Intrinsically Disordered Proteins: An Update

Keynote Paper of IEEE 7th BIBE, Oct. 14-17 Harvard Medical School Conference Center

A. Keith Dunker
Center for Computational Biology and Bioinformatics
Indiana University Schools of Medicine and Informatics
Indianapolis, IN 46202
kedunker@iupui.edu

Christopher J. Oldfield
Center for Computational Biology and Bioinformatics
Indiana University School of Informatics
Indianapolis, IN 46202

Jingwei Meng
Department of Biochemistry and Molecular Biology
Indiana University School of Medicine
Indianapolis, Indiana 46202

Pedro Romero
Center for Computational Biology and Bioinformatics
Indiana University School of Informatics
Indianapolis, IN 46202

Jack Y. Yang
Department of Radiation Oncology, Harvard Medical School and Massachusetts General Hospital
Boston, MA 02114

Zoran Obradovic
Center for Information Science and Technology
Temple University
Philadelphia, PA 19122

Vladimir N. Uversky
Center for Computational Biology and Bioinformatics
Indiana University School of Medicine
Indianapolis, IN 46202
and
Institute for Biological Instrumentation
Russian Academy of Sciences
142290 Puschino, Moscow Region, Russia
vuversky@iupui.edu

***Abstract*—Just over 10 years ago, in June, 1997, in the *Proceedings of the IEEE International Conference on Neural Networks*, we published our first predictor of intrinsically disordered protein [1]. Since then, we have substantially improved our predictors, and more than 20 other laboratory groups have joined in efforts to improve the prediction of protein disorder. At the algorithmic level, prediction of protein intrinsic disorder is similar to the prediction of secondary structure, but, at the structural level, secondary structure and intrinsic disorder are entirely different. The secondary structure class called random coil or irregular differs from intrinsic disorder due to very different dynamic properties, with the secondary structure class being much less mobile than the region of disorder. At the biological level, unlike the prediction of secondary structure, the prediction of intrinsic disorder has been revolutionary. That is, for many years, experimentalists have provided evidence that some proteins lack fixed structure or are disordered (or unfolded) under physiological conditions. Experimentalists further are showing that, for some proteins, functions depended on the unstructured rather than structured state. However, these examples have been mostly ignored. To our knowledge, not one disordered protein or disorder-associated function is discussed in any biochemistry textbook, even though such examples began to be discovered more than 50 years ago. Disorder prediction has been important for showing that the few experimentally characterized examples represent a very large cohort that is found all across all three domains of life. We now know that many significant biological functions depend directly on, or are importantly associated with, the unfolded or partially folded state. In this paper, we will briefly review some of the key discoveries that have occurred in the last decade, and, furthermore, will make a few highly speculative projections.**

Keywords - disorder prediction, cell signaling, regulation and control, protein-protein interactions, alternative splicing, and disorder-based drug discovery

I. Introduction

Speculation that antibody binding depends on unfolded rather than structured protein goes back more than seventy years [2, 3], when it was conjectured that high flexibility would enable one antibody molecule to bind to multiple antigens having different structures. The flexible antibody could randomly fluctuate among the different structures, with binding leading to the selection of the structure that fits with each different antigen [3]. The current body of evidence suggests that there are approximately two broad classes of antibodies, specific and non-specific. The sequence of a highly specific, high-affinity antibody folds into a specific structure that fits with its cognate antigen. On the other hand, the binding sites of low affinity, nonspecific antibodies are disordered in isolation but become differently folded when bound to different partners, thus confirming the early conjectures cited above (manuscript in preparation).

Experimental evidence supporting the view that some proteins remain unstructured, or incompletely structured, under physiological conditions began to be reported almost sixty years ago, with many additional papers as early as the next two decades [4-8]. Since the 1970s, an increased stream of disordered protein examples has been reported, and many of these are described in our database of intrinsically disordered proteins [9, 10]. In addition, this database contains a

1-4244-1509-8/07/$25.00 ©2007 IEEE

bibliography that shows explosive growth over the last few years, with more than 400 entries from 2006 alone.

One possibility is that the crowded conditions inside the cell cause intrinsically disordered proteins to fold into 3D structure. One test of this idea is to subject intrinsically disordered proteins to molecular crowding by adding high concentrations of agents such as glucose. Such in vitro molecular crowding experiments can induce folding of an acid-unfolded globular protein [11], but fail to induce folding in several intrinsically disordered proteins [11, 12].

Further recent studies along these lines suggest that some proteins remain unfolded even in the highly crowded environment inside the cell [13-15]. Another in-cell NMR report [16] involving some of the same authors was later retracted because protein leakage from the cells produced misleading data [17]. It is argued that the earlier experiments [13, 14] did not suffer from the same leakage problems, which appear to be specific for the protein used in the later studies [17]. Overall, these experiments provide some evidence that intrinsically disordered remain incompletely folded inside the cell, but clearly more experiments are needed to increase confidence in these studies.

A number of different terms have been used to describe these proteins, including natively denatured [18], natively unfolded [19], intrinsically unstructured [20], and several variants of disordered [1, 21, 22]. By now, several reviews on these proteins have appeared [11, 12, 23-27]

The development of a predictor of intrinsic protein disorder from amino acid sequence requires the prior assumption that disordered regions from different proteins have sequence features in common. Stated in another way, developing a predictor of protein disorder is in some sense testing the hypothesis that, just as the amino acid sequence codes for the 3D structure of a protein, the amino acid sequence codes also for the lack of 3D structure. If disorder is encoded in the amino acid sequence, then developing predictors of disorder provides a means to understand "the protein disorder code." In this regard, certain amino acids have been found to be highly "order-promoting" (namely cysteine, tryptophan, tyrosine, isoleucine, phenylalanine, valine, leucine, histidine, threonine, and asparagine) while others are highly "disorder-promoting" (namely aspartic acid, methionine, lysine, arginine, serine, glutamine, proline, and glutamic acid) [28].

A significant development was the inclusion of disorder prediction among the exercises in the Critical Assessment of Structure Prediction (known as CASP), beginning with the 5th CASP event and continuing in subsequent events [29, 30]. This has helped to stimulate the rapid development of at least 25 different predictors of protein disorder. A collection of links to many, if not most, of these is maintained at the Database of Disordered Protein website (www.disprot.org).

Several disordered protein predictors have been compared in recent publications [28, 31-36]. As more disordered proteins have been identified, and as more sophisticated machine learning methods have been applied, the per residue prediction accuracy has risen from ~70% to ~85%. A likely-to-be significant impediment to further improvement is the misclassification of the residues in the training sets.

Application of the disorder predictors to various organisms in the three domains of life, namely, prokaryotes, archaea, and eukaryotes, reveals a large increase in disorder among the eukaryotes compared to the other two types of organisms [31, 37, 38]. One speculation to account for this observation is that the increased disorder reflects the increased need for signaling and coordination among the various organelles in the more complex eukaryotic domain [39].

The recent explosion of papers on intrinsically disordered protein contains many new discoveries on these proteins by a large number of investigators. There is neither time nor space to adequately cover these important advances. We hope that other researchers in this field will not be offended by our focus on our own work for this paper, which has been written to accompany our lecture.

In the following, we will present four short stories that briefly review recent research on disordered proteins published by our group. These include the following: 1. a bioinformatics study of the relationship between disorder and function in the Swiss Protein Database [40-42]; 2. the mechanisms by which one disordered region can bind to many partners and by which many different disordered sequences can bind to one site on one protein partner [43] thereby contributing to the complex protein-protein interaction networks that are observed in nature; 3. the observation that regions of mRNA that undergo alternative splicing code for disordered protein much more often than they code for structured protein [44]; and 4. a novel method for drug discovery based on regions of disordered protein [45]. The novel drug discovery method suggests how the observations in the first three studies might be put to practical use.

II. INTRINSIC DISORDER AND PROTEIN FUNCTION

Our overall goal is to understand relationships between amino acid sequence and protein function so that, given a new sequence, possible functions could be suggested to interested experimentalists for laboratory testing. For proteins that form 3D structure, this is a well developed problem, but for intrinsically disordered proteins, work on this problem is just beginning. First we will very briefly review function prediction for structured proteins, and then we will compare and contrast the very limited amount of work in this area for intrinsically disordered proteins.

