in the dihydrolipoamide acetyltransferase component of the pyruvate dehydrogenase complex of Escherichia coli. *Protein Eng* 1993, **6**, 101-108.
198. Berger, J.M., Gamblin, S.J., Harrison, S.C., and Wang, J.C. Structure and mechanism of DNA topoisomerase II. *Nature* 1996, **379**, 225-232.
199. Reizer, J., Hoischen, C., Reizer, A., Pham, T.N., and Saier, M.H., Jr. Sequence analyses and evolutionary relationships among the energy-coupling proteins Enzyme I and HPr of the bacterial phosphoenolpyruvate: sugar phosphotransferase system. *Protein Sci.* 1993, **2**, 506-5521.
200. Shamoo, Y., Abdul-Manan, N., and Williams, K.R. Multiple RNA binding domains (RBDs) just don't add up. *Nucleic Acids Res* 1995, **23**, 725-728.
201. Zhou, H., McEvoy, M.M., Lowry, D.F., Swanson, R.V., Simon, M.I., and Dahlquist, F.W. Phosphotransfer and CheY-binding domains of the histidine autokinase CheA are joined by a flexible linker. *Biochemistry* 1996, **35**, 433-443.
202. Morais, M.C., Zhang, W., Baker, A.S., Zhang, G., Dunaway-Mariano, D., and Allen, K.N. The crystal structure of bacillus cereus phosphonoacetaldehyde hydrolase: insight into catalysis of phosphorus bond cleavage and catalytic diversification within the HAD enzyme superfamily.*Biochemistry* 2000, **39**, 10385-10396.
203. Wang, S., Wang, Z., Boise, L., Dent, P., and Grant, S. Loss of the bcl-2 phosphorylation loop domain increases resistance of human leukemia cells

(U937) to paclitaxel-mediated mitochondrial dysfunction and apoptosis. *Biochem. Biophys. Res. Commun.* 1999, **259**, 67-72.

204. Srivastava, R.K., Mi, Q.S., Hardwick, J.M., and Longo, D.L. Deletion of the loop region of Bcl-2 completely blocks paclitaxel- induced apoptosis. *Proc. Natl. Acad. Sci. USA* 1999, **96**, 3775-3780.
205. Cheng, E.H., Kirsch, D.G., Clem, R.J., Ravi, R., Kastan, M.B., Bedi, A., Ueno, K., and Hardwick, J.M. Conversion of Bcl-2 to a Bax-like death effector by caspases. *Science* 1997, **278**, 1966-1968.
206. Clem, R.J., Cheng, E.H., Karp, C.L., Kirsch, D.G., Ueno, K., Takahashi, A., Kastan, M.B., Griffin, D.E., Earnshaw, W.C., Veliuona, M.A., and Hardwick, J.M. Modulation of cell death by Bcl-x_L through caspase interaction. *Proc. Natl. Acad. Sci. USA* 1998, **95**, 554-559.
207. Chang, B.S., Minn, A.J., Muchmore, S.W., Fesik, S.W., and Thompson, C.B. Identification of a novel regulatory domain in Bcl-X(L) and Bcl-2. *Embo. J.* 1997, **16**, 968-977.
208. Yamamoto, K., Ichijo, H., and Korsmeyer, S.J. BCL-2 is phosphorylated and inactivated by an ASK1/Jun N-terminal protein kinase pathway normally activated at G(2)/M. *Mol. Cell. Biol.* 1999, **19**, 8469-8478.
209. Knighton, D.R., Zheng, J.H., Ten Eyck, L.F., Xuong, N.H., Taylor, S.S., and Sowadski, J.M. Structure of a peptide inhibitor bound to the catalytic subunit of cyclic adenosine monophosphate-dependent protein kinase. *Science* 1991, **253**, 414-420.
210. Hauer, J.A., Johnson, D.A., and Taylor, S.S. Binding-dependent disorder-order transition in PKI alpha: a fluorescence anisotropy study. *Biochemistry* 1999, **38**, 6774-6780.