A. Function prediction for structured proteins

For structured proteins, sequence homology, if obvious enough, can provide leads regarding protein function [46-48]. Attempts to improve sequence matching for function prediction have been carried out [49]. If no suggestive homologue can be found, an alternative approach is to determine the 3D structure and then to search structure for functional clues, such as residues positioned in space like the same or functionally similar residues in known active sites [50-52]. Often evolution within a family of related proteins can be helpful by means of the evolutionary trace approach [53]. Recent advances have

been made in the assessment of binding sites using both structural and sequence homology[54].

B. Function prediction for disordered proteins

Our first efforts to associate disorder with function were carried out by means of manual literature searches. In the development of our protein disorder predictors, we wanted to use disorder characterized by methods other than missing coordinates in X-ray structures, especially to test whether disorder identified by different methods was different at the amino acid sequence level [55]. Therefore, we had accumulated manuscripts describing disordered proteins and regions of disorder characterized by other methods such as NMR, circular dichroism, small angle X-ray scattering, and so on. In addition, we found many examples in which the disorder indicated by missing coordinates in X-ray crystal structures had been confirmed by other methods. Given these proteins and their associated manuscripts, we then carried out literature searches for functions associated with these well studied disordered protein examples. Out of about 100 disordered proteins and regions, these manual searches identified 27 different functions, and at least one (and commonly more than one) of these functions was found to be associated with > 80% of the disordered proteins or regions. Of course when a given disordered region or protein has no associated function, it is unclear whether the given disordered protein has no function or whether the function of the given disordered protein has simply not yet been found [56, 57].

For structured proteins, proteins can be grouped together if they display a common 3D fold as for example in the CATH [58] and SCOP [59] databases. Often these proteins with common folds have recognizable sequence similarity and so can be grouped into evolutionarily-related protein families. Sometimes, proteins have similar folds without recognizable sequence similarity [60].

Just as for structured proteins, disordered proteins can be grouped into related sets by sequence matching. However, probably due to the absence of structural constraints, disordered proteins tend to accumulate mutations at higher rates than do structured proteins [61] so sequence matching might easily miss relationships between two disordered proteins. In addition, the conservation of the functionally important residues within a disordered region would tend to be obscured due to the high overall mutation rate. In the absence of structure, can sequence features (rather than sequence matching) be used to organize disordered proteins and regions into functional sets?

Over the years we have tried various clustering algorithms to identify functionally related groups of disordered proteins but so far with little or no success. We developed an alternative approach that partitioned a set of disordered proteins and regions into groups based on predictions of disorder. For this approach, a group of disordered proteins was randomly partitioned into two sets and disorder predictors were developed for each group. The two predictors were then applied to all the proteins, and the proteins were redistributed according to which predictor was more accurate. Predictor training was done again on the two redistributed sets, the competition was repeated, and the redistribution was repeated. These steps were carried out iteratively until assigned partitions converged. To test for reproducibility, the original group was randomly divided into two sets several times, and the process was repeated for each new initial partition. The final sets of proteins were very similar for the different initializations, with just a few of the proteins changing their associations in the different repetitions. [62]

Next, the overall process was repeated for partitioning into three sets, into four sets, into five sets and into six sets. If the process were meaningful, one would expect improved agreement between disorder prediction and observation because the disordered proteins within a partitioned set would be more homogenous. Prediction clearly improved for partitions of two and three sets, but showed little or no improvement for increasing the number of partitions. These sets of disordered proteins were called flavors. For the division into three sets, the three distinct flavors were named V, C and S. [62]

Finally, the functions of the various proteins in each set were identified. Some association was found between these flavors and the observed functions, *e.g.* S was associated with protein binding, V was associated with RNA binding, and C was associated with posttranslational modification sites [62].

In our opinion, this approach should be revisited. There were several simplifications in this study that might have diminished the ability to discriminate different disorder flavors. In turn, the reduced ability to discriminate different flavors likely reduced the detection of flavor-function relationships.

A completely different approach is to search for the few, function-associated residues that remain conserved in the sea of changes among the surrounding disordered regions. Such conserved residues have been called Eukaryotic Linear Motifs (ELMs) and methods for their discovery from sequence, analogous to finding transcription factor binding sites, have been developed [63-65].

The overall idea is to search for overabundance of particular residues in regions of sequence that lie outside of Pfam domains. The sets of sequences to be tested typically bind to one specific partner. Thus, evidently the conserved residues represent a binding motif within a linker between (Pfam) structured domains or in a disordered tail at the carboxy or amino terminus of a (Pfam) structured domain [65].

We noticed several particular examples in which binding sites within disordered regions coincided with dips in our disorder prediction plots, especially PONDR VL-XT plots [66], so we developed a predictor of binding sites within disordered regions based on disorder prediction [67]. We suggested that these segments contain molecular recognition features or MoRFs. Experimentalists have successfully used our predictors to discover sites of protein-protein interactions that were subsequently confirmed in laboratory experiments [68, 69]. Other studies have independently verified similarly predicted interactions [70].

The ELM server predicts binding motifs based on overabundance of certain residues in regions known to bind to

a common partner. The MoRF predictor identifies sequence features commonly used for partner binding. Tompa and coworkers [71] showed that at least some MoRFs and ELMs have common characteristics. Developing a combination of the MoRF and ELM approaches might prove to be very useful.

The ELM approach excludes Pfam domains from analysis, which brings the focus most of the time to intrinsically disordered regions. However, a small fraction of the Pfam domains contains conserved regions of predicted disorder [72] and these disordered regions are implicated in biological functions [73], thus giving a set of disorder-associated functional regions not located by ELM analysis. Further study of these Pfam-associated regions of disorder is necessary.

More recently we carried out an analysis of the functional annotation over the entire SwissProt database from a structured-versus-disordered point of view [40-42]. The overall idea was to find keywords associated with 20 or more proteins in SwissProt. For each keyword-associated set, a length-matching set of random proteins was drawn from SwissProt. Order-disorder predictions were carried out for the keyword-associated sets and for the random sets. If a function described by a given keyword were carried out by a long region of disordered protein, one would expect the keyword-associated set to have a greater amount of predicted disorder compared to the random set. The keyword-associated set would have less prediction of disorder compared to the random set if the keyword-associated function were carried out by structured protein. Given the two sets of predictions for the pairs of sets, it is possible to calculate the p-values, where a p-value > 0.95 suggests a disorder-associated function, a p-value < 0.05 suggests an order-associated function, and intermediate p-values are ambiguous.

Out of 710 keywords each being assigned to at least 20 proteins, 310 had p-values < 0.05, suggesting order-associated functions, 238 had p-values > 0.95, suggesting disorder-associated functions, and the remainder, 170, gave intermediate p-values, yielding ambiguity in the likely function-structure associations [40-42] .

When the functional keywords were partitioned into eleven functional categories (Biological processes, cellular components, developmental stage, etc.) order-associated keywords were found for seven of the categories, but disorder-associated keywords were found for all eleven categories [40]. This observation supports a previous conjecture that the functional repertoire is larger for disordered proteins compared to that for structured proteins [21].

Considering the biological processes category, the order-associated keywords nearly all described processes carried out by (necessarily structured) enzymes (examples: amino acid biosynthesis, purine biosynthesis, lipid synthesis, etc.) or by (necessarily structured) integral membrane proteins (electron transport, sugar transport, ion transport). On the other hand, in this same category, the disorder-associated keywords described processes that typically involve control or regulation (differentiation, transcription, cell cycle, growth regulation, etc.). These observations slightly broaden an earlier conjecture that structured proteins are primarily associated with catalysis while disordered proteins are associated with signaling and regulation [21, 74].

Finally, it is interesting to compare the individual keywords associated with disorder prediction and with those associated with the absence of disorder prediction (which indicate structure-associated functions). Ribonucleoprotein and ribosomal protein are two disorder-associated keywords with the highest Z-scores (values of 22.1 and 20.6, respectively). Interestingly, the Z-scores drop off to values less than 10 after just a few proteins. Oxidoreductase and transferase are the order-associated keywords with the highest Z-scores (values of -29.5 and -24.5, respectively). Furthermore, the drop-off to values less than 10 occurs more slowly for the order-associated keywords. One possible explanation is that the structured regions for most of the proteins comprise most of the amino acid sequence for the given protein whereas the disordered region might comprise a small part of the entire sequence.