211. Johnson, L.N. and O'Reilly, M. Control by phosphorylation. *Curr. Opin. Struct. Biol.* 1996, **6**, 762-769.
212. Ohmori, T., Podack, E.R., Nishio, K., Takahashi, M., Miyahara, Y., Takeda, Y., Kubota, N., Funayama, Y., Ogasawara, H., Ohira, T., and et al. Apoptosis of lung cancer cells caused by some anti-cancer agents (MMC, CPT-11, ADM) is inhibited by bcl-2. *Biochem. Biophys. Res. Commun.* 1993, **192**, 30-36.
213. Tang, C., Willingham, M.C., Reed, J.C., Miyashita, T., Ray, S., Ponnathpur, V., Huang, Y., Mahoney, M.E., Bullock, G., and Bhalla, K. High levels of p26BCL-2 oncoprotein retard taxol-induced apoptosis in human pre-B leukemia cells. *Leukemia* 1994, **8**, 1960-1969.
214. Fang, G., Chang, B.S., Kim, C.N., Perkins, C., Thompson, C.B., and Bhalla, K.N. "Loop" domain is necessary for taxol-induced mobility shift and phosphorylation of Bcl-2 as well as for inhibiting taxol-induced cytosolic accumulation of cytochrome c and apoptosis. *Cancer Res* 1998, **58**, 3202-3208.
215. Rodi, D.J., Janes, R.W., Sanganee, H.J., Holton, R.A., Wallace, B.A., and Makowski, L. Screening of a library of phage-displayed peptides identifies human bcl- 2 as a taxol-binding protein. *J. Mol. Biol.* 1999, **285**, 197-203.

216. Rodi, D.J. and Makowski, L. Similarity between sequences of the taxol-selected peptides and the disordered loop of the anti-apoptotic protein, Bcl-2. *Pacific Symp. Biocomput.* 1999, **4**, 532-541.
217. Rice, S., Lin, A.W., Safer, D., Hart, C.L., Naber, N., Carragher, B.O., Cain, S.M., Pechatnikova, E., Wilson-Kubalek, E.M., Whittaker, M., Pate, E., Cooke, R., Taylor, E.W., Milligan, R.A., and Vale, R.D. A structural change in the kinesin motor protein that drives motility. *Nature* 1999, **402**, 778-784.
218. Kozielski, F., Sack, S., Marx, A., Thormahlen, M., Schonbrunn, E., Biou, V., Thompson, A., Mandelkow, E.M., and Mandelkow, E. The crystal structure of dimeric kinesin and implications for microtubule-dependent motility. *Cell* 1997, **91**, 985-994.
219. Kull, F.J., Sablin, E.P., Lau, R., Fletterick, R.J., and Vale, R.D. Crystal structure of the kinesin motor domain reveals a structural similarity to myosin. *Nature* 1996, **380**, 550-555.
220. Sack, S., Muller, J., Marx, A., Thormahlen, M., Mandelkow, E.M., Brady, S.T., and Mandelkow, E. X-ray structure of motor and neck domains from rat brain kinesin. *Biochemistry* 1997, **36**, 16155-16165.
221. Thomas, D.D., Ramachandran, S., Roopnarine, O., Hayden, D.W., and Ostap, E.M. The mechanism of force generation in myosin: a disorder-to-order transition, coupled to internal structural changes. *Biophys. J.* 1995, **68**, 135S-141S.
222. Houdusse, A., Kalabokis, V.N., Himmel, D., Szent-Gyorgyi, A.G., and Cohen, C. Atomic structure of scallop myosin subfragment S1 complexed with MgADP: a novel conformation of the myosin head. *Cell* 1999, **97**, 459-470.
223. Wahlstrom, J., Randall, A., Lawson, J.D., Lyons, D., Crouch, G., and Cremo, C.R. Sites of interaction between the regulatory light chains of dephosphorylated smooth muscle myosin. *Manuscript in preparation* 2000.