Another interesting feature of these data is that the top 20 order-associated keywords all end in "ase," indicating that all are enzymes of one type or another. This suggests that, for the order-associated keywords, the overall approach works rather well. Although some laboratory genetic engineering experiments have yielded molten globules with enzymatic activity [75], to our knowledge the currently known natural enzymes are structured proteins.

Further studies on the disorder-associated keywords involved ranking the proteins in each category by Z-score and then carrying out manual literature searches for evidence of association between disorder and function for the highest-ranking proteins. Indeed, for a significant fraction of the high Z-score proteins with functions predicted to be associated with disorder, an association between disorder and function was confirmed by these manual literature searches [41, 42].

The tedious work of confirming the associations between disorder and function needs to be carried out for more of the protein groups in this study. It would then be interesting to study these groups of proteins by the methods described above or by new methods to find sequence-function relationships for disorder-associated functions. Such work would provide the basis for enabling researchers to infer function from sequence.

III. Intrinsic Disorder and Protein-Protein Interaction Networks

Protein-protein interaction networks involve a few proteins with many partners (called hub proteins or hubs) and many proteins with a few partners. The architecture and evolution of these networks comprise a very active area of research, *e.g.* [76-78].

In a two-page News and Views article [79] commenting on a one-page article [80] on protein-protein interaction network architecture in the same issue of Nature, it was suggested that the ability of hub proteins to bind to many partners might depend on new principles. The question is, in essence, what feature of protein structure enables binding diversity?

We opened this paper with 70-year old conjectures that unfolded, dynamic protein ensembles could enable binding

diversity. Since those initial conjectures, several additional papers, including several based on experimental data, suggested that lack of structure (e.g. disorder) could enable binding diversity [81-84].

To further test the roles of disorder in protein-protein interaction networks, first we collected a set of structurally characterized hub proteins [85]. We found several hub proteins to be entirely disordered from one end to the other, and yet to be capable of binding large numbers of partners. Other hubs contained both ordered and disordered regions. For these hubs, many, but not all, of the interactions mapped to the regions of disorder. Two highly structured hubs were found. For both of these structured hubs, the partners were found to be entirely disordered.

Overall, our initial study suggested two primary mechanisms by which disorder is utilized in protein-protein interaction networks. Several groups have tested these overall ideas further via bioinformatics studies on collections of hub proteins. Several of these studies support one of the two primary mechanisms, namely the common use of disordered regions by hub proteins to bind to multiple partners [86-90]. Additional bioinformatics studies refine the analysis further with the suggestion that disorder is very commonly used for regions that bind sequentially to multiple partners (so called "date hubs" [90]).

Bioinformatics investigations of the binding partners of two mostly structured hubs, calmodulin and 14-3-3, suggest that the binding regions of their partners are very likely to be located in regions of disorder [91, 92] However, it has proven very difficult to globally test whether structured hubs bind to disordered partners. A difficulty with such studies is that the partners often contain both order and disorder, and the disordered regions typically comprise only small fractions of the partner sequences. Thus, without knowing the binding regions of the partners, it is difficult to estimate whether disorder is involved or not.

A search of PDB has revealed more than 2,500 short regions of one protein associated with a globular domain of a second protein. Further studies suggest that the short regions were very likely disordered before binding to their structured partners. Many of these short regions are related to each other, so the number reduces to several hundred families when they are grouped by sequence similarity. Most of these interactions are associated with signaling and regulation [93]. While we have not yet correlated these data with hub protein information, they show that disordered regions binding to structured partners is common, suggesting in turn that structured hub proteins may commonly bind to disordered partners.

Several years ago, without specific regard to networks, we considered possible roles of disorder in protein-protein interactions. We suggested that one disordered region could bind to many partners; we called this one-to-many signaling. We further suggested that flexibility would enable multiple disordered regions to bind to one site on one partner; we called this many to one signaling [84]. While papers back to 1936 suggest that flexibility could enable one protein to bind to many partners, we are unaware of a paper earlier than ours suggesting that flexibility would enable multiple disordered regions with different sequences to bind to single site on the partner protein.

To understand the structural principles in more detail, we recently studied carefully the structures of a one-to-many example (namely, the disordered regions in p53 binding to their many partners) and also the structures of a many-to-one example (namely many different disordered partners associating with the same binding site of 14-3-3).

For the one-to-many signaling example (using the structures currently in the PDB), a single disordered region of p53 is observed to form a helix when associating with one partner, a sheet with a second partner, an irregular structure with a third partner, and an irregular structure with a completely different trajectory with a fourth partner. The set of residues involved in these interactions exhibit a very high extent of overlap along the sequence [43].

The solvent accessible surface area (ASA) can be calculated from the three dimensional structure of a protein analytically [94] or numerically [95]. The amount ASA that becomes inaccessible upon complex formation is likewise easy to estimate by existing methods [96] and is expressed as the change in the ASA or as the ΔASA.

Plotting the ΔASA for each amino acid versus its position in the sequence gives a binding profile. The binding profiles for the single region of p53 bound to four different partners are completely different. It is as if the same sequence is "read" by the different partners in entirely different ways [43].

For a disordered region that binds to a partner, the binding profile comprises a highly localized set of amino acids. However, when the region of interaction is structured, the binding profile gives two (or more) localized sets separated by a considerable distance (or distances) along the sequence. The separated profile occurs because structured proteins bring together different regions of sequence to form the active site, which then leads to a binding profile that involves these separated regions of sequence. Interestingly, the DNA binding domain of p53 exhibits a complex but very distinctive binding profile when associating with DNA as compared to p53BP1 and p53BP2 [43] binding profiles. While both p53BP1 and p53BP2 profiles show similar localizations on the p53 sequence, their detailed structures are very different.

For a many-to-one signaling example (using structures currently in the PDB), five disordered sequences associated within a single binding groove in 14-3-3 were studied. As suggested previously [84], the flexibility of the disordered regions enabled them to fit into a common binding site. Not only backbone flexibility, but also side-chain flexibility is implicated in the movements needed for the different sequences to be able to fit into the common binding site [43].

What was not discussed in our earlier publication is the flexibility on the structured side of the complex (*e.g.* the flexibility in 14-3-3). In the 14-3-3 example, flexibility on the structured protein side of the complex also played a very important role in enabling the binding of many disordered segments to a single partner.

In 1958 Koshland proposed his now famous induced fit hypothesis. The original manuscript described a thought-experiment involving peptides of different sequences binding to a common site, thus requiring an "induced fit" to accommodate the different structures of the different sequences. While "induced fit" has been accepted and is described in current text books, the examples typically involve domain shifts. To our knowledge, our studies on the multiple peptides binding to 14-3-3 involve the first direct test of Koshland's original induced fit hypothesis. Further comparisons of these interactions confirm the Koshland's induced fit hypothesis with its original use of different sequences bound to a common partner and provide insight regarding the degree of structural change upon binding (manuscript in preparation).

IV. Intrinsic Disorder and Alternative Splicing

Alternative splicing is a process whereby multiple, mature mRNAs are produced from a single precursor pre-mRNA by the inclusion and omission of different segments [97, 98]. The joined segments leading to the mRNA are called exons and the omitted segments are called introns [99]. Alternative splicing is only prevalent in multicellular eukaryotes [100]. Current estimates are that 40-60% of human genes yield proteins via alternative splicing [101-103], and in many cases multiple proteins are produced from a single gene. These observations suggest that alternative splicing provides an important mechanism for enhancing the diversity of the proteome in multicellular eukaryotes [104].

Alternative splicing impacts many protein functions such as ligand binding, enzymatic activity, and protein-protein interactions [105-107]. Not surprisingly, abnormal alternative splicing has been associated with human diseases, including myotonic dystrophy [108], Axoospermia [109], Alzheimer's [110] and cancer [111].

Alternative splicing within a structured region of a protein would be expected to lead to significant problems with protein folding, thus leading simply to loss of function. In some cases, however, the alternatively spliced structured protein can maintain function.

Attempts have been made to predict the effects of alternative splicing on protein structure (and function) by homology modeling [112] and by a more sophisticated analysis that attempts to model the structural changes that are likely to result from the alternative splicing event [113]. This modeling could be attempted because the observed splicing alterations were of small size, were located on the protein surface, and were preferentially located in coil regions [113]. For these examples, the results suggested that the different splice variants folded into the same overall structure, with only slight, but perhaps functionally important, structural perturbations.

So far we have been able to find only five alternatively spliced isozyme pairs with structures determined for both partners [114-118]. Consistent with the ideas presented in the modeling paper [113], the protein isoforms did fold basically the same. These structures were not significantly perturbed because alternatively spliced segments were either short regions on the surface of the structure (in two cases) or disordered regions (in the remaining three cases). Of the two spliced surface segments the largest structural perturbation occurs where the RNA coding for a short helix was omitted in the shorter splice variant, which leads to a slight rearrangement when the adjacent secondary structure elements adjusted relative to each other due to the lack of the intervening, but short, helix. While this analysis would certainly benefit by additional structural data to test specific structural models, it suggests that alternatively spliced sites in ordered proteins are preferentially located in intrinsically disordered regions.

While the structural implications are interesting, it should be noted that only a small fraction of alternative splicing events clearly map to structured proteins. For example, while 40% to 60% of mammalian (human) genes are estimated to undergo alternative splicing, less than 20 individual proteins (out of over 6,000 structures of proteins from multicellular organisms in PDB[1]) are known to have splice variants that map to the regions of structure (unpublished observations). What about all the other splicing events that (at least so far) don't map to any currently known structure?

For disorder-associated splicing, we hypothesized that the structural problems discussed above would be solved if the mRNA regions that vary for different isoforms were to code for regions of intrinsically disordered protein, as is observed for the three of five alternatively spliced enzyme pairs that we could find [114-118]. For regions of disorder, there could be multiple splice variants and the spliced regions could be long because structural perturbation is simply not an issue.

To test the possible association of alternative splicing with disorder, we assembled a set of human proteins that contained structurally characterized regions of structure and regions of disorder. We then searched for alternative splicing for each of these proteins. At the time of this study, we could find 46 human proteins with 75 alternatively spliced segments that met these criteria [44]. Of these 75 alternatively spliced regions of RNA, 57% coded for entirely disordered protein, 24% coded for both ordered and disordered protein (with the splice boundaries very often in, or very near to, the disordered regions), and just 19% coded for fully structured regions [44]. These 75 disorder-associated alternatively spliced regions are much larger than the number of regions known by direct experiment to be directly associated with regions of structure. Nevertheless, it would be very useful to enlarge the dataset.

To enlarge the dataset, we collected all of the proteins in SwissProt with identified alternatively spliced isoforms, giving 558 proteins with 1,266 regions that are absent on one isoform due to alternative splicing. We predicted order/disorder for these alternatively spliced proteins as well as for the 46 proteins of known structure. For both sets of data, we plotted the frequency of observation versus 0-20% predicted disorder, 20-40% predicted disorder, etc. For the 75 alternatively spliced regions of known structure, the predictions and observations gave excellent agreement. For the 1,266 regions that were present or absent in different isoforms due to alternative splicing, the frequencies of the various percentages of disorder

[1] PDB contains 6,565 structures of metazoan proteins filtered for 95% sequence identity as of August 2007.

closely matched the corresponding frequencies for the set of 75 with known structure. These data provide evidence that a large fraction of alternative splicing occurs in regions of RNA that code for disordered protein.

About 50% of mammalian proteins have predicted disordered regions of 30 residues or longer [74]. This estimate is in the ball park of 40% to 60% of proteins undergoing alternative splicing. Thus, intrinsic disorder is certainly common enough for 80% of alternative splicing events to occur in regions of disorder.

Above we discuss the roles of disorder in various protein functions and in protein-protein interaction networks. Alternative splicing could readily alter such functions and could readily alter protein-protein interaction networks. We suggest that the linkage between alternative splicing and signaling by disordered region provides a novel and plausible mechanism for understanding the origins of cell differentiation, which ultimately gave rise to multicellular organisms in nature [44]. New studies are needed to test these ideas.

V. Intrinsic Disorder and Drug Discovery

Protein-protein interactions have long been regarded as a potential new source of drug targets. Systems biology approaches are mapping out the protein interactome, and a deeper understanding of these results is likely to indicate desirable drug targets [119]. However, attempts to develop drug molecules that block protein-protein interactions have generally not been successful [120, 121]. Indeed, our searches for drug molecules known to act by blocking protein-protein interactions have so far come up empty.

Upon this rather negative background, promising examples are creating a new optimism [122-124]. As described in these recent papers, several drug-like lead compounds act by blocking protein-protein interactions, and these are being actively investigated with the goal of developing new drugs.

We noticed that one interaction of interest, namely the interaction between p53 and Mdm2, has been the subject of numerous studies involving several molecules that could lead to a drug [125-127]. This particular example caught our attention because the same region of p53 involved in this interaction is known to be intrinsically disordered [128]. However, none of the papers discussing the promise of blocking this interaction even mention that the p53 side of the complex involves a disorder-to-order transition upon binding.

We used our bioinformatics and computational structural biology tools to investigate this interaction and discovered several features that explained why this region is so promising as a drug target. We then searched for analogous features (some of which are proprietary) over the entire human genome. The result was thousands of possible new drug targets that involve one disordered partner (and one structured partner) and that partition very nicely over the major diseases [45]. While much work needs to be done to turn these hypotheses into actual drug molecules, these discoveries certainly provide a new avenue for drug discovery that needs to be tested.

Here we point out that a protein-protein interaction involving one disordered partner and one structured partner is likely in many cases to be a good target for drug discovery. First, unlike the interactions between two globular proteins, the interaction surface involving one disordered partner is not flat. Typically the structured partner's surface has a groove, and the disordered region forms a helix with a hydrophobic face that nestles into the groove. We observe these features over and over in our MoRF dataset [93] and in the examples used to develop the MoRF predictor [67]. In the case of the p53-Mdm2 interaction, the Mdm2 forms the groove and the p53 binding site is predicted exactly as a MoRF.

Since one of the partners undergoes a disorder-to-order transition, some of the binding energy is spent to overcome the high entropy of the unfolded state. From the entropy point of view, such an interaction is likely to be weaker than an interaction between two structured proteins and thus will be easier to block with a small molecule competitor.

While protein disorder is not mentioned in any of the papers describing how a small molecule can block protein-protein interactions, we have found that 4 of the 8 examples described in the recent reviews [122, 124] involve one structured partner and one disordered partner, with 3 of the 4 disordered segments becoming helix upon binding. Thus, the p53-Mdm2 complex is not the only member of this class currently known to be blocked by a small drug-like molecule. We fully expect many more examples to appear shortly, and for some of these examples to lead to useful drug molecules.

VI. Summary Comments

If we link the concepts here with those in Section V, a very exciting possibility is that these approaches will lead to tissue-specific drugs via tissue-specific alternative splicing in disordered regions.

If we link the concepts here with those in Section III, we can see how one drug molecule could block one protein-protein interaction (for one-to-many signaling interactions) but could block many interactions (for many-to-one signaling interactions).

Finally, if we link the concepts here with those in Sections II and III, we can possibly find drugs aimed at a wide variety of signaling and regulatory functions.

We started applying bioinformatics to disordered proteins about 12 years ago, with our first paper just slightly more than 10 years ago. In these few short years our understanding of the biological importance of these proteins has increased markedly. We hope that the next 10 years will see practical outcomes (such as promising leads for new drug molecules) from this work.

Acknowledgment

We would like to thank NIH, NSF, DOE, and the Indiana Genomics Initiative (funded in part by the Lilly Endowment) for supporting our work. In addition, we would like to thank the many students, faculty members, and other scientists who have collaborated with us on this work over the past 10 years.

REFERENCES

[1] P. Romero, Z. Obradovic, C. Kissinger, J. E. Villafranca, and A. K. Dunker, "Identifying disordered regions in proteins from amino acid sequence," *Proceedings of the IEEE International Conference on Neural Networks,* vol. 1, pp. 90-95, 1997.

[2] K. Landsteiner, *The specificity of serological reactions*. Baltimore: C. C. Thomas, 1936.

[3] L. Pauling, "A Theory of the Structure and Process of Formation of Antibodies," *J. Am. Chem. Soc.,* vol. 62, pp. 2643-2657, 1940.