224. Lawson, J.D., Grammer, J.C., and Yount, R.G. Inter-head photoaffinity labeling in scallop myosin. *Manuscript in preparation* 2000.
225. Trombitas, K., Greaser, M., Labeit, S., Jin, J.P., Kellermayer, M., Helmes, M., and Granzier, H. Titin extensibility in situ: entropic elasticity of permanently folded and permanently unfolded molecular segments. *J. Cell Biol.* 1998, **140**, 853-859.
226. Armstrong, C.M. and Bezanilla, F. Inactivation of the sodium channel. II. Gating current experiments. *J. Gen. Physiol.* 1977, **70**, 567-590.
227. Antz, C., Geyer, M., Fakler, B., Schott, M.K., Guy, H.R., Frank, R., Ruppersberg, J.P., and Kalbitzer, H.R. NMR structure of inactivation gates from mammalian voltage-dependent potassium channels. *Nature* 1997, **385**, 272-275.
228. Lee, C.Y. On the activation-inactivation coupling in Shaker potassium channels. *FEBS Lett.* 1992, **306**, 95-97.

229. Brown, H.G. and Hoh, J.H. Entropic exclusion by neurofilament sidearms: a mechanism for maintaining interfilament spacing. *Biochemistry* 1997, **36**, 15035-15040.
230. Hoh, J.H. Functional protein domains from the thermally driven motion of polypeptide chains: a proposal. *Proteins* 1998, **32**, 223-228.
231. Lakoff, G., *Women, Fire and Dangerous Things: What categories Reveal About the Human Mind*, University of Chicago Press, Chicago, 1987.
232. Linderstrom-Lang, K. Structure and enzymatic break-down of proteins. *Lane Medical Lectures* 1952, **6**, 117-126.
233. Linderstrom-Lang, K.U. and Schellman, J.A., Protein structure and enzyme activity. In: *The Enzymes*, Boyer, P.D., Lardy, H., and Myrback, K., Eds., Academic Press, New York, 1959, pp. 443-510.
234. Blow, D.M. Hard facts on structure: hot air about mobility. *Nature* 1982, **297**, 454.
234A. Dunker, K.A., Fodor, S.P.A., and Williams, R.W., Lipid-dependent structural changes of an amphomorphic membrane protein. *Biophys. J.* 1981, **37**, 201-203.
235. Rosenblatt, M., Beaudette, N.V., and Fasman, G.D. Conformational studies of the synthetic precursor-specific region of preproparathyroid hormone. *Proc. Natl. Acad. Sci. USA* 1980, **77**, 3983-3987.
236. Dunker, A.K., Fodor, S.P.A., and Williams, R.W. Lipid dependent structural changes of an amphomorphic membrane protein. *Biophys. J.* 1982, **37**, 201-203.
237. Stein, P. and Chothia, C. Serpin tertiary structure transformation. *J. Mol. Biol.* 1991, **221**, 615-621.
238. Potempa, J., Korzus, E., and Travis, J. The serpin superfamily of proteinase inhibitors: structure, function, and regulation. *J. Biol. Chem.* 1994, **269**, 15957-15960.
239. Young, M., Kirshenbaum, K., Dill, K.A., and Highsmith, S. Predicting conformational switches in proteins. *Protein Sci.* 1999, **8**, 1752-1764.
240. Kirshenbaum, K., Young, M., and Highsmith, S. Predicting allosteric switches in myosins. *Protein Sci* 1999, **8**, 1806-1815.
241. Hobohm, U. and Sander, C. Enlarged representative set of protein structures. *Protein Sci.* 1994, **3**, 522-524.
242. Garner, E., Romero, P., Dunker, A.K., Brown, C., and Obradovic, Z. Predicting binding regions within disordered proteins. *Genome Informatics* 1999, **10**, 41-50.