[4] F. Karush, "Heterogeneity of the binding sites of bovine serum albumin," *Journal of the American Chemical Society,* vol. 72, pp. 2705-2713, 1950.

[5] T. L. McMeekin, "Milk proteins," *Journal of Food Protection,* vol. 15, pp. 57-63, 1952.

[6] M. Halwer, "Light-scattering study of effect of electrolytes on alpha- and beta-casein solutions," *Arch Biochem Biophys,* vol. 51, pp. 79-87, Jul 1954.

[7] B. Jirgensons, "Classification of proteins according to conformation," *Die Makromolekulare Chemie,* vol. 91, pp. 74-86, 1966.

[8] R. F. Doolittle, "Structural aspects of the fibrinogen to fibrin conversion," *Adv Protein Chem,* vol. 27, pp. 1-109, 1973.

[9] S. Vucetic, Z. Obradovic, V. Vacic, P. Radivojac, K. Peng, L. M. Iakoucheva, M. S. Cortese, J. D. Lawson, C. J. Brown, J. G. Sikes, C. D. Newton, and A. K. Dunker, "DisProt: a database of protein disorder," *Bioinformatics,* vol. 21, pp. 137-40, Jan 1 2005.

[10] M. Sickmeier, J. A. Hamilton, T. LeGall, V. Vacic, M. S. Cortese, A. Tantos, B. Szabo, P. Tompa, J. Chen, V. N. Uversky, Z. Obradovic, and A. K. Dunker, "DisProt: the Database of Disordered Proteins," *Nucleic Acids Res,* vol. 35, pp. D786-93, Jan 2007.

[11] A. S. Morar, A. Olteanu, G. B. Young, and G. J. Pielak, "Solvent-induced collapse of alpha-synuclein and acid-denatured cytochrome c," *Protein Sci,* vol. 10, pp. 2195-9, Nov 2001.

[12] S. L. Flaugh and K. J. Lumb, "Effects of macromolecular crowding on the intrinsically disordered proteins c-Fos and p27(Kip1)," *Biomacromolecules,* vol. 2, pp. 538-40, Summer 2001.

[13] M. M. Dedmon, C. N. Patel, G. B. Young, and G. J. Pielak, "FlgM gains structure in living cells," *Proc Natl Acad Sci U S A,* vol. 99, pp. 12681-4, Oct 1 2002.

[14] B. C. McNulty, G. B. Young, and G. J. Pielak, "Macromolecular crowding in the Escherichia coli periplasm maintains alpha-synuclein disorder," *J Mol Biol,* vol. 355, pp. 893-7, Feb 3 2006.

[15] P. Selenko and G. Wagner, "Looking into live cells with in-cell NMR spectroscopy," *J Struct Biol,* vol. 158, pp. 244-53, May 2007.

[16] J. E. Bryant, J. T. Lecomte, A. L. Lee, G. B. Young, and G. J. Pielak, "Protein dynamics in living cells," *Biochemistry,* vol. 44, pp. 9275-9, Jul 5 2005.

[17] J. E. Bryant, J. T. Lecomte, A. L. Lee, G. B. Young, and G. J. Pielak, "Retraction. Protein dynamics in living cells," *Biochemistry,* vol. 46, p. 8206, Jul 10 2007.

[18] O. Schweers, E. Schonbrunn-Hanebeck, A. Marx, and E. Mandelkow, "Structural studies of tau protein and Alzheimer paired helical filaments show no evidence for beta-structure," *J Biol Chem,* vol. 269, pp. 24290-7, Sep 30 1994.

[19] P. H. Weinreb, W. Zhen, A. W. Poon, K. A. Conway, and P. T. Lansbury, Jr., "NACP, a protein implicated in Alzheimer's disease and learning, is natively unfolded," *Biochemistry,* vol. 35, pp. 13709-15, Oct 29 1996.

[20] P. E. Wright and H. J. Dyson, "Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm," *J Mol Biol,* vol. 293, pp. 321-31, Oct 22 1999.

[21] G. W. Daughdrill, G. J. Pielak, V. N. Uversky, M. S. Cortese, and A. K. Dunker, "Natively Disordered Proteins," in *Protein Folding Handbook, Part II*, J. Buchner and T. Kiefhaber, Eds. Weinheim: Wiley-VCH, 2005, pp. 275-357.

[22] A. K. Dunker, "Disordered proteins," in *Encyclopedia of Life Sciences*: John Wiley & Sons, In Press.

[23] A. K. Dunker, J. D. Lawson, C. J. Brown, R. M. Williams, P. Romero, J. S. Oh, C. J. Oldfield, A. M. Campen, C. M. Ratliff, K. W. Hipps, J. Ausio, M. S. Nissen, R. Reeves, C. Kang, C. R. Kissinger, R. W. Bailey, M. D. Griswold, W. Chiu, E. C. Garner, and Z. Obradovic, "Intrinsically disordered protein," *J Mol Graph Model,* vol. 19, pp. 26-59, 2001.

[24] V. N. Uversky, "Natively unfolded proteins: a point where biology waits for physics," *Protein Sci,* vol. 11, pp. 739-56, Apr 2002.

[25] V. N. Uversky, C. J. Oldfield, and A. K. Dunker, "Showing your ID: intrinsic disorder as an ID for recognition, regulation and cell signaling," *J Mol Recognit,* vol. 18, pp. 343-84, Sep-Oct 2005.

[26] H. J. Dyson and P. E. Wright, "Intrinsically unstructured proteins and their functions," *Nat Rev Mol Cell Biol,* vol. 6, pp. 197-208, Mar 2005.

[27] C. Bracken, L. M. Iakoucheva, P. R. Romero, and A. K. Dunker, "Combining prediction, computation and experiment for the characterization of protein disorder," *Curr Opin Struct Biol,* vol. 14, pp. 570-6, Oct 2004.

[28] P. Radivojac, L. M. Iakoucheva, C. J. Oldfield, Z. Obradovic, V. N. Uversky, and A. K. Dunker, "Intrinsic disorder and functional proteomics," *Biophys J,* vol. 92, pp. 1439-56, Mar 1 2007.

[29] J. Moult, K. Fidelis, A. Zemla, and T. Hubbard, "Critical assessment of methods of protein structure prediction (CASP)-round V," *Proteins,* vol. 53 Suppl 6, pp. 334-9, 2003.

[30] J. Moult, K. Fidelis, B. Rost, T. Hubbard, and A. Tramontano, "Critical assessment of methods of protein structure prediction (CASP)--round 6," *Proteins,* vol. 61 Suppl 7, pp. 3-7, 2005.

[31] C. J. Oldfield, Y. Cheng, M. S. Cortese, C. J. Brown, V. N. Uversky, and A. K. Dunker, "Comparing and combining predictors of mostly disordered proteins," *Biochemistry,* vol. 44, pp. 1989-2000, Feb 15 2005.

[32] K. Peng, S. Vucetic, P. Radivojac, C. J. Brown, A. K. Dunker, and Z. Obradovic, "Optimizing long intrinsic disorder predictors with protein evolutionary information," *J Bioinform Comput Biol,* vol. 3, pp. 35-60, Feb 2005.

[33] K. Peng, P. Radivojac, S. Vucetic, A. K. Dunker, and Z. Obradovic, "Length-dependent prediction of protein intrinsic disorder," *BMC Bioinformatics,* vol. 7, p. 208, 2006.

[34] F. Ferron, S. Longhi, B. Canard, and D. Karlin, "A practical overview of protein disorder prediction methods," *Proteins,* vol. 65, pp. 1-14, Oct 1 2006.

[35] E. Melamud and J. Moult, "Evaluation of disorder predictions in CASP5," *Proteins,* vol. 53 Suppl 6, pp. 561-5, 2003.

[36] Y. Jin and R. L. Dunbrack, Jr., "Assessment of disorder predictions in CASP6," *Proteins,* vol. 61 Suppl 7, pp. 167-75, 2005.

[37] A. K. Dunker, Z. Obradovic, P. Romero, E. C. Garner, and C. J. Brown, "Intrinsic protein disorder in complete genomes," *Genome Inform Ser Workshop Genome Inform,* vol. 11, pp. 161-71, 2000.

[38] J. J. Ward, J. S. Sodhi, L. J. McGuffin, B. F. Buxton, and D. T. Jones, "Prediction and functional analysis of native disorder in proteins from the three kingdoms of life," *J Mol Biol,* vol. 337, pp. 635-45, Mar 26 2004.

[39] A. K. Dunker and Z. Obradovic, "The protein trinity--linking function and disorder," *Nat Biotechnol,* vol. 19, pp. 805-6, Sep 2001.

[40] H. Xie, S. Vucetic, L. M. Iakoucheva, C. J. Oldfield, A. K. Dunker, V. N. Uversky, and Z. Obradovic, "Functional anthology of intrinsic disorder. 1. Biological processes and functions of proteins with long disordered regions," *J Proteome Res,* vol. 6, pp. 1882-98, May 2007.

[41] S. Vucetic, H. Xie, L. M. Iakoucheva, C. J. Oldfield, A. K. Dunker, Z. Obradovic, and V. N. Uversky, "Functional anthology of intrinsic disorder. 2. Cellular components, domains, technical terms, developmental processes, and coding sequence diversities correlated with long disordered regions," *J Proteome Res,* vol. 6, pp. 1899-916, May 2007.

[42] H. Xie, S. Vucetic, L. M. Iakoucheva, C. J. Oldfield, A. K. Dunker, Z. Obradovic, and V. N. Uversky, "Functional anthology of intrinsic disorder. 3. Ligands, post-translational modifications, and diseases associated with intrinsically disordered proteins," *J Proteome Res,* vol. 6, pp. 1917-32, May 2007.

[43] C. J. Oldfield, J. Meng, J. Y. Yang, V. N. Uversky, and A. K. Dunker, "Intrinsic Disorder in Protein-Protein Interaction Networks: Case Studies of Complexes Involving p53 and 14-3-3," *Proceeding of BioComp '07,* vol. In Press, 2007.

[44] P. R. Romero, S. Zaidi, Y. Y. Fang, V. N. Uversky, P. Radivojac, C. J. Oldfield, M. S. Cortese, M. Sickmeier, T. LeGall, Z. Obradovic, and A. K. Dunker, "Alternative splicing in concert with protein intrinsic disorder enables increased functional diversity in multicellular organisms," *Proc Natl Acad Sci U S A,* vol. 103, pp. 8390-5, May 30 2006.

[45] Y. Cheng, T. LeGall, C. J. Oldfield, J. P. Mueller, Y. Y. Van, P. Romero, M. S. Cortese, V. N. Uversky, and A. K. Dunker, "Rational drug design via intrinsically disordered protein," *Trends Biotechnol,* vol. 24, pp. 435-42, Oct 2006.