243. Vihinen, M., Torkkila, E., and Riikonen, P. Accuracy of protein flexibility predictions. *Proteins* 1994, **19**, 141-149.
244. Williams, R.M., Obradovic, Z., Mathura, V., Braun, W., Garner, E.C., Young, J., Takayama, S., Brown, C.J., and Dunker, A.K. The protein non-folding problem: amino acid determinants of intrinsic order and disorder. *Pacific Symp. Biocomputing* 2001, **5**, In press.
245. Xie, Q., Arnold, G.E., Romero, P., Obradovic, Z., Garner, E., and Dunker, A.K. The sequence attribute method for determining relationships between sequence and protein disorder. *Genome Informatics* 1998, **9**, 193-200.

246. Li, X., Obradovic, Z., Brown, C.J., Garner, E.C., and Dunker, A.K. Comparing predictors of disordered protein. *Genome Informatics* 2000, **11**, In press.
247. Romero, P., Obradovic, Z., and Dunker, A.K. Sequence data analysis for long disordered regions prediction in the calcineurin family. *Genome Informatics* 1997, **8**, 110-124.
248. Garner, E., Cannon, P., Romero, P., Obradovic, Z., and Dunker, A.K. Predicting disordered regions from amino acid sequence: common themes despite differing structural characterization. *Genome Informatics* 1998, **9**, 201-213.
249. Li, X., Romero, P., Rani, M., Dunker, A.K., and Obradovic, Z. Predicting protein disorder for N-, C-, and internal regions. *Genome Informatics* 1999, **10**, 30-40.
250. Romero, P.Z., Obradovic, C., and Dunker, A.K. Intelligent data analysis for protein disorder prediction. *Artificial Intelligence Reviews* 2000, In press.
251. Wootton, J.C. Statistic of local complexity in amino acid sequences and sequence databases. *Computers Chem.* 1993, **17**, 149-163.
252. Wootton, J.C. Non-globular domains in protein sequences: automated segmentation using complexity measures. *Comput. Chem.* 1994, **18**, 269-285.
253. Wootton, J.C. Sequences with 'unusual' amino acid compositions. *Curr. Opin. Struct. Biol.* 1994, **4**, 413-421.
254. Wootton, J.C. and Federhen, S. Analysis of compositionally biased regions in sequence databases. *Methods in Enzymology* 1996, **266**, 554-571.
255. Karlin, S. and Brendel, V. Chance and statistical significance in protein and DNA sequence analysis. *Science* 1992, **257**, 39-49.
256. Karlin, S. Statistical significance of sequence patterns in proteins. *Curr. Opin. Struct. Biol.* 1995, **5**, 360-371.
257. Romero, P., Obradovic, Z., and Dunker, A.K. Folding minimal sequences: the lower bound for sequence complexity of globular proteins. *FEBS Lett.* 1999, **462**, 363-367.
258. Riddle, D.S., Santiago, J.V., Bray-Hall, S.T., Doshi, N., Grantcharova, V.P., Yi, Q., and Baker, D. Functional rapidly folding proteins from simplified amino acid sequences. *Nat. Struct. Biol.* 1997, **4**, 805-809.
259. Dunker, K.A., Obradovic, Z., Romero, P., Garner, E.C., and Brown, C.J. Intrinsic protein disorder in complete genomes. *Genome Informatics* 2000, **11**, In press.
260. Schoenheimer, R., *The Dynamic State of Body Constituents*, Harvard University Press, Cambridge, 1942.
261. Bond, J.S. and Beynon, R.J. Proteolysis and physiological regulation. *Molec. Aspects Med.* 1987, **9**, 173-287.
262. Rogers, S., Wells, R., and Rechsteiner, M. Amino acid sequences common to rapidly degraded proteins: the PEST hypothesis. *Science* 1986, **234**, 364-368.

263. Rechsteiner, M. and Rogers, S.W. PEST sequences and regulation by proteolysis. *Trends Biochem. Sci.* 1996, **21**, 267-271.