[46] T. Webster, H. Tsai, M. Kula, G. A. Mackie, and P. Schimmel, "Specific sequence homology and three-dimensional structure of an aminoacyl transfer RNA synthetase," *Science,* vol. 226, pp. 1315-7, Dec 14 1984.
[47] J. W. Thornton and R. DeSalle, "Gene family evolution and homology: genomics meets phylogenetics," *Annu Rev Genomics Hum Genet,* vol. 1, pp. 41-73, 2000.
[48] I. Friedberg, "Automated protein function prediction--the genomic challenge," *Brief Bioinform,* vol. 7, pp. 225-42, Sep 2006.
[49] Y. Ofran, M. Punta, R. Schneider, and B. Rost, "Beyond annotation transfer by homology: novel protein-function prediction methods to assist drug discovery," *Drug Discov Today,* vol. 10, pp. 1475-82, Nov 1 2005.
[50] S. C. Bagley and R. B. Altman, "Characterizing the microenvironment surrounding protein sites," *Protein Sci,* vol. 4, pp. 622-35, Apr 1995.
[51] S. D. Mooney, M. H. Liang, R. DeConde, and R. B. Altman, "Structural characterization of proteins using residue environments," *Proteins,* vol. 61, pp. 741-7, Dec 1 2005.
[52] J. S. Fetrow and J. Skolnick, "Method for prediction of protein function from sequence using the sequence-to-structure-to-function paradigm with application to glutaredoxins/thioredoxins and T1 ribonucleases," *J Mol Biol,* vol. 281, pp. 949-68, Sep 4 1998.
[53] O. Lichtarge, H. R. Bourne, and F. E. Cohen, "An evolutionary trace method defines binding surfaces common to protein families," *J Mol Biol,* vol. 257, pp. 342-58, Mar 29 1996.
[54] G. Lopez, A. Valencia, and M. L. Tress, "firestar--prediction of functionally important residues using structural templates and alignment reliability," *Nucleic Acids Res,* vol. 35, pp. W573-7, Jul 1 2007.
[55] E. Garner, P. Cannon, P. Romero, Z. Obradovic, and A. K. Dunker, "Predicting Disordered Regions from Amino Acid Sequence: Common Themes Despite Differing Structural Characterization," *Genome Inform Ser Workshop Genome Inform,* vol. 9, pp. 201-213, 1998.
[56] A. K. Dunker, C. J. Brown, J. D. Lawson, L. M. Iakoucheva, and Z. Obradovic, "Intrinsic disorder and protein function," *Biochemistry,* vol. 41, pp. 6573-82, May 28 2002.
[57] A. K. Dunker, C. J. Brown, and Z. Obradovic, "Identification and functions of usefully disordered proteins," *Adv Protein Chem,* vol. 62, pp. 25-49, 2002.
[58] L. H. Greene, T. E. Lewis, S. Addou, A. Cuff, T. Dallman, M. Dibley, O. Redfern, F. Pearl, R. Nambudiry, A. Reid, I. Sillitoe, C. Yeats, J. M. Thornton, and C. A. Orengo, "The CATH domain structure database: new protocols and classification levels give a more comprehensive resource for exploring evolution," *Nucleic Acids Res,* vol. 35, pp. D291-7, Jan 2007.
[59] A. Andreeva, D. Howorth, S. E. Brenner, T. J. Hubbard, C. Chothia, and A. G. Murzin, "SCOP database in 2004: refinements integrate structure and sequence family data," *Nucleic Acids Res,* vol. 32, pp. D226-9, Jan 1 2004.
[60] L. Holm, C. Ouzounis, C. Sander, G. Tuparev, and G. Vriend, "A database of protein structure families with common folding motifs," *Protein Sci,* vol. 1, pp. 1691-8, Dec 1992.
[61] C. J. Brown, S. Takayama, A. M. Campen, P. Vise, T. W. Marshall, C. J. Oldfield, C. J. Williams, and A. K. Dunker, "Evolutionary rate heterogeneity in proteins with long disordered regions," *J Mol Evol,* vol. 55, pp. 104-10, Jul 2002.
[62] S. Vucetic, C. J. Brown, A. K. Dunker, and Z. Obradovic, "Flavors of protein disorder," *Proteins,* vol. 52, pp. 573-84, Sep 1 2003.
[63] R. Aasland, C. Abrams, C. Ampe, L. J. Ball, M. T. Bedford, G. Cesareni, M. Gimona, J. H. Hurley, T. Jarchau, V. P. Lehto, M. A. Lemmon, R. Linding, B. J. Mayer, M. Nagai, M. Sudol, U. Walter, and S. J. Winder, "Normalization of nomenclature for peptide motifs as ligands of modular protein domains," *FEBS Lett,* vol. 513, pp. 141-4, Feb 20 2002.
[64] P. Puntervoll, R. Linding, C. Gemund, S. Chabanis-Davidson, M. Mattingsdal, S. Cameron, D. M. Martin, G. Ausiello, B. Brannetti, A. Costantini, F. Ferre, V. Maselli, A. Via, G. Cesareni, F. Diella, G. Superti-Furga, L. Wyrwicz, C. Ramu, C. McGuigan, R. Gudavalli, I. Letunic, P. Bork, L. Rychlewski, B. Kuster, M. Helmer-Citterich, W. N. Hunter, R. Aasland, and T. J. Gibson, "ELM server: A new resource for investigating short functional sites in modular eukaryotic proteins," *Nucleic Acids Res,* vol. 31, pp. 3625-30, Jul 1 2003.
[65] V. Neduva, R. Linding, I. Su-Angrand, A. Stark, F. de Masi, T. J. Gibson, J. Lewis, L. Serrano, and R. B. Russell, "Systematic discovery of new recognition peptides mediating protein interaction networks," *PLoS Biol,* vol. 3, p. e405, Dec 2005.
[66] E. Garner, P. Romero, A. K. Dunker, C. Brown, and Z. Obradovic, "Predicting Binding Regions within Disordered Proteins," *Genome Inform Ser Workshop Genome Inform,* vol. 10, pp. 41-50, 1999.
[67] C. J. Oldfield, Y. Cheng, M. S. Cortese, P. Romero, V. N. Uversky, and A. K. Dunker, "Coupled folding and binding with alpha-helix-forming molecular recognition elements," *Biochemistry,* vol. 44, pp. 12454-70, Sep 20 2005.
[68] A. J. Callaghan, J. P. Aurikko, L. L. Ilag, J. Gunter Grossmann, V. Chandran, K. Kuhnel, L. Poljak, A. J. Carpousis, C. V. Robinson, M. F. Symmons, and B. F. Luisi, "Studies of the RNA degradosome-organizing domain of the Escherichia coli ribonuclease RNase E," *J Mol Biol,* vol. 340, pp. 965-79, Jul 23 2004.
[69] J. M. Bourhis, K. Johansson, V. Receveur-Brechot, C. J. Oldfield, K. A. Dunker, B. Canard, and S. Longhi, "The C-terminal domain of measles virus nucleoprotein belongs to the class of intrinsically disordered proteins that fold upon binding to their physiological partner," *Virus Res,* vol. 99, pp. 157-67, Feb 2004.
[70] R. L. Kingston, D. J. Hamel, L. S. Gay, F. W. Dahlquist, and B. W. Matthews, "Structural basis for the attachment of a paramyxoviral polymerase to its template," *Proc Natl Acad Sci U S A,* vol. 101, pp. 8301-6, Jun 1 2004.
[71] M. Fuxreiter, P. Tompa, and I. Simon, "Local structural disorder imparts plasticity on linear motifs," *Bioinformatics,* vol. 23, pp. 950-6, Apr 15 2007.
[72] J. W. Chen, P. Romero, V. N. Uversky, and A. K. Dunker, "Conservation of intrinsic disorder in protein domains and families: I. A database of conserved predicted disordered regions," *J Proteome Res,* vol. 5, pp. 879-87, Apr 2006.
[73] J. W. Chen, P. Romero, V. N. Uversky, and A. K. Dunker, "Conservation of intrinsic disorder in protein domains and families: II. functions of conserved disorder," *J Proteome Res,* vol. 5, pp. 888-98, Apr 2006.
[74] L. M. Iakoucheva, C. J. Brown, J. D. Lawson, Z. Obradovic, and A. K. Dunker, "Intrinsic disorder in cell-signaling and cancer-associated proteins," *J Mol Biol,* vol. 323, pp. 573-84, Oct 25 2002.
[75] G. MacBeath, P. Kast, and D. Hilvert, "Redesigning enzyme topology by directed evolution," *Science,* vol. 279, pp. 1958-61, Mar 20 1998.
[76] J. D. Han, N. Bertin, T. Hao, D. S. Goldberg, G. F. Berriz, L. V. Zhang, D. Dupuy, A. J. Walhout, M. E. Cusick, F. P. Roth, and M. Vidal, "Evidence for dynamically organized modularity in the yeast protein-protein interaction network," *Nature,* vol. 430, pp. 88-93, Jul 1 2004.
[77] R. Jansen, H. Yu, D. Greenbaum, Y. Kluger, N. J. Krogan, S. Chung, A. Emili, M. Snyder, J. F. Greenblatt, and M. Gerstein, "A Bayesian networks approach for predicting protein-protein interactions from genomic data," *Science,* vol. 302, pp. 449-53, Oct 17 2003.
[78] D. R. Rhodes, S. A. Tomlins, S. Varambally, V. Mahavisno, T. Barrette, S. Kalyana-Sundaram, D. Ghosh, A. Pandey, and A. M. Chinnaiyan, "Probabilistic model of the human protein-protein interaction network," *Nat Biotechnol,* vol. 23, pp. 951-9, Aug 2005.
[79] J. Hasty and J. J. Collins, "Protein interactions. Unspinning the web," *Nature,* vol. 411, pp. 30-1, May 3 2001.
[80] H. Jeong, S. P. Mason, A. L. Barabasi, and Z. N. Oltvai, "Lethality and centrality in protein networks," *Nature,* vol. 411, pp. 41-2, May 3 2001.
[81] W. E. Meador, A. R. Means, and F. A. Quiocho, "Modulation of calmodulin plasticity in molecular recognition on the basis of x-ray structures," *Science,* vol. 262, pp. 1718-21, Dec 10 1993.
[82] R. W. Kriwacki, L. Hengst, L. Tennant, S. I. Reed, and P. E. Wright, "Structural studies of p21Waf1/Cip1/Sdi1 in the free and Cdk2-bound state: conformational disorder mediates binding diversity," *Proc Natl Acad Sci U S A,* vol. 93, pp. 11504-9, Oct 15 1996.
[83] V. N. Uversky, "A protein-chameleon: conformational plasticity of alpha-synuclein, a disordered protein involved in neurodegenerative disorders," *J Biomol Struct Dyn,* vol. 21, pp. 211-34, Oct 2003.
[84] A. K. Dunker, E. Garner, S. Guilliot, P. Romero, K. Albrecht, J. Hart, Z. Obradovic, C. Kissinger, and J. E. Villafranca, "Protein disorder and the evolution of molecular recognition: theory, predictions and observations," *Pac Symp Biocomput,* pp. 473-84, 1998.
[85] A. K. Dunker, M. S. Cortese, P. Romero, L. M. Iakoucheva, and V. N. Uversky, "Flexible nets. The roles of intrinsic disorder in protein interaction networks," *Febs J,* vol. 272, pp. 5129-48, Oct 2005.
[86] A. Patil and H. Nakamura, "Disordered domains and high surface charge confer hubs with the ability to interact with multiple proteins in interaction networks," *FEBS Lett,* vol. 580, pp. 2041-5, Apr 3 2006.