264. Ciechanover, A. The ubiquitin-proteasome proteolytic pathway. *Cell* 1994, **79**, 13-21.
265. Hochstrasser, M. Ubiquitin-dependent protein degradation. *Annu. Rev. Genet.* 1996, **30**, 405-439.
266. Dahlqvist-Edberg, U. and Ekman, P. Purification of a Ca2+-activated protease from rat erythrocytes and its possible effect on pyruvate kinase in vivo. *Biochim. Biophys. Acta.* 1981, **660**, 96-101.
267. Chan, S.L. and Mattson, M.P. Caspase and calpain substrates: roles in synaptic plasticity and cell death. *J. Neurosci. Res.* 1999, **58**, 167-190.
268. Murray, S.S., Grisanti, M.S., Bentley, G.V., Kahn, A.J., Urist, M.R., and Murray, E.J. The calpain-calpastatin system and cellular proliferation and differentiation in rodent osteoblastic cells. *Exp. Cell Res.* 1997, **233**, 297-309.
269. Neurath, H. and Walsh, K.A. Role of proteolytic enzymes in biological regulation (a review). *Proc. Natl. Acad. Sci. USA* 1976, **73**, 3825-3832.
270. Salama, S.R., Hendricks, K.B., and Thorner, J. G1 cyclin degradation: the PEST motif of yeast Cln2 is necessary, but not sufficient, for rapid protein turnover. *Mol. Cell Biol.* 1994, **14**, 7953-7966.
271. Santella, L. The role of calcium in the cell cycle: facts and hypotheses. *Biochem. Biophys. Res. Commun.* 1998, **244**, 317-324.
272. Kitamura, Y., Shimohama, S., Kamoshima, W., Ota, T., Matsuoka, Y., Nomura, Y., Smith, M.A., Perry, G., Whitehouse, P.J., and Taniguchi, T. Alteration of proteins regulating apoptosis, Bcl-2, Bcl-x, Bax, Bak, Bad, ICH-1 and CPP32, in Alzheimer's disease. *Brain Res.* 1998, **780**, 260-269.
273. Brown, M.S. and Goldstein, J.L. A proteolytic pathway that controls the cholesterol content of membranes, cells, and blood. *Proc. Natl. Acad. Sci. USA* 1999, **96**, 11041-11048.
274. Hubbard, S.J. The structural aspects of limited proteolysis of native proteins. *Biochim Biophys Acta* 1998, **1382**, 191-206.
275. Hemmingsen, S.M., Woolford, C., van der Vies, S.M., Tilly, K., Dennis, D.T., Georgopoulos, C.P., Hendrix, R.W., and Ellis, R.J. Homologous plant and bacterial proteins chaperone oligomeric protein assembly. *Nature* 1988, **333**, 330-334.
276. Hartl, F.U. Heat shock proteins in protein folding and membrane translocation. *Semin. Immunol.* 1991, **3**, 5-16.
277. Shapiro, L. and Lima, C.D. The Argonne Structural Genomics Workshop: Lamaze class for the birth of a new science. *Structure* 1998, **6**, 265-267.
278. Gaasterland, T. Structural genomics: Bioinformatics in the driver's seat. *Nat. Biotechnol.* 1998, **16**, 625-627.
279. Goldstein, D.J. An unacknowledged problem for structural genomics? *Nat. Biotechnol.* 1998, **16**, 696.
280. Sweet, R.M. and Eisenberg, D. Correlation of sequence hydrophobicities measures similarity in three- dimensional protein structure. *J. Mol. Biol.* 1983, **171**, 479-488.

281. Kyte, J. and Doolittle, R.F. A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.* 1982, **157**, 105-132.
282. Chechetkin, V.R. and Lobzin, V.V. Characterization and comparison of protein structures. Part I- characterization. *J. Theor. Biol.* 1999, **198**, 197-218.