[87] D. Ekman, S. Light, A. K. Bjorklund, and A. Elofsson, "What properties characterize the hub proteins of the protein-protein interaction network of Saccharomyces cerevisiae?," *Genome Biol,* vol. 7, p. R45, 2006.
[88] C. Haynes, C. J. Oldfield, F. Ji, N. Klitgord, M. E. Cusick, P. Radivojac, V. N. Uversky, M. Vidal, and L. M. Iakoucheva, "Intrinsic disorder is a common feature of hub proteins from four eukaryotic interactomes," *PLoS Comput Biol,* vol. 2, p. e100, Aug 4 2006.
[89] Z. Dosztanyi, J. Chen, A. K. Dunker, I. Simon, and P. Tompa, "Disorder and sequence repeats in hub proteins and their implications for network evolution," *J Proteome Res,* vol. 5, pp. 2985-95, Nov 2006.
[90] G. P. Singh and D. Dash, "Intrinsic disorder in yeast transcriptional regulatory network," *Proteins,* vol. 68, pp. 602-5, Aug 15 2007.
[91] P. Radivojac, S. Vucetic, T. R. O'Connor, V. N. Uversky, Z. Obradovic, and A. K. Dunker, "Calmodulin signaling: analysis and prediction of a disorder-dependent molecular recognition," *Proteins,* vol. 63, pp. 398-410, May 1 2006.
[92] D. M. Bustos and A. A. Iglesias, "Intrinsic disorder is a key characteristic in partners that bind 14-3-3 proteins," *Proteins,* vol. 63, pp. 35-42, Apr 1 2006.
[93] A. Mohan, C. J. Oldfield, P. Radivojac, V. Vacic, M. S. Cortese, A. K. Dunker, and V. N. Uversky, "Analysis of molecular recognition features (MoRFs)," *J Mol Biol,* vol. 362, pp. 1043-59, Oct 6 2006.
[94] M. L. Connolly, "The molecular surface package," *J Mol Graph,* vol. 11, pp. 139-41, Jun 1993.
[95] F. Eisenhaber, P. Lijnzaad, P. Argos, C. Sander, and M. Scharf, "The Double Cubic Lattice Method - Efficient Approaches to Numerical-Integration of Surface-Area and Volume and to Dot Surface Contouring of Molecular Assemblies," *Journal of Computational Chemistry,* vol. 16, pp. 273-284, Mar 1995.
[96] S. Jones and J. M. Thornton, "Analysis of protein-protein interaction sites using surface patches," *J Mol Biol,* vol. 272, pp. 121-32, Sep 12 1997.
[97] J. Sambrook, "Adenovirus amazes at Cold Spring Harbor," *Nature,* vol. 268, pp. 101-4, Jul 14 1977.
[98] D. L. Black, "Mechanisms of alternative pre-messenger RNA splicing," *Annu Rev Biochem,* vol. 72, pp. 291-336, 2003.
[99] W. Gilbert, "Why genes in pieces?," *Nature,* vol. 271, p. 501, Feb 9 1978.
[100] G. Ast, "How did alternative splicing evolve?," *Nat Rev Genet,* vol. 5, pp. 773-82, Oct 2004.
[101] S. Stamm, S. Ben-Ari, I. Rafalska, Y. Tang, Z. Zhang, D. Toiber, T. A. Thanaraj, and H. Soreq, "Function of alternative splicing," *Gene,* vol. 344, pp. 1-20, Jan 3 2005.
[102] D. Brett, J. Hanke, G. Lehmann, S. Haase, S. Delbruck, S. Krueger, J. Reich, and P. Bork, "EST comparison indicates 38% of human mRNAs contain possible alternative splice forms," *FEBS Lett,* vol. 474, pp. 83-6, May 26 2000.
[103] J. M. Johnson, J. Castle, P. Garrett-Engele, Z. Kan, P. M. Loerch, C. D. Armour, R. Santos, E. E. Schadt, R. Stoughton, and D. D. Shoemaker, "Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays," *Science,* vol. 302, pp. 2141-4, Dec 19 2003.
[104] B. R. Graveley, "Alternative splicing: increasing diversity in the proteomic world," *Trends Genet,* vol. 17, pp. 100-7, Feb 2001.
[105] K. P. Minneman, "Splice variants of G protein-coupled receptors," *Mol Interv,* vol. 1, pp. 108-16, Jun 2001.
[106] T. H. Thai and J. F. Kearney, "Distinct and opposite activities of human terminal deoxynucleotidyltransferase splice variants," *J Immunol,* vol. 173, pp. 4009-19, Sep 15 2004.
[107] W. Scheper, R. Zwart, and F. Baas, "Alternative splicing in the N-terminus of Alzheimer's presenilin 1," *Neurogenetics,* vol. 5, pp. 223-7, Dec 2004.
[108] R. Roberts, N. A. Timchenko, J. W. Miller, S. Reddy, C. T. Caskey, M. S. Swanson, and L. T. Timchenko, "Altered phosphorylation and intracellular distribution of a (CUG)n triplet repeat RNA-binding protein in patients with myotonic dystrophy and in myotonin protein kinase knockout mice," *Proc Natl Acad Sci U S A,* vol. 94, pp. 13221-6, Nov 25 1997.
[109] K. Ma, J. D. Inglis, A. Sharkey, W. A. Bickmore, R. E. Hill, E. J. Prosser, R. M. Speed, E. J. Thomson, M. Jobling, K. Taylor, and et al., "A Y chromosome gene family with RNA-binding protein homology: candidates for the azoospermia factor AZF controlling human spermatogenesis," *Cell,* vol. 75, pp. 1287-95, Dec 31 1993.
[110] S. Lovestone, C. H. Reynolds, D. Latimer, D. R. Davis, B. H. Anderton, J. M. Gallo, D. Hanger, S. Mulot, B. Marquardt, S. Stabel, and et al., "Alzheimer's disease-like phosphorylation of the microtubule-associated protein tau by glycogen synthase kinase-3 in transfected mammalian cells," *Curr Biol,* vol. 4, pp. 1077-86, Dec 1 1994.
[111] J. P. Venables, "Aberrant and alternative splicing in cancer," *Cancer Res,* vol. 64, pp. 7647-54, Nov 1 2004.
[112] N. Furnham, S. Ruffle, and C. Southan, "Splice variants: a homology modeling approach," *Proteins,* vol. 54, pp. 596-608, Feb 15 2004.
[113] P. Wang, B. Yan, J. T. Guo, C. Hicks, and Y. Xu, "Structural genomics analysis of alternative splicing and application to isoform structure modeling," *Proc Natl Acad Sci U S A,* vol. 102, pp. 18920-5, Dec 27 2005.
[114] A. J. Oakley, T. Harnnoi, R. Udomsinprasert, K. Jirajaroenrat, A. J. Ketterman, and M. C. Wilce, "The crystal structures of glutathione S-transferases isozymes 1-3 and 1-4 from Anopheles dirus species B," *Protein Sci,* vol. 10, pp. 2176-85, Nov 2001.
[115] S. G. Hymowitz, D. M. Compaan, M. Yan, H. J. Wallweber, V. M. Dixit, M. A. Starovasnik, and A. M. de Vos, "The crystal structures of EDA-A1 and EDA-A2: splice variants with distinct receptor specificity," *Structure,* vol. 11, pp. 1513-20, Dec 2003.
[116] C. Peneff, P. Ferrari, V. Charrier, Y. Taburet, C. Monnier, V. Zamboni, J. Winter, M. Harnois, F. Fassy, and Y. Bourne, "Crystal structures of two human pyrophosphorylase isoforms in complexes with UDPGlc(Gal)NAc: role of the alternatively spliced insert in the enzyme oligomeric assembly and active site architecture," *Embo J,* vol. 20, pp. 6191-202, Nov 15 2001.
[117] K. A. Lee, H. Fuda, Y. C. Lee, M. Negishi, C. A. Strott, and L. C. Pedersen, "Crystal structure of human cholesterol sulfotransferase (SULT2B1b) in the presence of pregnenolone and 3'-phosphoadenosine 5'-phosphate. Rationale for specificity differences between prototypical SULT2A1 and the SULT2BG1 isoforms," *J Biol Chem,* vol. 278, pp. 44593-9, Nov 7 2003.
[118] D. Fiegen, L. C. Haeusler, L. Blumenstein, U. Herbrand, R. Dvorsky, I. R. Vetter, and M. R. Ahmadian, "Alternative splicing of Rac1 generates Rac1b, a self-activating GTPase," *J Biol Chem,* vol. 279, pp. 4743-9, Feb 6 2004.
[119] E. Estrada, "Virtual identification of essential proteins within the protein interaction network of yeast," *Proteomics,* vol. 6, pp. 35-40, Jan 2006.
[120] J. Drews, "Drug discovery: a historical perspective," *Science,* vol. 287, pp. 1960-4, Mar 17 2000.
[121] A. G. Cochran, "Antagonists of protein-protein interactions," *Chem Biol,* vol. 7, pp. R85-94, Apr 2000.
[122] M. Arkin, "Protein-protein interactions and cancer: small molecules going in for the kill," *Curr Opin Chem Biol,* vol. 9, pp. 317-24, Jun 2005.
[123] D. C. Fry and L. T. Vassilev, "Targeting protein-protein interactions for cancer therapy," *J Mol Med,* vol. 83, pp. 955-63, Dec 2005.
[124] M. R. Arkin and J. A. Wells, "Small-molecule inhibitors of protein-protein interactions: progressing towards the dream," *Nat Rev Drug Discov,* vol. 3, pp. 301-17, Apr 2004.
[125] P. Chene, "Inhibition of the p53-MDM2 interaction: targeting a protein-protein interface," *Mol Cancer Res,* vol. 2, pp. 20-8, Jan 2004.
[126] A. Bottger, V. Bottger, A. Sparks, W. L. Liu, S. F. Howard, and D. P. Lane, "Design of a synthetic Mdm2-binding mini protein that activates the p53 response in vivo," *Curr Biol,* vol. 7, pp. 860-9, Nov 1 1997.
[127] L. T. Vassilev, "Small-molecule antagonists of p53-MDM2 binding: research tools and potential therapeutics," *Cell Cycle,* vol. 3, pp. 419-21, Apr 2004.
[128] R. Dawson, L. Muller, A. Dehner, C. Klein, H. Kessler, and J. Buchner, "The N-terminal domain of p53 is natively unfolded," *J Mol Biol,* vol. 332, pp. 1131-41, Oct 3 2003.